

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

Discovering the large-scale topological features in empirical networks is a crucial tool in understanding how complex systems function. However most existing methods used to obtain the modular structure of networks suffer from serious problems, such as the 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, but also for more principled ones based on statistical inference and model selection. Here we construct a nested generative model which, 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 by 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 which scales well for very large networks.

I. INTRODUCTION

The detection of communities and other large-scale structure in networks has become perhaps one of the largest undertakings in the science of networks [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 which 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 which has perhaps gathered the most widespread use is called modularity optimization [10] and consists in optimizing a quality function which favors partitions of nodes where 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 it fails to detect clusters with size below a given threshold [13, 14], 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. Additionally, results obtained with this method tend to be degenerate for large empirical networks [15], for which many different partitions can be found with modularity values very close to the global maximum. In these situations the method fails in giving a faithful representation of the actual large-scale structure present in the system. The method is also incapable of separating actual structure from those arising due to random fluctuations of the null model, and it even finds high scoring partitions in fully random graphs [16]. 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. This approach offers many advantages over many existing methods, including modularity, since it is more firmly grounded on well known principles and methods of statistical analysis. 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 [17–20]. Perhaps the most used generative model used for this purpose is the stochastic block model [21–36], which groups nodes in blocks with arbitrary probabilities of connections between them. This very simple definition already does away with the restriction of only considering purely assortative communities, and accommodates many different patterns, such as core-periphery structures and bipartite blocks, as well as a straightforward generalizations to directed graphs. The issue of 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) [20, 37] or Bayesian model selection (BMS) [38–42]. These approaches allow the selection of the most appropriate number of blocks, and avoid the detection of spurious communities. However, as it turns out, at least one of the limitations of modularity maximization is also present when

* tiago@itp.uni-bremen.de

doing model selection, namely the resolution limit mentioned above. As has been recently shown in Ref. [20], 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 define 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 non-hierarchical model to by much a smaller value which scales only logarithmically, enabling the detection of much smaller blocks in very large networks. Furthermore, the model 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 [43–48] 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 which fits the data, and is not subject to overfitting, and hence will not detect spurious modules in random networks. Finally, the method is fully non-parametric, and can be implemented efficiently, with a simple algorithm which scales well for very large networks.

We start in the next section 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. THE HIERARCHICAL MODEL

The original stochastic block model ensemble [21–24] 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 may differentiate between two very similar model variants: 1. The quantities e_{rs} are themselves the parameters of the model; 2. The parameters are the probabilities p_{rs} that an edge exists between two nodes of the respective blocks, such that the quantities $\langle e_{rs} \rangle = n_r n_s p_{rs}$ hold on average. Both are equally valid generative models, and as long as these quantities are sufficiently large, they should be equivalent (see [49] and below). Here we stick with the first variant, since it makes the following formu-

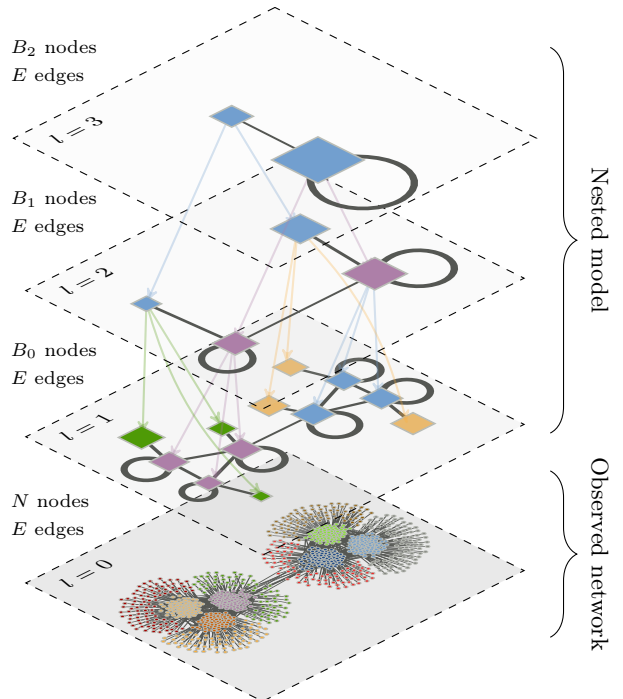


FIG. 1. Example of a nested stochastic block model with three levels, and a generated network at the bottom. The top-level structure describes a core-periphery network, which is further subdivided in the lower levels.

lation more convenient. We also consider a further variation called the degree-corrected block model [32], 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 [49]). The degree-corrected version, although being a relatively simple modification, yields much more convincing results on many empirical networks, since it is capable of incorporating degree variability inside each block.

The nested version which we define here is based on the simple fact that the edge counts e_{rs} themselves form a block graph, 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 chose 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).

Note that in order to describe the observed network which is generated by the deepest level of the hierarchy (e.g. $l = 1$ in Fig. 1), it is not necessary to involve information on the upper layers. Hence, in order to fit the nested block model to observed data, one may simply fit the flat, non-hierarchical model first, and obtain the

remaining hierarchy at later steps. Indeed this a simple way to proceed if one knows how many levels the hierarchy should have, as well as the number of blocks in each level. However, the full advantage of the nested approach lies precisely in the case when this type of information is not known, and all one has is the observed network. In this case — which is a much more likely scenario in practice — the nested model provides a way of detecting regularities in multiple scales. Although one is always operating with a flat model at the deepest layer, the knowledge of the upper layers allows one to infer deeper, and detect more detailed block structures, which would otherwise be discarded. This occurs by the mechanism of statistical model selection in the presence of prior information on the model parameters. The more complete is the available prior information on the model, further information on the parameters can be inferred from the data. The nested model exploits the fact that the model parameters themselves also form a multigraph, to which the same generative model can be fitted, so that the inference of an upper level serves as prior information to a lower level. This comes at the expense of the inclusion of hyperparameters describing the upper level itself, which in turn can be described by another generative model, forming a recursive nested block model hierarchy until no parameters are left to describe. In this way, one is able to improve the model selection process without introducing a single additional parameter. The result is a more refined model selection scheme which performs better than the non-hierarchical variant both in synthetic as well as in empirical data.

Despite its more elaborate formulation, this model remains tractable, and it is possible to apply it to very large networks, in a fully non-parametric manner, as discussed below. Furthermore, it generalizes cleanly the flat variants, which correspond simply to an hierarchy with only one level. It also does not impose any connection pattern (e.g. assortative or disassortative block structures), and is not restricted to any specific hierarchical form, such as binary trees or dendograms. 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 the supplemental material 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} occur

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 [49, 50] $\mathcal{S}(\{e_{rs}\}, \{n_r\}) = \ln \Omega(\{e_{rs}\}, \{n_r\})$.

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 [49]

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

for the traditional blockmodel ensemble and,

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

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 for the finding the block partition, since the others remain constant. However the full term is useful for comparing the models against each other when doing model selection, as discussed below.

For the upper level multigraphs the entropy can also be computed [49], and it takes a different form

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

where $\binom{n}{m} = \binom{n+m-1}{m}$ 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 become 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 $B_l \leq B_{l-1}$), where we set $B_{-1} \equiv N$. The edge counts at level l are denoted as e_{rs}^l , and the block sizes as n_r^l . Therefore we must have that $\sum_r n_r^l = B_{l-1}$ and $\sum_{rs} e_{rs}^l/2 = E$, 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 given then by

$$\mathcal{S}_n = \mathcal{S}_{t/c}(\{e_{rs}^0\}, \{n_r^0\}) + \sum_{l=1}^L \mathcal{S}_m(\{e_{rs}^l\}, \{n_r^l\}). \quad (4)$$

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 upper 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 which 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 which best describes the observed data without increasing its complexity. We present such an approach in the next section.

