Hierarchical Block Structures and High-Resolution Model Selection in Large Networks

Tiago P. Peixoto*

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

(Received 5 November 2013; published 24 March 2014)

Discovering and characterizing the large-scale topological features in empirical networks are crucial steps in understanding how complex systems function. However, most existing methods used to obtain the modular structure of networks suffer from serious problems, such as being oblivious to the statistical evidence supporting the discovered patterns, which results in the inability to separate actual structure from noise. In addition to this, one also observes a resolution limit on the size of communities, where smaller but well-defined clusters are not detectable when the network becomes large. This phenomenon occurs for the very popular approach of modularity optimization, which lacks built-in statistical validation, but also for more principled methods based on statistical inference and model selection, which do incorporate statistical validation in a formally correct way. Here, we construct a nested generative model that, through a complete description of the entire network hierarchy at multiple scales, is capable of avoiding this limitation and enables the detection of modular structure at levels far beyond those possible with current approaches. Even with this increased resolution, the method is based on the principle of parsimony, and is capable of separating signal from noise, and thus will not lead to the identification of spurious modules even on sparse networks. Furthermore, it fully generalizes other approaches in that it is not restricted to purely assortative mixing patterns, directed or undirected graphs, and *ad hoc* hierarchical structures such as binary trees. Despite its general character, the approach is tractable and can be combined with advanced techniques of community detection to yield an efficient algorithm that scales well for very large networks.

DOI: 10.1103/PhysRevX.4.011047

Subject Areas: Complex Systems, Interdisciplinary Physics, Statistical Physics

I. INTRODUCTION

The detection of communities and other large-scale structures in networks has become perhaps one of the largest undertakings in network science [1,2]. It is motivated by the desire to be able to characterize the most salient features in large biological [3–5], technological [6,7], and social systems [3,8,9], such that their building blocks become evident, potentially giving valuable insight into the central aspects governing their function and evolution. At its simplest level, the problem seems straightforward: Modules are groups of nodes in the network that have a similar connectivity pattern, often assumed to be assortative, i.e., connected mostly among themselves and less so with the rest of the network. However, when attempting to formalize this notion and develop methods to detect such structures, the combined effort of many researchers in recent years has spawned a great variety of competing approaches to the problem, with no clear, universally accepted outcome [2].

The method that has perhaps gathered the most widespread use is called modularity optimization [10] and consists in maximizing a quality function that favors partitions of nodes for which the fraction of internal edges inside each cluster is larger than expected given a null model, taken to be a random graph. This method is relatively easy to use and comprehend, works well in many accessible examples, and is capable of being applied in very large systems via efficient heuristics [11,12]. However it also suffers from serious drawbacks. In particular, despite measuring a deviation from a null model, it does not take into account the statistical evidence associated with this deviation, and as a result, it is incapable of separating actual structure from those arising simply of statistical fluctuations of the null model, and it even finds high-scoring partitions in fully random graphs [13]. This problem is not specific to modularity and is a characteristic shared by the vast majority of methods proposed for solving the same task [2]. In addition to the lack of statistical validation, modularity maximization fails to detect clusters with size below a given threshold [14,15], which increases with the size of the system as $\sim\sqrt{E}$, where E is the number of edges in the entire network. This limitation is independent of how salient these relatively smaller structures are, and makes this potentially very important information completely inaccessible. Furthermore,

*tiago@itp.uni-bremen.de

results obtained with this method tend to be degenerate for large empirical networks [16], for which many different partitions can be found with modularity values very close to the global maximum. In these common situations, the method fails in giving a faithful representation of the actual large-scale structure present in the system.

More recently, increasing effort has been spent on a different approach based on the statistical inference of generative models, which encode the modular structure of the network as model parameters [17–28]. This approach offers several advantages over a dominating fraction of existing methods, since it is more firmly grounded on well-known principles and methods of statistical analysis, which allows the incorporation of the statistical evidence present in the data in a formally correct manner. Under this general framework, one could hope to overcome some of the limitations existing in more *ad hoc* methods, or at least make any intrinsic limitations easier to understand in light of more robust concepts [29–32]. Perhaps the generative model most used for this purpose is the stochastic block model [17–28,33–36], which groups nodes in blocks with arbitrary probabilities of connections between them. This very simple definition already does away with the restriction of considering only purely assortative communities and accommodates many different patterns, such as core-periphery structures and bipartite blocks, as well as straightforward generalizations to directed graphs. In this context, the detectability of well-defined clusters amounts, in large part, to the issue of model selection based on principled criteria such as minimum description length (MDL) [32,37] or Bayesian model selection (BMS) [38–42]. These approaches allow the selection of the most appropriate number of blocks based on statistical evidence, and thus avoid the detection of spurious communities. However, frustratingly, at least one of the limitations of modularity maximization is also present when doing model selection, namely, the resolution limit mentioned above. As was recently shown in Ref. [32], when using MDL, the maximum number of detectable blocks scales with $\sqrt{N}$, where N is the number of nodes in the network, which is very similar to the modularity optimization limit. However, in this context, this limitation arises out of the lack of knowledge about the type of modular structure one is about to infer and the *a priori* assumption that all possibilities should occur with the same probability. Here, we develop a more refined method of model selection, which consists in a nested hierarchy of stochastic block models, where an upper level of the hierarchy serves as prior information to a lower level. This dramatically changes the resolution of the model selection procedure and replaces the characteristic block size of $\sqrt{N}$ in the nonhierarchical model by much a smaller value that scales only logarithmically with N, enabling the detection of much smaller blocks in very large networks. Furthermore, the approach provides a description of the network in many scales, in a complete model encapsulating its entire hierarchical structure at once. It generalizes previous methods of hierarchical community detection [43–49], in that it does not impose specific patterns such as dendograms or binary trees, in addition to allowing arbitrary modular structures as the usual stochastic block model, instead of purely assortative ones. Furthermore, despite its increased resolution, the approach attempts to find the simplest possible model that fits the data and is not subject to overfitting, and, hence, will not detect spurious modules in random networks. Finally, the method is fully nonparametric and can be implemented efficiently with a simple algorithm that scales well for very large networks.

In Sec. II, we start with the definition of the model and then we discuss the model-selection procedure based on MDL. We then move to the analysis of the resolution limit, and proceed to define an efficient algorithm for the inference of the nested model, and we finalize with the analysis of synthetic and empirical networks, where we demonstrate the quality of the approach. We then conclude with an overall discussion.

II. HIERARCHICAL MODEL

The original stochastic block model ensemble [33–36] is composed of N nodes, divided into B blocks, with e_{rs} edges between nodes of blocks r and s (or, for convenience of notation, twice that number if $r = s$). Here, we differentiate between two very similar model variants: (1) the edge counts e_{rs} are themselves the parameters of the model, and (2) the parameters are the probabilities p_{rs} that an edge exists between two nodes of the respective blocks, such that the edge counts $\langle e_{rs}\rangle = n_r n_s p_{rs}$ are constrained on average. Both are equally valid generative models, and as long as the edge counts are sufficiently large, they are fully equivalent (see Ref. [50] and Appendix A). Here, we stick with the first variant, since it makes the following formulation more convenient. We also consider a further variation called the degree-corrected block model [24], which is defined exactly as the traditional model(s) above, but one additionally specifies the degree sequence $\{k_i\}$ of the graph as an additional set of parameters (again, these values can be the parameters themselves, or they can be constrained on average [50]). The degree-corrected version, although it is a relatively simple modification, yields much more convincing results on many empirical networks, since it is capable of incorporating degree variability inside each block [24]. As will be seen below, it is, in general, also capable of providing a more compact description of arbitrary networks than the traditional version.

The nested version, which we define here, is based on the simple fact that the edge counts e_{rs} themselves form a block multigraph, where the nodes are the blocks and the edge counts are the edge multiplicities between each node pair (with self-loops allowed). This multigraph may also be constructed via a generative model of its own. If we choose

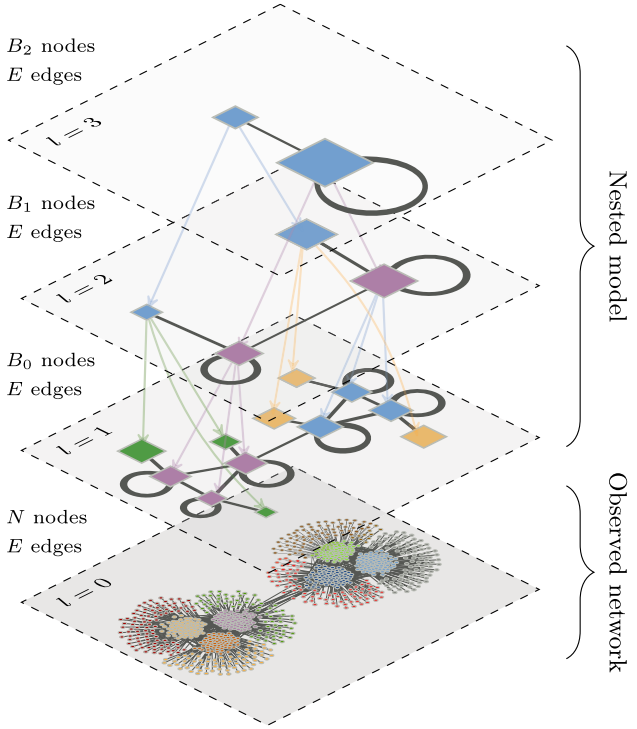


FIG. 1. Example of a nested stochastic block model with three levels, and a generated network at the bottom. The top-level structure describes a core-periphery network, which is further subdivided in the lower levels.

a stochastic block model again as a generative model, we obtain another smaller block multigraph as parameters at a higher level, and so on recursively, until we finally reach a model with only one block. This forms a nested stochastic block model hierarchy, which describes a given network at several resolution levels (see Fig. 1).

This approach provides an increased resolution when performing model selection, since the generative model inferred at an upper level serves as prior information to the one at a lower level. Despite its more elaborate formulation, this hierarchical model remains tractable, and it is possible to apply it to very large networks, in a fully nonparametric manner, as discussed below. Furthermore, it generalizes cleanly the flat variants, which correspond simply to a hierarchy with only one level. It also does not impose any preferred mixing pattern (e.g., assortative or dissortative block structures), and is not restricted to any specific hierarchical form, such as binary trees or dendograms [97]. In the following, we describe the maximum likelihood method to infer the multilevel partitions and the model selection process based on the minimum description length principle and compare it with Bayesian model selection.

In the analysis, we focus on undirected networks, but everything is straightforwardly applicable to directed networks as well. In Appendix C, we present a summary of the relevant expressions for the directed case.

A. Module inference

The inference approach consists in finding the best partition $\{b_i\}$ of the nodes, where $b_i \in [1, B]$ is the block membership of node i, in the observed network G, such that the posterior likelihood $\mathcal{P}(G|\{b_i\})$ is maximized. Since each graph with the same edge counts e_{rs} occurs with the same probability, the posterior likelihood is simply $\mathcal{P}(G|\{b_i\}) = 1/\Omega(\{e_{rs}\}, \{n_r\})$, where e_{rs} and n_r are the edge and node counts associated with the block partition $\{b_i\}$ and $\Omega(\{e_{rs}\}, \{n_r\})$ is the number of different network realizations. Hence, maximizing the likelihood is identical to minimizing the ensemble entropy [50,52] $\mathcal{S}(\{e_{rs}\}, \{n_r\}) = \ln \Omega(\{e_{rs}\}, \{n_r\})$.

For the lowest level of the hierarchy (which models directly the observed network), we have a simple graph, for which the entropies can be computed as [50]

$$\mathcal{S}_t = \frac{1}{2}\sum_{rs} n_r n_s H_b\left(\frac{e_{rs}}{n_r n_s}\right), \tag{1}$$

for the traditional block model ensemble, and

$$\mathcal{S}_c \simeq -E - \sum_k N_k \ln k! - \frac{1}{2}\sum_{rs} e_{rs} \ln\left(\frac{e_{rs}}{e_r e_s}\right), \tag{2}$$

for the degree-corrected variant, where $E = \sum_{rs} e_{rs}/2$ is the total number of edges, N_k is the total number of nodes with degree k, $e_r = \sum_s e_{rs}$ is the number of half-edges incident on block r, $H_b(x) = -x \ln x - (1-x)\ln(1-x)$ is the binary entropy function, and it was assumed that $n_r \gg 1$. Note that only the last term of Eq. (2) is, in fact, useful when finding the best block partition, since the other terms remain constant. However, the full expression is necessary when comparing the models against each other via model selection, as discussed below.

