Implicit models, latent compression, intrinsic biases, and cheap lunches in community detection

Tiago P. Peixoto*
Department of Network and Data Science, Central European University, Vienna, Austria

Alec Kirkley†
Institute of Data Science, University of Hong Kong, Hong Kong
Department of Urban Planning and Design, University of Hong Kong, Hong Kong and
Urban Systems Institute, University of Hong Kong, Hong Kong

The task of community detection, which aims to partition a network into clusters of nodes to summarize its large-scale structure, has spawned the development of many competing algorithms with varying objectives. Some community detection methods are inferential, explicitly deriving the clustering objective through a probabilistic generative model, while other methods are descriptive, dividing a network according to an objective motivated by a particular application, making it challenging to compare these methods on the same scale. Here we present a solution to this problem that associates any community detection objective, inferential or descriptive, with its corresponding implicit network generative model. This allows us to compute the description length of a network and its partition under arbitrary objectives, providing a principled measure to compare the performance of different algorithms without the need for "ground truth" labels. Our approach also gives access to instances of the community detection problem that are optimal to any given algorithm, and in this way reveals intrinsic biases in popular descriptive methods, explaining their tendency to overfit. Using our framework, we compare a number of community detection methods on artificial networks, and on a corpus of over 500 structurally diverse empirical networks. We find that more expressive community detection methods exhibit consistently superior compression performance on structured data instances, without having degraded performance on a minority of situations where more specialized algorithms perform optimally. Our results undermine the implications of the "no free lunch" theorem for community detection, both conceptually and in practice, since it is confined to unstructured data instances, unlike relevant community detection problems which are structured by requirement.

I. INTRODUCTION

Community detection methods [1] are a cornerstone of network data analysis. They fulfill the need to digest an otherwise intractable large-scale structure of a complex system into a simpler coarse-grained description, where groups of items are clustered together according to shared patterns of interactions. This methodological ansatz has proved useful in countless applications in biology, physics, engineering, computer science, the social sciences, and other fields.

The research on community detection has evolved substantially in the last 20 years [2], spawning a large variety of different approaches. Substantial effort in this area has been devoted to the development of methods that behave well in practice — both in the quality of results and algorithmic efficiency — as well as to our theoretical understanding of their behavior [3]. Despite these advances, what perhaps continues to be one of the biggest difficulties when employing community detection methods in practice is that the task itself is not uniquely defined: what constitutes a good coarse-graining of a network is intrinsically tied to an ultimate objective, of which there can be many [4], resulting in algorithms that yield different answers for the same network [5, 6].

Most methods agree qualitatively on what constitutes community structure — groups of nodes that are more connected with themselves than with the rest of the network, or more generally, groups of nodes that have the same tendency of connecting to other groups of nodes — but the context in which this concept is evoked and the resulting mathematical definitions can vary substantially, to the point where two algorithms can yield radically different partitions of the same network despite sharing an overall conceptual agreement [7].

In order to better understand the discrepancies and similarities between community detection methods, it is useful to divide them into two classes, according to their stated objectives: inferential and descriptive [8]. Inferential methods evoke explicitly the notion of probabilistic generative models, i.e. network formation mechanisms that define how a division of the network into groups affect the probability with which the nodes are connected. In this setting, the community detection task consists of assuming that an observed network is an instance of this generative procedure, and attempting to fit it to data in order to infer the hidden partition — or more generally, a set of partitions ranked according to their posterior plausibility [9]. In this scenario, it is possible to assess the statistical significance and uncertainties of our inferences, and to quantify precisely how parsimonious the obtained coarse-grained representation is, allowing us to detect overfitting and underfitting, as well as to perform model selection. Furthermore, from the fitted model it is possible to make statements about edge placement probabilities and to make generalizations about unobserved data [10, 11].

Descriptive methods, conversely, do not involve an explicit definition of a generative model, and divide the network into groups according to other, application-specific criteria. What are perhaps the oldest instances of this class of methods are the various algorithms for graph partitioning in computer sci-

* peixotot@ceu.edu
† alec.w.kirkley@gmail.com

ence [12], motivated in large part by circuit design and task scheduling problems, instead of data analysis. In this setting, the desired network division is the one that optimizes a task conditioned on a given network — such as the spatial placement of transistors or division of tasks among processors. Prominent descriptive methods also use network clustering to characterize the behavior of dynamical processes that run on the network, typically random walks. For example, the Infomap method [13] clusters nodes in a manner that minimizes the information required to encode a random walk taking place on a network, according to how often it leaves and enters individual groups. In this case, the network is a parameter of a dynamical process, and therefore its generation is not modelled directly [14]. Arguably the most popular community detection method, modularity maximization [15], can also be classified as descriptive. Although it was originally motivated according to an explicit inferential criterion — namely the deviation from a null model — it is inconsistent with this stated goal, since it notoriously finds spurious deviations on networks sampled from its own null model [16]. Despite an approximate equivalence with the parametric inference of a restricted version of the stochastic block model (SBM) [17], valid only when the true number of groups is known and the data obeys certain symmetries [18], this method lacks an explicit inferential interpretation in the nonparametric manner it is actually employed in practice. For these and other descriptive methods in general, the notions of uncertainty and statistical significance are not inherent or explicitly evoked.

Despite these clear differences in stated objectives, descriptive methods are often used in practice with inferential aims. For example, communities found with descriptive methods are frequently interpreted as being the result of a homophilic edge formation mechanism in social networks [19–23] and functional modules in biological networks [24–27], to name only a few analyses in which a concern for statistical significance is expressed. Furthermore, attempts to benchmark community detection methods against each other typically involve comparing their performance in terms of their ability at recovering known partitions in artificial random networks sampled from generative models — thus being clearly an inferential criterion — as is the case of the popular LFR benchmark [28, 29]. More recently, the tendency of algorithms to under- or overfit in a link prediction task was considered in Ref. [6], which relies on a manifestly inferential criterion. One also finds prominent claims in the literature [1, 3, 15] that it would be undesirable for an arbitrary community detection method to cluster a maximally random network sampled either from the Erdős-Rényi or configuration model into more than one group, since this division would unveil purely random fluctuations in the placement of the edges, and thus would amount to overfitting. Because of this, very often results of descriptive community detection methods are compared to what is obtained with randomized versions of the data, in an attempt to quantify statistical significance [30]. Such a comparison with a "null model" is evidently an inferential concern, since it amounts to assessing the generative process underlying the network formation.

We explore in this work the fact there is no formal mathematical distinction between inferential and descriptive methods, since, as we show, we can always find an implicit generative model which yields an inference procedure identical to any descriptive approach. The implicit generative models that we describe in this work allow us to perform Bayesian model comparisons and assess statistical significance, closing the gap between inferential and descriptive methods by evaluating descriptive methods from a generative perspective. These implicit generative models also unveil the intrinsic biases present in arbitrary community detection methods — in other words, what kind of structure they expect to encounter a priori, even when this is not explicitly articulated in the motivation of the method — which cause over- or underfitting of the data. Furthermore, we show how we can use our method to appropriately tune parameters of algorithms to mitigate these biases, simultaneously removing existing resolution limits and the identification of spurious communities in maximally random networks from arbitrary community detection methods.

We focus in particular on two descriptive community detection methods: modularity maximization [15] and Infomap [13]. Here we show that these, together with a wider class of methods, are equivalent to special cases of an assortative SBM with groups having uniform size and density (a.k.a. the planted partition model [31]), where the number of groups and assortativity strength are determined directly by the expected value of the quality function. We show that the prior distribution of the number of groups is typically bimodal, concentrating simultaneously on a low and a large value, where the latter is on the order of the number of nodes in the network. This bimodality induces discontinuous transitions in the statistical properties of typical problem instances, preventing networks with moderate community structure and a wide range of the number of groups from being generated. This a priori bias towards particular kinds of uniform, but strong, community structure gives new clarity to the observed behavior of these methods in practice, and their tendency to find communities of equal size and density and in maximally random networks.

Our framework allows us to perform a comparison between algorithms in their capacity of uncovering community structure sampled from instances that are optimal for a different method. In particular, we consider optimal instances generated by modularity and the nested stochastic block model (NSBM) — a more expressive, hierarchical parametrization of the SBM which is a priori agnostic about the actual mixing patterns between groups. We demonstrate that — according to compressibility and accuracy in community recovery — there are substantial asymmetries between methods, where the more general NSBM does just as well (but no better) for instances that are optimal for the other algorithms, but where other algorithms perform significantly worse on instances that are optimal for the NSBM. We also perform a systematic comparison of methods on a corpus of over 500 diverse empirical networks, finding that the NSBM provides a better compression for the vast majority of them. This provides evidence for "cheap lunches" in community detection — more versatile, but appropriately regularized approaches tend to yield systematically better results over structured problem instances. This

result reveals a practical and conceptual caveat to the "no free lunch" (NFL) theorem for community detection [32], which states that when averaged over all instances of community detection problems — primarily composed of *unstructured* instances where community labels have no correlation with network structure — all conceivable algorithms must yield the same performance.

This paper is divided as follows. We begin in Sec. II with a discussion of implicit models for community detection algorithms, describing how to compute their corresponding description lengths and implicit priors, then apply these methods to the modularity and Infomap community detection objectives. In Sec. III, we follow up on this discussion by demonstrating the correspondence between a broad class of community detection objectives and restricted instances of stochastic block models, showing that these methods implicitly assume networks with very limited structure. Next, in Sec. IV we discuss how our framework can provide insights into the shortcomings of the NFL theorem for community detection when applied to structured problem instances by revealing asymmetries in algorithm compression performance. Finally, in Sec. V we apply our method to compare the description lengths associated with fitting a range of community detection algorithms to a wide variety of empirical networks, finding that a small number of more expressive algorithms have systematically better performance, and that in the small minority of cases where a more specialized algorithm yields better performance, its result does not deviate substantially from what is obtained with the more general approach. We finalize in Sec. VI with a discussion.

II. GENERATIVE MODELS FROM COMMUNITY DETECTION METHODS

We begin by considering a general community detection algorithm based on the maximization of an arbitrary quality function $W(\boldsymbol{A},\boldsymbol{b}) \in \mathbb{R}$,

$$\hat{\boldsymbol{b}}(\boldsymbol{A}) = \arg\max_{\boldsymbol{b}} W(\boldsymbol{A},\boldsymbol{b}), \tag{1}$$

where $\boldsymbol{A} = \{A_{ij}\}$ is the adjacency matrix of an undirected simple graph of N nodes, with entries $A_{ij} \in \{0,1\}$, and $\boldsymbol{b} = \{b_i\}$ is a partition of the nodes into B groups, with $b_i \in [1,\ldots,B]$ being the group membership of node i.

A direct connection with an inference procedure is obtained by noting that the above optimization is equivalent to a maximum a posteriori (MAP) estimate of the following family of posterior distributions:

$$P(\boldsymbol{b}|\boldsymbol{A},g) = \frac{\mathrm{e}^{g(W(\boldsymbol{A},\boldsymbol{b}))}}{Z(\boldsymbol{A},g)}, \tag{2}$$

with $Z(\boldsymbol{A},g) = \sum_{\boldsymbol{b}} \mathrm{e}^{g(W(\boldsymbol{A},\boldsymbol{b}))}$ being a normalization constant, and where $g(x)$ is any function that preserves the optimization, i.e.

$$\arg\max_{\boldsymbol{b}} g(W(\boldsymbol{A},\boldsymbol{b})) = \arg\max_{\boldsymbol{b}} W(\boldsymbol{A},\boldsymbol{b}), \tag{3}$$

for every $\boldsymbol{A}$, which in general means that $g(x)$ needs to be invertible and strictly increasing. Going one step further, we observe that the above posterior can be obtained from a general joint distribution given by

$$P(\boldsymbol{A},\boldsymbol{b}|g,f) = \frac{\mathrm{e}^{g(W(\boldsymbol{A},\boldsymbol{b}))+f(\boldsymbol{A})}}{Z(g,f)}, \tag{4}$$

with $Z(g,f) = \sum_{\boldsymbol{A},\boldsymbol{b}} \mathrm{e}^{g(W(\boldsymbol{A},\boldsymbol{b}))+f(\boldsymbol{A})}$, and $f(\boldsymbol{A})$ being an arbitrary weight attributed to a given network, independent of how its nodes are partitioned.

The above shows us that, although the quality function $W(\boldsymbol{A},\boldsymbol{b})$ imposes very particular constraints on the generative models that are compatible with it — specifically how the partitions can affect the network structure — they are by no means unique, since they are constrained only up to an invertible function $g(x)$ and an arbitrary partition-independent weight $f(\boldsymbol{A})$. Therefore, both $g(x)$ and $f(\boldsymbol{A})$ are "free" modelling choices that are not directly specified by the quality function $W(\boldsymbol{A},\boldsymbol{b})$. This is analogous to how descriptive statistics on numeric data such as the population mean can serve as sufficient statistics for the estimation of parameters of different generative models, e.g. the mean of geometric and Poisson distributions in the case of non-negative integers.

In view of this degeneracy, a reasonable starting point is to consider all compatible generative models on equal footing. We can formalize this lack of additional information about the data generating process by employing the principle of maximum entropy [33], subject to a minimal set of constraints. Considering the expected value of the quality function itself as the only parameter of the model, i.e.

$$\sum_{\boldsymbol{A},\boldsymbol{b}} W(\boldsymbol{A},\boldsymbol{b}) P(\boldsymbol{A},\boldsymbol{b}|g,f) = \langle W \rangle, \tag{5}$$

and maximizing the entropy $-\sum_{\boldsymbol{A},\boldsymbol{b}} P(\boldsymbol{A},\boldsymbol{b}) \ln P(\boldsymbol{A},\boldsymbol{b})$ subject to the above constraint, we obtain

$$P(\boldsymbol{A},\boldsymbol{b}|\beta) = \frac{\mathrm{e}^{\beta W(\boldsymbol{A},\boldsymbol{b})}}{Z(\beta)}, \tag{6}$$

with $Z(\beta) = \sum_{\boldsymbol{A},\boldsymbol{b}} \mathrm{e}^{\beta W(\boldsymbol{A},\boldsymbol{b})}$, and β being an "inverse temperature" Lagrange multiplier. Thus, the maximum entropy ansatz amounts to a choice $g(x) = \beta x$ and $f(\boldsymbol{A})$ being an arbitrary constant.

The above joint distribution yields a posterior probability for partitions,

$$P(\boldsymbol{b}|\boldsymbol{A},\beta) = \frac{\mathrm{e}^{\beta W(\boldsymbol{A},\boldsymbol{b})}}{\sum_{\boldsymbol{b}'} \mathrm{e}^{\beta W(\boldsymbol{A},\boldsymbol{b}')}}, \tag{7}$$

which has been used before by Massen and Doye [34] and Zhang and Moore [35], for the particular case of modularity, to investigate the ensemble of all competing partitions, rather than the single one that optimizes the quality function. Here we are more directly interested in the joint distribution of Eq. 6, for two reasons. The first one is that it generates problem instances for which the original community detection

method is optimal. More specifically, if we consider an estimator $\hat{\boldsymbol{b}}(\boldsymbol{A})$ for the partition of a network $\boldsymbol{A}$, and the average of the error $\epsilon(\boldsymbol{b}', \boldsymbol{b})$ between the true and inferred partitions over all problem instances,

$$\Lambda = \sum_{\boldsymbol{A},\boldsymbol{b}} \epsilon(\boldsymbol{b}, \hat{\boldsymbol{b}}(\boldsymbol{A}))P(\boldsymbol{A}, \boldsymbol{b}|\beta), \tag{8}$$

then the estimator is optimal if it minimizes Λ, in which case it must correspond to

$$\hat{\boldsymbol{b}}(\boldsymbol{A}) = \operatorname*{argmin}_{\boldsymbol{b}} \sum_{\boldsymbol{b}'} \epsilon(\boldsymbol{b}, \boldsymbol{b}')P(\boldsymbol{b}'|\boldsymbol{A}, \beta), \tag{9}$$

which is the estimate that minimizes the error over the posterior distribution conditioned on $\boldsymbol{A}$. In particular, for the "zero-one" error, $\epsilon(\boldsymbol{b}, \boldsymbol{b}') = 1 - \prod_i \delta_{b_i, b'_i}$, which simply identifies the correct answer and ignores all other ones, we recover the original optimization

$$\hat{\boldsymbol{b}}(\boldsymbol{A}) = \arg\max_{\boldsymbol{b}} P(\boldsymbol{b}|\boldsymbol{A}, \beta) \tag{10}$$

$$= \arg\max_{\boldsymbol{b}} W(\boldsymbol{A}, \boldsymbol{b}). \tag{11}$$

Therefore, according to this error criterion [36], for the problem instances sampled from Eq. 6 there exists no algorithm that can perform on average better than one that corresponds to the optimization of Eq. 1 (although it is still possible for alternative algorithms to perform just as well on the same instances). This gives us access to problem instances for which, in a formal sense, the results obtained with an arbitrary community detection algorithm are maximally correct. As we will show, we can use this information to investigate the implicit expected instances of arbitrary community detection algorithms.

A. The description length

In addition to the above, our second reason to focus on the joint distribution of Eq. 6 is that can be used to assess the overall statistical evidence for a particular partition of the network, and to enable comparison with alternative models. More precisely, from Eq. 6 we can compute the so-called *description length* [37, 38] of the data, defined as

$$\Sigma(\boldsymbol{A}, \boldsymbol{b}|\beta) = -\log_2 P(\boldsymbol{A}, \boldsymbol{b}|\beta) \tag{12}$$

$$= \underbrace{-\log_2 P(\boldsymbol{A}|\boldsymbol{b}, \beta)}_{\mathcal{S}} \underbrace{-\log_2 P(\boldsymbol{b}|\beta)}_{\mathcal{L}}. \tag{13}$$

The description length measures the size of the shortest binary message required to transmit both the partition $\boldsymbol{b}$ (with length $\mathcal{L}$) and network $\boldsymbol{A}$ (with length $\mathcal{S}$) over a noiseless channel, in such a manner that they can both be decoded from the message without errors, and assuming that the value of β is already known to the decoder. This connection exposes a fundamental equivalence between inference and compression, where the most likely model [largest $P(\boldsymbol{A}, \boldsymbol{b}|\beta)$] is also the most compressive [smallest $\Sigma(\boldsymbol{A}, \boldsymbol{b}|\beta)$]. The description length measures the degree of parsimony of the obtained network partition, allowing us to compare with alternative ones in what amounts to a formalization of Occam's razor. In the context of the SBM, the description length has been used as a criterion to perform order [39, 40] and model [41–43] selection, and here we extend this concept to arbitrary community detection algorithms.