B. Model selection

A method which directly formalizes the Occam's razor principle is known as minimum description length (MDL) [51, 52], where one specifies the *total* amount of information necessary to describe the data, which includes not only the sample but the model parameters as well. The description length for the model above is

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

where $\mathcal{L}_{t/c}$ is the amount of information necessary to describe the model. In a given level l of the hierarchy, the information required to describe the model parameters $\{e_{rs}^l\}$ is given by the entropy S_m (Eq. 3) of the model in level $l + 1$. The only missing information is how to partition the nodes of the current level into B_l blocks. The total number of partitions with the same block sizes $\{n_r^l\}$ is given by $B_{l-1}! / \prod_r n_r^l!$, and the total number of different block sizes is $\binom{B_l}{B_{l-1}}$. Hence the total amount of information necessary to describe the block partition of level l is

$$\mathcal{L}_t^l = \ln \left(\binom{B_l}{B_{l-1}} \right) + \ln B_{l-1}! - \sum_r \ln n_r^l!, \quad (6)$$

and the total description length is $\mathcal{L}_t = \sum_{l=0}^L \mathcal{L}_t^l$. Note that this is different than the choice made in Refs. [20, 37] which considered all possible $B_l^{B_{l-1}}$ partitions to be equally likely, and hence the necessary amount of information as $B_{l-1} \ln B_l$. This choice implicitly assumes that all blocks have equal sizes, and offers worse description when this is not the case. Note that for $B_{l-1} \gg 1$ we have

$$\mathcal{L}_t^l \simeq B_{l-1} H(\{n_r^l / B_{l-1}\}), \quad (7)$$

where $H(\{p_i\}) = -\sum_i p_i \ln p_i$ is the entropy of the distribution $\{p_i\}$. Therefore, for uniform blocks $n_r^l = B_{l-1} / B_l$ we recover asymptotically the value $\mathcal{L}_t^l \simeq B_{l-1} \ln B_l$. However the value of Eq. 6 can be much smaller for non-uniform 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,

$$\mathcal{L}_c = \mathcal{L}_t - N \sum_k p_k \ln p_k. \quad (8)$$

Since the degree distribution of the observed network is a constant during the inference process, this extra term can be discarded. However, the full value is useful when comparing both model variants against each other, or with other classes of models.

It is easy to see that if one has a flat $L = 1$ hierarchy, with $\{B_l\} = \{B, 1\}$, the description length of the non-hierarchical model is recovered [20], e.g. for the traditional model we have $\Sigma_{L=1} = \mathcal{L}_{L=1} + \mathcal{S}_t$, with

$$\mathcal{L}_{L=1} = \ln \left(\binom{B}{E} \right) + \ln N! - \sum_r \ln n_r!, \quad (9)$$

where the only difference is that here we are using the improved partition description length of Eq. 6. Therefore the nested generalization fully encapsulates the flat version, such that $\min \Sigma \leq \min \Sigma_{L=1}$, i.e. the nested model can only provide 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 [29, 38, 39, 41, 42, 53], log-likelihood ratios [54] or more approximative methods such as Bayesian information criterion (BIC) [55] and Akaike information criterion (AIC) [56]. If any two of such methods are derived from equivalent assumptions, one would expect them to deliver compatible results. In the following we make a comparison of the MDL approach with Bayesian model selection via integrated likelihood (BMS), since it is non-approximative and can be computed exactly for the stochastic block model. We show that under compatible assumptions these two methods deliver the exact same results. We compare the results obtained with non-hierarchical MDL/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 non-hierarchical approaches.

1. Bayesian Model Selection (BMS)

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

$$\mathcal{P}(G|\{b_i\}, \{p_{rs}\}, B) = \prod_{rs} p_{rs}^{\frac{e_{rs}}{2}} (1 - p_{rs})^{\frac{n_r n_r - e_{rs}}{2}}. \quad (10)$$

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. 10 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 larger number of blocks B will invariably possess 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 $\{p_r\}$ which control the probability of each partition $\{b_i\}$ being observed, which is given by

$$\mathcal{P}(\{b_i\}|\{p_r\}, B) = \prod_r p_r^{n_r}. \quad (11)$$

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 which maximize this quantity, we compute the *integrated likelihood* [38, 42, 53],

$$\mathcal{P}(G, \{b_i\}|B) = \int dp_{rs} dp_r \mathcal{P}(G, \{b_i\}, \{p_{rs}\}, \{p_r\}|B) \quad (12)$$

$$= \int dp_{rs} \mathcal{P}(G|\{b_i\}, \{p_{rs}\}, B) \mathcal{P}(\{p_{rs}\}|B) \times \int dp_r \mathcal{P}(\{b_i\}|\{p_r\}) \mathcal{P}(\{p_r\}|B) \quad (13)$$

$$= \mathcal{P}(G|\{b_i\}, B) \times \mathcal{P}(\{b_i\}|B). \quad (14)$$

By maximizing $\mathcal{P}(G, \{b_i\}|B)$, instead of Eq. 10, one should avoid overfitting the data, since the larger models with many parameters are dominated by a majority of choices with fit the data very badly, and hence have a smaller contribution in the integral of Eq. 12. 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 to MDL. However, in practice things are more nuanced, since the value of Eq. 12 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 $\mathcal{P}(\{p_r\}|B) = \text{Dirichlet}(\{\alpha_r\})$, with $\alpha_r = 1$ so that all

counts are equally likely. The integral of Eq. 13 is then computed as

$$\ln \mathcal{P}(\{b_i\}|B) = -\ln \left(\binom{B}{N} \right) - \ln N! + \sum_r \ln n_r!, \quad (15)$$

which is identical to the partition description length of Eq. 6, 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 $\mathcal{P}(\{p_{rs}\}|B) = 1$ [31, 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 [31, 42],

$$\begin{aligned} \ln \mathcal{P}(G|\{b_i\}, B) &= -\sum_{r>s} \ln \binom{n_r n_s}{e_{rs}} + \ln(n_r n_s + 1) \\ &\quad - \sum_r \ln \binom{n_r^2}{e_{rr}/2} + \ln(n_r^2/2 + 1) \end{aligned} \quad (16)$$

$$\simeq -\frac{1}{2} \sum_{rs} n_r n_s H_b \left(\frac{e_{rs}}{n_r n_s} \right) - (B+1) \sum_r \ln n_r, \quad (17)$$

where the approximation in Eq. 17 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 which 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 constraint the average degree to something closer to what is observed in the data, but at the same time being otherwise non-informative 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. 13 becomes difficult to solve. Instead, an easier approach is to modify the model slightly, 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 that $\sum_{r \geq s} q_{rs} = 1$. The full posterior likelihood of this model is

$$\mathcal{P}(G|\{b_i\}, \{q_{rs}\}, E, B) = \frac{E!}{\Omega(\{e_{rs}\}, \{n_r\})} \frac{\prod_{r \geq s} q_{rs}^{m_{rs}}}{\prod_{r \geq s} m_{rs}!}, \quad (18)$$

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. 18 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 re-parametrization, the average degree remains fixed independently of the choice of prior. Therefore we may finally use a flat prior $\mathcal{P}(\{q_{rs}\}|B) = \text{Dirichlet}(\{\alpha_{rs} = 1\})$, without the risk of the graphs becoming inadvertently dense, and again the integrated likelihood can be computed exactly,

$$\mathcal{P}(G|\{b_i\}, B) = \int dq_{rs} \mathcal{P}(G|\{b_i\}, \{q_{rs}\}, E, B) \mathcal{P}(\{q_{rs}\}|B) \quad (19)$$

$$= \left[\Omega(\{e_{rs}\}, \{n_r\}) \times \left(\left(\binom{B}{2} \right) \right)^{-1} \right]. \quad (20)$$

By inserting Eq. 20 into Eq. 14, and comparing with equation Eq. 9, 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 an 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. What remains to be seen, is to what extent these different choices can affect the results of the model selection in practical situations. This analysis, which is presented in the next session, allows us also to derive limits on the detectability of modular structures according to each method, and tackle the so-called “resolution limit” problem, in which perceivedly well-pronounced blocks are merged together during model selection.