For the upper-level multigraphs, the entropy can also be computed [50], and it takes a different form

$$\mathcal{S}_m = \sum_{r>s} \ln \left(\!\!\binom{n_r n_s}{e_{rs}}\!\!\right) + \sum_r \ln \left(\!\!\binom{\left(\!\binom{n_r}{2}\!\right)}{e_{rr}/2}\!\!\right), \tag{3}$$

where $\left(\!\binom{n}{m}\!\right) = \binom{n+m-1}{m}$ is the number of m-combinations with repetitions from a set of size n. Note that we no longer assume that $n_r \gg 1$, since at the upper levels the number of nodes becomes arbitrarily small.

At each level $l \in [0, L]$ in the hierarchy there are B_{l-1} nodes, which are divided into B_l blocks (with $B_l \leq B_{l-1}$), were we set $B_{-1} \equiv N$. The edge counts at level l are denoted e^l_{rs} and the block sizes n^l_r. Therefore, we must have that $\sum_r n^l_r = B_{l-1}$ and $\sum_{rs} e^l_{rs}/2 = E$; i.e., the total number of nodes decreases in the upper levels, but the total number of edges remains the same. The combined entropy of all layers is then given by

$$\mathcal{S}_n = \mathcal{S}_{t/c}(\{e^0_{rs}\},\{n^0_r\}) + \sum_{l=1}^{L} S_m(\{e^l_{rs}\},\{n^l_r\}). \quad (4)$$

The full generative model corresponds to a nested sequence of network ensembles, where each sample from a given level generates another ensemble at a lower level. The entropy in Eq. (4) represents the amount of information necessary to encode the decision sequence, which, starting from the topmost model, selects the observed network among all possible branches in the lower levels.

Whenever both the number of levels and the number of blocks B_l of each level is known, the best multilevel partition is the one that minimizes $\mathcal{S}_n$. However, such information regarding the size of the model is most often not available and needs to be inferred from the data as well. Using Eq. (4) for this purpose is not appropriate, since minimizing it across all possible hierarchies leads to a trivial and meaningless result where $B_l = N$ for all l. Instead, one must employ some form of Occam's razor and select the simplest possible model that best describes the observed data without increasing its complexity. We present such an approach in the next section.

B. Model selection

A method that directly formalizes Occam's razor principle is known as minimum description length [53,54], where one specifies the *total* amount of information necessary to described the data, which includes not only the sample but the model parameters as well. The description length for the model above is

$$\Sigma = \mathcal{L}_{t/c} + \mathcal{S}_{t/c}, \quad (5)$$

where $\mathcal{L}_{t/c}$ is the amount of information necessary to fully describe the model and $\mathcal{S}_{t/c}$ corresponds to entropy of the lowest level $l = 0$ of the hierarchy. In a given level l of the hierarchy, the information required to describe the model parameters $\{e^l_{rs}\}$ is given by the entropy S_m [Eq. (3)] of the model in level $l+1$, so that we may write

$$\mathcal{L}_t = \sum_{l=1}^{L} S_m(\{e^l_{rs}\},\{n^l_r\}) + \mathcal{L}^{l-1}_t. \quad (6)$$

The only missing information is how to partition the nodes of the current level into B_l blocks, which corresponds to the term $\mathcal{L}^l_t$ in the equation above. The total number of partitions with the same block sizes $\{n^l_r\}$ is given by $B_{l-1}!/\prod_r n^l_r!$, and the total number of different block sizes is $\left(\!\binom{B_l}{B_{l-1}}\!\right)$. Hence, the amount of information necessary to describe the block partition of level l is

$$\mathcal{L}^l_t = \ln\left(\!\binom{B_l}{B_{l-1}}\!\right) + \ln B_{l-1}! - \sum_r \ln n^l_r!. \quad (7)$$

Note that this is different from the choice made in Refs. [32,37], which considered all possible $B_l^{B_{l-1}}$ partitions to be equally likely, and, hence, the necessary amount of information was computed as $B_{l-1}\ln B_l$. This choice implicitly assumes that all blocks have equal sizes and offers a worse description when this is not the case. Note that for $B_{l-1} \gg 1$, we have

$$\mathcal{L}^l_t \simeq B_{l-1} H(\{n^l_r/B_{l-1}\}), \quad (8)$$

where $H(\{p_i\}) = -\sum_i p_i \ln p_i$ is the entropy of the distribution $\{p_i\}$. Therefore, for uniform blocks $n^l_r = B_{l-1}/B_l$, we recover asymptotically the value $\mathcal{L}^l_t \simeq B_{l-1}\ln B_l$. However, the value of Eq. (7) can be much smaller for nonuniform partitions. This choice has important consequences for the resolution of relatively small blocks, as will be seen below.

For the degree-corrected version, we still need to include the information necessary to describe the degrees at the lowest level,

$$\mathcal{L}_c = \mathcal{L}_t + \sum_r n_r H(\{p^r_k\}), \quad (9)$$

where $\{p^r_k\}$ is the degree distribution of nodes belonging to block r [98]. It is worth noting that, if a network is sampled from the traditional block model ensemble, so that p^r_k is a Poisson with average e_r/n_r, Eq. (9) becomes $\mathcal{L}_c = \mathcal{L}_t + 2E - \sum_r e_r \ln e_r/n_r + \sum_k N_k \ln k!$, which means that $\mathcal{S}_c + \mathcal{L}_c = \mathcal{S}_t + \mathcal{L}_t$, i.e., the total description length is identical for both the traditional and the degree-corrected models in this case, and, therefore, both models describe the same network equally well [99]. However, if the distributions $\{p^r_k\}$ deviate from Poissons, the degree-corrected variant will provide, in general, a shorter description length.

It is easy to see that, if one has a flat $L = 1$ hierarchy, with $\{B_l\} = \{B, 1\}$, the description length of the nonhierarchical model is recovered [32]; e.g., for the traditional model, we have $\Sigma_{L=1} = \mathcal{L}_{L=1} + \mathcal{S}_t$, with

$$\mathcal{L}_{L=1} = \ln\left(\!\binom{\binom{B}{2}}{E}\!\right) + \ln\left(\!\binom{B}{N}\!\right) + \ln N! - \sum_r \ln n_r!, \quad (10)$$

where the only difference in comparison to Ref. [32] is that here we are using the improved partition description length of Eq. (7). Therefore, the nested generalization fully encapsulates the flat version, such that $\min\Sigma \leq \min\Sigma_{L=1}$; i.e., the nested model can provide only a shorter or equal description length of the observed network.

The MDL principle predicates that whenever the hierarchy size itself needs to be inferred, one should minimize Eq. (5), instead of Eq. (4) directly. However, MDL is one of the many principled methods one could use to do model selection, which include, e.g., Bayesian model selection via

integrated likelihood [21,38,39,41,42,55], likelihood ratios [56], or more approximative methods, such as Bayesian information criterion [57] and Akaike information criterion [58]. If any two of these methods are derived from equivalent assumptions, one should expect them to deliver compatible results. In Appendix A, we make a comparison of the MDL approach with Bayesian model selection via integrated likelihood (BMS), since it is nonapproximative and can be computed exactly for the stochastic block model, where we show that under compatible assumptions, these two methods deliver the exact same results. In the following, we compare the results obtained with nonhierarchical MDL (or BMS) and the nested model presented, and show that it yields a higher quality model selection criterion, which detects the correct number of blocks for sparse networks, without being overconfident. Based on this analysis, we are capable of deriving the optimum number of blocks given a network size, and we show that the nested model does not suffer from the resolution limit, which hinders the nonhierarchical approach.

1. Module detectability and the "resolution limit"

The general problem of module detectability can be formulated as follows: Suppose we generate a network with a given parameter set. To what extent can we recover the planted parameters by observing this single sample from the model? The answer is conditional on the amount of one's prior knowledge. If the number of blocks B is known beforehand, the remaining task is simply to classify the nodes in one of these B classes. This problem has been shown to exhibit a detectability-indetectability phase transition [29,30,59,60]: If the existing block structure is too weak, it becomes impossible to infer the correct partition with any method, despite the fact that the model parameters deviate from that of a fully random graph. On the other hand, if the block structure is sufficiently strong, it is possible to detect the correct partition with a precision that increases as the block structure becomes stronger. Another situation is when one does not know the correct number B, which is arguably more relevant in practice. In this case, in addition to the node classification, one needs to perform model selection. Ideally, one would like to find the correct B value whenever the corresponding partition is detectable. However, in situations where the correct partition is only *partially* detectable, i.e., the inferred partition is positively but weakly correlated with the true model, an application of Occam's razor may actually choose a simpler model, with smaller B, with a comparable correlation with the true partition. Hence, if we lack knowledge of the model size B, there will be situations where the true partition will be more poorly detected, when compared to the case where we have this information. This can be clearly illustrated with a very simple example known as the planted partition (PP) model [61]. It corresponds to an assortative block structure given by $e_{rs} = 2E[\delta_{rs}c/B + (1-\delta_{rs})(1-c)/B(B-1)]$, $n_r = N/B$, and $c \in [0,1]$ is a free parameter that controls the assortativity strength. For this model, if we have that $N/B \gg 1$, it can be shown that the detectable phase exists for $\langle k\rangle > [(B-1)/(cB-1)]^2$ [29–31]. Let us make the situation even simpler and consider the strongest possible block structure with $c = 1$, i.e., B perfectly isolated assortative communities with N/B nodes. In this case, the detectability threshold lies at $\langle k\rangle = 1$. Therefore, for any $\langle k\rangle > 1$, we should be able to detect all B blocks, with a precision increasing with $\langle k\rangle$, if we know we have B blocks to begin with. If we do not know this, we must apply a model-selection criterion as described above to obtain the best value of B. For simplicity, let us assume that, for the correct value of $B \equiv B_{\text{true}}$, the true partition is perfectly detected, such that $\mathcal{S}_t \simeq -E\ln B$, ignoring additive constants, which are irrelevant at this point. If a value of $B > B_{\text{true}}$ is used, we assume that the inferred partition corresponds to regular subdivisions of the planted one, such that the entropy remains approximately unchanged $\mathcal{S}_t \simeq -E\ln B_{\text{true}}$. For $B < B_{\text{true}}$, the blocks are uniformly merged together, so that $\mathcal{S}_t \simeq -E\ln B$. Hence, we may write the expected value of the minimum description length in the nonhierarchical model by summing $\mathcal{S}_t = -E\ln\min(B, B_{\text{true}})$ with Eq. (10). For the nested version of the model, we assume a regular hierarchy tree of depth L and with a fixed branching ratio σ, i.e., $B_l = \sigma^{L-l}$, so that Eq. (5) becomes

$$\Sigma \simeq \left(\!\!\binom{\sigma}{2}\!\!\right)\frac{B}{\sigma-1}\ln E + \frac{\sigma}{2}B\ln B + N\ln B - E\ln\min(B, B_{\text{true}}), \tag{11}$$

where $B_l \gg \sigma$ was assumed, together with $L \gg 1$, and $B \equiv B_0$. One may compare these criteria against each other in their capacity of recovering the planted value of B, by finding the extremum of each function. In Fig. 2, we show the optimum values of B for a model with $N = 10^4$ and $B_{\text{true}} = 100$, as well as the results for the direct minimization of the corresponding exact quantities for actual network realizations, and a comparison of the obtained partitions using the normalized mutual information (NMI) [100]. We also include the comparison with a dense BMS criterion (see Appendix A) both in its full form [Eqs. (A5) and (A8)] and with the partition likelihood term omitted, i.e., $\mathcal{P}(\{b_i\}|B) = 1$, as was done in Refs. [23,40]. We see that the dense BMS criterion fails to detect the correct model size for sparser networks, which is in accordance with its inadequacy in this region. The hierarchical model provides, as expected, the best results and detects the correct model for the sparsest networks. The incomplete BMS criterion is clearly overconfident for sparse networks and detects $B > 1$ structures even when the model lies below the detectability threshold $\langle k\rangle = 1$; hence, this shows that the partition likelihood should not simply be

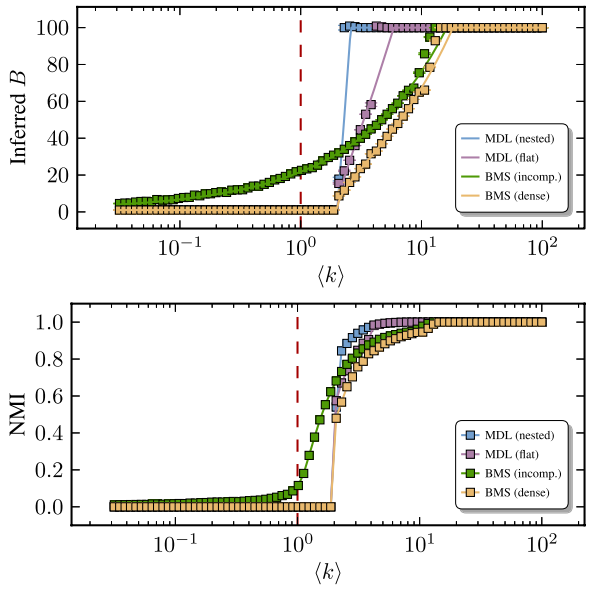


FIG. 2. Model selection results for a PP model with $N = 10^4$, $B_{\text{true}} = 100$, and fully isolated blocks ($c = 1$), using the model-selection criteria described in the text. The top panel shows the inferred value of B versus the average degree $\langle k \rangle$ in the network. The solid lines show the theoretical value according to each criterion, and the data points are direct optimization of the corresponding quantities for actual generated network, averaged over 40 independent realizations. The bottom panel shows the normalized mutual information (NMI) between the inferred and planted partitions. The dashed line marks the threshold $\langle k \rangle = 1$ where inference becomes impossible for $N \to \infty$.

discarded [101]. Both MDL and dense BMS fail to detect anything for $\langle k \rangle < 2$, which corresponds to a strong threshold [102], which interestingly lies above the strict detectability limit at $\langle k \rangle = 1$. This corresponds to a region where detectability is possible, but only if the true value of B is known (or if a more refined model-selection criterion exists). Note that the incomplete BMS criterion performs better in the region $1 < \langle k \rangle < 2$, but this is perhaps better interpreted as a by-product of its overall overconfidence for very sparse networks. Note that all criteria eventually agree on the correct value if $\langle k \rangle$ is made sufficiently large, which corresponds to the intuitive notion that the detection problem becomes much easier for dense networks.