From Eq. 6, we can obtain the description length for an arbitrary $W(\boldsymbol{A}, \boldsymbol{b})$ as follows (for convenience of notation, we will henceforth compute the description length using the natural base instead of base two, yielding values in nats instead of bits):

$$\Sigma(\boldsymbol{A}, \boldsymbol{b}|\beta) = -\beta W(\boldsymbol{A}, \boldsymbol{b}) + \ln Z(\beta) \tag{14}$$

$$= -\beta W(\boldsymbol{A}, \boldsymbol{b}) + \ln \sum_{\boldsymbol{A}',\boldsymbol{b}'} e^{\beta W(\boldsymbol{A}', \boldsymbol{b}')}. \tag{15}$$

(Note that we will always have $\Sigma(\boldsymbol{A}, \boldsymbol{b}|\beta) > 0$, regardless of our choice of $W(\boldsymbol{A}, \boldsymbol{b})$ and β.) The difficulty in obtaining $\Sigma(\boldsymbol{A}, \boldsymbol{b})$ lies in computing $Z(\beta)$, which is in general intractable, since it involves a sum over all networks and partitions. However, differently from the normalization of Eq. 7, we can make progress whenever the quality function $W(\boldsymbol{A}, \boldsymbol{b})$ can be expressed as a function of the edge counts between groups and group sizes, i.e.

$$W(\boldsymbol{A}, \boldsymbol{b}) = W(\boldsymbol{e}, \boldsymbol{n}), \tag{16}$$

with $e_{rs} = \sum_{ij} A_{ij}\delta_{b_i,r}\delta_{b_j,s}$ and $n_r = \sum_i \delta_{b_i,r}$. In this case we can perform the following change of variables,

$$\sum_{\boldsymbol{A},\boldsymbol{b}} \mathrm{e}^{\beta W(\boldsymbol{A},\boldsymbol{b})} = \sum_{B,\boldsymbol{e},\boldsymbol{n}} \mathrm{e}^{\beta W(\boldsymbol{e},\boldsymbol{n})}\Omega(\boldsymbol{e}, \boldsymbol{n}, B), \tag{17}$$

with $\Omega(\boldsymbol{e}, \boldsymbol{n}, B)$ being the microcanonical partition function of the SBM [44]

$$\Omega(\boldsymbol{e}, \boldsymbol{n}, B) = \sum_{\boldsymbol{A},\boldsymbol{b}} \prod_{r\le s} \delta_{\sum_{ij} A_{ij}\delta_{b_i,r}\delta_{b_j,s}, e_{rs}} \times \prod_r \delta_{\sum_i \delta_{b_i,r}, n_r} \tag{18}$$

$$= \prod_{r<s} \binom{n_r n_s}{e_{rs}} \prod_r \binom{\binom{n_r}{2}}{e_{rr}/2} \times \frac{N!}{\prod_r n_r!}. \tag{19}$$

The above computation makes it clear that whenever Eq. 16 holds, which happens to be true for many popular quality functions, then the overall approach can be seen as equivalent to the inference of a particular version of the SBM, with a specific weighting factor given by $W(\boldsymbol{e}, \boldsymbol{n})$.

Based on this parametrization, we can now decompose $Z(\beta)$ as

$$Z(\beta) = \sum_{B,\boldsymbol{e},\boldsymbol{n}} \mathrm{e}^{\beta W(\boldsymbol{e},\boldsymbol{n})}\Omega(\boldsymbol{e}, \boldsymbol{n}, B) \tag{20}$$

$$= \int \mathrm{e}^{\beta W}\Xi(W)\,\mathrm{d}W, \tag{21}$$

with

$$\Xi(W) = \sum_B \Xi(W, B), \tag{22}$$

being the β-independent entropic density of the quality function (which we call the "density of states"), where

$$\Xi(W, B) = \sum_{\boldsymbol{e},\boldsymbol{n}} \Omega(\boldsymbol{e}, \boldsymbol{n}, B)\delta(W(\boldsymbol{e}, \boldsymbol{n}) - W), \tag{23}$$

is the contribution for a particular number of groups B.

With $\Xi(W)$ at hand, the description length is then computed as

$$\Sigma(\boldsymbol{A}, \boldsymbol{b}|\beta) = -\beta W(\boldsymbol{A}, \boldsymbol{b}) + \ln \int \mathrm{e}^{\beta W} \Xi(W)\, \mathrm{d}W. \tag{24}$$

The parameter β is important since it determines the expected value of the quality function, so we will consider its optimal value with

$$\Sigma(\boldsymbol{A}, \boldsymbol{b}) = \min_{\beta} \Sigma(\boldsymbol{A}, \boldsymbol{b}|\beta). \tag{25}$$

(Strictly speaking, for the description length to be complete we would need to include the amount of information required to transmit the value of β up to a desired precision as well — but since this is a single global parameter, this will amount to an overall small constant that we can neglect.)

The computation above allows us to ascribe a description length to an arbitrary quality function $W(\boldsymbol{A}, \boldsymbol{b})$, and hence compare it with any other generative model in its relative ability to provide a plausible account for the data.

We note that if the quality function being used is already the joint log-likelihood of a generative model, i.e.

$$W(\boldsymbol{A}, \boldsymbol{b}) = \ln P(\boldsymbol{A}, \boldsymbol{b}), \tag{26}$$

then the above procedure will recover the original description length $\Sigma(\boldsymbol{A}, \boldsymbol{b}) = -\ln P(\boldsymbol{A}, \boldsymbol{b})$ for $\beta = 1$. The optimization of the parameter β may yield a marginal compression, which will vanish asymptotically if the data happens to be sampled from the same model.

B. Partition-independent compression

Any given posterior distribution $P(\boldsymbol{b}|\boldsymbol{A})$ is not uniquely associated with a description length, since the latter depends also on modelling choices that are independent of the relationship between network and partition. In fact, for any generative model $P(\boldsymbol{A}, \boldsymbol{b})$, we can devise an entire family of model alternatives determined up to an arbitrary exponential weight $f(\boldsymbol{A})$, i.e.

$$P'(\boldsymbol{A}, \boldsymbol{b}) = \frac{P(\boldsymbol{A}, \boldsymbol{b})\, \mathrm{e}^{f(\boldsymbol{A})}}{\sum_{\boldsymbol{A}',\boldsymbol{b}'} P(\boldsymbol{A}', \boldsymbol{b}')\, \mathrm{e}^{f(\boldsymbol{A}')}}, \tag{27}$$

all of which will result in the same posterior distribution for the partitions, $P(\boldsymbol{b}|\boldsymbol{A}) = P'(\boldsymbol{A}, \boldsymbol{b})/P'(\boldsymbol{A}) = P(\boldsymbol{A}, \boldsymbol{b})/P(\boldsymbol{A})$, independent of $f(\boldsymbol{A})$. Therefore, the choice of $f(\boldsymbol{A})$ will affect the description length (as well as predictive tasks such as link prediction [6, 10, 11]), but not the posterior for the node partitions, despite the corresponding model generating different networks. It is important to emphasize that the choice of $f(\boldsymbol{A})$ cannot significantly alter the community structure of the networks generated. We can see this by formulating the sampling of an instance of the model of Eq. 27 with the following rejection algorithm:

1. A pair $(\boldsymbol{A}, \boldsymbol{b})$ is sampled from the original $P(\boldsymbol{A}, \boldsymbol{b})$.
2. With probability $\mathrm{e}^{f(\boldsymbol{A})-f^*}$, where $f^* = \max_{\boldsymbol{A}} f(\boldsymbol{A})$, the sample is accepted, otherwise it is rejected and we go to step 1.

Therefore, the re-weighting of Ref. 27 will only suppress networks from the original ensemble in a manner that cannot take into account the node partition $\boldsymbol{b}$.

Although all models in the above family generate networks with the same kind of community structure, they can deviate with respect to other attributes that are uncoupled from this property. If these attributes happen to match more closely an observed network, this can be used to compress it further.

In the calculation of the previous section we used the principle of maximum entropy to fill this modelling gap, which yielded a constant value for $f(\boldsymbol{A})$. However, it is possible to deviate from this principle, and improve the description length by including properties we know to be ubiquitous. For example, we can introduce the exact number of edges E as an additional hard constraint,

$$\sum_{\boldsymbol{A},\boldsymbol{b}} P(\boldsymbol{A}, \boldsymbol{b}|g, f)\delta_{\sum_{i<j} A_{ij}, M} = \delta_{M,E}., \tag{28}$$

which if added to the entropy maximization yields

$$P(\boldsymbol{A}, \boldsymbol{b}|\beta, E) = \frac{\mathrm{e}^{\beta W(\boldsymbol{A},\boldsymbol{b})}\delta_{\sum_{i<j} A_{ij}, E}}{Z(\beta, E)}, \tag{29}$$

with $Z(\beta, E) = \sum_{\boldsymbol{A},\boldsymbol{b}} \mathrm{e}^{\beta W(\boldsymbol{A},\boldsymbol{b})}\delta_{\sum_{i<j} A_{ij}, E}$. To remove the parameter E we must introduce a uniform prior,

$$P(E) = \frac{1}{\binom{N}{2} + 1}, \tag{30}$$

obtaining thus an alternative joint likelihood via marginalization,

$$P(\boldsymbol{A}, \boldsymbol{b}|\beta) = \sum_{E} P(\boldsymbol{A}, \boldsymbol{b}|\beta, E)P(E) \tag{31}$$

$$= \frac{\mathrm{e}^{\beta W(\boldsymbol{A},\boldsymbol{b})}}{Z(\beta, \sum_{i<j} A_{ij})\left[\binom{N}{2} + 1\right]}. \tag{32}$$

The description length obtained with the joint distribution above will almost always be significantly shorter than what is obtained with Eq. 6, since the latter will sample networks which will tend to be dense — as long as the values of $W(\boldsymbol{A}, \boldsymbol{b})$ are not affected directly by the network density. We will use Eq. 31 in our ensuing analysis, instead of Eq. 6, since we will be considering only sparse networks. The density of

states in this case is computed in the same manner as before, but keeping the total number of edges fixed,

$$\Xi(W, B, E) = \sum_{\boldsymbol{e},\boldsymbol{n}} \Omega(\boldsymbol{e}, \boldsymbol{n}, B)\delta(W(\boldsymbol{e}, \boldsymbol{n}) - W)\delta_{\sum_{rs} e_{rs},2E}. \tag{33}$$

We can follow this route further and seek additional constraints that condition $f(\boldsymbol{A})$ to favor network patterns that are more likely to be encountered. For example, instead of constraining only the total number of edges, we can fix the entire degree sequence $\boldsymbol{k} = \{k_i\}$, where $k_i = \sum_j A_{ij}$ is degree of node i, i.e.

$$\sum_{\boldsymbol{A},\boldsymbol{b}} P(\boldsymbol{A}, \boldsymbol{b}|g, f)\delta_{\sum_j A_{ij},m_i} = \delta_{m_i,k_i}, \tag{34}$$

which will lead to

$$P(\boldsymbol{A}, \boldsymbol{b}|\beta, \boldsymbol{k}) = \frac{\mathrm{e}^{\beta W(\boldsymbol{A},\boldsymbol{b})}\prod_i \delta_{\sum_j A_{ij},k_i}}{Z(\beta, \boldsymbol{k})}, \tag{35}$$

with $Z(\beta, \boldsymbol{k}) = \sum_{\boldsymbol{A},\boldsymbol{b}} \mathrm{e}^{\beta W(\boldsymbol{A},\boldsymbol{b})}\prod_i \delta_{\sum_j A_{ij},k_i}$. Note that, now, instead of a single parameter, we have $N + 1$. To retain the same number of parameters as before, we need a prior for the degree sequence $\boldsymbol{k}$. One choice is a uniform model with

$$P(\boldsymbol{k}|E) = \left(\!\!\binom{N}{E}\!\!\right)^{-1}, \tag{36}$$

where $\left(\!\binom{n}{m}\!\right) = \binom{n+m-1}{m}$ is the number of n-tuples of non-negative integers whose sum is m. Another choice is a deeper Bayesian hierarchy with

$$P(\boldsymbol{k}|E) = P(\boldsymbol{k}|\boldsymbol{\eta})P(\boldsymbol{\eta}|E), \tag{37}$$

where $\boldsymbol{\eta} = \{\eta_k\}$ are the degree counts, i.e. $\eta_k = \sum_i \delta_{k_i,k}$, such that

$$P(\boldsymbol{k}|\boldsymbol{\eta}) = \frac{\prod_k \eta_k!}{N!}, \quad P(\boldsymbol{\eta}|E) = q(2E, N)^{-1}, \tag{38}$$

where $q(m, n)$ is the number of possible partitions of integer m into at most n parts, which can be calculated exactly via a recursion, or approximated accurately for large arguments, as described in Ref. [45]. The latter choice tends to provide a more parsimonious model for most empirical degree sequences, as long as they deviate sufficiently from a geometric degree distribution, which is (marginally) better described by Eq. 36 (see Ref. [45] for a discussion). With this prior in place, the final joint distribution becomes,

$$P(\boldsymbol{A}, \boldsymbol{b}|\beta) = \sum_{\boldsymbol{k},E} P(\boldsymbol{A}, \boldsymbol{b}|\beta, \boldsymbol{k})P(\boldsymbol{k}|E)P(E), \tag{39}$$

$$= \frac{\mathrm{e}^{\beta W(\boldsymbol{A},\boldsymbol{b})}\prod_k \hat{\eta}_k!}{Z(\beta, \hat{\boldsymbol{k}})q(\sum_{ij} A_{ij}, N)\left[\binom{N}{2} + 1\right]N!}, \tag{40}$$

where $\hat{k}_i = \sum_j A_{ij}$ and $\hat{\eta}_k = \sum_i \delta_{\hat{k}_i,k}$. In this case the SBM partition function is given by

$$\Omega(\boldsymbol{e}, \boldsymbol{n}, \boldsymbol{k}, B) = \sum_{\boldsymbol{A},\boldsymbol{b}} \prod_{r\le s} \delta_{\sum_{ij} A_{ij}\delta_{b_i,r}\delta_{b_j,s},e_{rs}} \times \prod_r \delta_{\sum_i \delta_{b_i,r},n_r} \times \prod_i \delta_{\sum_j A_{ij},k_i}, \tag{41}$$

which is unfortunately intractable to compute exactly [46]. However, it can be approximated by counting configurations [44, 45],

$$\Omega(\boldsymbol{e}, \boldsymbol{n}, \boldsymbol{k}, B) \approx \frac{\prod_r e_r!}{\prod_{r<s} e_{rs}! \prod_r e_{rr}!! \prod_i k_i!} \times \frac{N!}{\prod_r n_r!}. \tag{42}$$

which will yield an asymptotically exact enumeration as long as $k_i \ll \sqrt{N/B}$, and a still useful approximation otherwise.

The above alternative yields "degree-corrected" variants for the description length, which we will use in our analysis as well. Note that the above modification is different from the degree correction of the SBM [47], which correlates the degrees with the group memberships, and hence alters the posterior distribution [45]. The correction above changes the description length, but not the posterior distribution of partitions — all of the variations above remain fully equivalent to the original community detection ansatz of Eq. 1

One could in principle proceed indefinitely with adding partition-independent constraints that influence $f(\boldsymbol{A})$, together with prior distributions that keep the final distribution nonparametric — however, these quickly become very difficult to compute as soon as higher-order structures are considered. The above choices already amount to a careful effort to extract the largest amount of compression compatible with an arbitrary quality function, in accordance with what is typically done with state-of-the-art inferential methods based on the SBM [9].

C. Implicit priors and the role of the inverse temperature

From the joint distribution of Eq. 6 we can recover implicit priors via marginalization. For example, the marginal distribution for the value of the quality function is

$$P(W|\beta) = \sum_{\boldsymbol{A},\boldsymbol{b}} \delta(W(\boldsymbol{A}, \boldsymbol{b}) - W)P(\boldsymbol{A}, \boldsymbol{b}|\beta) \tag{43}$$

$$= \frac{\mathrm{e}^{\beta W}\Xi(W)}{Z(\beta)}. \tag{44}$$

Likewise, the prior for the number of groups can be obtained via

$$P(B|\beta) = \sum_{\boldsymbol{A},\boldsymbol{b}} \delta_{B(\boldsymbol{b}),B}P(\boldsymbol{A}, \boldsymbol{b}|\beta) \tag{45}$$

$$= \frac{\int \mathrm{e}^{\beta W}\Xi(W, B)\,\mathrm{d}W}{Z(\beta)}. \tag{46}$$

From the above equations we see that the inverse temperature β will influence both the expected number of groups, as well

as the values of the quality function. Notably, the conditional prior

$$P(B|W) = \frac{P(W,B|\beta)}{P(W|\beta)} = \frac{\Xi(W,B)}{\Xi(W)} \tag{47}$$

is β-independent.

For inferential methods based on the SBM [9] the priors above are set explicitly, usually in a non-informative manner to avoid biases during inference. Instead, for a given $W(\boldsymbol{A},\boldsymbol{b})$ these need to be back-engineered via the above computations.

We proceed now to the application of the above method for particular quality functions.

D. Modularity maximization

The generalized modularity quality function is given by [48]

$$Q(\boldsymbol{A},\boldsymbol{b},\gamma) = \frac{1}{2E}\sum_r e_{rr} - \gamma\frac{e_r^2}{2E}, \tag{48}$$

where γ is the so-called resolution parameter. The method of modularity maximization [15] consists of finding the partition that maximizes this quantity, typically with $\gamma = 1$.