2. 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 to the amount of prior knowledge one has. If the number of blocks B is known beforehand, the remaining task is simply to classify the nodes in one these B classes. This problem has been shown to exhibit a detectability-indetectability phase transition [17, 18]: 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 which 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 [57]. 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 which 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 $\langle k \rangle > [(B-1)/(cB-1)]^2$ [17–19]. 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 knew we had B blocks to begin with. If we do not know this, we may apply any of the model selection methods 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 non-hierarchical model by summing $\mathcal{S}_t = -E \ln \min(B, B_{\text{true}})$ with Eq. 9. Similarly, an expression for the dense integrated likelihood is obtained by including the same entropy expression as the first term of Eq. 17 and $n_r = N/B$ in the second. For the nested version of model, we assume a regular hierarchical tree, with a fixed branching ratio σ , with $B_l = \sigma^{L-l}$, so that Eq. 5 becomes

$$\Sigma \simeq \left(\binom{\sigma}{2} \right) \frac{B}{\sigma - 1} \ln E + \frac{\sigma}{2} B \ln B + N \ln B - E \ln \min(B, B_{\text{true}}), \quad (21)$$

where $B_l \gg \sigma$ was assumed, 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 it is shown the optimum values of B for a model with $N = 10^4$ and $B = 100$, as well as the results for the direct minimization of the corresponding exact quantities for actual network realizations. We also include the comparison with a dense BMS version where the partition likelihood term is omitted in Eq. 14, i.e. $\mathcal{P}(\{b_i\}|B) = 1$, as was done in Refs. [31, 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 be simply discarded [58]. Both MDL and dense BMS fail to detect anything for $\langle k \rangle < 2$, which corresponds to a strong threshold [59], 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 byproduct 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 $c \rightarrow 1$, but with the additional restriction that the graph remains connected, it has been shown [13] that modules are merged together as long as $B > \sqrt{E}$. This phenomenon is considered counter-intuitive, and has been called the “resolution limit” of community detection via this method. 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 non-hierarchical 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 [20]. 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 no longer can be used to compress the data. This is a direct result of the assumption that

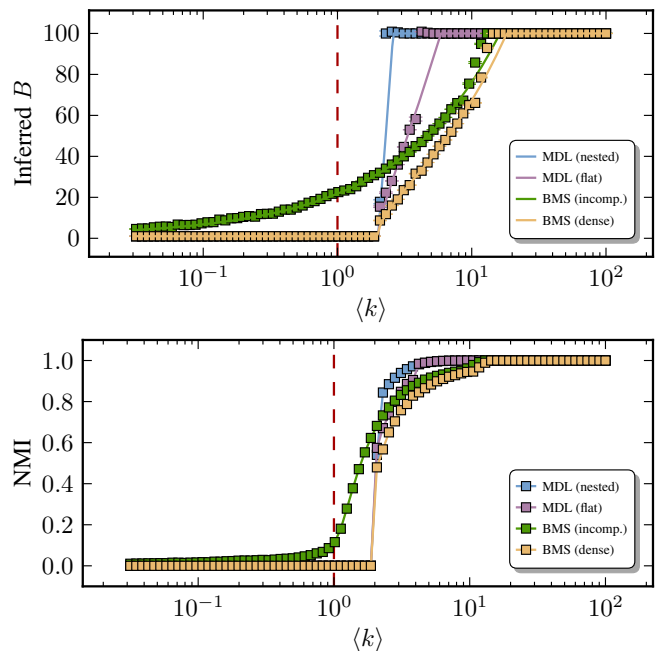


FIG. 2. Model selection results for a PP model with $N = 10^4$, $B = 100$ and fully isolated blocks ($c = 1$), using the model selection criteria described in the text. The top panel shows the inferred value of B versus the average degree $\langle k \rangle$ in the network. The solid lines show the theoretical 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 $N \rightarrow \infty$.

all possible block structures of a given size are equally possible, and the number of such models become 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. 21 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

$$B^* \sim \frac{N}{\ln N + \ln \ln N} \sim \frac{N}{\ln N}, \quad (22)$$

for sufficiently large N . This is a significant improvement, since the maximum number of detectable blocks

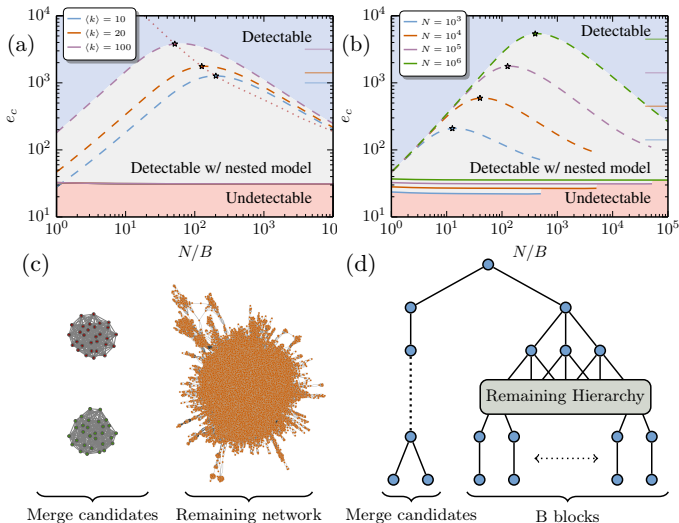


FIG. 3. Parameter region where two isolated blocks with $e_c/2$ internal nodes are detectable as separate blocks [shown schematically in panel (c)], as a function of the average block size N/B , and depending on (a) the average degree $\langle k \rangle$ with $N = 10^5$ and (b) the number of nodes N with $\langle k \rangle = 20$. The dashed curves show the boundaries for the non-hierarchical 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 [13], $e_c^* = \sqrt{2E}$. The points marked with stars (\star) correspond to the maximum value of B which is detectable in the remaining network with the non-hierarchical model, and the dotted line shows the same quantity for various $\langle k \rangle$ values. (d) The hierarchical construction used to decide if the two isolated blocks are merged together with the nested model.

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 more precisely the origin of the improvement by considering a related problem, which is the detection of blocks which are much smaller than the remaining network. With the modularity approach, another facet of the resolution limit manifests itself when two 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 [13], 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 $\Delta \mathcal{S}_t = e_r \ln 2$ (where we assume $e_r \ll n_r^2$, but the dense 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. 9. 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 Fig. 3(a) and (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 is, 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_c^* 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 [60]), 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 which 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. By computing the difference via Eq. 4 and Eq. 6, 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 from each other. This means that in many practical situations one can fully overcome the resolution limit, without abandoning a global model which 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 Monte Carlo [20, 40], simulated annealing [61], or belief prop-

agation [17, 18, 54]. Here we use the method described in Ref. [62], which is an agglomerative heuristic which provides high-quality results, while being unbiased with respect to the types of block structure which are inferred, and is also very efficient, with an algorithmic complexity of $O(N \ln^2 N)$. 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 $l = 1$, 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 cannot promise 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 which 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 only depends 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.
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 blocks of level $l - 1$ are merged 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 in starting with some initial hierarchy, and keeping track of whether or not each level is “done” or “not done”. One marks initially 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 one cannot guarantee that the global minimum is found. However we found that in the majority of cases this algorithm succeeds in finding the same or very similar hierarchies, independently of the initial choice, which can be simply $\{B_l\} = \{1\}$. 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 on 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 succeeded in obtaining reliable results with this algorithm for networks in excess of 10^7 edges, hence it is suitable for large scale systems [63].