A prominent problem in the detectability of block structures via other methods, such as modularity optimization [10], is when modules are merged together, regardless of how strong the community structure is perceived to be. For the modularity-based approach, when considering a maximally modular network, similar to the PP model with $c \to 1$, but with the additional restriction that the graph remains connected, it has been shown [14] that modules are merged together as long as $B > \sqrt{E}$. This phenomenon is considered counterintuitive, and has been called the "resolution limit" of community detection via this method [103]. As it happens, this problem does not only occur for modularity-based methods, but also if one does statistical inference based on MDL. For the nonhierarchical model, it can be shown that, according to this criterion, the optimal number of blocks scales as $B^* \simeq \mu(\langle k \rangle)\sqrt{N}$, where $\mu(x)$ is an increasing function [32]. Therefore, if the planted number exceeds this threshold, blocks will be merged together, despite the fact that the block structure is detectable with arbitrary precision if one knows the correct value of B, and it sufficiently exceeds the detectability threshold $\langle k \rangle > 1$ of the PP model. This means that the true parameters of the model can no longer be used to compress the generated data. This is a direct result of the assumption that all possible block structures of a given size are equally possible, and the number of such models becomes very large, with a model description length scaling roughly with $\sim B^2 \ln E + N \ln B$. In the presence of additional assumptions about the model, such as the fact that one is dealing with the PP model instead of a more general block structure, this can, in principle, be improved. However, in most practical situations, such assumptions cannot be made. One main advantage of the nested model is that this limit can be overcome *without* requiring such prior knowledge. The description length via the nested model for the maximally modular network above is given by Eq. (11), with $B_{\text{true}} = B$. As can be seen, this equation has only log-linear dependencies on the model size B, instead of the quadratic one present in the flat MDL. The result of this is that, if one finds the value of B^*, which minimizes the nested description length, one obtains the scaling

$$B^* \propto \frac{N}{\ln N}, \tag{12}$$

for sufficiently large N. This is a significant improvement, since the maximum number of detectable blocks grows almost linearly with the number of nodes. Thus, a characteristic detectable block size $N/B^* \sim \sqrt{N}$ is replaced by a much smaller value $N/B^* \sim \ln N$, which allows for a precise assessment of small communities even in very large networks.

It is possible to understand in more detail the origin of the improvement by considering a related problem, which is the detection of specific blocks that are much smaller than the remaining network. Another facet of the resolution limit manifests itself when two such blocks are merged together, despite the fact that if they are considered in isolation they would be kept separate. Here, we consider this problem by using a slightly modified scenario than the one proposed in Ref. [14], which is a network composed of two fully isolated blocks, each with $e_c/2$ internal edges and n_c nodes, and a remaining network with N nodes, E edges, average degree $\langle k \rangle = 2E/N$, and an arbitrary topology [see Fig. 3(c)]. We may decide if these blocks are merged together by considering the difference in the description length. The entropy difference for the merge is simply $\Delta\mathcal{S}_t = e_c \ln 2$ (where we assume $e_c \ll n_c^2$, but the dense

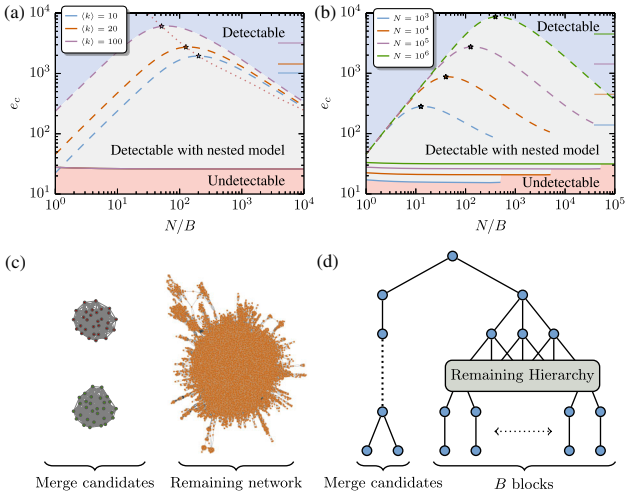


FIG. 3. Parameter region where two isolated blocks with $e_c/2$ internal edges and $n_c = e_c/5$ nodes are detectable as separate blocks [shown schematically in panel (c)], as a function of the average block size N/B, and depending on (a) the average degree $\langle k\rangle$ with $N = 10^5$ and (b) the number of nodes N with $\langle k\rangle = 20$. The dashed curves show the boundaries for the nonhierarchical block model, and the solid lines for the hierarchical variant. The line segments on the right-hand side of the plots show the detectability threshold for modularity [14], $e_c^* = \sqrt{2E}$. The points marked with stars ($\star$) correspond to the maximum value of B that is detectable in the remaining network with the nonhierarchical model, and the dotted line shows the same quantity for various $\langle k\rangle$ values (i.e., the region on the left of this curve corresponds to an overfitting of the remaining network, according to the non-hierarchical criterion). (d) The hierarchical construction used to decide if the two isolated blocks are merged together with the nested model.

case can be computed as well, with no significant difference in the result). For the flat block model, we have $\Delta\mathcal{L}_{\text{flat}} = \mathcal{L}_{L=1}(E + e_c, N + 2n_c, B - 1, \{n_r\} \cup \{2n_c\}) - \mathcal{L}_{L=1}(E + e_c, N + 2n_c, B, \{n_r\} \cup \{n_c, n_c\})$, computed using Eq. (10). For this case, the point at which the merge happens, $\Delta\mathcal{L}_{\text{flat}} + \Delta\mathcal{S}_t = 0$, will depend not only on the values of E and N, but also on the average block size N/B of the remaining network, as can be seen in Figs. 3(a) and 3(b). As the number of blocks in the remaining network approaches the maximum detectable value, $B^* \sim \sqrt{N}$, the more difficult it becomes to resolve the smaller blocks. The detectable region recedes further with increasing $\langle k\rangle$, and also with the number of nodes in the remaining network as $e_c^* \sim \sqrt{N}$. Hence, the denser or larger the remaining network, the harder it becomes to detect the smaller blocks with the flat variant of the model. In Fig. 3 are also shown the values of e_r^* for which modularity also fails to separate the blocks (if one considers that they are connected to themselves and to the rest of the network by single edges [104]), which are overall compatible with the flat MDL criterion. The situation changes significantly with the nested model. To consider the merge, we assume an optimal block hierarchy that splits at the top into two branches, the left one containing the two smaller blocks and the right one containing the remaining network and its arbitrary hierarchical structure [see Fig. 3(d)]. To consider the merge, we need to compute the description length only at the lowest level, since the rest remains unchanged after the merge. By computing the difference via Eq. (5), after some manipulations we obtain $\Delta\Sigma_{\text{nested}} = \Delta\mathcal{S}_t + \ln n_c - \ln(\binom{3}{e_c}) + \ln(B+1) - \ln(B+N-1) - \ln(B_1+B+2)$, with $B = B_0$. Note that this expression is independent of E, and, hence, the density of the remaining network cannot influence the merging decision. Since $B_1 \leq B$, and assuming $B \gg 1$, we obtain $\Delta\Sigma_{\text{nested}} \simeq \Delta\mathcal{S}_t + \ln n_c - \ln[(e_c+2)(e_c+1)] - \ln(B+N)$, and, hence, the dependence on either N or B is again only logarithmic, $e_c^* \approx [\ln(B+N) - \ln n_c]/\ln 2$, as shown in Fig. 3(b). With this example, one can notice that the nested model is capable of compartmentalizing the network at the upper levels, such that the lower-level branches can become almost independent of each other. This means that, in many practical situations, one can sufficiently overcome the resolution limit without abandoning a global model that describes the whole network at once.

In the following section, we specify an efficient algorithm to infer the parameters of the nested block model in arbitrary networks, and we test its efficacy in uncovering the multilevel structure of synthetic as well as empirical networks.

III. INFERENCE ALGORITHM

Individually, any specific level l of the hierarchical structure is a regular block model, and, hence, the classification of the B_{l-1} nodes of this level into B_l blocks can be done via well-established methods, such as the Monte Carlo method [32,40], simulated annealing, or belief propagation [29,30,56]. Here, we use the method described in Ref. [63], which is an agglomerative heuristic that provides high-quality results, while being unbiased with respect to the types of block structure that are inferred, and is also very efficient, with an algorithmic complexity of $O(N\ln^2 N)$, independent of the number of blocks B. If one knows the depth L of the hierarchy, and all $\{B_l\}$ values, the multilevel partitions can be obtained by starting from the lowest level $l = 0$ and progressing upwards to $l = L$. However, this cannot be done when the number and sizes of the hierarchical levels are unknown. Although it is relatively simple to heuristically impose such patterns as binary trees or dendograms, these are not satisfactory given the general character of the model, which accommodates arbitrary branching patterns. However, traversing all possible hierarchies is not feasible for moderate or large networks; thus, one must settle with approximative methods. Here, we propose a very simple greedy heuristic, which, given any starting hierarchy, performs a series of local moves to obtain the optimal branching. Although this

algorithm is not guaranteed to find the global optimum, we have found it to perform very well for many synthetic and empirical networks, and it tends to find consistent hierarchies, independently of the starting estimate. It is also efficient enough not to hinder its application to very large networks, since it does not significantly change the overall algorithmic complexity of the inference procedure. The algorithm is based on the following local moves at a given hierarchy level l.

(1) *Resize.*—A new partition of the B_{l-1} nodes into a newly chosen number of blocks B_l is obtained. This is done via the agglomerative heuristic mentioned previously, with the modification that it must not invalidate the partition at the level $l+1$; i.e., no nodes that belong to different blocks at the upper level can be merged together in the current level. This restriction enables the difference in Σ [Eq. (5)] to be computed easily, since it depends only on the modifications made in the current and upper levels, l and $l+1$. The actual new value of B_l is chosen via progressive bisection of the range $B_l \in [B_{l-1}, B_{l+1}]$, so that the minimum of Σ is bracketed, and for each value of B_l attempted, the best partition is found with the algorithm of Ref. [63].

(2) *Insert.*—A new level is inserted at position l. Its size and partition are chosen exactly as in the resize move above.

(3) *Delete.*—The model in level l is removed from the hierarchy; i.e., the nodes of level $l-1$ are grouped together directly as described in level $l+1$.