As is required for our computation, modularity can be written solely as a function of the microcanonical SBM parameters, i.e. $Q(\boldsymbol{A},\boldsymbol{b},\gamma) = Q(\boldsymbol{e},\boldsymbol{n},\gamma)$, and we are interested in obtaining the density of states,

$$\Xi(Q,E) = \sum_{\boldsymbol{e},\boldsymbol{n}}\Omega(\boldsymbol{e},\boldsymbol{n})\delta(Q(\boldsymbol{e},\boldsymbol{n},\gamma)-Q)\delta_{2E,\sum_{rs}e_{rs}}. \tag{49}$$

An asymptotic estimate of this value can be obtained by rewriting

$$Q(\boldsymbol{e},\boldsymbol{n},\gamma) = \frac{E_{\text{in}}}{E} - \gamma\sum_r\frac{e_r^2}{(2E)^2}, \tag{50}$$

with $E_{\text{in}} = \sum_r e_{rr}/2$ being the edges internal to communities. Now we note that if we distribute edges uniformly between pairs of groups then, while keeping E_{in} fixed, we have that the values of e_r will be distributed according to a uniform multinomial, leading to an average value

$$\langle Q\rangle = \frac{E_{\text{in}}}{E} - \frac{\gamma}{B}\left(1+\frac{B-1}{2E}\right) \tag{51}$$

$$\approx \frac{E_{\text{in}}}{E} - \frac{\gamma}{B}, \tag{52}$$

for $E \gg 1$ and variance $\sigma_Q^2 \sim \frac{\gamma^2}{EB}$ to leading order in E and B, which can therefore be neglected in the same limit. Based on this, we can ignore fluctuations in Q and approximate the sum in Eq. 49 with (see Appendix A for details),

$$\Xi(Q,E) \approx \sum_B \Omega(E, E_{\text{in}}(Q,E,B,\gamma),B), \tag{53}$$

with

$$E_{\text{in}}(Q,E,B,\gamma) = E(Q+\gamma/B), \tag{54}$$

where Eq. 53 accounts for the number of partitioned networks with exactly E_{in} edges between nodes of the same groups, which can be computed as

$$\Omega(E,E_{\text{in}},B) = \sum_{\boldsymbol{e},\boldsymbol{n}}\Omega(\boldsymbol{e},\boldsymbol{n})\delta_{\sum_r e_{rr}/2,E_{\text{in}}}\delta_{\sum_{r<s}e_{rs},E-E_{\text{in}}} \tag{55}$$

$$\geq \sum_{\boldsymbol{e}}\left[\prod_{r<s}\binom{N^2/B^2}{e_{rs}}\prod_r\binom{\binom{N/B}{2}}{e_{rr}/2} \times\frac{N!}{[(N/B)!]^B} \times\delta_{\sum_r e_{rr}/2,E_{\text{in}}}\delta_{\sum_{r<s}e_{rs},E-E_{\text{in}}}\right] \tag{56}$$

$$= \binom{B\binom{N/B}{2}}{E_{\text{in}}}\binom{\frac{N^2}{B^2}\binom{B}{2}}{E-E_{\text{in}}}\frac{N!}{[(N/B)!]^B}. \tag{57}$$

Eq. 56 is a strict lower bound on the total sum, since it accounts only for partitions with equal size, however will asymptotically dominate it for large networks. To obtain Eq. 57 from Eq. 56 we simply used the generalized Vandermonde's identity,

$$\sum_{k_1+\cdots+k_p=m}\binom{n_1}{k_1}\cdots\binom{n_p}{k_p} = \binom{n_1+\cdots+n_p}{m}. \tag{58}$$

For the degree-corrected version of modularity we have instead

$$\Omega(E,E_{\text{in}},\boldsymbol{k},B) = \sum_{\boldsymbol{e},\boldsymbol{n}}\Omega(\boldsymbol{e},\boldsymbol{n},\boldsymbol{k})\delta_{\sum_r e_{rr}/2,E_{\text{in}}}\delta_{\sum_{r<s}e_{rs},E-E_{\text{in}}} \tag{59}$$

$$\gtrsim \sum_{\boldsymbol{e}}\left[\frac{[(2E/B)!]^B}{\prod_{r<s}e_{rs}!\prod_r e_{rr}!!\prod_i k_i!} \times\frac{N!}{[(N/B)!]^B} \times\delta_{\sum_r e_{rr}/2,E_{\text{in}}}\delta_{\sum_{r<s}e_{rs},E-E_{\text{in}}}\right] \tag{60}$$

$$= \frac{[(2E/B)!]^B B^{E_{\text{in}}}\binom{B}{2}^{E-E_{\text{in}}}N!}{(2E_{\text{in}})!!(E-E_{\text{in}})![(N/B)!]^B\prod_i k_i!}, \tag{61}$$

where in the last step we have used the multinomial theorem,

$$\sum_{k_1+\cdots+k_p=m}\frac{m!}{\prod_{i=1}^p k_i!}\prod_{i=1}^p x_i^{k_i} = \left(\sum_{i=1}^p x_i\right)^m. \tag{62}$$

With the density of states at hand, we can obtain the description length according to Eq. 24, which involves a sum

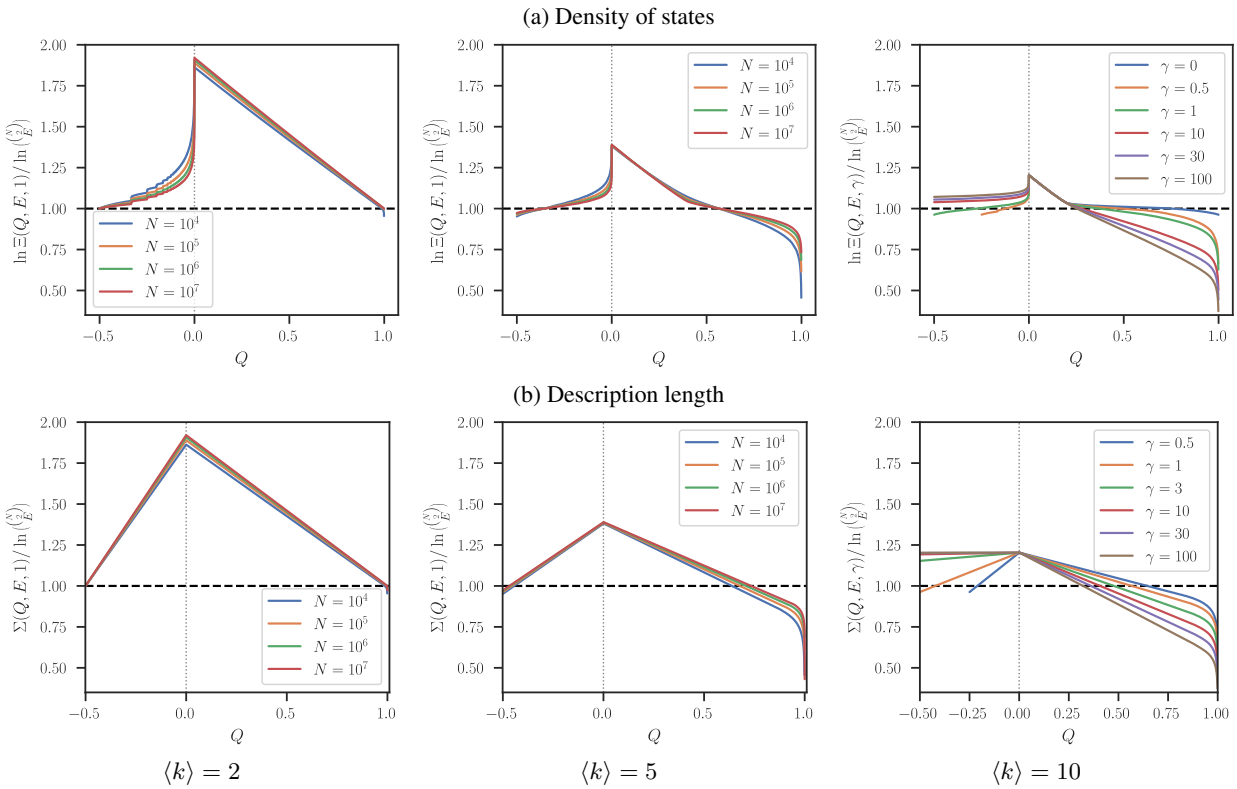

Figure 1. (a) Density of states $\Xi(Q, E, \gamma)$ and (b) description length $\Sigma(Q, E, \gamma)$, as a function of the value of modularity Q, for different number of nodes N and average degree values, $\langle k \rangle = 2, 5, 10$, from left to right. The values are shown relative to the ER baseline. The description length in particular tells us what should be considered a statistically significant modularity value.

over B in Eq. 53 and an integral over $W = Q$, both of which can be done efficiently numerically, up to an arbitrary precision.

In Fig. 1 we see the result of the above computation for some network sizes and densities. [We focus for the moment on the non-degree-corrected version, although the degree-corrected variants are qualitatively very similar (not shown).] It shows the density of states and description length values relative to the ER baseline

$$\Sigma_{ER} = \ln \binom{\binom{N}{2}}{E}. \tag{63}$$

Therefore, a value smaller than this would amount to a compression relative to a fully random model, pointing thus to statistically significant structure. The values shown on the bottom row of Fig. 1 offer us an important mapping from Q values — which by themselves cannot be interpreted statistically — to description length values. The latter quantities allow for an information-theoretical evaluation of the statistical significance and degree parsimony for Q values obtained with modularity maximization algorithms. As can be seen in Fig. 1, we often obtain *inflation* for intermediary values of Q — which therefore would indicate overfitting — and compression only for relatively high values. The compression region becomes larger for denser networks (for $\langle k \rangle = 2$ compression is impossible for most Q values), which is also anticipated by higher values of the resolution parameter γ.

An important aspect of our analysis is that they allow us to understand the implicit prior assumptions that are intrinsic to modularity maximization, as we show in Fig. 2. As seen in panels (a) and (b), both the prior for the modularity value, $P(Q|\beta)$, and the number of groups, $P(B|\beta)$, are extremely informative and bimodal, concentrating very strongly on particular high and low values. The value of β determines which mode dominates, inducing a discontinuous transition at a particular value β^* for the mean values $\langle Q \rangle$ and $\langle B \rangle$, as we can see in panels (c) and (d). This kind of transition is reminiscent of the degeneracy encountered in exponential random graphs models [49, 50], where the ensemble mean of an enforced constraint results in bimodal distributions, where no typical sample from the ensemble obeys the enforced constraint. Importantly, this kind of prior assumption is hardly justified in most applications in the absence of substantial additional evidence supporting it. The case of strict modularity maximization, where we are interested only in the partition that maximizes the posterior of Eq. 7, amounts to the situation $\beta \to \infty$, where prior modularity values concentrate on $Q = 1$ and $B \propto N$, explaining the tendency of the method to overfit, which is only avoided only if the evidence in the data is sufficiently strong to contradict the prior assumptions.

We can further understand the behavior of modularity maximization via the conditional prior $P(B|Q)$, which is β-independent, seen in Fig. 2e. The range of large Q values shows an intuitive behavior: as Q increases, so does the ex-

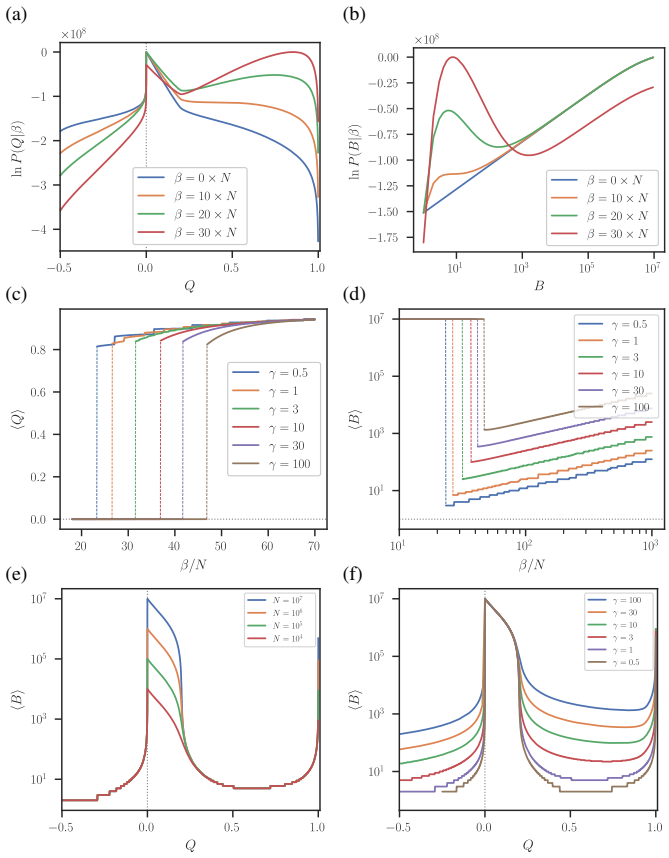

Figure 2. Implicit priors for the method of modularity maximization. Top row: (a) Implicit prior distribution for the value of modularity Q, and (b) number of groups B, for different values of β, $\gamma = 1$, $N = 10^7$ and $\langle k \rangle = 10$. Middle row: Average values of (c) Q and (d) B, as a function of β, and different values of γ. Bottom row: (e) Average value of B as a function of Q for different values of N and $\gamma = 1$, and (f) the same as (e) but with $N = 10^7$ only and different values of γ.

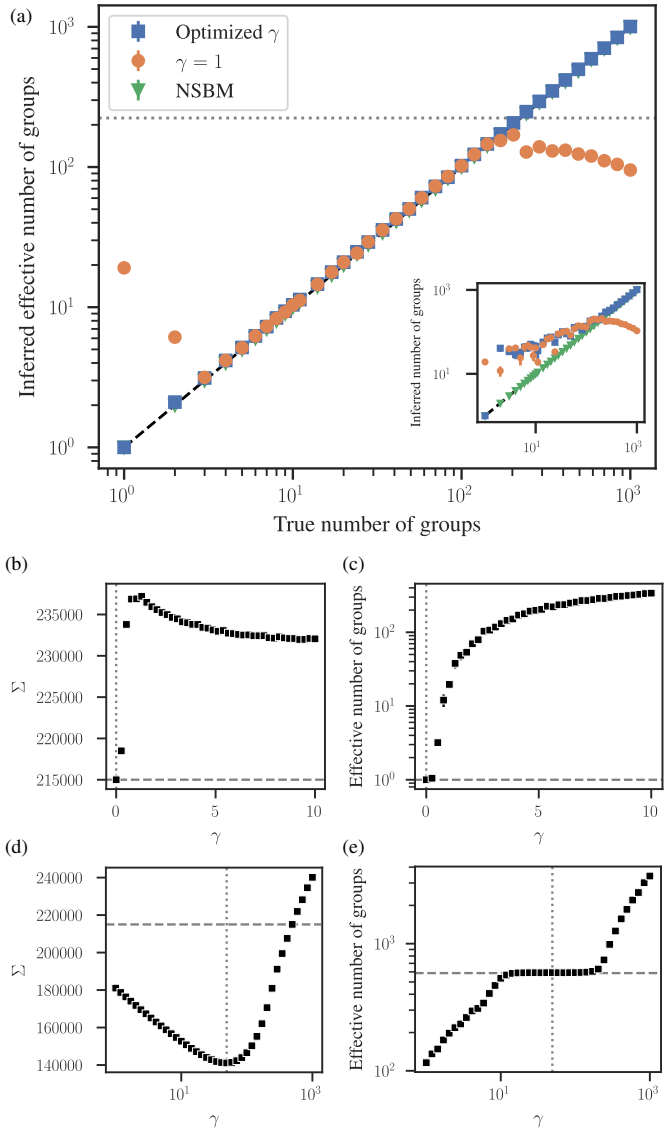

Figure 3. Computing the description length can alleviate the overfitting and underfitting (resolution limit) of the modularity maximization method. Panel (a) shows the inferred effective number of groups $B_e = \exp(-\sum_r \frac{n_r}{N} \ln \frac{n_r}{N})$, as a function of the true number of groups B, for networks sampled from a PP model with uniform group sizes, $E_{\text{in}} = E - (B-1)\langle k \rangle$, for $N = 10^4$ and $\langle k \rangle = 5$, obtained using modularity maximization with $\gamma = 1$ and the value of γ that minimizes the description length, as shown in the legend. It also shows the results obtained with the NSBM. The horizontal dashed line marks the value $\sqrt{2E}$. The inset shows the inferred number of non-empty groups, instead of the effective number. (b) Description length versus γ for $B = 1$, for the same networks as in (a). The dashed vertical line marks the minimum value, and the horizontal line the description length of the ER model. (c) Effective number of groups for the same networks as in (b). The horizontal line marks the planted value. The panels (d) and (e) are analogous to (b) and (c), but with $B = 587$.

pected number of groups. However, the same happens for low Q values approaching zero. This contradicts the intuition that low Q values, specially $Q = 0$, would amount to small or negligible community structure. What is occurring here is that for low Q the density of states is dominated by the contribution of the node partitions, which is largest for $B = O(N)$, since there are many networks that admit a low Q with an arbitrary partition. As soon as Q increases, the contribution of the actual network structure dominates instead, since relatively fewer networks allow for a high Q partition, and forces the number of groups to decrease, before increasing again. This tension between the partition and network entropic contributions also explains the transitions between the low Q and divergent B, and high Q and finite B regimes observed as a function of β.