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 planted partition model (PP) [57], inspired by similar constructions done in [64, 65]. We define a seed structure with B_0 blocks and $[\mathbf{m}_1]_{rs} = \delta_{rs}c/B_0 + (1 - \delta_{rs})(1 - c)/B_0(B_0 - 1)$, and construct a nested matrix of depth $L - 1$ via $\mathbf{m}_l = \mathbf{m}_{l-1} \otimes \mathbf{m}_{l-1}$ where \otimes denotes the Kronecker product, and $l \in [1, L - 1]$. The parameters of the model at level l are $e_{rs}^l = 2Em_{rs}$, and all $B = B_0^{L-1}$ blocks have the same number of nodes. Via spectral methods [66] 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 $B = B_0$ [17–19, 67].

In Fig. 4 are shown 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 become 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 which 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. The shallower hierarchies which are inferred correspond to

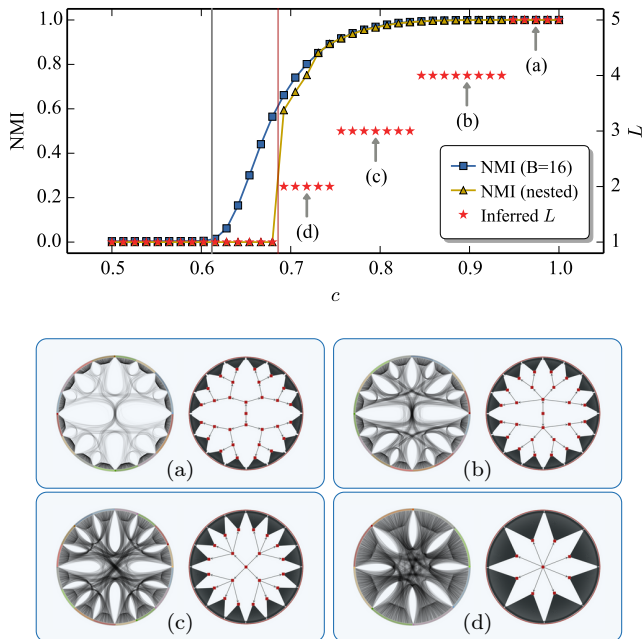


FIG. 4. Top: Normalized mutual information between the inferred and true partitions for network realizations of the nested PP model described in the text with $B_1 = 2$, $L = 5$, $\langle k \rangle = 20$ and $N = 10^4$, as a function of the assortativity strength c , both via the standard stochastic 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 grey 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 image shows the network realization itself, and the right one the hierarchical structure [the planted hierarchy corresponds to the one in (a)].

identical representations of the same e_{rs} matrix, 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 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 which are actually found, since despite the increased resolution capabilities it does not tend to find spurious hierarchies.

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 used 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 US elections. This is a network compiled by Adamic et al [68] of political blogs during the 2004 presidential elections in the USA. The nodes are $N = 1222$ individual blogs, and $E = 19,027$ 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 matches closely to the accepted division (see Fig. 5). This partition is also obtained with the non-hierarchical stochastic block model if one imposes $B = 2$ [32]. However, the nested version reveals a much more complete picture of the network, where these two partitions possess a detailed internal structure, culminating in $B_0 = 10$ subgroups with quite different connection patterns. For instance, one can see that each of the two higher level groups possesses one subgroup composed mainly of peripheral nodes, i.e. blogs which cite other blogs, but are not themselves cited as often. Conversely, both factions possess subgroups which tend to be cited by most other groups, and others which are cited predominantly by specific groups. It is also interesting to notice that a large fraction of the connections between the larger factions are concentrated between two of these 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 (AS) topology of the Internet. Autonomous Systems (AS) are intermediary building blocks of the internet topology. They represent organizational units, which are used to control the routing of packets in the network. A single AS usually corresponds to a network of its own, and 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 [70]. 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 comprised 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.

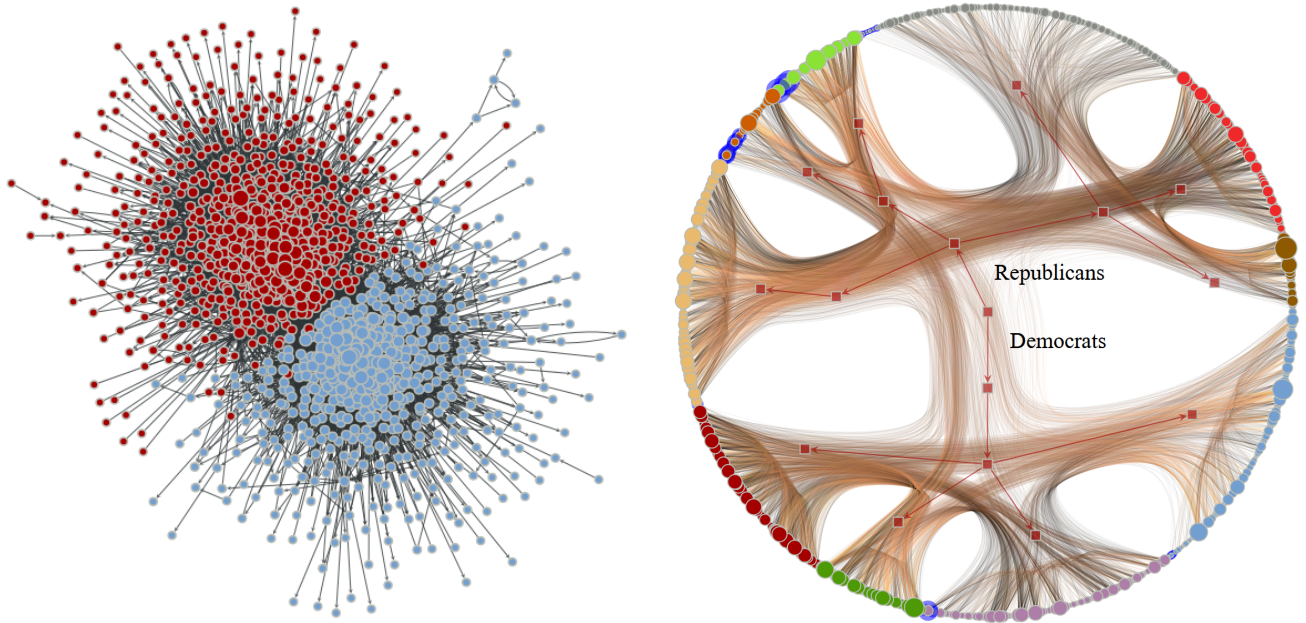


FIG. 5. The political blog network of Adamic et al [68]. Left: Top-most partition of the hierarchy inferred with the nested model. Right: The same network, using a circular layout, with edge bundling following the inferred hierarchy [69] (indicated also by the square nodes, and the node colors). The size of the nodes correspond 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 top-most level, according to the accepted partition.