Through repeated applications of these moves, it is possible to construct any hierarchy. The actual greedy optimization consists of starting with some initial hierarchy and keeping track of whether or not each level is "done" or "not done." One initially marks all levels as not done and starts at the top level $l = L$. For the current level l, if it is marked done, it is skipped and one moves to the level $l-1$. Otherwise, all three moves are attempted. If any of the moves succeeds in decreasing the description length Σ, one marks the levels $l-1$ and $l+1$ (if they exist) as not done, the level l as done, and one proceeds (if possible) to the upper level $l+1$, and repeats the procedure. If no improvement is possible, the level l is marked as done and one proceeds to the lower level $l-1$. If the lowest level $l = 0$ is reached and cannot be improved, the algorithm ends. Note that, in order to keep the description length complete, we must impose that $B_L = 1$ throughout the above process. The final hierarchy will, in general, depend on the starting hierarchy, and as was mentioned above, one cannot guarantee that the global minimum is always found. However, we find that, in the majority of cases, this algorithm succeeds in finding the same or very similar hierarchies, independently of the initial choice, which can simply be $\{B_l\} = \{1\}$. However, the actual time it takes to reach the optimum will depend on how close the initial tree was to the final one, and, hence, it is difficult to give an estimate of the total number of moves necessary. However, the slowest move is the resize operation, which completes in $O(B_{l-1}\ln^2 B_{l-1})$ steps, and, hence, most of the time is spent at the lowest level $l = 0$ with $B_{-1} = N$, which scales well for very large networks. We have succeed in obtaining reliable results with this algorithm for networks in excess of 10^7 edges; hence, it is suitable for large-scale systems [105].

IV. SYNTHETIC BENCHMARKS

Here, we consider the performance of the nested block model inference procedure on artificially constructed networks. Here, we use a nested version of the usual PP model [61], inspired by similar constructions done in Refs. [51,64]. We define a seed structure with B_0 blocks and $[\boldsymbol{m}_1]_{rs} = \delta_{rs}c/B_0 + (1-\delta_{rs})(1-c)/B_0(B_0-1)$, and construct a nested matrix of depth $L-1$ via $\boldsymbol{m}_l = \boldsymbol{m}_{l-1} \otimes \boldsymbol{m}_{l-1}$, where $\otimes$ denotes the Kronecker product and $l \in [1, L-1]$. The parameters of the model at level l are $e^l_{rs} = 2Em_{rs}$, and all $B = B_0^{L-1}$ blocks have the same number of nodes. Via spectral methods [65], one can show that the detectability transition happens at $\langle k\rangle = [(B_0-1)/(cB_0-1)]^2$, which is the same as the regular PP model with $B = B_0$ [29–31,66].

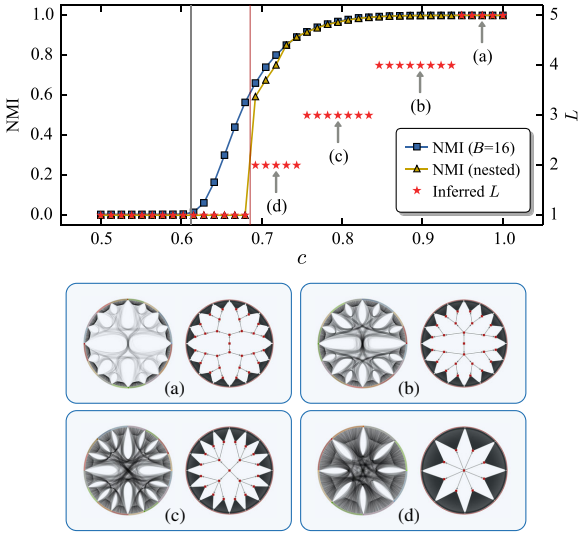


FIG. 4. Top: NMI between the inferred and true partitions for network realizations of the nested PP model described in the text with $B_1 = 2$, $L = 5$, $\langle k\rangle = 20$, and $N = 10^4$, as a function of the assortativity strength c, both via the standard stochastic block model with $B = 16$ and the nested variant with unspecified parameters. The star symbols ($\star$) show the value of L for the inferred hierarchy. All points are averaged over 20 independent realizations. The gray vertical line marks the detectability threshold when B is predetermined, and the red line when the nested model fails to detect any structure. Bottom: Example hierarchies inferred for the values of c indicated in the top panel. The left-hand image shows the network realization itself, and the right-hand one the hierarchical structure [the planted hierarchy corresponds to the one in (a)].

In Fig. 4, we show the results of the inference procedure for a generated model with $B_0 = 2$ and $L = 5$, $N = 10^4$ nodes, and $\langle k \rangle = 20$. The correct number of blocks is detected up to a given value of $c > c^*$, where c^* is the detectability threshold. The hierarchy itself matches the nested PP model exactly only for higher values of c, and becomes progressively simplified for lower values. Note that, for a large fraction of c values, the correct lower-level partition is detected with a very high precision, but the hierarchy that is inferred is simpler than the planted one. In these cases, however, both the inferred hierarchy as well as the planted model are fully equivalent; i.e., they generate the same networks. In other words, the shallower hierarchies that are inferred correspond to identical representations of the same e_{rs} matrix at the lowest level, which require less information to be described, in comparison to the sequence of Kronecker products used in the model specification, and, hence, cannot really be seen as a failure of the inference method, since it simply manages to compress the original model. Before the value of c reaches the detectability threshold, the inference method settles on a fully random $L = 1$, $B = 1$ structure, corresponding once again to a parameter region where the block detection is only possible with limited precision and if one knows the correct model size. As predicated by the MDL criterion, the inferred models tend to be as simple as possible, with the hierarchies becoming shallower as one approaches a random graph. The approach is, therefore, conservative, which brings confidence to the blocks and hierarchies that are actually found, since, despite the increased resolution capabilities, it does not tend to find spurious hierarchies.

In Appendix B, we also include a comparison of the method with other algorithms for community detection that are not based on statistical inference.

V. EMPIRICAL NETWORKS

Here, we present a detailed analysis of some selected empirical networks, as well as a meta-analysis of several networks, spanning different domains and size scales. In all cases, we use the degree-corrected stochastic block model at the lowest hierarchical level, instead of the traditional model, since it almost always provides better results.

Political blogs of the 2004 U.S. election.—This is a network compiled by Adamic and Glance [67] of political blogs during the 2004 presidential election in the U.S. The nodes are $N = 1222$ individual blogs, and $E = 19027$ directed edges exists between pairs of blogs, if one blog cites the other. This network is often used as an empirical example of community structure, since it displays a division along political lines, with two clearly distinct groups representing those aligned with the Republican and the Democratic parties. Indeed, if one applies the nested block model to this network, the topmost division in the hierarchy corresponds exactly to this bimodal partition, which closely matches the accepted division (see Fig. 5). This partition is also obtained with the nonhierarchical stochastic block model if one imposes $B = 2$ [24].

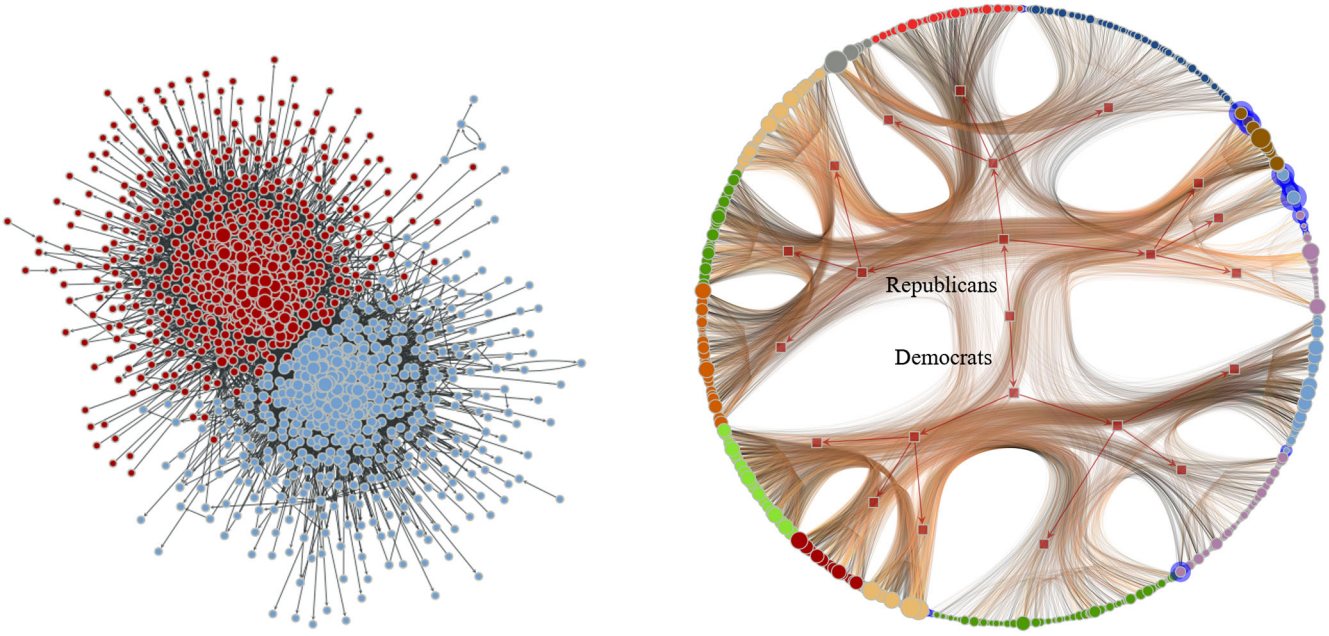


FIG. 5. The political blog network of Adamic and Glance [67]. Left: Topmost partition of the hierarchy inferred with the nested model. Right: The same network, using a circular layout, with edge bundling following the inferred hierarchy [68] (indicated also by the square nodes and the node colors). The size of the nodes corresponds to the total degree, and the edge color indicates its direction (from dark to light). Nodes marked with a blue halo were incorrectly classified at the topmost level, according to the accepted partition in Ref. [67].

However, the nested version reveals a much more complete picture of the network, where these two partitions possess a detailed internal structure, culminating in $B_0 = 15$ subgroups with quite heterogeneous connection patterns. For instance, one can see that each of the two higher-level groups possesses one or more subgroups composed mainly of peripheral nodes, i.e., blogs that cite other blogs but are not themselves cited as often. Conversely, both factions possess subgroups that tend to be cited by most other groups, and others which are cited predominantly by specific groups. It is also interesting to note that a large fraction of the connections between the two top-level groups are concentrated between only two specific subgroups, which, therefore, act as bridges between the larger groups.

This example shows that the model is capable of revealing the structure of the network at multiple scales, which reveal simultaneously the existence of the bimodal large-scale division, as well the lower-level subdivisions.

The autonomous systems topology of the Internet.—Autonomous systems (AS) are intermediary building blocks of the Internet topology. They represent organizational units that are used to control the routing of packets in the network. A single AS often corresponds to a network of its own, which is usually owned by a private company or a government body. The network analyzed here corresponds to the traffic of information between the AS nodes, as measured by the CAIDA project [106]. Each node in the network is an AS, and a directed link exists between two nodes if direct traffic has been observed between the two AS. As of September 2013, the network is composed of $N = 52\,104$ AS nodes and $E = 399\,625$ direct connections between them. The application of the nested block model to this network yields the hierarchy seen in Fig. 6, with $B = 191$ blocks at the lowest level. The most prominent feature observed is a strong core-periphery structure, where most connections go through a relatively small group of nodes, which act as hubs in the network. The groups both in the core and in the periphery seem strongly correlated to geographical location. However, the nodes of the core groups are not confined to a single geographical location, and are instead spread all over the globe (see inset of Fig. 6 and the Supplemental Material [96]).

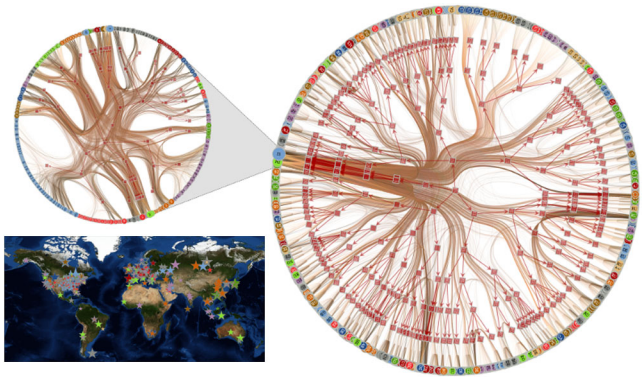


FIG. 6. Large-scale structure of the Internet at the autonomous systems level, as obtained by the nested stochastic block model, displaying a prominent core-periphery architecture. The magnification shows the nodes that belong to the "core" top-level branch, containing AS nodes spread all over the globe, as shown in the map inset. See the Supplemental Material [96] for a higher-resolution version of this figure.