The behavior above also explains the tendency of the modularity method to simultaneously overfit (i.e. when it finds spurious communities) and underfit, i.e. when the number of groups exceeds the $\sqrt{\gamma 2E}$ resolution limit [51] it merges groups together. In fact, we can use the value of description length to correct for both these effects via the param-

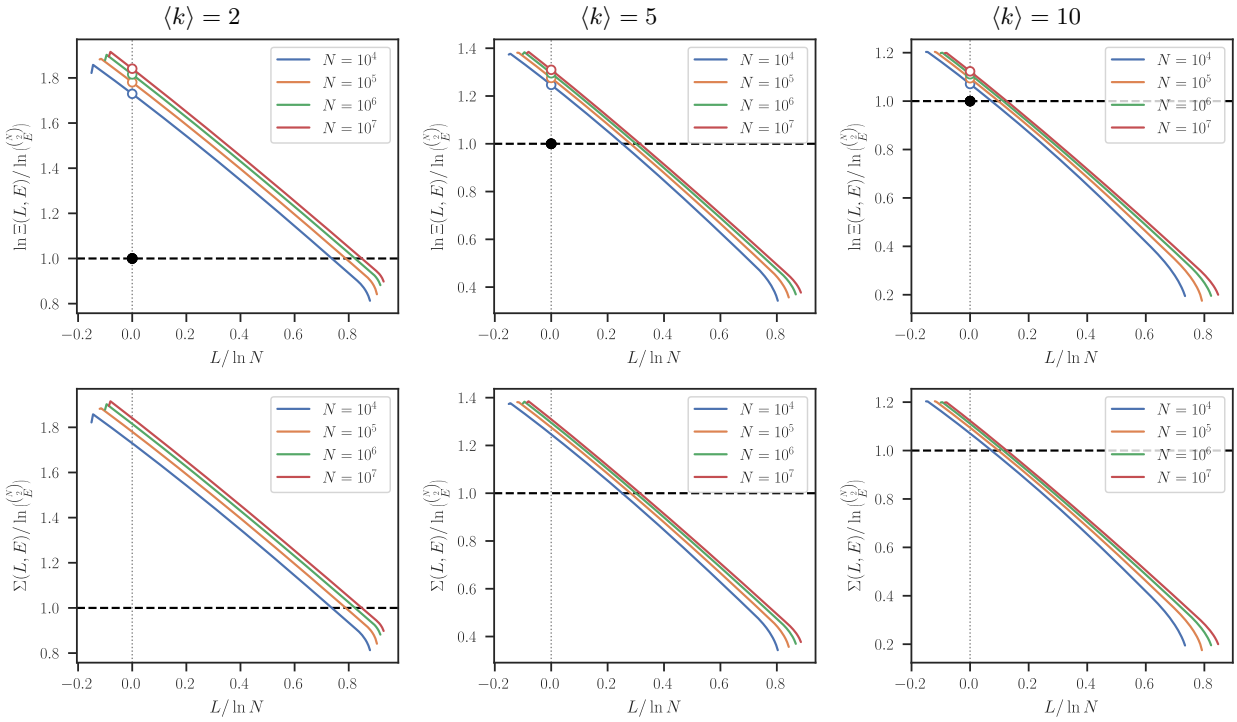

Figure 4. Density of states $\Xi(L, E)$ (top row) and description length $\Sigma(L, E)$ (bottom row) as a function of value of Infomap score L, for different number of nodes N and average degree values, $\langle k\rangle = 2, 5, 10$, from left to right. The values are shown relative to the ER baseline.

eter γ by choosing the value that most compresses the network, as shown in Fig. 3. While using a value of $\gamma = 1$ finds spurious groups whenever the true number of planted groups is small, and too few groups whenever the true number lies above $\sqrt{2E}$, the most compressive γ values reveals the correct number throughout the entire range, thus removing a long-standing limitation of this method.

Although the above approach serves as principled, unified and non-parametric solution to the overfitting and resolution limit problems of modularity maximization, we emphasize that are other problems intrinsic to the method that remains. In particular, optimizing γ yields an effective number of groups, computed as

$$B_e = \exp\left(-\sum_r \frac{n_r}{N}\ln\frac{n_r}{N}\right), \tag{64}$$

which lies very close to the true value, but the actual number of inferred groups is often larger, as shown in the inset of Fig. 3a. This is because the value of Q, and as a consequence its description length encoding, are insensitive to the existence of very small groups, therefore some marginal amount of overfitting cannot be fully removed. More importantly, the method will still enforce a characteristic scale for the community sizes, and will not behave well when communities of unequal sizes exist [52]. The computation and minimization of the description length can be seen as a "post-processing" of the results obtained with modularity maximization, and it can only influence the intrinsic biases of the method via a free parameter like γ. A more direct strategy to tackle the vices of the method involves a more appropriate formulation the prior assumptions, precisely as is done with the SBM-based approaches [9, 18]. In Fig. 3 we show the result obtained with the nested stochastic block model (NSBM) [45, 53], discussed in more detail in Sec. IV A, which has no difficulty in finding not only the effective number of groups, but also its nominal value.

E. Infomap

The Infomap quality function [13] is given by

$$\begin{aligned} L(\boldsymbol{A}, \boldsymbol{b}) = &-\left(1-\sum_r \frac{e_{rr}}{2E}\right)\ln\left(1-\sum_r \frac{e_{rr}}{2E}\right) \\ &+ 2\sum_r \frac{e_r - e_{rr}}{2E}\ln\left(\frac{e_r - e_{rr}}{2E}\right) - H(\boldsymbol{k}) \\ &- \sum_r \left(\frac{2e_r - e_{rr}}{2E}\right)\ln\left(\frac{2e_r - e_{rr}}{2E}\right), \end{aligned} \tag{65}$$

where $H(\boldsymbol{k})$ is the entropy of the normalized degree distribution. (We have flipped the sign so that the optimal partition is obtained through maximizing the objective, consistent with Eq. 1.) Since it only depends on $\boldsymbol{A}$ but not $\boldsymbol{b}$, we can ignore this degree entropy term, since it will disappear when doing the normalization, to obtain an objective that only depends on the SBM parameters.

The computation of the density of states is analogous to modularity (see Appendix A), with the only difference that

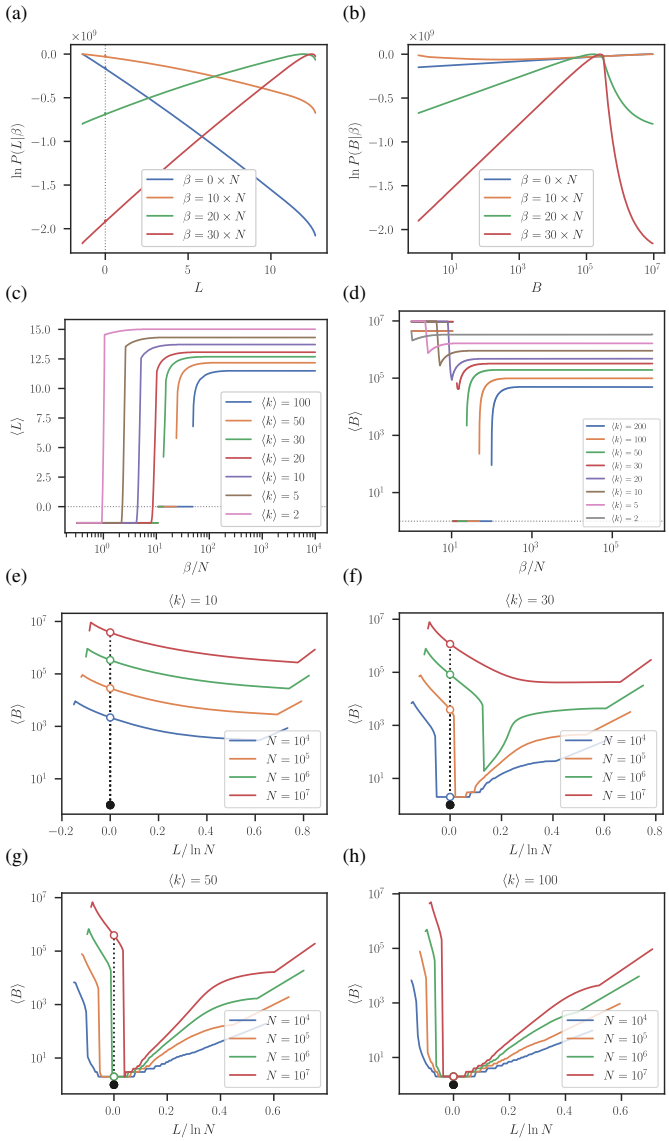

Figure 5. Top row: Implicit prior distribution for the value of Infomap quality function L (a), and number of groups B (b), for different values of β, $N = 10^7$ and $\langle k \rangle = 30$. Second row: Average values of L (c) and B (d) as a function of β, and different average degrees $\langle k \rangle$. Bottom rows (e) to (h): Average value of B as a function of Q for different values of N and $\langle k \rangle$.

for the planted partition we have:

$$L(E, E_{\text{in}}, B) = -\frac{E - E_{\text{in}}}{E} \ln \frac{E - E_{\text{in}}}{E} + 2\frac{E - E_{\text{in}}}{E} \ln \frac{E - E_{\text{in}}}{EB} - \frac{2E - E_{\text{in}}}{E} \ln \frac{2E - E_{\text{in}}}{EB}. \quad (66)$$

Following the same procedure as before, we can now invert Eq. 66 for E_{in} to give $E_{\text{in}}(E, L, B)$, which can be inserted into either Eq. 57 or Eq. 61 to obtain the description length according to Eq. 24. Unlike the modularity function, the inversion of Eq. 66 cannot be done in closed form, so it needs to be performed numerically.

Fig. 4 shows the density of states and description length values as a function of L. Unlike modularity, the relationship between these two quantities is almost linear. In Fig. 5 we see the implicit priors for L and the number of groups B — we also observe a transition from low to high values with β, which although abrupt is continuous, unlike what is obtained for modularity. In the case of Infomap, what is noteworthy is a qualitative dependence on the network density — only if the average degree is sufficiently large does the prior for B allows for finite values, as seen in Fig. 5d, otherwise the mean is always at $\langle B \rangle = O(N)$ (the precise value of $\langle k \rangle$ at which this transition happens is size-dependent). This transition is reflected in the expected value of B as a function of L, which displays a minimum at $L = 0$ only for sufficiently dense networks, besides a discontinuity at $L = 0$, since for this value only a partition in $B = 1$ groups is allowed. This overall picture is entirely consistent with the observed tendency of the method to find spurious groups in fully random networks whenever they are sufficiently sparse [29, 54].

III. OPTIMAL PROBLEM INSTANCES

As discussed previously, problem instances $(\boldsymbol{A}, \boldsymbol{b})$ sampled from the distribution

$$P(\boldsymbol{A}, \boldsymbol{b}|\beta) = \frac{\mathrm{e}^{\beta W(\boldsymbol{A},\boldsymbol{b})}}{Z(\beta, \sum_{i<j} A_{ij}) \left[\binom{N}{2} + 1\right]}, \quad (67)$$

are optimal for a community detection algorithm that maximizes the quality function $W(\boldsymbol{A}, \boldsymbol{b})$, since no other algorithm can achieve better average performance on those instances. If a quality function can be written in terms of the microcanonical SBM parameters $W(\boldsymbol{A}, \boldsymbol{b}) = W(\boldsymbol{e}, \boldsymbol{n})$, then it can be interpreted as being proportional to the log-likelihood of a particular constrained version of the SBM. We can see this by approximating

$$Z(\beta, E) = \int e^{\beta W} \Xi(W, E)\, \mathrm{d}W \approx e^{\beta W^*} \Xi(W^*, E), \quad (68)$$

with $W^* = \arg\max_W e^{\beta W} \Xi(W, E)$, such that

$$P(\boldsymbol{A}, \boldsymbol{b}|\beta) \approx \frac{\mathrm{e}^{\beta[W(\boldsymbol{e},\boldsymbol{n}) - W^*]}}{\Xi(W^*, \sum_{i<j} A_{ij})}. \quad (69)$$

Approximating further

$$\Xi(W^*, E) = \sum_B \Xi(W^*, B, E) \approx \Xi(W^*, B^*, E) \quad (70)$$

with $B^* = \arg\max_B \Xi(W^*, B, E)$, and neglecting finite-size fluctuations around the most typical samples with $W(\boldsymbol{e}, \boldsymbol{n}) = W^*$, we can write the likelihood as

$$P(\boldsymbol{A}, \boldsymbol{b}|\beta) \approx \frac{\delta_{W(\sum_r e_{rr}/2, \sum_{i<j} A_{ij}, B^*), W^*}}{\Xi(W^*, B^*, \sum_{i<j} A_{ij}) \left[\binom{N}{2} + 1\right]}. \quad (71)$$

where $W(E_{\text{in}}, E, B)$ is the value of the quality function for exactly E_{in} edges internal to equal-sized communities. Rearranging, we have

$$P(\boldsymbol{A}, \boldsymbol{b}|\beta) \approx P(\boldsymbol{A}|E^*_{\text{in}}, E, \boldsymbol{b})P(\boldsymbol{b}|B^*)P(E), \tag{72}$$

where E^*_{in} is the solution of

$$W(E_{\text{in}}, E, B^*) = W^*, \tag{73}$$

and

$$P(\boldsymbol{A}|E_{\text{in}}, E, \boldsymbol{b}) = \frac{\delta_{\sum_{i<j} A_{ij}\delta_{b_i,b_j},E_{\text{in}}}\delta_{\sum_{i<j} A_{ij},E}}{\left(\sum_r \binom{n_r}{2}_{E_{\text{in}}}\right)\left(\sum_{r<s} n_r n_s \atop E-E_{\text{in}}\right)} \tag{74}$$

is the likelihood of a microcanonical planted partition SBM with exactly E_{in} edges internal to communities, and

$$P(\boldsymbol{b}|B) = \frac{\prod_r \delta_{n_r,N/B}}{N!/[(N/B)!]^B} \tag{75}$$

is the likelihood of a random partition into B groups of the same size, and finally $P(E) = \left[\binom{N}{2} + 1\right]^{-1}$. The values of W^* and B^* are uniquely determined by β with

$$W^* = \arg\max_W \; e^{\beta W}\Xi(W, E) \tag{76}$$

$$B^* = \arg\max_B \; \Xi(W^*, B, E). \tag{77}$$

Therefore, the model of Eq. 67 is asymptotically equivalent to sampling a network from a planted partition SBM with the number of groups and assortativity strength determined by the same β parameter.

The above equivalence is a more general, but compatible nonparametric version of the approximate one shown for modularity in Ref. [17]. That work showed that if both the number of groups and the planted partition mixing parameter are known and fixed, and if the partitions have equal size and density [18], then the maximum likelihood of the degree-corrected planted partition model is approximately the same as the maximum modularity one with a particular value of γ. In contrast, the model we derive above is nonparametric, i.e. generates in addition to the network also the number of groups, partition, and mixing strength, and does not rely on any assumptions on the data. Crucially, unlike the model of Ref. [17], from ours we can compute the description length of the data.

In Fig. 6 we show some example networks sampled from the optimal model for modularity maximization, for various values of β. As discussed previously, for a small value of β the model concentrates on low Q values with diverging $B \propto N$, and undergoes a discontinuous transition at value $\beta = \beta^*$, after which it concentrates on high Q values with a finite B. An example of this transition is shown in Fig. 7 via the joint probability $P(E_{\text{in}}, B|\beta) = \mathrm{e}^{\beta Q(E_{\text{in}},E,\gamma,B)}\Xi(W, B, E)/Z(\beta)$.

Note that for a single value of γ there is no way to independently control the number of groups and strength of community structure. However, we might imagine that setting the value of the resolution parameter γ would allow for a precise tuning of the strength of assortativity E_{in} together with any

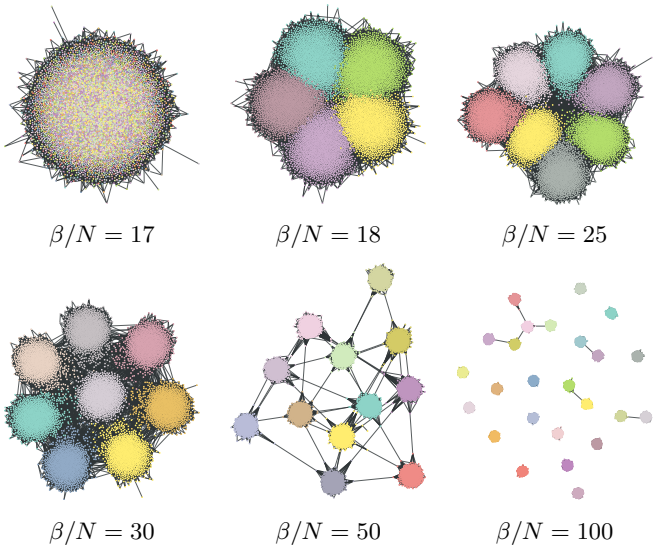

Figure 6. Samples from the implicit generative model behind modularity maximization with $\gamma = 1$, for different inverse temperature values β, $N = 10^4$ and $\langle k\rangle = 10$. The colors indicate the sampled network partitions. For these problem instances, the method of modularity maximization is Bayes-optimal.

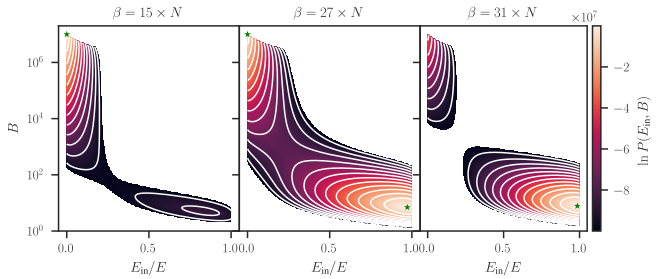

Figure 7. Joint probability $P(E_{\text{in}}, B|\beta)$ for the modularity model, with $N = 10^7$, $\langle k\rangle = 10$, $\gamma = 1$, and different values of β. The global maxima of the distribution are marked with star symbols. As the value of β increases, the global maximum changes abruptly from a value close to $(E_{\text{in}}, B) = (0, N)$ to a value with large E_{in} and finite B.

arbitrary number of groups B — in other words, we could expect a bijection between (β, γ) and (E_{in}, B), up to discretization. In reality, however, a wide range of (E_{in}, B) values is not achievable for any combination of (β, γ), as we show in Fig. 8. Indeed, the model is only capable of generating networks with quite strong community structure, far away from the detectability threshold of the plated partition model, which lies at

$$\frac{E^\star_{\text{in}}}{E} = \frac{1}{B} + \frac{B-1}{B\sqrt{\langle k\rangle}}. \tag{78}$$

For any network sampled from the PP model with $E_{\text{in}} < E^\star_{\text{in}}$, it is not possible with any algorithm to recover any information about the true partition [55]. As we see in Fig. 8, the optimal model for modularity only generates networks with E_{in} much larger than $E^\star_{\text{in}}$ — except for a small fraction of (β, γ)

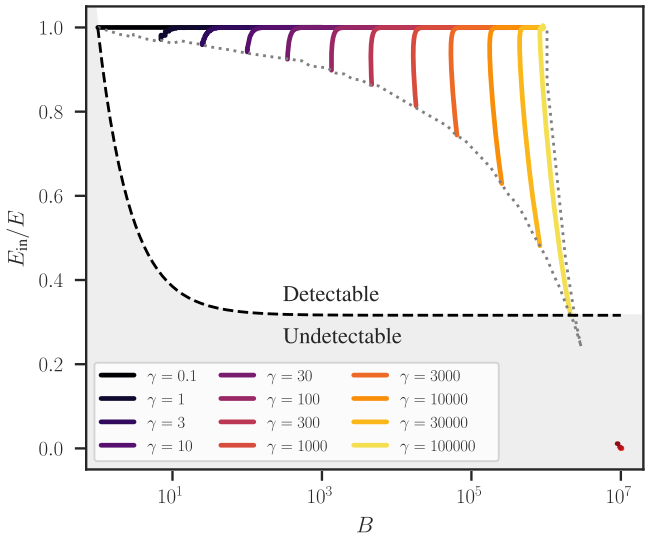

Figure 8. Feasible realizations of the modularity model. Each curve corresponds to the $(E_{\rm in}, B)$ values achieved with β in the range $[0, \infty]$ for a specific value of γ, as indicated in the legend, $N = 10^7$, and $\langle k \rangle = 10$. The dotted line delineates the feasible region for any parameter value. The dashed line marks the detectability transition of Eq. 78.