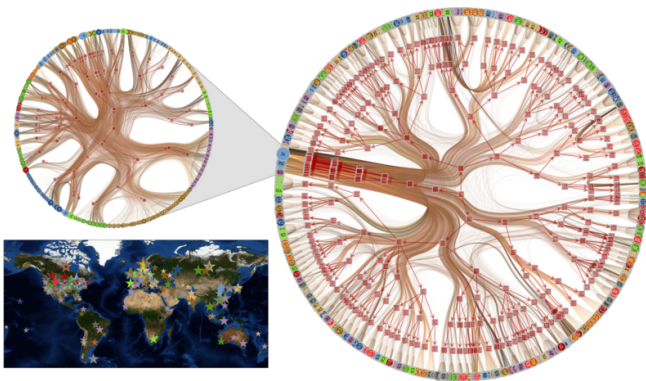


FIG. 6. Large-scale structure of the Internet at the autonomous systems level, as obtained by the nested stochastic block model, displaying a prominent core-periphery architecture. The “blow up” shows the nodes which belong to the “core” top-level branch, containing AS nodes spread all over the globe, as shown in the map inset.

The most prominent feature observed is a strong core-periphery structure, where most connections go through a 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 all the nodes of the core groups are not confined to a single geographical location, and are instead spread all over the

globe.

The Film-Actor Network. This network is compiled by extracting information available in the Internet Movie Database (IMDB) [71], and it contain 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 top-most 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 non-hierarchical model [20], but one finds $B = 717$ blocks, instead of $B = 332$ with the flat version.

Meta-analysis of several empirical networks. We performed 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 used both the non-hierarchical stochastic block model, as well as the nested variant. In Fig. 8(a) and (b) are shown the average block sizes N/B for all networks using both models. For the non-hierarchical 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) are shown the results for the nested model, where such trend can no longer be observed, and the smallest average block sizes no longer

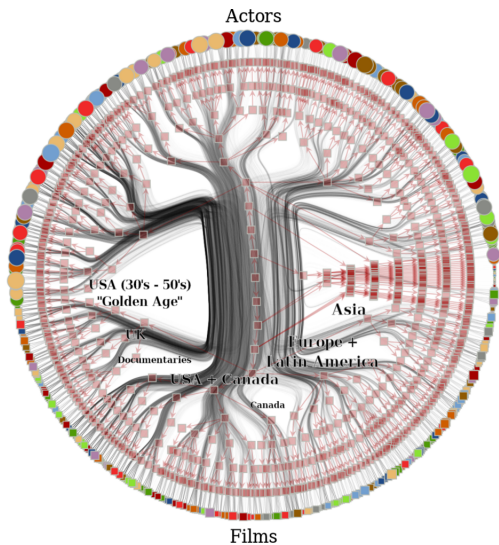
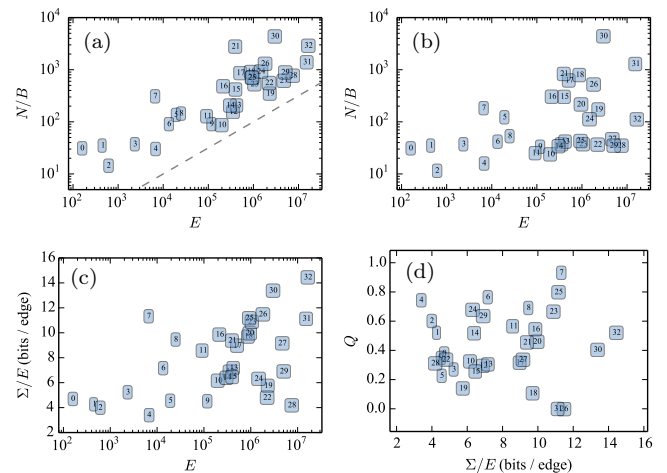


FIG. 7. Large-scale structure of the IMDB film-actor network. Each node in this graph represents a lowest-level block in the hierarchy, instead of individual nodes in the graph. The size of the nodes indicates the number of nodes in each group. The hierarchy branch at the top are the actors, and at the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

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 length themselves are also distributed in a seemingly non-organized manner [See Fig. 8(c)], i.e. no general tendency for larger networks can be observed, other than an increased range of possible values. Any difference observed seems to be due to the actual topological organization, rather than intrinsic constraints imposed by the method. We also computed the modularity of the inferred block structures, $Q = \sum_r e_{rr}/2E - e_r^2/(2E)^2$, which measures how assortative is the topology. 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 which matches the data. In Fig. 8 it is shown 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



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	2,345	Yes	14	39,796	301,498	Yes	25	334,863	925,872	No
4	903	6,760	No	15	52,104	399,625	Yes	26	372,547	1,812,312	No
5	1,222	19,021	Yes	16	56,739	212,945	No	27	449,087	4,690,321	Yes
6	4,158	13,422	No	17	75,877	508,836	Yes	28	654,782	7,499,425	Yes
7	4,941	6,594	No	18	82,168	870,161	Yes	29	855,802	5,066,842	Yes
8	8,638	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 [72]	11	arXiv Co-Authors (cond-mat) [73]	22	Web graph of stanford.edu [74]
1	Political Books ^a	12	arXiv Citations (hep-th) [73, 75]	23	DBLP collaboration [76]
2	American Football [3, 77]	13	arXiv Citations (hep-ph) [73, 75]	24	WWW [6]
3	C. Elegans Neurons [78]	14	PGP [79]	25	Amazon product network [76]
4	Disease Genes [80]	15	Internet AS (Caida) ^b	26	IMDB film-actor ^c [20] (bipartite)
5	Political Blogs [68]	16	Brightkite social network [81]	27	APS citations ^d
6	arXiv Co-Authors (gr-qc) [73]	17	Epinions.com trust network [82]	28	Berkeley/Stanford web graph [74]
7	Power Grid [78]	18	Slashdot [83]	29	Google web graph [74]
8	arXiv Co-Authors (hep-th) [73]	19	Flickr [84]	30	Youtube social network [76]
9	arXiv Co-Authors (hep-ph) [73]	20	Gowalla social network [81]	31	Yahoo groups ^e (bipartite)
10	arXiv Co-Authors (astro-ph) [73]	21	EU email [73]	32	US patent citations [85]

^a V. Krebs, unpublished.

^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 non-hierarchical model, as a function of E , for the empirical networks listed in the bottom table. 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.

to find partitions which lead to much higher Q values. These partitions would, on the other hand, invariably correspond to a block model ensemble 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 inferences possess clear advantages, such as the lack of resolution limit, model selection guarantees, and the overall more principled nature of the approach.

VI. DISCUSSION

We have presented 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 [43–48], 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 dendrograms. We have shown that a major advantage of this approach is that it breaks the so-called resolution limit of approaches like modularity and non-hierarchical 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 much a smaller logarithmic dependence, making it virtually non-existent 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 have shown 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, 86–88], or link communities [33, 89]. We also predict it should serve as more refined method of detecting missing information in networks [31, 43], as well as for the prediction of the network evolution [90], determining the more salient topological features [91, 92], or as a large-scale functional summary of the network topology [93].

