


ON MODEL SELECTION FOR DENSE STOCHASTIC BLOCK MODELS

ILKKA NORROS ^{*} *University of Helsinki*
HANNU REITTU ^{**} *VTT Technical Research Centre of Finland*
FÜLÖP BAZSÓ, ^{***} *Wigner Research Centre for Physics*

Abstract

This paper studies estimation of stochastic block models with Rissanen’s minimum description length (MDL) principle in the dense graph asymptotics. We focus on the problem of model specification, i.e., identification of the number of blocks. Refinements of the true partition always decrease the code part corresponding to the edge placement, and thus a respective increase of the code part specifying the model should overweight that gain in order to yield a minimum at the true partition. The balance between these effects turns out to be delicate. We show that the MDL principle identifies the true partition among models whose relative block sizes are bounded away from zero. The results are extended to models with Poisson-distributed edge weights.

Keywords: Minimum description length principle; random graph; Szemerédi’s regularity lemma; regular decomposition

2020 Mathematics Subject Classification: Primary 60B20
Secondary 62B10

1. Introduction

Let us define a *stochastic block model* (SBM) as a random graph $G = (V, E)$, whose structure is defined by a partition $\xi = \{A_1, \dots, A_k\}$ of V and by a symmetric $k \times k$ matrix $D = (d_{ij})_{i,j=1}^k$ of real numbers $d_{ij} \in [0, 1]$ as follows: for every pair $\{v, w\}$ of distinct vertices of V such that $v \in A_i$, $w \in A_j$, $\{v, w\} \in E$ with probability d_{ij} , and all Bernoulli random variables $e_{vw} = e_{wv} = 1_{\{\{v,w\} \in E\}}$ are independent. For unambiguity, we also set the *irreducibility condition* that no two rows of D are equal, i.e.,

$$\text{for all } i, j, i < j, \text{ there is } q_{ij} \in \{1, \dots, k\} \text{ such that } d_{iq_{ij}} \neq d_{jq_{ij}}. \quad (1.1)$$

An SBM is said to be irreducible if its partition ξ and matrix D satisfy (1.1). This condition is necessary for identifiability, because a random graph generated by an irreducible block model with (ξ, D) can be equivalently generated by any reducible model obtained by *refining* ξ , i.e., splitting one or more of its blocks into smaller sets.

Received 27 January 2020; revision received 20 May 2021; accepted 27 May 2021.

^{*} Postal address: Department of Mathematics and Statistics, P.O. Box 64, FI-00014 University of Helsinki, Finland. Email address: ilkka.norros@elisanet.fi

^{**} Postal address: VTT Technical Research Centre of Finland, Espoo, 02044 VTT, Finland. Email address: hannu.reittu@vtt.fi

^{***} Postal address: Department of Computational Sciences, Institute for Particle and Nuclear Physics, Wigner Research Centre for Physics, P.O. Box 49, H-1525 Budapest, Hungary. Email address: bazso.fulop@wigner.hu

© The Author(s), 2022. Published by Cambridge University Press on behalf of Applied Probability Trust.

A slightly different variant of SBMs has been defined by first drawing the sizes of blocks $A_i^{(n)}$ as independent Poisson random variables and then proceeding with the matrix D as before. SBMs have also been called generalized random graphs; the case of trivial partition $\xi = \{V\}$ yields a random graph with edge probability d_{11} .

SBMs were first defined for community modeling in [8], and community detection has remained their main application as well as a source of research problems—see e.g. [7, 10, 13]. However, large networks appear now in almost all imaginable application areas, and there are grounds to consider SBMs as a rather generic form of the separation of structure and randomness in large real-world systems. Indeed, Szemerédi’s regularity lemma [22], a fundamental result in graph theory, states that *any* large enough graph can be approximated by an SBM-like block structure, where the randomness is replaced by pseudo-randomness in terms of so-called ϵ -regularity, with a number of blocks k that depends only on ϵ . This remarkable combinatorial fact has been utilized to solve hard problems in many areas of mathematics (see e.g. [1, 9]). Looking for SBM-like structures in large real-world graphs and matrices is a generic way to obtain low-dimensional approximations and compressed representations of them.

The typical case of SBM estimation is that the graph $G = (V, E)$ is observed, but the partition ξ is unobserved. For any fixed candidate partition, a corresponding estimate of the matrix D of edge densities is obtained immediately as consisting of empirical means.

If the number of blocks k is known (and the irreducibility condition (1.1) holds, as we always assume), the partition can be identified accurately by several methods, when the graph is sufficiently dense. In mathematical theorems, one considers asymptotics, and dense graph asymptotics means that the densities d_{ij} remain the same when the graph size $n = |V|$ grows to infinity, the relative block sizes staying fixed. The main estimation techniques are maximum likelihood fitting, where the optimization may utilize expectation maximization (e.g. [3]), simulated annealing or Markov chain Monte Carlo algorithms, and spectral methods (e.g. [2, 10]). In sparse graph asymptotics, where all densities decrease to zero while keeping their proportions, identification of the partition may be possible or impossible even with known k , depending on the proportions of the densities; see [7, 10]. Gao *et al.* [5] consider achieving the minimal possible fraction of errors in SBM estimation. Their work is mostly relevant for the sparse case and when k is known.