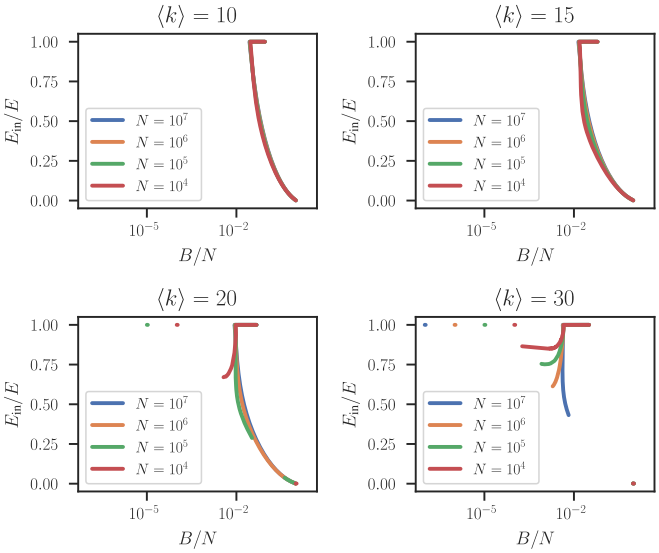

Figure 9. Feasible realizations of the Infomap model. Each curve corresponds to the $(E_{\rm in}, B)$ values achieved with β in the range $[0, \infty]$ for a specific value of N, as indicated in the legend, and various values of the average degree $\langle k \rangle = 2E/N$. The isolated points correspond to discontinuous transitions both for high and low values of β.

combinations that lead to very large B values. However, the undetectable regime (and hence also the detectability transition) only exists in the limit $B/N \to 0$, and the values of B for which we obtain $E_{\rm in} < E^\star_{\rm in}$ scale proportionally with N as it increases (not shown). Therefore, it is not possible to generate an undetectable community structure with this model, other than by setting $\beta < \beta^*$, in which case the networks generated are maximally random and uncorrelated with the node partitions.

The result above is not entirely surprising, since it is known that modularity maximization is not an optimal algorithm for networks sampled close to the detectability transition of the PP model, since it already fails for easier problem instances [56]. If it were possible to generate such hard realizations with the above optimal model for modularity, it would lead to a contradiction.

For Infomap, the situation is comparable. As we show in Fig. 9, the lack of an additional parameter analogous to the resolution γ of modularity means that the value of β can only select values on a line in the $(E_{\rm in}, B)$ plane. We can observe two regimes: 1. For sufficiently sparse networks, although a wide range of $E_{\rm in}$ can be reached, we cannot meaningfully talk about an undetectable regime because B is proportional to N — all instances are easy; 2. For denser networks, a discontinuous transition is observed between a $(E_{\rm in}, B) = (E, 1)$ value and another range of values far away from the detectability transition. The transition between these regimes is size dependent, such that as the number of nodes increases, then even denser networks are required for the transition between the above two regimes to be seen.

Overall, we see that the optimal instances for both methods considered — modularity maximization and Infomap — are quite contrived, and composed of unrealistically strong community structure, resulting in relatively easy labelling tasks, as we will see in the following section. They are also unrealistic in their regularity, with a maximally homogeneous community structure composed of equal-sized groups that also have the same density. Although the respective optimization algorithms are optimal for these instances, it is very likely that other algorithms will work just as well for them too. In the following, we demonstrate that more general algorithms indeed perform just as well in these instances, but the opposite is not true: neither Infomap nor modularity maximization perform well with instances that are optimal to the more general algorithm.

IV. "CHEAP LUNCHES"

Recently, the notion of universal algorithms for community detection has been challenged by a "no free lunch" (NFL) theorem [32], which states that when averaged over all instances of community detection problems, all conceivable algorithms must yield the same performance. This would mean, therefore, that no algorithm can be truly universal, and that for one algorithm to behave better than another on a subset of the problem instances, then it must do worse on the remaining instances in a complementary fashion. However, digging only slightly below the surface of the statement of the NFL theorem of Ref. [32] reveals that it in fact tells us very little about the kind of problems that virtually any community detection method attempts to solve. As stated previously, despite their different mathematical definitions, most methods

attempt to divide networks into groups of nodes with more internal than external connections, or more generally, according to arbitrary preferences of connection between groups. In spite of this, the class of problems considered in Ref. [32] completely violates this qualitative constraint, and considers instead as equally valid instances of a community detection problem any arbitrary pairing of a network and a true node partition that an algorithm needs to find to be maximally accurate — regardless of how the nodes are actually divided in this partition and how this division relates to the structure of the network. In fact, most such problem instances are *unstructured*, in a formal sense, since they correspond to maximally random networks with nodes divided in equally maximally random partitions, in violent disagreement with almost every notion of community structure in the entire literature on the topic.

In more detail, the NFL theorem states that, given an arbitrary deterministic community detection algorithm indexed by f which ascribes a partition $\hat{\boldsymbol{b}}_f(\boldsymbol{A})$ to a network $\boldsymbol{A}$, and an appropriately chosen error function $\epsilon(\boldsymbol{b}, \boldsymbol{b}')$, then we must have

$$\sum_{\boldsymbol{A},\boldsymbol{b}} \epsilon(\hat{\boldsymbol{b}}_f(\boldsymbol{A}), \boldsymbol{b}) = \Lambda_\epsilon, \tag{79}$$

where Λ_ϵ is a constant that does not depend on the chosen algorithm f, only on the error function $\epsilon(\boldsymbol{b}, \boldsymbol{b}')$. In other words, when summed over all possible pairs $(\boldsymbol{A}, \boldsymbol{b})$, all algorithms must have the same performance. Crucially, the sum above does not necessarily involve pairs $(\boldsymbol{A}, \boldsymbol{b})$ which correspond to a partitioned network with any actual community structure — regardless of how one defines it — they are entirely arbitrary. In fact, we can re-write the statement of the theorem using a probabilistic language, thus

$$\sum_{\boldsymbol{A},\boldsymbol{b}} P(\boldsymbol{A}, \boldsymbol{b}) \epsilon(\hat{\boldsymbol{b}}_f(\boldsymbol{A}), \boldsymbol{b}) \propto \Lambda_\epsilon, \tag{80}$$

where the joint probability is trivially uniform and hence uncorrelated, i.e.

$$P(\boldsymbol{A}, \boldsymbol{b}) = P(\boldsymbol{A})P(\boldsymbol{b}), \tag{81}$$

$$P(\boldsymbol{A}) \propto 1, \tag{82}$$

$$P(\boldsymbol{b}) \propto 1. \tag{83}$$

Indeed, in this situation an uniformity between algorithms is entirely unsurprising, since the posterior distribution is maximally uniform $P(\boldsymbol{b}|\boldsymbol{A}) = P(\boldsymbol{b}) \propto 1$, and the Bayes-optimal algorithm amounts to simply selecting a random partition uniformly at random, ignoring the network altogether. The best possible algorithm will achieve a minimal accuracy corresponding to a blind random guess, and hence Λ_ϵ will correspond to the maximal possible value for every algorithm. Since all algorithms perform maximally poorly, there is no actual trade-off between them in this scenario [8] — in contrast to how the NFL theorem is sometimes interpreted [6, 57].

The vast majority of problem instances sampled from the uniform distribution are incompressible, i.e. cannot be described using fewer bits than what is prescribed by the uniform distribution, and hence correspond to *unstructured* problem instances. Crucially, the subset of structured problem instances, i.e. a network with actual community structure — again, regardless of how one precisely defines it — has an asymptotic measure of zero with respect to the set of all instances, i.e. the probability of encountering them when sampling from the uniform distribution will vanish rapidly as the size of the data increases [58]. Therefore, the statement of Eq. 79 tells us very little about actual community detection problems, which in order to be structured, need to be compressible. (The same can be said about other kinds of NFL theorems, outside of community detection [59–64].)

Importantly, the NFL theorem does not imply that there is a performance equivalence between algorithms when they are faced with structured problem instances. Using our understanding of the connection between descriptive community detection objectives and implicit network generative models, here we address this issue and demonstrate that for structured problem instances, there are asymmetries where more general approaches can outperform more specialized ones, without degrading the performance in more specific instances.

Let us consider two alternative distributions of problem instances, $P(\boldsymbol{A}, \boldsymbol{b})$ and $Q(\boldsymbol{A}, \boldsymbol{b})$. We can quantify the ability of model $Q(\boldsymbol{A}, \boldsymbol{b})$ to capture the structure of instances sampled from a model $P(\boldsymbol{A}, \boldsymbol{b})$ via the Kullback-Leibler (KL) divergence from Q to P,

$$D_{\mathrm{KL}}(P||Q) = \sum_{\boldsymbol{A},\boldsymbol{b}} P(\boldsymbol{A}, \boldsymbol{b}) \ln \frac{P(\boldsymbol{A}, \boldsymbol{b})}{Q(\boldsymbol{A}, \boldsymbol{b})} \tag{84}$$

$$= \sum_{\boldsymbol{A},\boldsymbol{b}} P(\boldsymbol{A}, \boldsymbol{b}) \left[\Sigma_Q(\boldsymbol{A}, \boldsymbol{b}) - \Sigma_P(\boldsymbol{A}, \boldsymbol{b})\right], \tag{85}$$

which in this context measures the average description length difference according to models Q and P, for problem instances sampled from P. Note that the KL divergence is strictly positive, $D_{\mathrm{KL}}(P||Q) \geq 0$, with the equality attainable only for $P = Q$. Therefore, it is not possible on average to obtain improved compression with a code optimized for Q if the instances come from $P \neq Q$. Crucially, the KL divergence is in general asymmetric, i.e. $D_{\mathrm{KL}}(P||Q) \neq D_{\mathrm{KL}}(Q||P)$. Therefore, the amount of information "wasted" by encoding data from P with model Q is not the same as encoding from Q with P. Indeed, this indicates the possibility of more general models which not only compress their own instances optimally (as every model does), but also do very well for instances of other models, while the converse is not true. A concrete example of this is a general mixture given by

$$Q(\boldsymbol{A}, \boldsymbol{b}) = \sum_{m=1}^{M} P_m(\boldsymbol{A}, \boldsymbol{b}) P(m), \tag{86}$$

where the individual components $P_m(\boldsymbol{A}, \boldsymbol{b})$ are entirely arbitrary. In this case, we have $\Sigma_Q(\boldsymbol{A}, \boldsymbol{b}) \leq \Sigma_m(\boldsymbol{A}, \boldsymbol{b}) - \ln P(m)$ for every m, and hence

$$D_{\mathrm{KL}}(P_m||Q) \leq -\ln P(m), \tag{87}$$

where $-\ln P(m) = O(\ln M)$ if the mixtures have similar probability, while the reverse $D_{\mathrm{KL}}(Q||P_m)$ can be arbitrar-

ily large. In our context, we can speak of a good alternative code Q for P if $D_{\mathrm{KL}}(P||Q) = O(\ln N)$, since in this case the encoding "penalty" of using Q instead of P will be much smaller than the optimal Σ_P, which tends to scale as $O(N \ln N)$. Therefore, in the uniform case $P(m) = 1/M$, the general mixture will provide a good description for any of its components even if their number M grows as any polynomial in N.

Since the intrinsic models behind modularity maximization and Infomap amount to particular parametrizations of the SBM, we can therefore posit that a more general mixture will have a superior performance in most cases, while still performing very well for instances that are optimal for these more specialized algorithms. Here we review one such mixture, the nested stochastic block model (NSBM) [45, 53], and demonstrate that it indeed possesses this property.

A. The nested stochastic block model (NSBM)

The NSBM is based on a parametric formulation of the microcanonical SBM, which is defined by a likelihood

$$P(\boldsymbol{A}|\boldsymbol{e},\boldsymbol{b}), \tag{88}$$

where $\boldsymbol{e} = \{e_{rs}\}$ is again the matrix of edge counts between groups. The matrix $\boldsymbol{e}$ determines the mixing patterns between groups, which is a free parameter. Clearly, we can realize optimal instances of modularity and Infomap by choosing $\boldsymbol{e}$ accordingly. The NSBM consists of introducing a parametric prior for $\boldsymbol{e}$ which depends on a partition $\boldsymbol{b}_2$ of the *groups* of $\boldsymbol{b}$, and another matrix of edge counts $\boldsymbol{e}_2 = \{e^{(2)}_{tu}\}$, with elements $e^{(2)}_{tu}$ determining the number of edges between groups of groups. As a result, we have a marginal likelihood

$$P(\boldsymbol{A}|\boldsymbol{e}_2,\boldsymbol{b},\boldsymbol{b}_2) = \sum_{\boldsymbol{e}} P(\boldsymbol{A}|\boldsymbol{e},\boldsymbol{b})P(\boldsymbol{e}|\boldsymbol{e}_2,\boldsymbol{b}_2), \tag{89}$$

with the sum having trivially only one non-zero summand, due to the hard constraints imposed. Clearly, we can proceed indefinitely up to L hierarchical levels, where we enforce that on the last level $L+1$ there is a trivial partition into one group, leading to a marginal likelihood

$$P(\boldsymbol{A}|\boldsymbol{b},\boldsymbol{b}_2,\ldots,\boldsymbol{b}_L). \tag{90}$$

Choosing priors $P(\boldsymbol{b}_l)$ for the partitions leads to a nonparametric joint distribution $P(\boldsymbol{A},\boldsymbol{b},\boldsymbol{b}_2,\ldots,\boldsymbol{b}_L)$ and a description length for the hierarchical partition given by

$$\Sigma(\boldsymbol{A},\boldsymbol{b},\boldsymbol{b}_2,\ldots,\boldsymbol{b}_L) = -\ln P(\boldsymbol{A},\boldsymbol{b},\boldsymbol{b}_2,\ldots,\boldsymbol{b}_L). \tag{91}$$

For further details on the derivation of the likelihoods, including the degree-corrected variation (DC-NSBM), we refer to Refs [45, 53]. The description length for the first-level partition is obtained by marginalization,

$$\Sigma_{\mathrm{NSBM}}(\boldsymbol{A},\boldsymbol{b}) = -\ln \sum_{\boldsymbol{b}_2,\boldsymbol{b}_3,\ldots,\boldsymbol{b}_L} P(\boldsymbol{A},\boldsymbol{b},\boldsymbol{b}_2,\boldsymbol{b}_3,\ldots,\boldsymbol{b}_L) \tag{92}$$

$$\leq -\ln P(\boldsymbol{A},\boldsymbol{b},\boldsymbol{b}^*_2,\boldsymbol{b}^*_3,\ldots,\boldsymbol{b}^*_L). \tag{93}$$

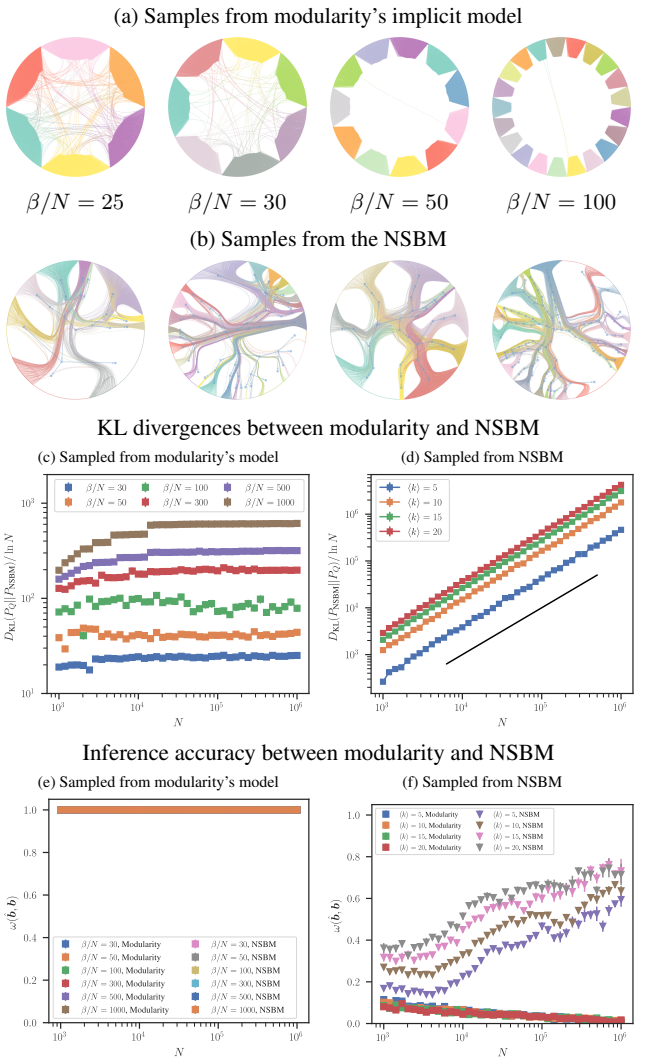

Figure 10. There is no appreciable trade-off between the NSBM and the implicit model behind modularity maximization. In panel (a) we show samples from modularity's model for different values of β, and in (b) we show samples from the NSBM (which is nonparametric) — in both cases visualized as chord diagrams. In (b) the corresponding hierarchical partitions are overlaid. In (c) and (d) we show the KL divergences, $D_{\mathrm{KL}}(P_Q||P_{\mathrm{NSBM}})$ and $D_{\mathrm{KL}}(P_{\mathrm{NSBM}}||P_Q)$ respectively, in both cases divided by $\ln N$, as a function of the number of nodes N. The solid lines shows the linear slope. In (c) the networks are sampled from modularity's model with $\langle k\rangle = 10$, and various values of β as indicated in the legend. In (d) the networks are sampled from the NSBM, for various $\langle k\rangle$ as indicated in the legend. In (e) and (f) are shown the overlaps between the inferred and true partitions, for the same problem instances in (c) and (d), respectively, when inferred with modularity maximization and with the NSBM, as indicated in the legend.