-
- [1] M. E. J. Newman, *Nat Phys* **8**, 25 (2011).
 - [2] S. Fortunato, *Physics Reports* **486**, 75 (2010).
 - [3] M. Girvan and M. E. J. Newman, *Proceedings of the National Academy of Sciences* **99**, 7821 (2002).
 - [4] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabási, *Science* **297**, 1551 (2002), PMID: 12202830.
 - [5] R. J. Fletcher Jr, A. Revell, B. E. Reichert, W. M. Kitchens, J. D. Dixon, and J. D. Austin, *Nature Communications* **4** (2013), 10.1038/ncomms3572.
 - [6] R. Albert, H. Jeong, and A.-L. Barabási, *Nature* **401**, 130 (1999).
 - [7] S.-H. Yook, H. Jeong, and A.-L. Barabási, *Proceedings of the National Academy of Sciences* **99**, 13382 (2002), PMID: 12368484.
 - [8] Y. Zhao, E. Levina, and J. Zhu, *Proceedings of the National Academy of Sciences* **108**, 7321 (2011).
 - [9] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, *Nature* **435**, 814 (2005).
 - [10] M. E. J. Newman and M. Girvan, *Physical Review E* **69**, 026113 (2004).
 - [11] A. Clauset, M. E. J. Newman, and C. Moore, *Physical Review E* **70**, 066111 (2004).
 - [12] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, 0803.0476 (2008).
 - [13] S. Fortunato and M. Barthélemy, *Proceedings of the National Academy of Sciences* **104**, 36 (2007).
 - [14] A. Lancichinetti and S. Fortunato, 1107.1155 (2011).
 - [15] B. H. Good, Y.-A. de Montjoye, and A. Clauset, *Physical Review E* **81**, 046106 (2010).
 - [16] R. Guimerà, M. Sales-Pardo, and L. A. N. Amaral, *Physical Review E* **70**, 025101 (2004).
 - [17] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová, *Physical Review Letters* **107**, 065701 (2011).
 - [18] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová, *Physical Review E* **84**, 066106 (2011).
 - [19] E. Mossel, J. Neeman, and A. Sly, arXiv:1202.1499 (2012).
 - [20] T. P. Peixoto, *Physical Review Letters* **110**, 148701 (2013).
 - [21] P. W. Holland, K. B. Laskey, and S. Leinhardt, *Social Networks* **5**, 109 (1983).
 - [22] S. E. Fienberg, M. M. Meyer, and S. S. Wasserman, *Journal of the American Statistical Association* **80**, 51 (1985).
 - [23] K. Faust and S. Wasserman, *Social Networks* **14**, 5 (1992).
 - [24] C. J. Anderson, S. Wasserman, and K. Faust, *Social Networks* **14**, 137 (1992).
 - [25] M. B. Hastings, *Physical Review E* **74**, 035102 (2006).
 - [26] D. Garlaschelli and M. I. Loffredo, *Physical Review E* **78**, 015101 (2008).
 - [27] M. E. J. Newman and E. A. Leicht, *Proceedings of the National Academy of Sciences* **104**, 9564 (2007).
 - [28] J. Reichardt and D. R. White, *The European Physical Journal B* **60**, 217 (2007).
 - [29] J. M. Hofman and C. H. Wiggins, *Physical Review Letters* **100**, 258701 (2008).
 - [30] P. J. Bickel and A. Chen, *Proceedings of the National Academy of Sciences* **106**, 21068 (2009).
 - [31] R. Guimerà and M. Sales-Pardo, *Proceedings of the National Academy of Sciences* **106**, 22073 (2009).
 - [32] B. Karrer and M. E. J. Newman, *Physical Review E* **83**, 016107 (2011).
 - [33] B. Ball, B. Karrer, and M. E. J. Newman, *Physical Review E* **84**, 036103 (2011).
 - [34] J. Reichardt, R. Alaminio, and D. Saad, *PLoS ONE* **6**, e21282 (2011).
 - [35] Y. Zhu, X. Yan, and C. Moore, arXiv:1205.7009 (2012).
 - [36] E. B. Baskerville, A. P. Dobson, T. Bedford, S. Allesina, T. M. Anderson, and M. Pascual, *PLoS Comput Biol* **7**, e1002321 (2011).
 - [37] M. Rosvall and C. T. Bergstrom, *Proceedings of the National Academy of Sciences* **104**, 7327 (2007).
 - [38] J.-J. Daudin, F. Picard, and S. Robin, *Statistics and Computing* **18**, 173 (2008).
 - [39] M. Mariadassou, S. Robin, and C. Vacher, *The Annals of Applied Statistics* **4**, 715 (2010), mathematical Reviews number (MathSciNet): MR2758646.
 - [40] C. Moore, X. Yan, Y. Zhu, J.-B. Rouquier, and T. Lane,

- in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11 (ACM, New York, NY, USA, 2011) p. 841–849.
- [41] P. Latouche, E. Birmele, and C. Ambroise, *Statistical Modelling* **12**, 93–115 (2012).
 - [42] E. Côme and P. Latouche, *Model selection and clustering in stochastic block models with the exact integrated complete data likelihood*, arXiv e-print 1303.2962 (2013).
 - [43] A. Clauset, C. Moore, and M. E. J. Newman, *Nature* **453**, 98 (2008).
 - [44] M. Rosvall and C. T. Bergstrom, *PLoS ONE* **6**, e18209 (2011).
 - [45] M. Sales-Pardo, R. Guimerà, A. A. Moreira, and L. A. N. Amaral, *Proceedings of the National Academy of Sciences* **104**, 15224 (2007), PMID: 17881571.
 - [46] P. Ronhovde and Z. Nussinov, *Physical Review E* **80**, 016109 (2009).
 - [47] I. A. Kovács, R. Palotai, M. S. Szalay, and P. Csermely, *PLoS ONE* **5**, e12528 (2010).
 - [48] Y. Park, C. Moore, and J. S. Bader, *PLoS ONE* **5**, e8118 (2010).
 - [49] T. P. Peixoto, *Physical Review E* **85**, 056122 (2012).
 - [50] G. Bianconi, *Physical Review E* **79**, 036114 (2009).
 - [51] P. D. Grünwald, *The Minimum Description Length Principle* (The MIT Press, 2007).
 - [52] J. Rissanen, *Information and Complexity in Statistical Modeling*, 1st ed. (Springer, 2010).
 - [53] C. Biernacki, G. Celeux, and G. Govaert, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**, 719 (2000).
 - [54] X. Yan, J. E. Jensen, F. Krzakala, C. Moore, C. R. Shalizi, L. Zdeborova, P. Zhang, and Y. Zhu, arXiv:1207.3994 (2012).
 - [55] G. Schwarz, *The Annals of Statistics* **6**, 461 (1978), mathematical Reviews number (MathSciNet): MR468014; Zentralblatt MATH identifier: 0379.62005.
 - [56] H. Akaike, *IEEE Transactions on Automatic Control* **19**, 716 (1974).
 - [57] A. Condon and R. M. Karp, *Random Structures & Algorithms* **18**, 116–140 (2001).
 - [58] 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 $N \rightarrow \infty$. On the other hand, according to this criterion, the inferred value of B in this region increases as N becomes larger.
 - [59] This threshold corresponds simply 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 $\langle k \rangle = 2$.
 - [60] Note that 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.
 - [61] S. Kirkpatrick, C. D. Gelatt Jr, and M. P. Vecchi, *Science* **220**, 671 (1983).
 - [62] T. P. Peixoto, arXiv: (pending) (2013).
 - [63] An efficient C++ implementation of the algorithm described here is freely available as part of the graph-tool Python library at <http://graph-tool.skewed.de>.
 - [64] J. Leskovec, D. Chakrabarti, J. Kleinberg, C. Faloutsos, and Z. Ghahramani, arXiv:0812.4905 (2008).
 - [65] G. Palla, L. Lovász, and T. Vicsek, *Proceedings of the National Academy of Sciences* **107**, 7640 (2010).
 - [66] T. P. Peixoto, *Physical Review Letters* **111**, 098701 (2013).
 - [67] R. R. Nadakuditi and M. E. J. Newman, *Physical Review Letters* **108**, 188701 (2012).
 - [68] L. A. Adamic and N. Glance, in *Proceedings of the 3rd international workshop on Link discovery*, LinkKDD '05 (ACM, New York, NY, USA, 2005) p. 36–43.
 - [69] D. Holten, *IEEE Transactions on Visualization and Computer Graphics* **12**, 741 (2006).
 - [70] The IPv4 Routed /24 AS Links Dataset, http://www.caida.org/data/active/ipv4_routed_topology_aslinks_dataset.xml.
 - [71] <http://www.imdb.com>.
 - [72] D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson, *Behavioral Ecology and Sociobiology* **54**, 396 (2003).
 - [73] J. Leskovec, J. Kleinberg, and C. Faloutsos, *ACM Trans. Knowl. Discov. Data* **1** (2007), 10.1145/1217299.1217301.
 - [74] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney, *Internet Mathematics* **6**, 29 (2009).
 - [75] J. Gehrke, P. Ginsparg, and J. Kleinberg, *SIGKDD Explor. Newsl.* **5**, 149–151 (2003).
 - [76] J. Yang and J. Leskovec, in *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*, MDS '12 (ACM, New York, NY, USA, 2012) p. 3:1–3:8.
 - [77] T. S. Evans, FigShare (2012), 10.6084/m9.figshare.93179.
 - [78] D. J. Watts and S. H. Strogatz, *Nature* **393**, 409 (1998).
 - [79] O. Richters and T. P. Peixoto, *PLoS ONE* **6**, e18384 (2011).
 - [80] K. I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A. L. Barabási, *Proceedings of the National Academy of Sciences* **104**, 8685 (2007).
 - [81] E. Cho, S. A. Myers, and J. Leskovec, in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11 (ACM, New York, NY, USA, 2011) p. 1082–1090.
 - [82] M. Richardson, R. Agrawal, and P. Domingos, in *The Semantic Web - ISWC 2003*, Lecture Notes in Computer Science No. 2870, edited by D. Fensel, K. Sycara, and J. Mylopoulos (Springer Berlin Heidelberg, 2003) pp. 351–368.
 - [83] J. Leskovec, D. Huttenlocher, and J. Kleinberg, in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10 (ACM, New York, NY, USA, 2010) p. 1361–1370.
 - [84] J. McAuley and J. Leskovec, in *Computer Vision – ECCV 2012*, Lecture Notes in Computer Science No. 7575, edited by A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid (Springer Berlin Heidelberg, 2012) pp. 828–841.
 - [85] J. Leskovec, J. Kleinberg, and C. Faloutsos, in *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, KDD '05 (ACM, New York, NY, USA, 2005) p. 177–187.
 - [86] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing, *J. Mach. Learn. Res.* **9**, 1981–2014 (2008).
 - [87] A. Lancichinetti, S. Fortunato, and J. Kertész, *New Journal of Physics* **11**, 033015 (2009).
 - [88] P. K. Gopalan and D. M. Blei, *Proceedings of the National Academy of Sciences* **110**, 14534 (2013), PMID: 23950224.