We focus on the case called model selection, where k is unknown, and on the dense graph setting. In practical tasks of grouping real-world data, it is often a major challenge to select the ‘right’ partition size k . Intuitively, the optimal choice of k should strike a balance between the degree of pseudo-randomness of the edge-level structures and the model complexity specified by the size of the partition. A popular device has been the Akaike information criterion (AIC), which has been applied to block model estimation in, e.g., [11]. The AIC simply adds the ‘number of model parameters’ to the maximal log-likelihood and chooses the model that minimizes the sum.

We apply a more refined information-theoretic approach, Rissanen’s minimum description length (MDL) principle (see [6]), according to which the quality of a model should be measured by the length of the bit string that it yields for unique encoding of the data. In our case, the data consist of the graph G , and the model consists of the partition ξ and the density matrix D . Rosvall and Bergstrom [21] pointed out information theory as a natural choice for community structure analysis, grouping redundant elements into a few communities. The MDL principle, in particular, has been applied to SBMs at least by Peixoto [14]. He derived interesting estimates which indicate that there is an upper limit on the number of communities that can

be detected, $k \sim \sqrt{n}$ as a function of the number of vertices n . However, he did not consider the exact identification of the number of blocks k , which is the focus of the present paper.

Wang and Bickel [23] also used information theory for likelihood-based model selection for SBMs. Their conclusions are similar to ours. In addition, they show the validity of results in a sparse case when the average degree grows at a polylog (polynomial in $\log n$) rate and in the case of degree-corrected block models. They use asymptotic distributions instead of the exact ones that we use. The algorithmic part of their work is also different from ours. They end up in a likelihood-based Bayesian information criterion (BIC) that is asymptotically consistent. Along with a term that corresponds to log-likelihood, there is a term proportional to $k^2 n \log n$ with some tuning coefficient that has to be defined separately in every specific case. So defined, the BIC has a minimum at the right value of k . Instead of such a term, we prefer to use MDL model complexity that is not case-sensitive. Our techniques are different as well.

The main contributions of this paper are the following: (i) the consequent application of the MDL approach, providing (ii) basically simple and transparent techniques, building on Chernoff bounds, to (iii) prove three theorems on model specification. A crucial role is played by Theorem 2, which states that all refinements of ξ increase an MDL-based objective function with high probability. We show that the MDL principle identifies the true partition with high probability among models whose relative block sizes are bounded away from zero. The results are also shown to hold with Poissonian edge weights. We have applied Poissonian block models as a heuristic for finding regular structures in large real-world data matrices, in a method we call *regular decomposition* [11, 12, 15, 17, 18].

The paper is structured as follows. Section 2 defines the main notions: SBMs and the MDL principle. Section 3 presents our main results. The proofs are given in Section 2; some of the auxiliary propositions may be of independent interest. Finally, the results are discussed in Section 3. The proofs apply simple information-theoretic tools collected in Appendix A.

2. Basics and definitions

2.1. Stochastic block models

If $G(V, E)$ is a graph, where V is the set of vertices and E is the set of edges, the *link density* of a nonempty vertex set $X \subseteq V$ is defined as

$$d(X) = \frac{|e(X)|}{\binom{|X|}{2}}, \quad \text{where } e(X) = \{\{v, w\} \in E : v, w \in X\},$$

and $|\cdot|$ denotes the cardinality of a set. Similarly, the link density between two disjoint nonempty vertex sets $X, Y \subseteq V$ is defined as

$$d(X, Y) := \frac{|e(X, Y)|}{|X| |Y|}, \quad \text{where } e(X, Y) = \{\{v, w\} \in E : v \in X, w \in Y\}.$$

Definition 2.1. Let $\epsilon > 0$. A pair of disjoint sets $A, B \subseteq V$ of $G(V, E)$ is called ϵ -regular if, for every $X \subseteq A$ and $Y \subseteq B$ such that $|X| > \epsilon |A|$ and $|Y| > \epsilon |B|$, we have

$$|d(X, Y) - d(A, B)| < \epsilon.$$

The notion of a (binary) stochastic block model (SBM) was defined at the very beginning of this paper. A *Poissonian block model* is defined similarly except that the elements of the matrix D are numbers in $[0, \infty)$, and the random variables e_{ij} have Poisson distributions with

respective parameters d_{ij} . Thanks to the independence assumption, the sums $\sum_{u \in X} \sum_{v \in Y} e_{uv}$ are Poisson-distributed for any disjoint $X, Y \subset V$.

A sequence of random graphs $G_n = (V_n, E_n)$ presenting copies of the same SBM in different sizes can, for definiteness, be constructed as follows.

Construction 2.1. Let $\gamma_1, \dots, \gamma_k$ be positive, distinct real numbers such that $\sum_{i=1}^k \gamma_i = 1$.