The film-actor network.—This network is compiled by extracting information available in the Internet Movie Database (IMDB), which contains each cast member and film as distinct nodes, and an undirected edge exists between a film and each of its cast members. If nodes with a single connection are recursively removed, a network of $N = 372\,447$ and $E = 1\,812\,312$ remains (as of late 2012). As can be seen in Fig. 7, the nested block model fully captures the bipartite nature of the network and separates movies and actors at the topmost hierarchical level, and proceeds to separate them in geographical, temporal, and topical (genre) lines. The observed partition is similar to the one obtained via the nonhierarchical model [32], but one finds $B = 971$ blocks, instead of $B = 332$ with the flat version.

Meta-analysis of several empirical networks.—We perform an analysis of several empirical networks shown in Fig. 8, which belong to a wide variety of domains and are distributed across many size scales. We use the nonhierarchical stochastic block model as well as the nested variant. In Figs. 8(a) and 8(b), we show the average block

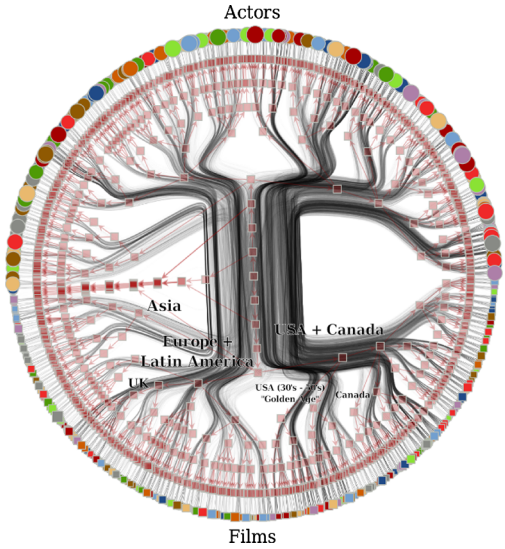
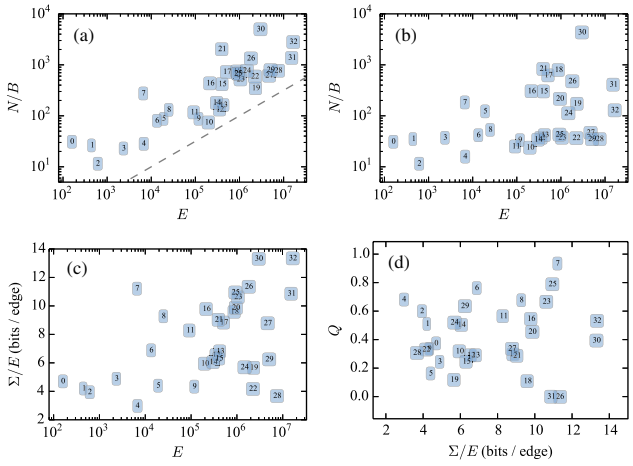


FIG. 7. Large-scale structure of the IMDB film-actor network. Each node in this graph represents a lowest-level block in the hierarchy, instead of individual nodes in the graph. The size of the nodes indicates the number of nodes in each group. The hierarchy branch at the top are the actors, and at the bottom are the films. The labels classify each branch according to the most prominent geographical and temporal characteristics found in the database. See the Supplemental Material [96] for a higher-resolution version of this figure.

sizes N/B for all networks using both models. For the nonhierarchical version, a clear $N/B \sim \sqrt{E}$ trend is observed, which corresponds to the resolution limit present with this method and other approaches as well. In Fig. 8(b), we show the results for the nested model, where such a trend can no longer be observed, and the smallest average block sizes no longer seem to depend on the size of the network, which serves as an empirical demonstration of the lack of resolution limit shown previously. The values of the description lengths themselves are also distributed in a seemingly nonorganized manner [see Fig. 8(c)], i.e., no general tendency for larger networks can be observed, other than an increased range of possible values for larger E values. Any difference observed seems to be due to the actual topological organization, rather than intrinsic constraints imposed by the method. We also compute the modularity of the inferred block structures, $Q = \sum_r e_{rr}/2E - e_r^2/(2E)^2$, which measures how assortative the topology is. Higher values of Q close to 1 indicate the existence of densely connected communities. The value of Q is the most common quantity used to detect blocks in networks, and it presumes that such assortative connections are present. In contrast, by fitting a general stochastic block model, no specific pattern is assumed, and the partition found corresponds to the least random model that matches the data. In Fig. 8, we show the values of Q obtained for the analyzed networks. Indeed, some networks are modular, with high values of Q. However, one does not observe any strong correlation of the description length and the modularity values. Hence, the most structured networks do not necessarily possess much larger Q values, which indicate that the building blocks of their topological organization are not predominantly assortative communities (this is clear in some of the examples considered previously, such as the Internet AS topology and the IMDB network). However, for many of these networks, it is probably possible to find partitions that lead to much higher Q values. These partitions would, on the other hand, correspond to block model ensembles with a larger entropy than those inferred via maximum likelihood. Therefore, the maximization of Q in these cases would invariably discard topological information present in the network and provide a much simplified and possibly misleading picture of the large-scale structure of the network. Hence, it seems more appropriate to confine modularity maximization only to cases where the assortative structure is known to be the dominating pattern. However, even in these cases, methods based on statistical inference possess clear advantages, such as the lack of resolution limit, model selection guarantees, and the overall more principled nature of the approach.



No.	N	E	Dir.	No.	N	E	Dir.	No.	N	E	Dir.
0	62	159	No	11	21 363	91 286	No	22	255 265	2 234 572	Yes
1	105	441	No	12	27 400	352 504	Yes	23	317 080	1 049 866	No
2	115	613	No	13	34 401	421 441	Yes	24	325 729	1 469 679	Yes
3	297	2345	Yes	14	39 796	301 498	Yes	25	334 863	925 872	No
4	903	6760	No	15	52 104	399 625	Yes	26	372 547	1 812 312	No
5	1222	19 021	Yes	16	56 739	212 945	No	27	449 087	4 690 321	Yes
6	4158	13 422	No	17	75 877	508 836	Yes	28	654 782	7 499 425	Yes
7	4941	6594	No	18	82 168	870 161	Yes	29	855 802	5 066 842	Yes
8	8638	24 806	No	19	105 628	2 299 623	No	30	1 134 890	2 987 624	No
9	11 204	117 619	No	20	196 591	950 327	No	31	1 637 868	15 205 016	No
10	17 903	196 972	No	21	224 832	394 400	Yes	32	3 764 117	16 511 740	Yes

No.	Network	No.	Network	No.	Network
0	Dolphins [69]	11	arXiv Co-Authors (cond-mat) [70]	22	Web graph of `stanford.edu`. [71]
1	Political Books[a]	12	arXiv Citations (hep-th) [70, 72]	23	DBLP collaboration [73]
2	American Football [3, 74]	13	arXiv Citations (hep-ph) [70, 72]	24	WWW [6]
3	C. Elegans Neurons [75]	14	PGP [76]	25	Amazon product network [73]
4	Disease Genes [77]	15	Internet AS (Caida)[b]	26	IMDB film-actor[c] [32] (bipartite)
5	Political Blogs [67]	16	Brightkite social network [78]	27	APS citations[d]
6	arXiv Co-Authors (gr-qc) [70]	17	Epinions.com trust network [79]	28	Berkeley/Stanford web graph [71]
7	Power Grid [75]	18	Slashdot [80]	29	Google web graph [71]
8	arXiv Co-Authors (hep-th) [70]	19	Flickr [81]	30	Youtube social network [73]
9	arXiv Co-Authors (hep-ph) [70]	20	Gowalla social network [78]	31	Yahoo groups[e] (bipartite)
10	arXiv Co-Authors (astro-ph) [70]	21	EU email [70]	32	US patent citations [82]

[a] V. Krebs, retrieved from `http://www-personal.umich.edu/~mejn/netdata/`
[b] Retrieved from `http://www.caida.org`.
[c] Retrieved from `http://www.imdb.com/interfaces`.
[d] Retrieved from `http://publish.aps.org/dataset`.
[e] Retrieved from `http://webscope.sandbox.yahoo.com`.

FIG. 8. (a) The average block size N/B obtained using the nonhierarchical model, as a function of E, for the empirical networks listed in the bottom table (the Dir. column specifies whether a given network is directed). The dashed line shows a $\sqrt{E}$ slope. (b) The same as (a) but with the nested model. (c) The description length Σ/E for the nested model as a function of E. (d) The value of modularity Q as function of Σ/E, for the nested model.

VI. DISCUSSION

In this paper, we present a principled method to detect hierarchical structures in networks via a nested stochastic block model. This method fully generalizes previous approaches for the detection of hierarchical community structures [43–49], since it makes no assumptions either on the actual types of large-scale structures possible (assortative, dissortative, or any arbitrary mixture) or on the hierarchical form, which is not confined to binary trees or dendograms. We show that a major advantage of this approach is that it breaks the so-called resolution limit of approaches, such as modularity optimization and nonhierarchical model inference, where modules smaller than a characteristic size scaling with $\sqrt{N}$ cannot be resolved. With the nested model presented, this characteristic scale is replaced by a much smaller logarithmic dependence, making it, in practice, nonexistent for many applications. This increased resolution comes as a result of robust model selection principles, and is integrated with the desirable capacity of differentiating between noise and actual

structure, and, therefore, it is not susceptible to the detection of spurious communities. We show that the model is capable of inferring the large-scale features of empirical networks in significant detail, even for very large networks.

This type of approach should, in principle, also be applicable to other model classes, such as those based on overlapping [9,83–85] or link communities [25,86]. We also predict that it should serve as a more refined method of detecting missing information in networks [23,44], as well as for the prediction of the network evolution [87], determining the more salient topological features [88,89] or large-scale functional summaries of the network topology [90].

ACKNOWLEDGMENTS

The author would like to thank Sebastian Krause for a careful reading of the manuscript, as well as Cris Moore and Lenka Zdeborová for useful comments. This work was funded by the University of Bremen, under the funding line ZF04.

APPENDIX A: BAYESIAN MODEL SELECTION

In the following, we compare Bayesian model selection via integrated likelihood with the MDL approach considered in the main text, and we show that they lead to the same criterion if the model constraints are equivalent.

For the purpose of performing BMS, we evoke the most usual definition of the stochastic block model ensemble, where one defines as parameters the probabilities p_{rs} that an edge exists between two nodes belonging to blocks r and s. The posterior likelihood of observing a given graph with a block partition $\{b_i\}$ and model parameters $\{p_{rs}\}$ is

$$\mathcal{P}(G|\{b_i\},\{p_{rs}\},B) = \prod_{rs} p_{rs}^{e_{rs}/2}(1-p_{rs})^{(n_r n_r - e_{rs})/2}. \tag{A1}$$

The inference procedure consists in, as before, maximizing this quantity with respect to the parameters $\{p_{rs}\}$ and the block partition $\{b_i\}$. It is easy to see that if one maximizes Eq. (A1) with respect to $\{p_{rs}\}$, one recovers $\max_{\{p_{rs}\}} \ln \mathcal{P}(G|\{b_i\},\{p_{rs}\},B) = -\mathcal{S}_t$, given in Eq. (1), so indeed these models are equivalent. However, this does not provide a means for model selection, since models with a larger number of blocks B will invariably posses a larger likelihood. Instead, the Bayesian model selection approach is to consider the joint probability $\mathcal{P}(G,\{b_i\},\{p_{rs}\},\{p_r\}|B)$ of observing not only the graph but also the partition $\{b_i\}$, the model parameters $\{p_{rs}\}$, as well as the parameters $\{p_r\}$ that control the probability of each partition $\{b_i\}$ being observed, which is given by

$$\mathcal{P}(\{b_i\}|\{p_r\},B) = \prod_r p_r^{n_r}. \tag{A2}$$

This invariably leads to the inclusion of prior probabilities of observing the model parameters, $\mathcal{P}(\{p_{rs}\}|B)$ and $\mathcal{P}(\{p_r\}|B)$. Now, instead of finding the model parameters that maximize this quantity, we compute the *integrated likelihood* [38,42,55],

$$\mathcal{P}(G,\{b_i\}|B) = \int dp_{rs}dp_r \mathcal{P}(G,\{b_i\},\{p_{rs}\},\{p_r\}|B) \tag{A3}$$

$$= \int dp_{rs}\mathcal{P}(G|\{b_i\},\{p_{rs}\},B)\mathcal{P}(\{p_{rs}\}|B) \times \int dp_r \mathcal{P}(\{b_i\}|\{p_r\})\mathcal{P}(\{p_r\}|B) \tag{A4}$$

$$= \mathcal{P}(G|\{b_i\},B) \times \mathcal{P}(\{b_i\}|B). \tag{A5}$$

By maximizing $\mathcal{P}(G,\{b_i\}|B)$, instead of Eq. (A1), one should avoid overfitting the data, since the larger models with many parameters are dominated by a majority of choices that fit the data very badly, and, hence, have a smaller contribution in the integral of Eq. (A3). Therefore, the maximization of the integrated likelihood also corresponds to an application of Occam's razor, and one should expect it to deliver results compatible with MDL [53]. However, in practice, things are more nuanced, since the value of Eq. (A3) is heavily dependent on the choice of priors $\mathcal{P}(\{p_{rs}\}|B)$ and $\mathcal{P}(\{p_r\}|B)$. For the block partitions themselves, this choice is more straightforward. Since one wants to be agnostic with respect to what block sizes are possible, one should choose a flat prior $\mathcal{P}(\{p_r\}|B) = \text{Dirichlet}(\{p_r\}|\{\alpha_r\})$, with $\alpha_r = 1$, so that all counts are equally likely. The integral of Eq. (A4) is then computed as

$$\ln \mathcal{P}(\{b_i\}|B) = -\ln\left(\!\!\binom{B}{N}\!\!\right) - \ln N! + \sum_r \ln n_r!, \tag{A6}$$

which is identical to the partition description length of Eq. (7), i.e., $\ln \mathcal{P}(\{b_i\}|B) = -\mathcal{L}_t^0$.