- [89] Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann, *Nature* **466**, 761 (2010).
- [90] D. Liben-Nowell and J. Kleinberg, *Journal of the American Society for Information Science and Technology* **58**, 1019–1031 (2007).
- [91] D. Grady, C. Thiemann, and D. Brockmann, *Nature Communications* **3**, 864 (2012).
- [92] G. Bianconi, P. Pin, and M. Marsili, *Proceedings of the National Academy of Sciences* **106**, 11433 (2009), PMID: 19571013.
- [93] R. Guimerà and L. A. Nunes Amaral, *Nature* **433**, 895 (2005).

Supporting Information: 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

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 [1],

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

while for the directed case it reads,

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

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,

$$\mathcal{S}_t \cong E - \frac{1}{2} \sum_{rs} e_{rs} \ln \left(\frac{e_{rs}}{n_r n_s} \right), \quad (3)$$

$$\mathcal{S}_t^d \cong E - \sum_{rs} e_{rs} \ln \left(\frac{e_{rs}}{n_r n_s} \right). \quad (4)$$

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

$$\mathcal{S}_c \cong -E - \sum_k N_k \ln k! - \frac{1}{2} \sum_{rs} e_{rs} \ln \left(\frac{e_{rs}}{e_r e_s} \right), \quad (5)$$

$$\begin{aligned} \mathcal{S}_c^d \cong & -E - \sum_{k^+} N_{k^+} \ln k^+! - \sum_{k^-} N_{k^-} \ln k^-! \\ & - \sum_{rs} e_{rs} \ln \left(\frac{e_{rs}}{e_r^+ e_s^-} \right), \end{aligned} \quad (6)$$

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,

$$e_{rs} \frac{\langle k^2 \rangle_r - \langle k \rangle_r \langle k^2 \rangle_s - \langle k \rangle_s}{\langle k \rangle_r^2} \ll n_r n_s, \quad (7)$$

where $\langle k^l \rangle_r = \sum_{i \in r} k_i^l / n_r$ (for the directed case we simply replace $\langle k^l \rangle_r \rightarrow \langle (k^+)^l \rangle_r$ and $\langle k^l \rangle_s \rightarrow \langle (k^-)^l \rangle_s$ in the equation above). Unfortunately there is no closed-form expression for the entropy outside the sparse limit, unlike the traditional variant [1].

For the upper level multigraphs the entropies are [1],

$$\mathcal{S}_m = \sum_{r>s} \ln \left(\binom{n_r n_s}{e_{rs}} \right) + \sum_r \ln \left(\binom{\binom{n_r}{2}}{e_{rr}/2} \right) \quad (8)$$

$$\mathcal{S}_m^d = \sum_{rs} \ln \left(\binom{n_r n_s}{e_{rs}} \right), \quad (9)$$

where $\binom{n}{m} = \binom{n+m-1}{m}$ is the number of m -combinations with repetitions from a set of size n .

Empirical networks

In Fig. 1 is shown a higher resolution version of Fig. 6 in the main text, with additional information on the correlation with geographical location. As can be seen, at the topmost level is separated in three blocks, comprising: 1. The “core” AS nodes; 2. The Americas, Africa, South Asia and Oceania; 3. Europe, Russia and the middle east. These get further subdivided in the lower levels.

In Fig. 2 is shown a higher resolution version of Fig. 7 in the main text, containing the large-scale structure of the IMDB network. See also Ref. [2] for more details on the type of information which can be associated with each block.

* tiago@itp.uni-bremen.de

[1] T. P. Peixoto, Physical Review E **85**, 056122 (2012).

[2] T. P. Peixoto, Physical Review Letters **110**, 148701 (2013).