Although the sum over the higher-level partitions is intractable, the marginal description length is upper bounded by any particular choice $\{\boldsymbol{b}^*_l\}$, as shown in the last line of the above equation. This gives us an upper bound for $D_{\mathrm{KL}}(P_{\mathrm{NSBM}}||P_Q)$ and a lower bound for $D_{\mathrm{KL}}(P_Q||P_{\mathrm{NSBM}})$, which are sufficient for our analysis.

In Fig. 10(a) and (b) we compare samples from modularity's implicit model and the NSBM. Contrary to the former, the NSBM is completely nonparametric and yields more realistic problem instances that combine structure with disorder at several scales. Although they have an extremely varied number and composition of groups, and mixing patterns between them, the sampled instances always deviate from a maximally random graph — they are always compressive. Indeed, the structural regularity of a lower level of the hierarchy is generated with some amount of randomness and regularity from the level above, and so on recursively, attributing the samples with a mixture or randomness and regularity at multiple scales. This larger diversity of samples from the NSBM comes precisely from its more agnostic character when it is used for inference, since in this case we make fewer commitments about the structure of the data — with respect to the number of groups, how uniformly distributed they are and the preference of connections between them — before the data is actually seen. Importantly, as we will shortly demonstrate, once these patterns are actually identified, the resulting description length tends to be very close to the optimal one [45].

Due to its more general character, the NSBM generates the kind of regular community structure expected by modularity only with a relatively low probability, and hence provides a strictly sub-optimal encoding for networks that are sampled from this model. However, as Fig. 10(c) shows, the KL divergence $D_{\mathrm{KL}}(P_Q||P_{\mathrm{NSBM}})$ grows only logarithmically with N, meaning that it can nevertheless efficiently describe networks sampled from this model. The opposite situation, however, is quite different: As Fig. 10(d) shows, the reversed KL divergence $D_{\mathrm{KL}}(P_{\mathrm{NSBM}}||P_Q)$ grows log-linearly with N, meaning that modularity's model is very inefficient at encoding samples from the NSBM.

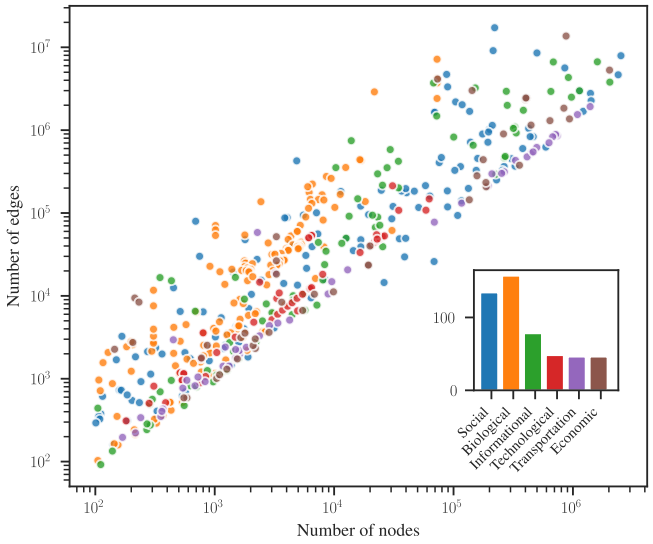

Figure 11. Number of nodes and edges, as well as distribution of domains (inset), for the 509 empirical networks considered in this work, avaliable from the Netzschleuder repository [65]. The symbol colors correspond to the network domain, as shown in the inset.

It is important to remember that, instead of compression directly, typically the primary objective in community detection is simply to uncover latent community assignments. Although these objectives are intimately related — as we already discussed, the optimal accuracy is always obtained with the true generative model, which is also the only one that can achieve maximal compression — a method might still be maximally successful at uncovering the correct community labels while providing strictly inferior compression. We show this in Fig. 10(e), with the maximum overlap $\omega(\hat{\boldsymbol{b}}, \boldsymbol{b})$ between the inferred and true partitions, $\hat{\boldsymbol{b}}$ and $\boldsymbol{b}$ respectively, defined as

$$\omega(\hat{\boldsymbol{b}}, \boldsymbol{b}) = \max_{\mu} \frac{1}{N}\sum_{i} \delta_{\hat{b}_i,\mu(b_i)}, \tag{94}$$

where $\mu(r)$ is a bijection between the labels of $\hat{\boldsymbol{b}}$ and $\boldsymbol{b}$, for problem instances sampled from modularity's model, and inferred both with modularity maximization and the NSBM. In all cases (which consist only of $\beta > \beta^*$, otherwise the overlap is always zero) the overlap is maximal with $\omega(\hat{\boldsymbol{b}}, \boldsymbol{b}) = 1$, showing that both methods uncover the exact same partition for these easy instances. Again, the opposite situation is quite different: with problem instances sampled from the NSBM, the accuracy of modularity maximization tends to zero, while the NSBM performs significantly better; although not perfectly — there is no guarantee of perfect recovery in these harder instances, only optimality.

It is not surprising that modularity maximization can neither compress nor correctly uncover the true assignments of samples from the NSBM, since those will not necessarily correspond to assortative communities. Our central point here is there is a lack of trade-off: the NSBM performs just as well for obvious assortative instances, while still being able to accommodate more general structures that are harder to detect.

Note that in the discussion above we did not have to make any reference to particular domains of application. The lack of trade-off is a general principle that must hold for mixtures in general, and can be articulated simply using fundamental concepts of mixing patterns between groups. Although one could expect networks belonging to different domains having different kinds of mixing patterns, the above arguments tell us that the superiority of hierarchical mixtures should transcend various domains. We evaluate this hypothesis in the following.

V. EMPIRICAL NETWORKS

The arguments above mean that we should expect that methods that are optimal for general mixtures of models should perform just as well as those that are specialized for any of the mixture components. However, when encountering networks in the real world, we can confidently assume that they are not in fact sampled from any model we can articulate exactly — even though it is often easy to determine that they are structured (e.g. either via statistical tests designed to reject the uniform null model, or simply by compressing it with any model). In these structured "out-of-distribution" cases, we are, strictly speaking, simultaneously out of scope

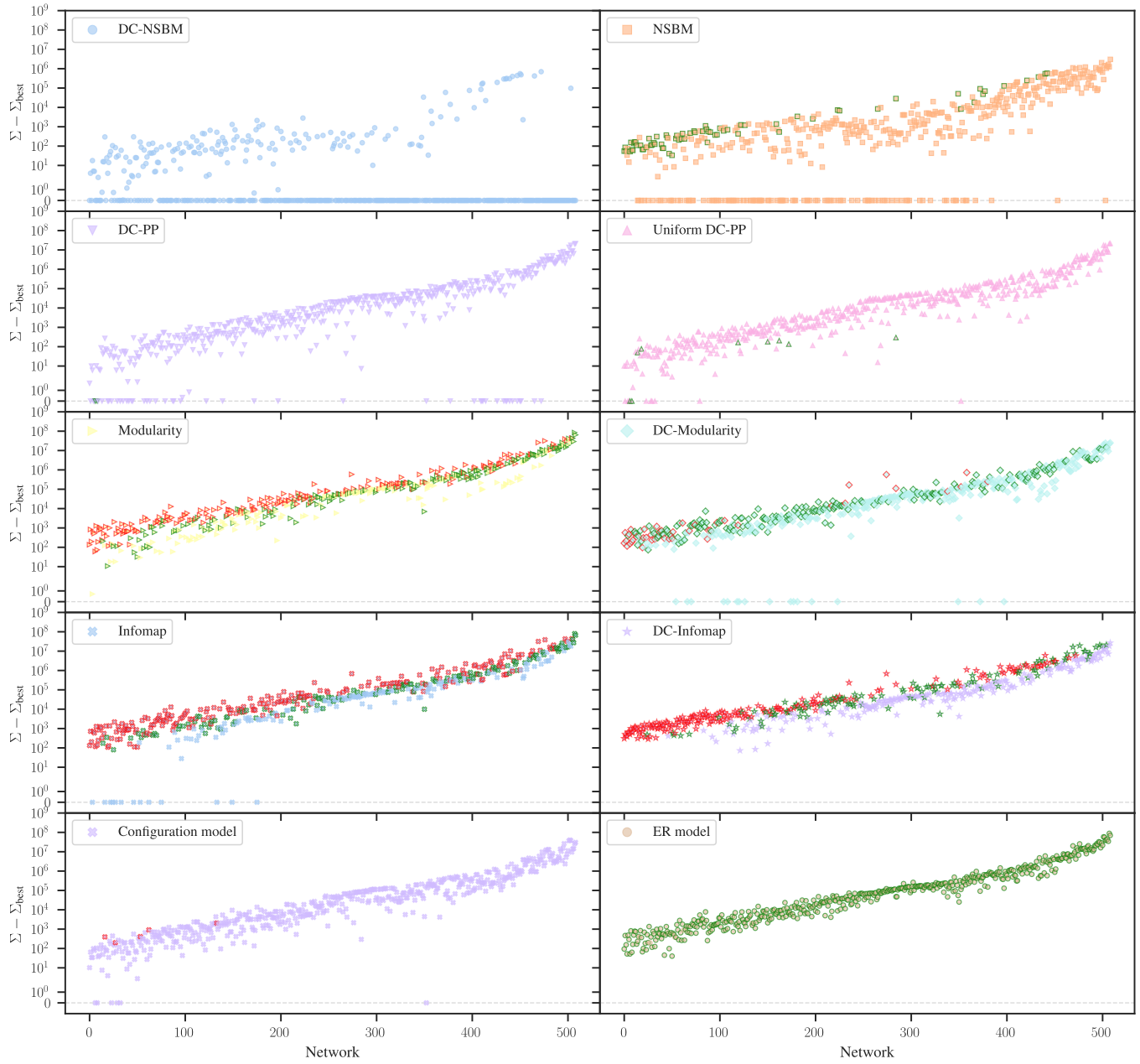

Figure 12. Difference in description length values according to the best model, obtained with several models for 509 empirical networks, ordered according to number of edges. A value of zero indicates that the respective model is the most compressive for the particular network. Symbols highlighted in red (green) correspond to description length values that are larger than the Erdős-Rényi model (configuration model).

of the NFL theorem and of the situation considered previously, where the sample comes from one of the models being considered.

Despite this, we should expect to be much closer to the scenario considered in the previous section than that of the NFL theorem, as soon as our models under considerations can serve as reasonable approximations of the data [67]. Here we test this hypothesis on a corpus of 509 structurally diverse empirical networks, from many domains of science, and across several orders of magnitude in size, as summarized in Fig. 11.

For each of these networks, we find the partition according to maximum modularity, Infomap, and well as various versions of the SBM: the NSBM, its degree-corrected version (DC-SBM), the non-uniform degree-corrected planted partition model (DC-PP), and its uniform version [18]. For Infomap and modularity we then compute their implicit description lengths, using also the degree-corrected alternatives. We also compute the description length for the configuration and Erdős-Rényi models as baselines.

In Fig. 12 we show for each model and network the difference in description length according to the best model for each network — a value of zero thus means that the specific

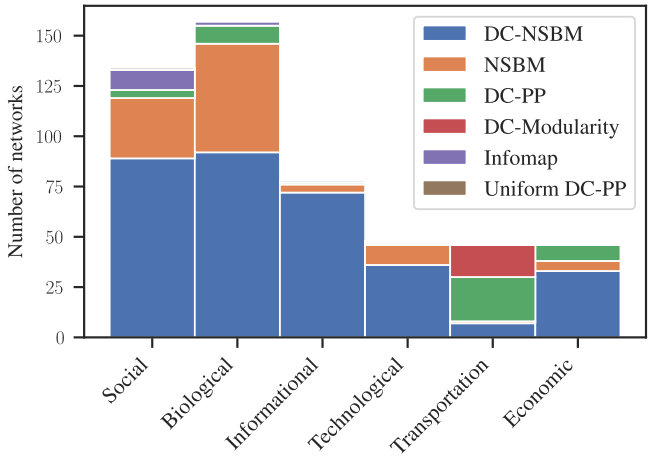

Figure 13. Number of networks for which each model provides the smallest description length, as indicated in the legend, across the different domains.

model is the best one for that network. We can see clear performance gains for the SBM variants, with the DC-NSBM and the NSBM having the best compression in the majority of cases, and the DC-PP also performing well, primarily on small networks. As shown in Fig. 13, this is also true when each domain is considered separately — with the exception of transportation networks, where the DC-PP provides an improved compression than the DC-NSBM for a larger fraction of cases. Most cases where other algorithms achieve the best compression are smaller networks, which may be due to the fact that we have used lower bounds to estimate the partition function for these alternative models, giving them a slight advantage. Alternatively, the communication cost for more complex models in these cases may outweigh the corresponding improvements in fit to the data if these happen to be better described by the more specific constraints of the implicit generative models of either modularity or Infomap.

We compare the relative compression of the models in a different manner in Fig. 12, plotting the fraction of cases where a given model $\mathcal{M}_1$ achieves equal or better compression than the alternative model $\mathcal{M}_2$, as well as their average compression ratio for all networks. Based on these pairwise comparisons we ranked each model according to the SpringRank [66] algorithm, which is reflected in the ordering of Fig. 14. We can also see here that the SBM variants are much more compressive than the other algorithms, even for the models that do not achieve the lowest overall compression. The row of the heatmap labelled "Best SBM" takes the best compression among all SBM variants for each network, which is almost completely unmatched in its compression when compared to all other algorithms, performing the worst relative to the degree-corrected modularity, where superior compression is achieved for 96% of the empirical networks. We can also see that the configuration and Erdős-Rényi models provide superior compression to modularity and Infomap in a large fraction of instances. This inflated description in comparison to a maximally random baseline indicates a massive amount of overfitting in the results produced by these algorithms — i.e. the structures found are better justified by being the outcome of purely random fluctuations.

Although the SBM variations, and in particular the NSBM, provide a description superior to the alternatives for the large majority of networks considered, there are in fact a few exceptions where either modularity or Infomap do provide a better description. As we discussed in the previous section, this is expected when the networks are closer to the typical ones generated by the implicit generative models of these methods, which have a specific relationship between the number of groups and the strength of the community structure. Note that we should not be tempted to attribute the existence of these minority cases as a necessary outcome of a supposed trade-off that comes as an unavoidable consequence of the NFL theorem, as suggested in Refs. [6, 57]. As discussed previously, the NFL theorem is only valid when problem instances are sampled uniformly at random, resulting almost exclusively in incompressible networks — a hypothesis that we can confidently reject for all networks considered in our corpus. Furthermore, even when considering the maximally uniform case, the NFL theorem does not imply any actual trade-off, only that all algorithms must perform equally poorly in the asymptotic totality of instances. Besides, the negation that all algorithms perform equally well when averaged over all cases does not necessarily imply that a single algorithm must perform strictly better in all of them — it would be sufficient that some algorithms perform better than others on average, precisely as our results and those of Refs. [6, 57] show.

Indeed, we can see evidence of an systematic hierarchy between community detection algorithms when we compare the description lengths with the actual partitions found. In Fig. 15 we show for every network in our corpus the difference between the best description length per node found with any version of the SBM and the one found with either modularity or Infomap (the best from the degree-corrected and non-degree-corrected versions) together with the adjusted mutual information (AMI) [68] between their respective partitions. In both cases, we see that for networks where either modularity or Infomap provide a better description (which are often relatively small or very sparse networks), they yield partitions that are very similar to the SBM inference. Examples of such instances can be seen in Eq. 16, where we can see that both methods tend to agree substantially on the network divisions.

The fact that Infomap and modularity tend to agree with the SBM whenever they yield compressive answers is also a statement about the partial similarities between these algorithms. Indeed, as we argued previously, both modularity and Infomap are approximately equivalent to the inference of versions of the SBM with very particular constraints imposed on its parameters. Therefore, neither algorithm can exploit features in the network that deviate from the same underlying SBM assumption. When we compare them a posteriori, we can only tell which SBM parametrization is relatively better justified according to the evidence in the data.