For the block probabilities, on the other hand, the situation is more subtle. A common choice is the flat prior $\mathcal{P}(\{p_{rs}\}|B) = 1$ [23,38,40–42]. This choice is agnostic with respect to what block structures are expected, and it is also practical, since the integral can be evaluated exactly [23,42],

$$\ln \mathcal{P}(G|\{b_i\},B) = -\sum_{r>s} \ln \binom{n_r n_s}{e_{rs}} + \ln(n_r n_s + 1) - \sum_r \ln \binom{n_r^2}{e_{rr}/2} + \ln(n_r^2/2 + 1) \tag{A7}$$

$$\simeq -\frac{1}{2}\sum_{rs} n_r n_s H_b\left(\frac{e_{rs}}{n_r n_s}\right) - (B+1)\sum_r \ln n_r, \tag{A8}$$

where the approximation in Eq. (A8) was made assuming $n_r \gg 1$, and $H_b(x)$ is the binary entropy function. However, there is one important issue with this approach. Namely, there is a strong discrepancy between the models generated by the flat prior $\mathcal{P}(\{p_{rs}\}|B) = 1$ and most observed empirical networks. Specifically, typical parameters with $p_{rs} = 1/2$ sampled by this prior will result in *dense* networks with average degree $\langle k\rangle = \sum_{rs} p_{rs} n_r n_s/N = N/2$. However, most large empirical networks tend to be *sparse*, with an average degree that is many orders of magnitude smaller than N. Hence, as N becomes large, most observed networks will lie in a vanishingly small portion of the parameter space produced by this prior. A better choice would constrain the average degree to something closer to what is observed in the data, but at the same time being otherwise noninformative regarding the block structure. A choice such as $\mathcal{P}(\{p_{rs}\}|B) \propto \delta(\sum_{rs} p_{rs} n_r n_s - 2E)$, where E is the number of edges in the observed network, seems appropriate, but the integral in Eq. (A4) becomes difficult to solve. Instead, an easier approach is to modify the model sightly, so that the average degree is implicitly constrained. Here, we consider the model variant where the number of edges E is a fixed parameter, and each sampled edge may land between any two nodes belonging to blocks r and s with probability q_{rs}, and we have, therefore, $\sum_{r\geq s} q_{rs} = 1$. The full posterior likelihood of this model is

$$\mathcal{P}(G|\{b_i\},\{q_{rs}\},E,B) = \frac{E!}{\Omega(\{e_{rs}\},\{n_r\})}\frac{\prod_{r\geq s} q_{rs}^{m_{rs}}}{\prod_{r\geq s} m_{rs}!}, \tag{A9}$$

where $\Omega(\{e_{rs}\},\{n_r\})$ is, as before, the number of different graphs with the same block partition and edge counts, and $m_{rs} = e_{rs}$ if $r \neq s$ or $e_{rr}/2$ otherwise. By maximizing Eq. (A9) with respect to $\{q_{rs}\}$, one obtains $\max_{\{q_{rs}\}} \ln \mathcal{P}(G|\{b_i\},\{q_{rs}\},E,B) \simeq -\ln \Omega(\{e_{rs}\},\{n_r\}) = -\mathcal{S}_t$, as long as $m_{rs} \gg 1$ or $m_{rs} = 0$, so it also is equivalent to the previous models in this limit. With this reparametrization, the average degree remains fixed independently of the choice of prior. Therefore, we may finally use a flat prior $\mathcal{P}(\{q_{rs}\}|B) = \text{Dirichlet}(\{q_{rs}\}|\{\alpha_{rs} = 1\})$, without the risk of the graphs becoming inadvertently dense, and again the integrated likelihood can be computed exactly,

$$\mathcal{P}(G|\{b_i\},B) = \int dq_{rs}\mathcal{P}(G|\{b_i\},\{q_{rs}\},E,B)\mathcal{P}(\{q_{rs}\}|B) \tag{A10}$$

$$= \left[\Omega(\{e_{rs}\},\{n_r\}) \times \left(\!\!\binom{\binom{B}{2}}{E}\!\!\right)\right]^{-1}. \tag{A11}$$

By inserting Eq. (A11) into Eq. (A5) and comparing with Eq. (10), we see that $\ln \mathcal{P}(G,\{b_i\}|B) = -\Sigma_{L=1}$, and we conclude reassuringly that the MDL approach is fully equivalent to BMS when all model constraints are compatible. In fact, even in the dense case, although not quite the same, the (dense) BMS and MDL penalties are very similar. If one assumes $N \gg B^2$, $E \propto N^2$, and equal block sizes $n_r = N/B$, both penalties become $\sim B(B+1)\ln N + N \ln B$. Therefore, it seems that whatever differences arising from the two approaches stem simply from nuances in the choice of prior probabilities. This comparison also allows us to interpret the nested block model as a hierarchical Bayesian approach, where the priors $\mathcal{P}(\{q_{rs}\}|B)$ are replaced by a nested sequence of priors and hyperpriors, so that their integrated likelihood matches the description length defined previously.

APPENDIX B: COMPARISON WITH OTHER COMMUNITY DETECTION METHODS

In this section, we compare results obtained for synthetic networks with popular community detection methods that are not based on statistical inference. Here, we focus not only on the capacity of the method of finding a partition correlated with the planted one, but also on the number of blocks detected. We concentrate on two methods that have been reported to provide good results in synthetic benchmarks [91], namely, the Louvain method [12], based on modularity optimization, and the Infomod method [45,92,93], based on compression of random walks. We make use of the LFR benchmark [94], which corresponds to a specific parametrization of the degree-corrected stochastic block model [24], where both the degree distribution and the block size distribution follow truncated power laws. Here, we employ a parametrization similar to Ref. [91], with a degree distribution following a power law with exponent -2 and a minimum degree $k_{\min} = 5$, and a community size distribution also following a power law, but with exponent -1, and minimum block size of 50. We also impose the following additional restrictions: The total number of blocks is always fixed at $B = 100$, and for every node i belonging to block r, its degree k_i cannot exceed $\sqrt{n_r}$, to avoid intrinsic degree-degree correlations [50]. With this parameter choice, the networks generated with $N = 2 \times 10^4$ possess an average degree $\langle k\rangle \simeq 7.8$. The actual block structure is parametrized as $e_{rs} = (1-c)e_r e_s/2E + \delta_{rs} c e_r$, where c controls the assortativity: For $c = 1$, all edges connect nodes of the same block, and for $c = 0$, we have a fully random configuration model [107].

Because the different methods result in quite different numbers of detected blocks, the normalized mutual information is not the most appropriate measure of the overlap between partitions in this case. This is due to the fact that, if the number of nodes is kept fixed, the NMI values tend to be larger simply if the number of blocks is increased, even if this larger partition is in no other way more strongly correlated to the true one. Another measure that is less susceptible to this problem is the variation of information (VI) [95], defined as

$$\mathrm{VI}(\{x_i\},\{y_i\}) = H(\{x_i\}) + H(\{y_i\}) - 2I(\{x_i\},\{y_i\}), \tag{B1}$$

where $H(\{x_i\})$ is the entropy of the partition $\{x_i\}$ and $I(\{x_i\},\{x_i\})$ is the (non-normalized) mutual information between $\{x_i\}$ and $\{y_i\}$. A value of VI equal to zero means that the partitions are identical, whereas any positive value indicates a reduced overlap between them.

The VI values between the planted partitions and those obtained with different methods for several network realizations of the above model are shown in Fig. 9, together with the obtained number of blocks. By observing the VI values for the inference method with a fixed number of blocks $B = 100$, we conclude that the strict detectability transition (when the value of B is known) lies somewhere slightly above $c \approx 0.2$. However, the model-selection procedure based on the nested stochastic block model presented in the main text discards any structure below the $c \approx 0.4$ range, and decides on a fully random $B = 1$ structure. Above this value, the inferred value of B increases from $B = 1$ until agreeing with the planted value for sufficiently large c values. As can also be seen in Fig. 9, the Louvain method exhibits the "worst of both worlds," i.e., it fails to find the correct partition for all values except $c = 1$, finding systematically smaller values of B, while at the same time finding spurious partitions below the detectability threshold, even when the network is completely random ($c = 0$). The Infomod method, on the other hand, seems to find partitions that are largely compatible with the planted one, at least for the parameter region above $c \approx 0.6$. However, for even larger values of c, this method detects a number of blocks that is significantly larger than the planted value, which increases steadily as c decreases. Hence, this method is also incapable of separating structure from noise, and finds spurious partitions far below the detectability threshold. Thus, from the three methods analyzed, the one described in the main text is the only one that combines the following three desirable properties: (1) optimal inference in the detectable range, (2) guarantee against overfitting and detection of spurious modules, and (3) fully nonparametric implementation.

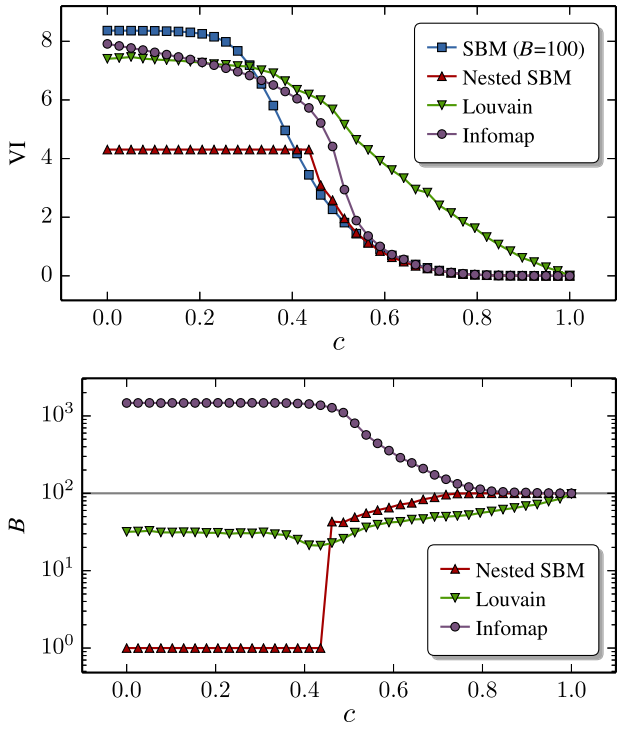


FIG. 9. Top: Variation of information (VI) between the planted and obtained partitions as a function of the assortativity parameter c, for networks with $N = 2 \times 10^4$, generated as described in the text. The legend indicates results obtained with different methods: Fitting the degree-corrected stochastic block model with a fixed number of blocks $B = 100$ (SBM), performing model selection with the nested stochastic block model (Nested SBM), the Louvain modularity maximization method [12], and the Infomod method [45,92,93]. Bottom: The obtained number of blocks B as a function of c, for the same methods as in the top panel. The gray horizontal line marks the planted $B = 100$ value. All results were obtained by averaging over 20 network realizations.