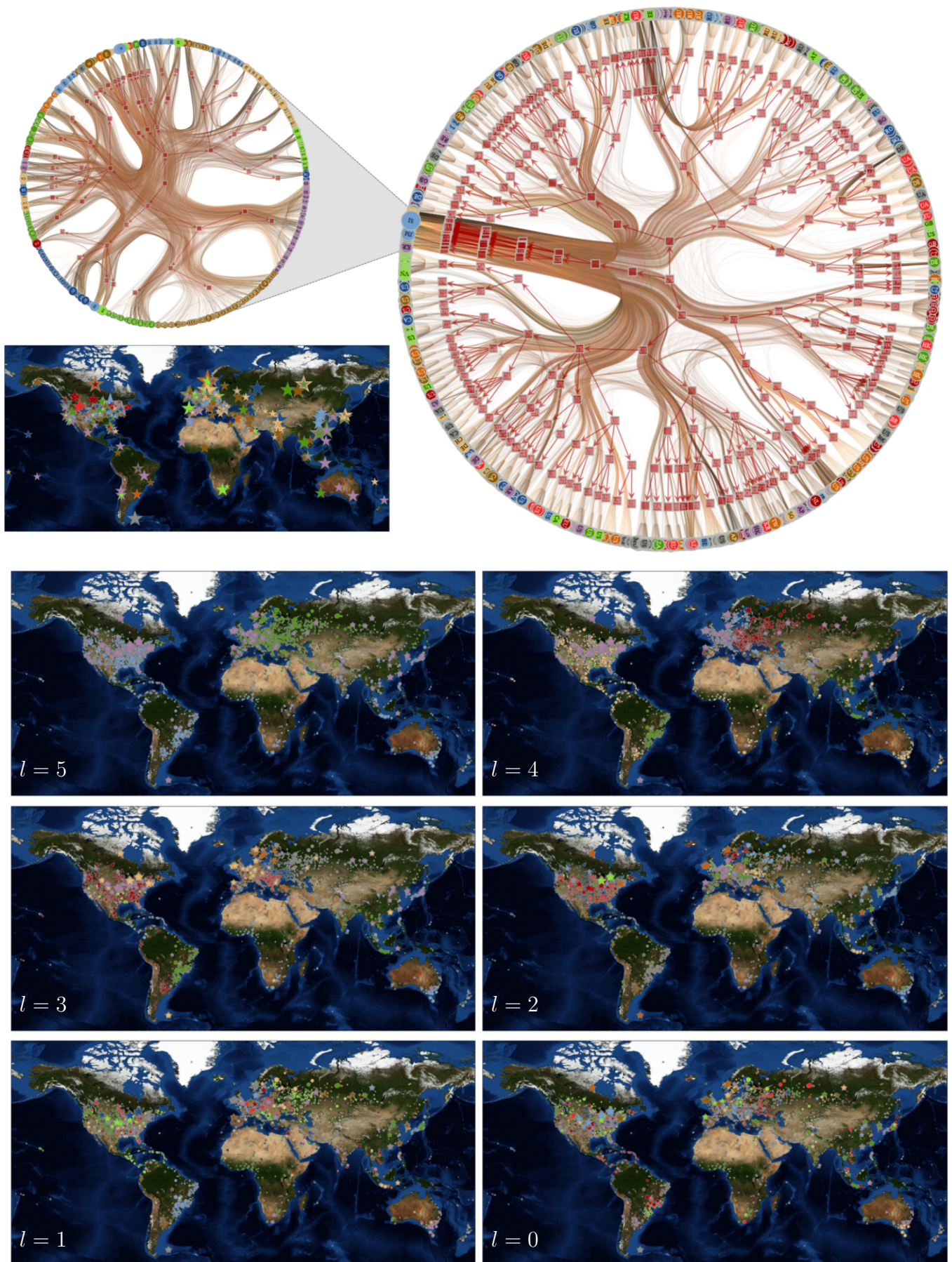


FIG. 1. Large-scale structure of the Internet at the autonomous systems level, as obtained by the nested stochastic block model, displaying a prominent core-periphery architecture. The “blow up” shows the nodes which belong to the “core” top-level branch, containing AS nodes spread all over the globe, as shown in the map inset below it. The maps at the bottom show the network partitions at different hierarchical levels, from top to bottom, showing the strong correlation with geographical divisions. The stars (\star) are the “core” nodes, and the circles are regular AS nodes.

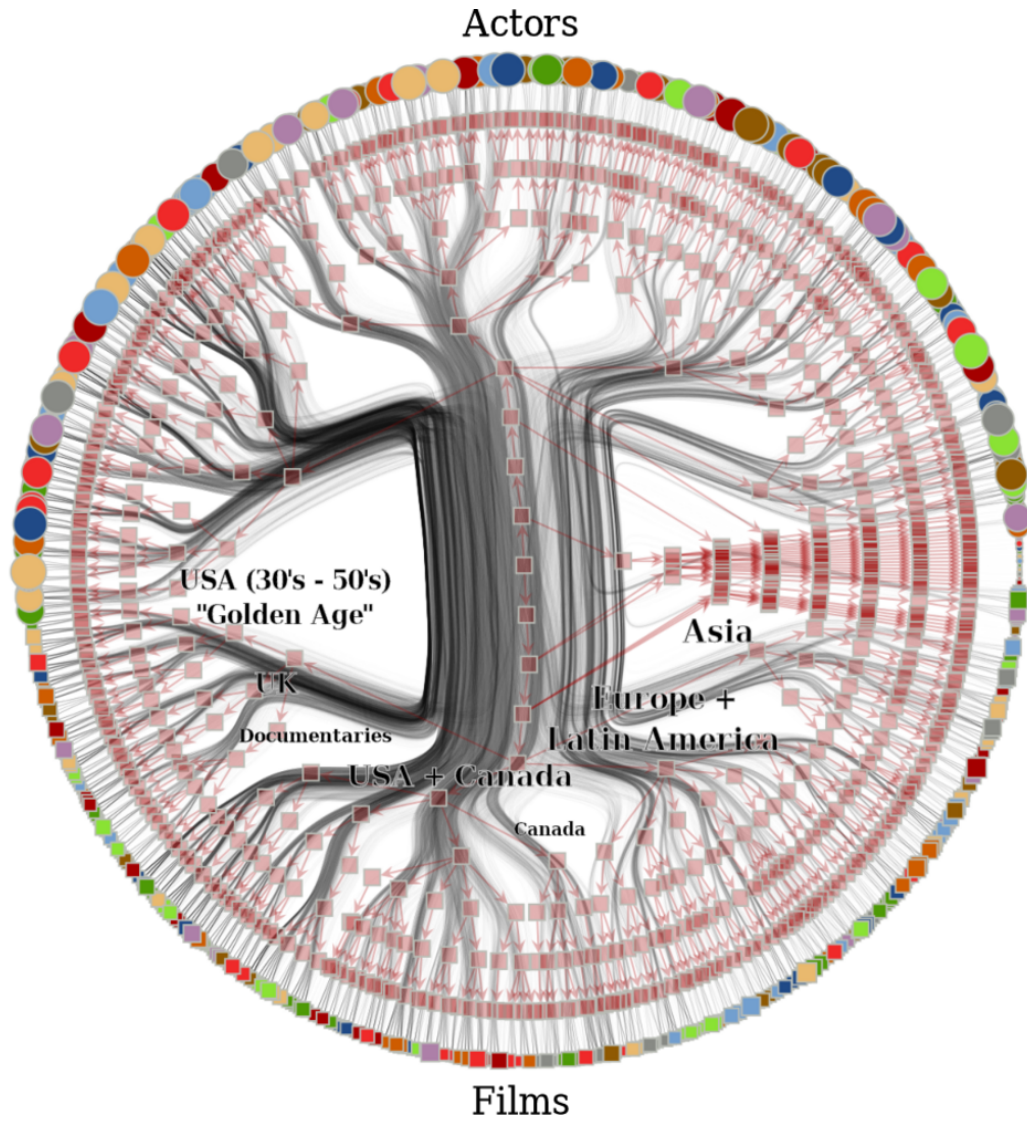


FIG. 2. Large-scale structure of the IMDB actor-film network. Each node in this graph represents a lowest-level block in the hierarchy, instead of individual nodes in the graph. The size of the nodes indicates the number of nodes in each group. The hierarchy branch at the top are the actors, and at the bottom are the films. The labels classify each branch according to the most prominent geographical, temporal or genre characteristics found in the database.