Clearly, the fact that Infomap and modularity amount to particular SBM parametrizations should not be used as a justification for their use as reliable inference methods. The im-

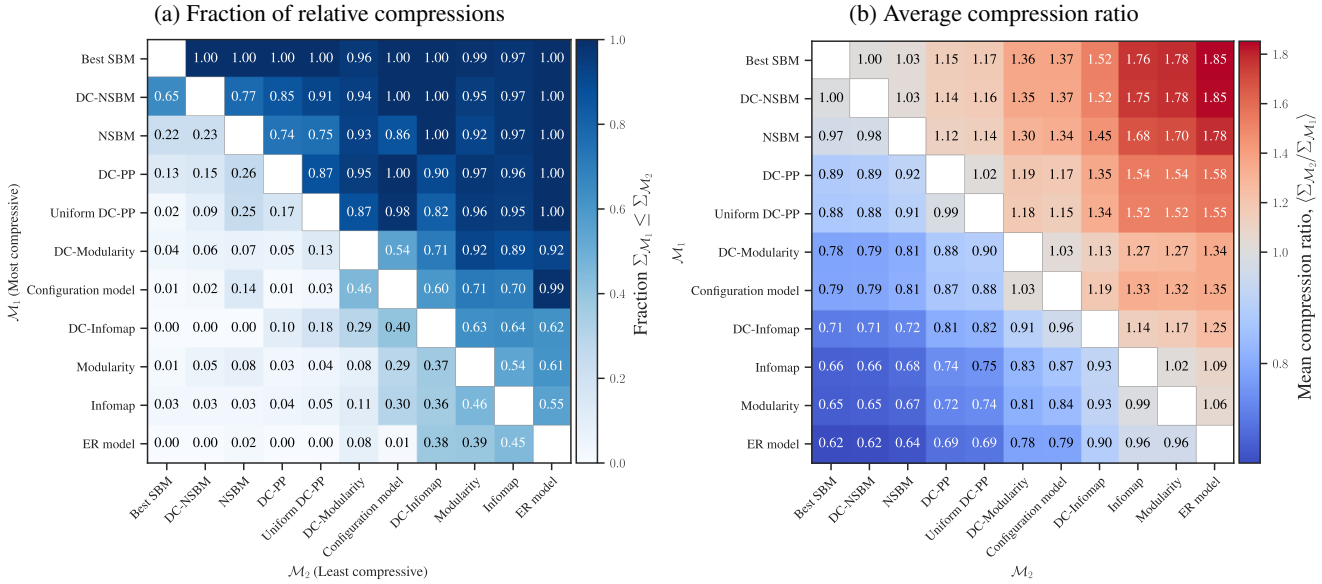

Figure 14. (a) Fraction of networks in our corpus where a given model $\mathcal{M}_1$ (vertical axis) achieves equal or better compression than the alternative model $\mathcal{M}_2$ (horizontal axis). (b) Average compression ratio between models $\langle \Sigma_{\mathcal{M}_2} / \Sigma_{\mathcal{M}_1} \rangle$ across all networks. In both (a) and (b) the order of models corresponds to the SpringRank [66], computed using the respective pairwise comparisons.

plicit priors are so strongly committed to particular patterns that they will be dredged out of pure randomness, resulting in description lengths that are not only most of the time significantly larger than the properly agnostic SBMs, but very often even larger than maximally random networks.

We can investigate further the tendency of Infomap and modularity to overfit by comparing how many groups are obtained with each method, as shown in Fig. 17. For modularity maximization, we can observe its tendency of both overfit and underfit depending on the circumstance, since most networks have a number of groups smaller than the resolution limit, i.e. $B < \sqrt{2E}$ — except those with more than one component, where this limit does not apply. Despite this limitation, a large fraction of the results are less compressive than the maximally random baselines, indicating substantial overfitting. For Infomap the overfitting is more extreme, with the number of groups found scaling linearly with the number of nodes. This corresponds exactly to the implicit prior for the number of groups in Infomap which strongly prefers a characteristic group size that is independent of the number of nodes, as shown in Figs. 5 and 9.

It is important to emphasize that even when the description lengths of Infomap and modularity are smaller than one of the maximally random baselines, this does not necessarily mean that method is not overfitting, since the partition found can still amount to a substantial amount of randomness. We can assess this by comparing the number of groups obtained with the model version that yields the smallest description length, as shown in the bottom row of Fig. 17. Indeed we can see that modularity tends to both under- and overfit for a comparable fraction of the networks, although the larger tendency is to underfit, while with Infomap the overwhelming tendency is to overfit, and return a much larger number of groups than the most compressive partition.

We observe also that all SBM flavors manage to find a number of groups in a range that does not necessarily conform to a $\sqrt{E}$ or $\sqrt{N}$ scaling — a lack of constraint that is theoretically prescribed [18, 53]. This dispels the notion that this scaling is a fundamental limitation of community detection methods in general, as suggested in Ref. [6]. Importantly, this lack of resolution limit of the NSBM and PP models comes together with a regularization against overfitting, unlike what we observe for Infomap.

VI. DISCUSSION

In this paper we have presented a framework for identifying the implicit generative model associated with an arbitrary community detection algorithm, allowing us to compare descriptive and inferential methods on the same scale by computing their associated description lengths for a network and corresponding partition. This method also allows us to compute the implicit priors on the objective value and number of groups associated with a community detection objective, giving insights into the intrinsic biases in existing algorithms such as modularity and Infomap, which we demonstrate are biased towards overfitting due to strong priors favoring high objective values. We also find that the implicit models for a wide range of methods, including modularity and Infomap, correspond asymptotically to restricted instances of the stochastic block model (SBM). By exploiting the latent compression associated with community detection algorithms, we were able to compare these methods on real and synthetic data, demonstrating that in these structured problem instances certain algorithms (more expressive variants of the

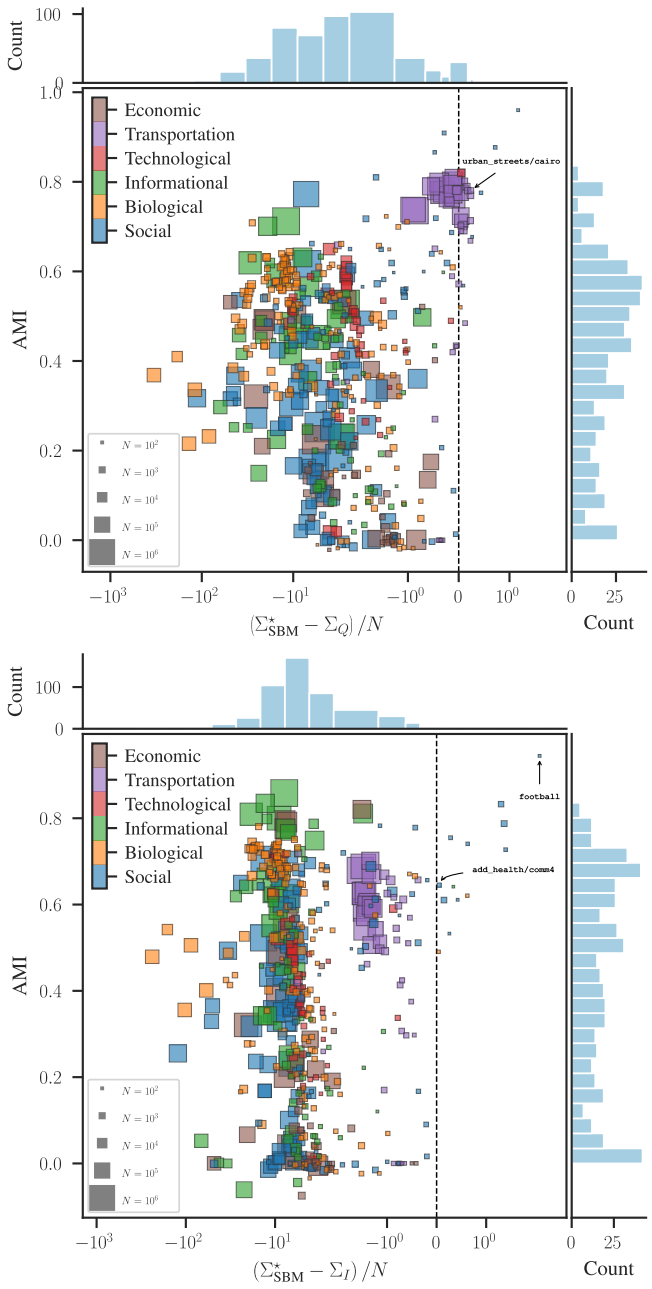

Figure 15. Adjusted mutual information (AMI) between the partitions inferred with the best fitting SBM and either modularity (top) or Infomap (bottom) as a function of the description length difference between models, $\Sigma^{\star}_{\mathrm{SBM}}$, and Σ_Q or Σ_I, divided by the number of nodes in the network. The symbol colors indicate the domain, and the size the number of nodes in the network, as shown in the legends. The text annotations refer to the networks shown in Fig. 16

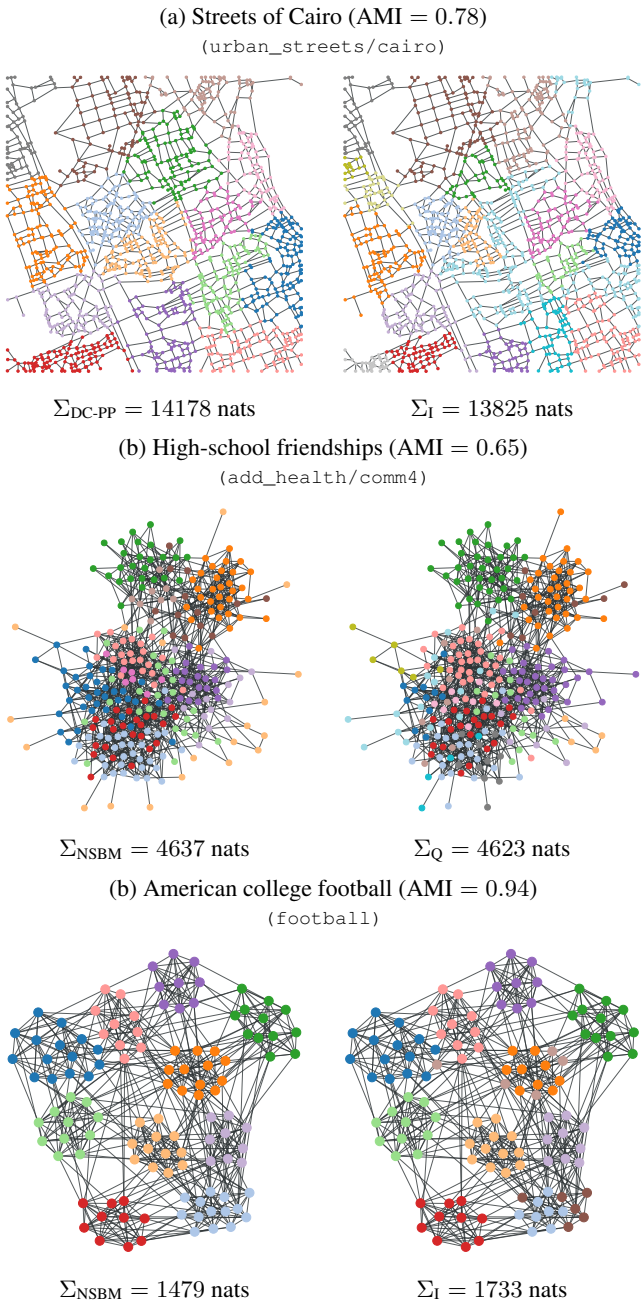

Figure 16. Examples of exceptional networks where either modularity maximization or infomap yield description length values, Σ_Q and Σ_I respectively, that are smaller than what is obtained with any of the SBM variants. In all cases the partitions obtained are shown as node colors, and the adjusted mutual information (AMI) between them is given in the panel title, which also shows the corresponding Netzschleuder [65] codename used in Fig. 15.

SBM) are systematically favored over others (variants of modularity and Infomap).

Since it provides a universal scale on which we can assess the capacity of a model to capture structural regularities in network data, the description length provides a principled measure to compare the performance of community detection algorithms without the need for "ground truth" labels — unknowable information for empirical networks [32]. The empirical experiments here show that by evaluating algorithms using this measure we can reveal a clear breakdown of the implications of the NFL theorem for real, structured problem instances. This weakens the practical and conceptual pertinence of the NFL theorem, which equates all possible community

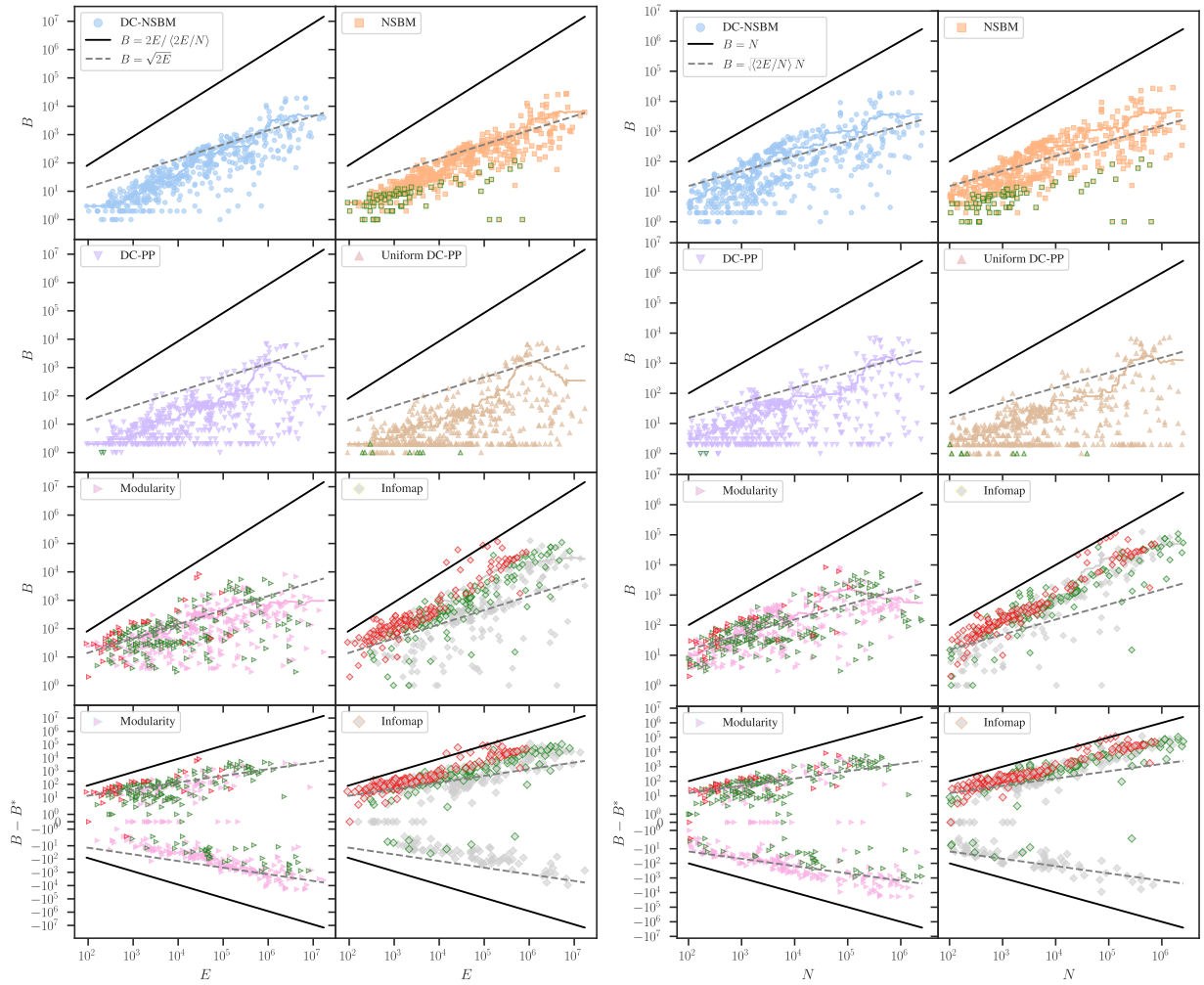

Figure 17. Number of groups B as function of number of edges E (left panel) and number of nodes N (right panel) according to each method as indicated by the legend, for every network in our corpus. The symbols in red indicate partitions for which the description length value is larger than the Erdős-Rényi model, and likewise for those in green for the configuration model. The solid lines correspond to moving averages, and the black solid and dashed lines are slopes as indicated in the legend.

detection algorithms in terms of performance, but applies only to unstructured problem instances.

Part of the results in this work confirm what has been found by Ghasemian et al [6] with respect to modularity maximization and Infomap overfitting in a link prediction task for a diverse set of smaller networks, and the regularized SBM performing better on average (although Ref. [6] omitted the NSBM, which can be shown to perform strictly better, and substantially so for larger networks [45]). This is not unexpected, since it is known that algorithmic learning procedures where the objective is to obtain a succinct representation of data (called broadly "Occam learning" in the machine learning theory literature) are in general equivalent to learning procedures where the objective is to choose a predictive model with low generalization error [known as "probably approximately correct" (PAC) learning] [69]. Because of this, we can in principle expect a MDL approach to yield compatible results with link prediction in suitable limits. However, there is an important caveat that prevents this equivalence from being exact; namely, the nominal task would correspond to predicting an entire new network from past observations of a complete network. Instead, in a more realistic link prediction scenario one attempts to predict a subset of the possible edges by observing the remaining network, which is commonly sparse. In this situation we cannot guarantee that a sufficient data limit exists, regardless of how large the network is — the removal of a fraction of the edges always destroys important information which could be used to improve the detection of the community labels. Because of this, discrepancies between both approaches can exist, with link prediction having a tendency to overfit when used as a model selection criterion [70]. Therefore, the results we present in this work have a more definitive character than those of Ref. [6], since ours make use of the whole data.

We have applied our method for computing the description length of modularity and Infomap by exploiting the fact that

their objective functions can be written in terms of the microcanonical SBM parameters. Our calculations are possible for a much wider set of objective functions that can also be described in the same manner. We speculate that a significant fraction of community detection algorithms proposed in the literature are, like Infomap and modularity, also equivalent to the inference of constrained versions of the SBM, as has been suggested by others [71]. This would have wider implications for the systematic superiority of SBM-based approaches. It remains to be determined to what extent this is true, and what other modelling avenues are possible for community detection that step significantly outside of the SBM framework.

[1] S. Fortunato, Community detection in graphs, Physics Reports **486**, 75 (2010).

[2] S. Fortunato and M. E. J. Newman, 20 years of network community detection, Nature Physics **18**, 848 (2022).

[3] S. Fortunato and D. Hric, Community detection in networks: A user guide, Physics Reports 10.1016/j.physrep.2016.09.002 (2016).

[4] M. T. Schaub, J.-C. Delvenne, M. Rosvall, and R. Lambiotte, The many facets of community detection in complex networks, Applied Network Science **2**, 1 (2017).