The suboptimal behavior of the modularity-based method is simply a combination of the resolution limit [14] and lack of built-in model selection based on statistical evidence [13]. It is not currently known if the Infomod method suffers from problems similar to the resolution limit, but clearly it lacks guarantees against detection of spurious modules. Although it is also based on the principle of parsimony, it tries to compress random walks taking place on the network, instead of the network itself. Apparently, the method cannot distinguish between the actual planted block structure and quenched topological fluctuations—both of which will affect random walks—and gradually transitions between the two properties in order to best describe the network dynamics. (As has been shown in Ref. [91], this problem diminishes if the average degree of the network is made sufficiently large, in which case the method finally settles in a $B = 1$ partition for fully random graphs.) On the other hand, the method in the main text is based on maximizing the likelihood of the *exact same* generative process that was used to construct the network, which puts it in clear advantage over the other two (and, in fact, many other methods, including all those analyzed in Refs. [91,94]), in addition to including a robust and formally motivated model-selection procedure.

APPENDIX C: DIRECTED AND UNDIRECTED NETWORKS

As mentioned in the main text, the model described is easily generalized for directed graphs. For the ensemble entropies, we have for the undirected case [50]

$$\mathcal{S}_t = \frac{1}{2}\sum_{rs} n_r n_s H_b\left(\frac{e_{rs}}{n_r n_s}\right), \tag{C1}$$

while for the directed case it reads

$$\mathcal{S}_t^d = \sum_{rs} n_r n_s H_b\left(\frac{e_{rs}}{n_r n_s}\right), \tag{C2}$$

where $H_b(x) = -x\ln x - (1-x)\ln(1-x)$ is the binary entropy function. In both cases, e_{rs} is the number of edges from block r to s (or the number of half-edges for the undirected case when $r = s$), and n_r is the number of nodes in block r. In the sparse limit, $e_{rs} \ll n_r n_s$, these expressions may be written approximately as

$$\mathcal{S}_t \cong E - \frac{1}{2}\sum_{rs} e_{rs}\ln\left(\frac{e_{rs}}{n_r n_s}\right), \tag{C3}$$

$$\mathcal{S}_t^d \cong E - \sum_{rs} e_{rs}\ln\left(\frac{e_{rs}}{n_r n_s}\right). \tag{C4}$$

For the degree-corrected variant with "hard" degree constraints, we have

$$\mathcal{S}_c \cong -E - \sum_k N_k \ln k! - \frac{1}{2}\sum_{rs} e_{rs}\ln\left(\frac{e_{rs}}{e_r e_s}\right), \tag{C5}$$

$$\mathcal{S}_c^d \cong -E - \sum_{k^+} N_{k^+}\ln k^+! - \sum_{k^-} N_{k^-}\ln k^-! - \sum_{rs} e_{rs}\ln\left(\frac{e_{rs}}{e_r^+ e_s^-}\right), \tag{C6}$$

where $e_r = \sum_s e_{rs}$ is the number of half-edges incident on block r, and $e_r^+ = \sum_s e_{rs}$ and $e_r^- = \sum_s e_{sr}$ are the number of out and in edges adjacent to block r, respectively. These expressions are also only valid in the sparse limit, which in this case amounts to the following conditions,

$$e_{rs}\frac{\langle k^2\rangle_r - \langle k\rangle_r}{\langle k\rangle_r^2}\frac{\langle k^2\rangle_s - \langle k\rangle_s}{\langle k\rangle_s^2} \ll n_r n_s, \tag{C7}$$

where $\langle k^l\rangle_r = \sum_{i\in r} k_i^l/n_r$ [for the directed case, we simply replace $\langle k^l\rangle_r \to \langle (k^+)^l\rangle_r$ and $\langle k^l\rangle_s \to \langle (k^-)^l\rangle_s$ in the equation above]. Unfortunately, there is no closed-form expression for the entropy outside the sparse limit, unlike the traditional variant [50].

For the upper-level multigraphs, the entropies are [50]

$$\mathcal{S}_m = \sum_{r>s}\ln\left(\!\!\binom{n_r n_s}{e_{rs}}\!\!\right) + \sum_r \ln\left(\!\!\binom{\binom{n_r}{2}}{e_{rr}/2}\!\!\right), \tag{C8}$$

$$\mathcal{S}_m^d = \sum_{rs}\ln\left(\!\!\binom{n_r n_s}{e_{rs}}\!\!\right), \tag{C9}$$

where, as before, $\left(\!\binom{n}{m}\!\right) = \binom{n+m-1}{m}$ is the number of m-combinations with repetitions from a set of size n.

For the degree-corrected model, the description length needs to be augmented with the information necessary to describe the degree sequence, analogously to Eq. (9) for the undirected case,

$$\mathcal{L}_c = \mathcal{L}_t + \sum_r n_r H(\{p_{k^-,k^+}^r\}), \tag{C10}$$

where $\{p_{k^-,k^+}^r\}$ is the joint (in, out)-degree distribution of nodes belonging to block r.

Note that other generalizations for the directed case are possible [27], and it should be straightforward to adapt the nested model for them as well.

[1] M. E. J. Newman, *Communities, Modules, and Large-Scale Structure in Networks*, Nat. Phys. **8**, 25 (2011).

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

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

[4] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabási, *Hierarchical Organization of Modularity in Metabolic Networks*, Science **297**, 1551 (2002).

[5] R. J. Fletcher, Jr, A. Revell, B. E. Reichert, W. M. Kitchens, J. D. Dixon, and J. D. Austin, *Network Modularity Reveals Critical Scales for Connectivity in Ecology and Evolution*, Nat. Commun. **4**, 2572 (2013).

[6] R. Albert, H. Jeong, and A.-L. Barabási, *Internet: Diameter of the World Wide Web*, Nature (London) **401**, 130 (1999).

[7] S.-H. Yook, H. Jeong, and A.-L. Barabási, *Modeling the Internet's Large-Scale Topology*, Proc. Natl. Acad. Sci. U.S.A. **99**, 13 382 (2002).

[8] Y. Zhao, E. Levina, and J. Zhu, *Community Extraction for Social Networks*, Proc. Natl. Acad. Sci. U.S.A. **108**, 7321 (2011).

[9] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, *Uncovering the Overlapping Community Structure of Complex Networks in Nature and Society*, Nature (London) **435**, 814 (2005).

[10] M. E. J. Newman and M. Girvan, *Finding and Evaluating Community Structure in Networks*, Phys. Rev. E **69**, 026113 (2004).

[11] A. Clauset, M. E. J. Newman, and C. Moore, *Finding Community Structure in Very Large Networks*, Phys. Rev. E **70**, 066111 (2004).

[12] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, *Fast Unfolding of Communities in Large Networks*, J. Stat. Mech. (2008) P10008.

[13] R. Guimerà, M. Sales-Pardo, and L. A. N. Amaral, *Modularity from Fluctuations in Random Graphs and Complex Networks*, Phys. Rev. E **70**, 025101 (2004).

[14] S. Fortunato and M. Barthélemy, *Resolution Limit in Community Detection*, Proc. Natl. Acad. Sci. U.S.A. **104**, 36 (2007).
[15] A. Lancichinetti and S. Fortunato, *Limits of Modularity Maximization in Community Detection*, Phys. Rev. E **84**, 066122 (2011).
[16] B. H. Good, Y.-A. de Montjoye, and A. Clauset, *Performance of Modularity Maximization in Practical Contexts*, Phys. Rev. E **81**, 046106 (2010).
[17] M. B. Hastings, *Community Detection as an Inference Problem*, Phys. Rev. E **74**, 035102 (2006).
[18] D. Garlaschelli and M. I. Loffredo, *Maximum Likelihood: Extracting Unbiased Information from Complex Networks*, Phys. Rev. E **78**, 015101 (2008).
[19] M. E. J. Newman and E. A. Leicht, *Mixture Models and Exploratory Analysis in Networks*, Proc. Natl. Acad. Sci. U.S.A. **104**, 9564 (2007).
[20] J. Reichardt and D. R. White, *Role Models for Complex Networks*, Eur. Phys. J. B **60**, 217 (2007).
[21] J. M. Hofman and C. H. Wiggins, *Bayesian Approach to Network Modularity*, Phys. Rev. Lett. **100**, 258701 (2008).
[22] P. J. Bickel and A. Chen, *A Nonparametric View of Network Models and Newman-Girvan and Other Modularities*, Proc. Natl. Acad. Sci. U.S.A. **106**, 21068 (2009).
[23] R. Guimerà and M. Sales-Pardo, *Missing and Spurious Interactions and the Reconstruction of Complex Networks*, Proc. Natl. Acad. Sci. U.S.A. **106**, 22 073 (2009).
[24] B. Karrer and M. E. J. Newman, *Stochastic Block Models and Community Structure in Networks*, Phys. Rev. E **83**, 016107 (2011).
[25] B. Ball, B. Karrer, and M. E. J. Newman, *Efficient and Principled Method for Detecting Communities in Networks*, Phys. Rev. E **84**, 036103 (2011).
[26] J. Reichardt, R. Alamino, and D. Saad, *The Interplay between Microscopic and Mesoscopic Structures in Complex Networks*, PLoS One **6**, e21282 (2011).
[27] Y. Zhu, X. Yan, and C. Moore, *Oriented and Degree-Generated Block Models: Generating and Inferring Communities with Inhomogeneous Degree Distributions*, J. Complex Netw. **2**, 1 (2014).
[28] E. B. Baskerville, A. P. Dobson, T. Bedford, S. Allesina, T. Michael Anderson, and M. Pascual, *Spatial Guilds in the Serengeti Food Web Revealed by a Bayesian Group Model*, PLoS Comput. Biol. **7**, e1002321 (2011).
[29] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová, *Inference and Phase Transitions in the Detection of Modules in Sparse Networks*, Phys. Rev. Lett. **107**, 065701 (2011).
[30] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová, *Asymptotic Analysis of the Stochastic Block Model for Modular Networks and Its Algorithmic Applications*, Phys. Rev. E **84**, 066106 (2011).
[31] E. Mossel, J. Neeman, and A. Sly, *Stochastic Block Models and Reconstruction*, arXiv:1202.1499.
[32] T. P. Peixoto, *Parsimonious Module Inference in Large Networks*, Phys. Rev. Lett. **110**, 148701 (2013).
[33] P. W. Holland, K. B. Laskey, and S. Leinhardt, *Stochastic Block Models: First Steps*, Soc. Networks **5**, 109 (1983).
[34] S. E. Fienberg, M. M. Meyer, and S. S. Wasserman, *Statistical Analysis of Multiple Sociometric Relations*, J. Am. Stat. Assoc. **80**, 51 (1985).
[35] K. Faust and S. Wasserman, *Block Models: Interpretation and Evaluation*, Soc. Networks **14**, 5 (1992).
[36] C. J. Anderson, S. Wasserman, and K. Faust, *Building Stochastic Block Models*, Soc. Networks **14**, 137 (1992).
[37] M. Rosvall and C. T. Bergstrom, *An Information-Theoretic Framework for Resolving Community Structure in Complex Networks*, Proc. Natl. Acad. Sci. U.S.A. **104**, 7327 (2007).
[38] J.-J. Daudin, F. Picard, and S. Robin, *A Mixture Model for Rndom Graphs*, Stat. Comput. **18**, 173 (2008).
[39] M. Mariadassou, S. Robin, and C. Vacher, *Uncovering Latent Structure in Valued Graphs: A Variational Approach*, Ann. Appl. Stat. **4**, 715 (2010).
[40] C. Moore, X. Yan, Y. Zhu, J.-B. Rouquier, and T. Lane, *Active Learning for Node Classification in Assortative and Disassortative Networks*, in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '11)* (ACM, New York, 2011), pp. 841–849.
[41] P. Latouche, E. Birmele, and C. Ambroise, *Variational Bayesian Inference and Complexity Control for Stochastic Block Models*, Stat. Model. **12**, 93 (2012).
[42] E. Côme and P. Latouche, *Model Selection and Clustering in Stochastic Bock Models with the Exact Integrated Complete Data Likelihood*, arXiv:1303.2962.
[43] A. Clauset, C. Moore, and M. E. J. Newman, *Statistical Network Analysis: Models, Issues, and New Directions*, Lecture Notes in Computer Science Vol. 4503, edited by E. Airoldi, D. M. Blei, S. E. Fienberg, A. Goldenberg, E. P. Xing, and A. X. Zheng (Springer, Berlin, 2007), pp. 1–13.
[44] A. Clauset, C. Moore, and M. E. J. Newman, *Hierarchical Structure and the Prediction of Missing Links in Networks*, Nature (London) **453**, 98 (2008).
[45] M. Rosvall and C. T. Bergstrom, *Multilevel Compression of Random Walks on Networks Reveals Hierarchical Organization in Large Integrated Systems*, PLoS One **6**, e18209 (2011).
[46] M. Sales-Pardo, R. Guimerà, A. A. Moreira, and L. A. N. Amaral, *Extracting the Hierarchical Organization of Complex Systems*, Proc. Natl. Acad. Sci. U.S.A. **104**, 15 224 (2007).
[47] P. Ronhovde and Z. Nussinov, *Multiresolution Community Detection for Megascale Networks by Information-Based Replica Correlations*, Phys. Rev. E **80**, 016109 (2009).
[48] I. A. Kovács, R. Palotai, M. S. Szalay, and P. Csermely, *Community Landscapes: An Integrative Approach to Determine Overlapping Network Module Hierarchy, Identify Key Nodes, and Predict Network Dynamics*, PLoS One **5**, e12528 (2010).
[49] Y. Park, C. Moore, and J. S. Bader, *Dynamic Networks from Hierarchical Bayesian Graph Clustering*, PLoS One **5**, e8118 (2010).
[50] T. P. Peixoto, *Entropy of Stochastic Block Model Ensembles*, Phys. Rev. E **85**, 056122 (2012).
[51] J. Leskovec, D. Chakrabarti, J. Kleinberg, C. Faloutsos, and Z. Ghahramani, *Kronecker Graphs: An Approach to*

Modeling Networks, J. Mach. Learn. Res. **11**, 9851042 (2010).
[52] G. Bianconi, *Entropy of Network Ensembles*, Phys. Rev. E **79**, 036114 (2009).
[53] P. D. Gràwald, *The Minimum Description Length Principle* (MIT Press, Cambridge, MA, 2007).
[54] J. Rissanen, *Information and Complexity in Statistical Modeling*, (Springer, New York, 2010), 1st ed.
[55] C. Biernacki, G. Celeux, and G. Govaert, *Assessing a Mixture Model for Clustering with the Integrated Completed Likelihood*, IEEE Trans. Pattern Anal. Mach. Intell. **22**, 719 (2000).
[56] X. Yan, J. E. Jensen, F. Krzakala, C. Moore, C. R. Shalizi, L. Zdeborova, P. Zhang, and Y. Zhu, *Model Selection for Degree-Corrected Block Models*, arXiv:1207.3994.
[57] G. Schwarz, *Estimating the Dimension of a Model*, Ann. Stat. **6**, 461 (1978).
[58] H. Akaike, *A New Look at the Statistical Model Identification*, IEEE Trans. Autom. Control **19**, 716 (1974).
[59] J. Reichardt and M. Leone, *(Un)detectable Cluster Structure in Sparse Networks*, Phys. Rev. Lett. **101**, 078701 (2008).
[60] D. Hu, P. Ronhovde, and Z. Nussinov, *Phase Transitions in Random Potts Systems and the Community Detection Problem: Spin-Glass Type and Dynamic Perspectives*, Philos. Mag. **92**, 406 (2012).
[61] A. Condon and R. M. Karp, *Algorithms for Graph Partitioning on the Planted Partition Model*, Random Struct. Algorithms **18**, 116 (2001).
[62] J. Xiang, X. G. Hu, X. Y. Zhang, J. F. Fan, X. L. Zeng, G. Y. Fu, K. Deng, and K. Hu, *Multiresolution Modularity Methods and Their Limitations in Community Detection*, Eur. Phys. J. B **85**, 1 (2012).
[63] T. P. Peixoto, *Efficient Monte Carlo and Greedy Heuristic for the Inference of Stochastic Block Models*, Phys. Rev. E **89**, 012804 (2014).
[64] G. Palla, L. Lovász, and T. Vicsek, *Multifractal Network Generator*, Proc. Natl. Acad. Sci. U.S.A. **107**, 7640 (2010).
[65] T. P. Peixoto, *Eigenvalue Spectra of Modular Networks*, Phys. Rev. Lett. **111**, 098701 (2013).
[66] R. R. Nadakuditi and M. E. J. Newman, *Graph Spectra and the Detectability of Community Structure in Networks*, Phys. Rev. Lett. **108**, 188701 (2012).
[67] L. A. Adamic and N. Glance, *The Political Blogosphere and the 2004 U.S. Election: Divided They Blog*, in *Proceedings of the 3rd International Workshop on Link Discovery (LinkKDD '05)* (ACM, New York, 2005), pp. 36–43.
[68] D. Holten, *Hierarchical Edge Bundles: Visualization of Adjacency Relations in Hierarchical Data*, IEEE Trans. Visual. Comput. Graph. **12**, 741 (2006).
[69] D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson, *The Bottlenose Dolphin Community of Doubtful Sound Features a Large Proportion of Long-Lasting Associations*, Behav. Ecol. Sociobiol., **54**, 396 (2003).
[70] J. Leskovec, J. Kleinberg, and C. Faloutsos, *Graph Evolution: Densification and Shrinking Diameters*, ACM Trans. Knowl. Discov. Data **1**, 2 (2007).
[71] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney, *Community Structure in Large Networks: Natural Cluster Sizes and the Absence of Large Well-Defined Clusters*, arXiv:0810.1355.
[72] J. Gehrke, P. Ginsparg, and J. Kleinberg, *Overview of the 2003 KDD Cup*, SIGKDD Explor. Newsletter **5**, 149 (2003).
[73] J. Yang and J. Leskovec, *Defining and Evaluating Network Communities Based on Ground Truth*, in *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics (MDS '12)* (ACM, New York, 2012) p. 3:1–3:8.
[74] T. S. Evans, American College Football Network Files, (FigShare), http://figshare.com/articles/American_College_Football_Network_Files/93179.
[75] D. J. Watts and S. H. Strogatz, *Collective dynamics of "Small-World" Networks*, Nature (London) **393**, 440 (1998).
[76] O. Richters and T. P. Peixoto, *Trust Transitivity in Social Networks*, PLoS One **6**, e18384 (2011).
[77] K. I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A. L. Barabási, *The Human Disease Network*, Proc. Natl. Acad. Sci. U.S.A. **104**, 8685 (2007).
[78] E. Cho, S. A. Myers, and J. Leskovec, *Friendship and Mobility: User Movement in Location-Based Social Networks*, in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, California, USA, 2011 (KDD '11)* (ACM, New York, NY, 2011), (Ref. [40]), pp. 1082–1090.
[79] M. Richardson, R. Agrawal, and P. Domingos, *The Semantic Web–ISWC 2003*, in Lecture Notes in Computer Science Vol. 2870, edited by D. Fensel, K. Sycara, and J. Mylopoulos (Springer, Berlin, 2003), pp. 351–368.
[80] J. Leskovec, D. Huttenlocher, and J. Kleinberg, *Signed Networks in Social Media*, in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)* (ACM, New York, 2010) pp. 1361–1370.
[81] J. McAuley and J. Leskovec, *Computer Vision ECCV 2012*, in Lecture Notes in Computer Science Vol. 7575, edited by A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid (Springer, Berlin, 2012), pp. 828–841.
[82] J. Leskovec, J. Kleinberg, and C. Faloutsos, *Graphs Over Time: Densification Laws, Shrinking diameters, and Possible Explanations*, in *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, Chicago, Illinois, USA, 2005 (KDD '05)* (ACM, New York, NY, 2005), (Ref. [67]), pp. 177–187.
[83] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing, *Mixed Membership stochastic Block Models*, J. Mach. Learn. Res. **9**, 1981 (2008).
[84] A. Lancichinetti, S. Fortunato, and J. Kertész, *Detecting the Overlapping and Hierarchical Community Structure in Complex Networks*, New J. Phys. **11**, 033015 (2009).
[85] P. K. Gopalan and D. M. Blei, *Efficient discovery of Overlapping Communities in Massive Networks*, Proc. Natl. Acad. Sci. U.S.A. **110**, 14 534 (2013).
[86] Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann, *Link Communities Reveal Multiscale Complexity in Networks*, Nature (London) **466**, 761 (2010).

[87] D. Liben-Nowell and J. Kleinberg, *The Link-Prediction Problem for Social Networks*, J. Am. Soc. Inf. Sci. Technol. **58**, 1019 (2007).

[88] D. Grady, C. Thiemann, and D. Brockmann, *Robust Classification of Salient Links in Complex Networks*, Nat. Commun. **3**, 864 (2012).

[89] G. Bianconi, P. Pin, and M. Marsili, *Assessing the Relevance of Node Features for Network Structure*, Proc. Natl. Acad. Sci. U.S.A. **106**, 11 433 (2009).

[90] R. Guimerà and L. A. N. Amaral, *Functional Cartography of Complex Metabolic Networks*, Nature (London) **433**, 895 (2005).

[91] A. Lancichinetti and S. Fortunato, *Community Detection Algorithms: A Comparative Analysis*, Phys. Rev. E **80**, 056117 (2009).

[92] M. Rosvall and C. T. Bergstrom, *Maps of Random Walks on Complex Networks Reveal Community Structure*, Proc. Natl. Acad. Sci. U.S.A. **105**, 1118 (2008).

[93] M. Rosvall, D. Axelsson, and C. T. Bergstrom, *The Map Equation*, Eur. Phys. J. Spec. Top. **178**, 13 (2009).

[94] A. Lancichinetti, S. Fortunato, and F. Radicchi, *Benchmark Graphs for Testing Community Detection Algorithms*, Phys. Rev. E **78**, 046110 (2008).

[95] M. Meilă, *Learning Theory and Kernel Machines*, in Lecture Notes in Computer Science Vol. 2777, edited by B. Schölkopf and M. K. Warmuth (Springer, Berlin, 2003), pp. 173–187.

[96] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevX.4.011047 for high resolution versions of Figs. 6 and 7, and additional information on the internet data.

[97] This specification generalizes other hierarchical constructions in a straightforward manner. For instance, the generative model of Refs. [43,44] can be recovered as a special case by forcing a binary tree hierarchy, terminating at the individual nodes, and a strictly assortative modular structure. A similar argument holds for the variant of Ref. [51] as well.

[98] In Ref. [32] the degree sequence entropy was taken to be $NH(\{p_k\})$, with $p_k = \sum_r n_r p_k^r/N$, which implicitly assumed that the degrees are uncorrelated with the block partitions, and, hence, should be interpreted only as an upper bound to the actual description length given by Eq. (9).

[99] Note that MDL can still be used to select the simpler model in this case: Although the complete description length Σ will be asymptotically the same with both models for networks sampled from the traditional block model, we still have that $\mathcal{L}_t < \mathcal{L}_c$, since the degree-corrected version still needs to include the information on the degree sequence, as in Eq. (9).

[100] The normalized mutual information (NMI) is defined as $2I(\{b_i\},\{c_i\})/[H(\{b_i\}) + H(\{c_i\})]$, where $I(\{b_i\},\{c_i\}) = \sum_{rs} p_{bc}(r,s) \ln [p_{bc}(r,s)/p_b(r)p_c(r)]$ and $H(\{x_i\}) = -\sum_r p_x(r) \ln p_x(r)$, where $\{b_i\}$ and $\{c_i\}$ are two partitions of the network.

[101] The fact that the NMI between the true and inferred partitions remains slightly above zero in Fig. 2 for $\langle k\rangle < 1$ with the incomplete BMS criterion is a finite size effect, as it tends increasingly to zero as $N \to \infty$. On the other hand, according to this criterion, the inferred value of B in this region increases as N becomes larger.

[102] This threshold corresponds simply to the point where it becomes impossible to fully encode the block partition in the network structure, i.e., for uniform blocks $-E \ln B + N \ln B = 0$, which leads to $E = N$ and, hence, $\langle k\rangle = 2$.

[103] This limit cannot be significantly changed even if one introduces scale parameters to the definition of modularity [15,62].

[104] In the model selection context, adding a single edge between the blocks is not a necessary condition for the observation of the resolution limit, and has a negligible effect, differently from the modularity approach, where it is a deciding factor.

[105] An efficient and fully documented C++ implementation of the algorithm described here is freely available as part of the graph-tool Python library at http://graph-tool.skewed.de.

[106] IPv4 Routed /24 AS Links Dataset, http://www.caida.org/data/active/ipv4-routed-topology-aslinks-dataset.xml.

[107] Note that this is slightly different than in Ref. [94], which parametrized the fraction of internal and external degrees via a local mixing parameter μ, which is the same for all communities. That choice corresponds to a different parametrization of the degree-corrected block model than the one used here. However, since the blocks have different sizes, and the degrees are approximately the same in all blocks, in general, there is no choice of μ that would allow one to recover the fully random configuration model, since the intrinsic mixing would be different for each block in this case. Because of this, we have opted for the parametrization used here; however, this should not alter the interpretation of the benchmark and the comparison with Ref. [94] in a significant way.