[5] D. Hric, R. K. Darst, and S. Fortunato, Community detection in networks: Structural communities versus ground truth, Physical Review E **90**, 062805 (2014).

[6] A. Ghasemian, H. Hosseinmardi, and A. Clauset, Evaluating Overfit and Underfit in Models of Network Community Structure, IEEE Transactions on Knowledge and Data Engineering , 1 (2019).

[7] In principle, this should not constitute an obstacle, as one would need only to match the most appropriate algorithm to a given objective supplied by the practitioner within the context of a particular application. However, due to their qualitative similarity, users often expect universal algorithms that work well independently of context — an attitude which is also reflected on a variety of works that benchmark competing methods against the same criterion, such as recovering planted community structure in artificial networks [28, 72], prediction of node covariates [5], or of missing links [6], regardless of their divergences in motivation.

[8] T. P. Peixoto, Descriptive vs. inferential community detection in networks: pitfalls, myths and half-truths, arXiv:2112.00183 [physics, stat] (2022).

[9] T. P. Peixoto, Bayesian Stochastic Blockmodeling, in *Advances in Network Clustering and Blockmodeling* (John Wiley & Sons, Ltd, 2019) pp. 289–332.

[10] R. Guimerà and M. Sales-Pardo, Missing and spurious interactions and the reconstruction of complex networks, Proceedings of the National Academy of Sciences **106**, 22073 (2009).

[11] T. P. Peixoto, Reconstructing Networks with Unknown and Heterogeneous Errors, Physical Review X **8**, 041011 (2018).

[12] Ü. V. Çatalyürek, K. D. Devine, M. F. Faraj, L. Gottesbüren, T. Heuer, H. Meyerhenke, P. Sanders, S. Schlag, C. Schulz, D. Seemaier, and D. Wagner, More Recent Advances in (Hyper)Graph Partitioning (2022).

[13] M. Rosvall and C. T. Bergstrom, Maps of random walks on complex networks reveal community structure, Proceedings of the National Academy of Sciences **105**, 1118 (2008).

[14] Therefore, if we take its stated objective at face value, when a method such as Infomap clusters a maximally random network into many groups, as it is prone to do [29, 54], it is not meaningful to describe this as overfitting, since no model fit is nominally being attempted. Indeed, if a random graph is sufficiently sparse, then a random walk may genuinely get trapped into quenched random structures, such as groups of nodes that are more internally connected by chance alone [16], or other structures such as dangling trees [56], which could be well characterized by the network division found. This is precisely what these methods set out to identify, and whatever consternation this may cause in a particular application likely indicates a mismatch between the stated objective of the method and what is in fact desired or more appropriate in context, instead of a problem with the method itself.

[15] M. E. J. Newman, Modularity and community structure in networks, Proceedings of the National Academy of Sciences **103**, 8577 (2006).

[16] R. Guimerà, M. Sales-Pardo, and L. A. N. Amaral, Modularity from fluctuations in random graphs and complex networks, Physical Review E **70**, 025101 (2004).

[17] M. E. J. Newman, Equivalence between modularity optimization and maximum likelihood methods for community detection, Physical Review E **94**, 10.1103/PhysRevE.94.052315 (2016).

[18] L. Zhang and T. P. Peixoto, Statistical inference of assortative community structures, Physical Review Research **2**, 043271 (2020).

[19] J. R. Tyler, D. M. Wilkinson, and B. A. Huberman, Email as Spectroscopy: Automated Discovery of Community Structure within Organizations, in *Communities and Technologies*, edited by M. Huysman, E. Wenger, and V. Wulf (Springer Netherlands, Dordrecht, 2003) pp. 81–96.

[20] K. Yuta, N. Ono, and Y. Fujiwara, A Gap in the Community-Size Distribution of a Large-Scale Social Networking Site (2007).

[21] Y. Zhang, A. J. Friend, A. L. Traud, M. A. Porter, J. H. Fowler, and P. J. Mucha, Community structure in Congressional cosponsorship networks, Physica A: Statistical Mechanics and its Applications **387**, 1705 (2008).

[22] V. Red, E. D. Kelsic, P. J. Mucha, and M. A. Porter, Comparing Community Structure to Characteristics in Online Collegiate Social Networks, SIAM Review **53**, 526 (2011).

[23] A. L. Traud, P. J. Mucha, and M. A. Porter, Social structure of Facebook networks, Physica A: Statistical Mechanics and its Applications **391**, 4165 (2012).

[24] V. Spirin and L. A. Mirny, Protein complexes and functional modules in molecular networks, Proceedings of the National Academy of Sciences **100**, 12123 (2003).

[25] J. Chen and B. Yuan, Detecting functional modules in the yeast protein–protein interaction network, Bioinformatics **22**, 2283 (2006).

[26] A. C. F. Lewis, N. S. Jones, M. A. Porter, and C. M. Deane, The Function of Communities in Protein Interaction Networks, 0904.0989 (2009).

[27] D. M. Wilkinson and B. A. Huberman, A method for finding communities of related genes, Proceedings of the National Academy of Sciences **101**, 5241 (2004).

[28] A. Lancichinetti, S. Fortunato, and F. Radicchi, Benchmark graphs for testing community detection algorithms, Physical

Review E **78**, 046110 (2008).
[29] A. Lancichinetti and S. Fortunato, Community detection algorithms: A comparative analysis, Physical Review E **80**, 056117 (2009).
[30] J. Reichardt and S. Bornholdt, When are networks truly modular?, Physica D: Nonlinear Phenomena **224**, 20 (2006).
[31] A. Condon and R. M. Karp, Algorithms for graph partitioning on the planted partition model, Random Structures & Algorithms **18**, 116 (2001).
[32] L. Peel, D. B. Larremore, and A. Clauset, The ground truth about metadata and community detection in networks, Science Advances **3**, e1602548 (2017).
[33] E. T. Jaynes, *Probability Theory: The Logic of Science*, edited by G. L. Bretthorst (Cambridge University Press, Cambridge, UK ; New York, NY, 2003).
[34] C. P. Massen and J. P. K. Doye, Thermodynamics of Community Structure, arXiv:cond-mat/0610077 (2006).
[35] P. Zhang and C. Moore, Scalable detection of statistically significant communities and hierarchies, using message passing for modularity, Proceedings of the National Academy of Sciences **111**, 18144 (2014).
[36] The optimal estimate of the partition will always depend on which criterion we use to judge performance, namely the particular choice of the error function $\epsilon(\boldsymbol{b}, \boldsymbol{b}')$. The choice of error function is an application-dependent decision, and other choices will lead to estimates that are different from Eq. 1 [73], although they will always involve the posterior of Eq. 7.
[37] J. Rissanen, *Information and Complexity in Statistical Modeling*, 1st ed. (Springer, 2010).
[38] P. D. Grünwald, *The Minimum Description Length Principle* (The MIT Press, 2007).
[39] M. Rosvall and C. T. Bergstrom, An information-theoretic framework for resolving community structure in complex networks, Proceedings of the National Academy of Sciences **104**, 7327 (2007).
[40] T. P. Peixoto, Parsimonious Module Inference in Large Networks, Physical Review Letters **110**, 148701 (2013).
[41] T. P. Peixoto, Model Selection and Hypothesis Testing for Large-Scale Network Models with Overlapping Groups, Physical Review X **5**, 011033 (2015).
[42] T. P. Peixoto and L. Zhang, Statistical inference of assortative community structures, Physical Review Research **2**, 043271 (2020).
[43] T. P. Peixoto, Disentangling Homophily, Community Structure, and Triadic Closure in Networks, Physical Review X **12**, 011004 (2022).
[44] T. P. Peixoto, Entropy of stochastic blockmodel ensembles, Physical Review E **85**, 056122 (2012).
[45] T. P. Peixoto, Nonparametric Bayesian inference of the microcanonical stochastic block model, Physical Review E **95**, 012317 (2017).
[46] E. A. Bender and E. R. Canfield, The asymptotic number of labeled graphs with given degree sequences, Journal of Combinatorial Theory, Series A **24**, 296 (1978).
[47] B. Karrer and M. E. J. Newman, Stochastic blockmodels and community structure in networks, Physical Review E **83**, 016107 (2011).
[48] J. Reichardt and S. Bornholdt, Statistical mechanics of community detection, Physical Review E **74**, 016110 (2006).
[49] J. Park and M. E. J. Newman, Solution of the two-star model of a network, Physical Review E **70**, 066146 (2004).
[50] J. Park and M. E. J. Newman, Solution for the properties of a clustered network, Physical Review E **72**, 026136 (2005).
[51] S. Fortunato and M. Barthélemy, Resolution limit in community detection, Proceedings of the National Academy of Sciences **104**, 36 (2007).
[52] A. Lancichinetti and S. Fortunato, Limits of modularity maximization in community detection, Physical Review E **84**, 066122 (2011).
[53] T. P. Peixoto, Hierarchical Block Structures and High-Resolution Model Selection in Large Networks, Physical Review X **4**, 011047 (2014).
[54] T. Kawamoto and Y. Kabashima, Comparative analysis on the selection of number of clusters in community detection, Physical Review E **97**, 022315 (2018).
[55] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová, Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications, Physical Review E **84**, 066106 (2011).
[56] F. Krzakala, C. Moore, E. Mossel, J. Neeman, A. Sly, L. Zdeborová, and P. Zhang, Spectral redemption in clustering sparse networks, Proceedings of the National Academy of Sciences , 201312486 (2013).
[57] A. Ghasemian, H. Hosseinmardi, A. Galstyan, E. M. Airoldi, and A. Clauset, Stacking models for nearly optimal link prediction in complex networks, Proceedings of the National Academy of Sciences **117**, 23393 (2020).
[58] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 99th ed. (Wiley-Interscience, 1991).
[59] M. J. Streeter, Two Broad Classes of Functions for Which a No Free Lunch Result Does Not Hold, in *Genetic and Evolutionary Computation — GECCO 2003*, Lecture Notes in Computer Science, edited by E. Cantú-Paz, J. A. Foster, K. Deb, L. D. Davis, R. Roy, U.-M. O'Reilly, H.-G. Beyer, R. Standish, G. Kendall, S. Wilson, M. Harman, J. Wegener, D. Dasgupta, M. A. Potter, A. C. Schultz, K. A. Dowsland, N. Jonoska, and J. Miller (Springer, Berlin, Heidelberg, 2003) pp. 1418–1430.
[60] S. McGregor, No free lunch and algorithmic randomness, in *GECCO*, Vol. 6 (2006) pp. 2–4.
[61] T. Everitt, Universal induction and optimisation: No free lunch? (2013).
[62] T. Lattimore and M. Hutter, No Free Lunch versus Occam's Razor in Supervised Learning, in *Algorithmic Probability and Friends. Bayesian Prediction and Artificial Intelligence: Papers from the Ray Solomonoff 85th Memorial Conference, Melbourne, VIC, Australia, November 30 – December 2, 2011*, Lecture Notes in Computer Science, edited by D. L. Dowe (Springer, Berlin, Heidelberg, 2013) pp. 223–235.
[63] G. Schurz, *Hume's Problem Solved: The Optimality of Meta-Induction*, illustrated edition ed. (The MIT Press, Cambridge, Massachusetts, 2019).
[64] M. Hutter, On universal prediction and Bayesian confirmation, Theoretical Computer Science Theory and Applications of Models of Computation, **384**, 33 (2007).
[65] T. P. Peixoto, The Netzschleuder network catalogue and repository. (2020), accessible at `https://networks.skewed.de`.
[66] C. De Bacco, D. B. Larremore, and C. Moore, A physical model for efficient ranking in networks, Science Advances **4**, eaar8260 (2018).
[67] S. C. Olhede and P. J. Wolfe, Network histograms and universality of blockmodel approximation, Proceedings of the National Academy of Sciences **111**, 14722 (2014).
[68] N. X. Vinh, J. Epps, and J. Bailey, Information theoretic measures for clusterings comparison: is a correction for chance necessary?, in *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09 (ACM, New York, NY, USA, 2009) pp. 1073–1080.

[69] M. J. Kearns and U. Vazirani, *An Introduction to Computational Learning Theory* (MIT Press, 1994).
[70] T. Vallès-Català, T. P. Peixoto, M. Sales-Pardo, and R. Guimerà, Consistencies and inconsistencies between model selection and link prediction in networks, Physical Review E **97**, 062316 (2018).
[71] J.-G. Young, G. St-Onge, P. Desrosiers, and L. J. Dubé, Universality of the stochastic block model, Physical Review E **98**, 032309 (2018).
[72] A. Lancichinetti and S. Fortunato, Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities, Physical Review E **80**, 016118 (2009).
[73] T. P. Peixoto, Revealing Consensus and Dissensus between Network Partitions, Physical Review X **11**, 021003 (2021).
[74] F. Ouimet, General Formulas for the Central and Non-Central Moments of the Multinomial Distribution, Stats **4**, 18 (2021).

Appendix A: Approximations of the density of states

In this Appendix we discuss the approximation used in Eq. 53 for the density of states in Eq. 49, as well as the corresponding approximation for the Infomap objective of Eq. 66.

If we fix the number of groups B and number of within-group edges E_{in}, as well as allow for multi-edges (which will give asymptotically identical results to the single-edge case), the distribution of the marginal sums $\{e_r\}$ for the ensemble of matrices $\boldsymbol{e}$ whose entries sum to $2E$ will follow the multinomial distribution with mean $2E/B$ and probabilities $\{1/B\}$ for the bins $\{e_r\}$. We therefore expect the distribution of Q over this ensemble to have mean

$$\begin{aligned}\langle Q\rangle &= \frac{E_{in}}{E} - \frac{\gamma}{(2E)^2}\left\langle\sum_r e_r^2\right\rangle \\ &= \frac{E_{in}}{E} - \frac{\gamma}{(2E)^2}\sum_r\langle e_r^2\rangle \\ &= \frac{E_{in}}{E} - \frac{\gamma}{(2E)^2}\times B\times\left[\frac{2E(B-1)}{B^2}+\frac{(2E)^2}{B^2}\right] \\ &= \frac{E_{in}}{E} - \frac{\gamma}{B}\left[1+\frac{B-1}{2E}\right] \\ &\approx \frac{E_{in}}{E} - \frac{\gamma}{B},\end{aligned} \tag{A1}$$

where we take the limit $E \gg B$. The variance σ_Q^2 of the modularity over this ensemble is given by

$$\begin{aligned}\sigma_Q^2 &= \langle Q^2\rangle - \langle Q\rangle^2 \\ &= \frac{\gamma^2}{(2E)^4}\sum_{rs}\left[\langle e_r^2e_s^2\rangle - \langle e_r^2\rangle\langle e_s^2\rangle\right] \\ &= \frac{\gamma^2}{16E^4}\sum_{rs}\langle e_r^2e_s^2\rangle - \frac{\gamma^2}{B^2} \\ &= \frac{\gamma^2B}{16E^4}\langle e_4^4\rangle + \frac{\gamma^2B(B-1)}{16E^4}\langle e_r^2e_s^2\rangle_{r\neq s} - \frac{\gamma^2}{B^2}.\end{aligned} \tag{A2}$$

Using the identity [74]

$$\left\langle\prod_{r=1}^B e_r^{z_r}\right\rangle = \sum_{t_1=0}^{z_1}\cdots\sum_{t_B=0}^{z_B}(2E)^{\sum_{r=1}^B t_r}\prod_{r=1}^B\left\{\begin{matrix}z_r\\t_r\end{matrix}\right\}\left(\frac{1}{B}\right)^{t_r}, \tag{A3}$$

we have

$$\langle e_r^4\rangle = \frac{2E}{B}+\frac{28E^2}{B^2}+\frac{48E^3}{B^3}+\frac{16E^4}{B^4}, \tag{A4}$$

$$\langle e_r^2e_s^2\rangle = \frac{4E^2}{B^2}+\frac{16E^3}{B^3}+\frac{16E^4}{B^4}. \tag{A5}$$

Plugging these expressions back into Eq. A2, we find that to leading order in E and B we have

$$\sigma_Q^2 \sim \frac{\gamma^2}{EB}, \tag{A6}$$

which vanishes for $E \gg 1$.

In this regime we can then approximate the distribution of Q over the ensemble of $B\times B$ mixing matrices $\boldsymbol{e}$ with fixed sum $2E$ and E_{in} within-group edges as a delta function centered at $Q = \frac{E_{in}}{E} - \frac{\gamma}{B}$. We thus have a correspondence between Q and E_{in} for fixed B, E, γ, allowing us to approximate the density of states in Eq. 49 with the restricted version in Eq. 53. Similar reasoning can be applied for the Infomap objective in Eq. 65, but the variance in this case is not analytically tractable, so we resort to numerical simulations to demonstrate that the variance vanishes for $E \gg 1$ (see Fig. 18).

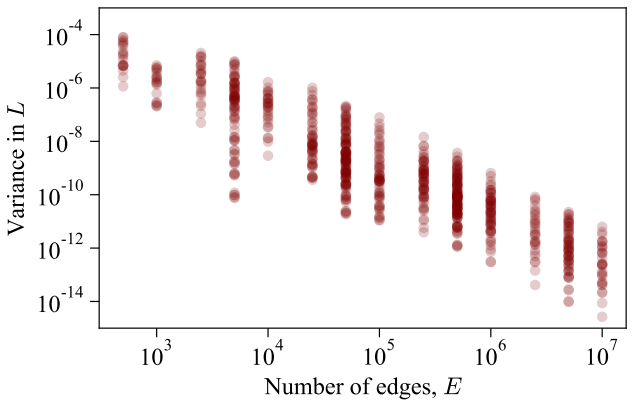

Figure 18. Variance of Infomap objective L (Eq. 65) over mixing matrices $\boldsymbol{e}$ sampled from the ensemble with fixed sum, with other ensemble parameters scanned over the ranges $N \in [10^2, 10^5]$, $B \in [2, 200]$, $\langle k\rangle \in [5, 100]$, $E_{in}/E \in [0.05, 0.95]$. Each point represents the variance over 100 simulated instances of the mixing matrix $\boldsymbol{e}$. We can see that the variance in L vanishes with E.