1. Divide the interval $(0,1]$ into k segments

$$I_1 = (0, \gamma_1], I_2 = (\gamma_1, \gamma_1 + \gamma_2], \dots, I_k = \left(\sum_{i=1}^{k-1} \gamma_i, 1 \right],$$

and define $\Gamma = \{I_1, \dots, I_k\}$. For $n = 1, 2, \dots$, let the vertices of G_n be

$$V_n = \left\{ \frac{i}{n} : i \in \{1, \dots, n\} \right\}.$$

For each n , let ξ_n be the partition of V_n into the blocks

$$A_i^{(n)} = I_i \cap V_n, \quad i = 1, \dots, k.$$

For small n , we may obtain several empty copies of the empty set numbered as blocks. However, from some n_0 on, all blocks are nonempty and $\xi_n = \{A_1^{(n)}, \dots, A_k^{(n)}\}$ is a genuine partition of V_n . We can then generate a sequence of SBMs based on the sequence (V_n, ξ_n, D) .

2.2. The minimum description length principle

The minimum description length (MDL) principle was introduced by J. Rissanen, inspired by Kolmogorov’s complexity theory (see [6, 20]; our general reference on coding and information theory is [4]). The basic idea is the following: a set \mathcal{D} of data is optimally explained by a model \mathcal{M} when a combined unique encoding of (i) the model and (ii) the data as interpreted in this model is as short as possible. An encoding means here a bijective mapping of objects to a binary prefix code, i.e., a code where no code word is the beginning of another code word, and any sequence of them is thus uniquely decodable.

The principle is best illustrated by our actual case, that of simple graphs. A graph $G = (V, E)$ with $|V| = n$ can always be encoded as a binary string of length $\binom{n}{2} = n(n-1)/2$, where each binary variable corresponds to a vertex pair, and a value 1 (resp. 0) indicates an edge (resp. absence of an edge). Thus, the MDL of G is always at most $\binom{n}{2}$. However, G may have a structure whose disclosure would allow a much shorter description. For example, let G be a bipartite graph consisting of two blocks with cardinality $n/2$. Knowing this, it suffices to code the edges between the blocks, which requires at most $(n/2)^2$ bits, just half of the previous number. However, the partition must still be specified. It requires n additional bits to say which block each vertex belongs to. On the other hand, if the overall density of G is a number $d = h/\binom{n}{2} \neq 1/2$, the edges can be encoded using only $\binom{n}{2}H(d)$ bits, where $H(d)$ denotes the Shannon entropy $H(d) = -d \log_2 d - (1-d) \log_2 (1-d)$. (There are $\binom{N}{K}$ binary sequences of length N that have exactly K 1s, and $\binom{N}{K} \leq \exp(NH(K/N))$; see e.g. [24, IV.2].) However, the integer h must be encoded, and the shortest-length prefix coding for integers requires

$$l^*(h) = \max(0, \log_2(h)) + \max(0, \log_2 \log_2(h)) + \dots \tag{2.1}$$

bits [6, 19].

Definition 2.2. Denote by $\mathcal{M}_{n/k}$ the set of all irreducible (see the condition (1.1)) SBMs (V, ξ, D) with $|V| = \{1, \dots, n\}$ and $\xi = \{V_1, \dots, V_k\}$ such that, for $i, j \in \{1, \dots, k\}$,

$$d_{ij} = \frac{h_{ij}}{|V_i||V_j|}, \quad d_{ii} = \frac{h_{ii}}{\binom{|V_i|}{2}}, \quad h_{ij} \in \{0, \dots, |V_i||V_j|\}, \quad h_{ii} \in \left\{0, \dots, \binom{|V_i|}{2}\right\}.$$

$\mathcal{M}_{n/k}$ is called the modeling space of SBMs with n vertices and k blocks, and

$$\mathcal{M}_n := \bigcup_{1 \leq k \leq n} \mathcal{M}_{n/k} \tag{2.2}$$

is called the modeling space of all SBMs with n vertices.

The condition that the h_{ij} be integers entails that the modeling space \mathcal{M}_n be finite. The models in $\mathcal{M}_{n/k}$ are parameterized by $\Theta = (\xi, D)$. Note that without the irreducibility condition (1.1) there would not be a bijection between SBMs and their parameterizations.

A good model $\Theta \in \mathcal{M}_{n/k}$ for a graph G is one that gives the largest probability for G , and it is called the maximum likelihood model. Denote the parameter of this model by

$$\hat{\Theta}_k(G) := \arg \max_{\Theta \in \mathcal{M}_{n/k}} P(G | \Theta), \tag{2.3}$$

where $P(G | \Theta)$ denotes the probability that the probabilistic model specified by Θ produces G . Part of likelihood optimization is trivial: when a partition $\eta = \{A_1, \dots, A_k\}$ is considered for a given graph G , the optimal edge probabilities are the empirical edge densities,

$$d_{ij} = \frac{|e(A_i, A_j)|}{|A_i||A_j|}, \quad i \neq j, \quad d_{ii} = \frac{|e(A_i)|}{\binom{|A_i|}{2}}. \tag{2.4}$$

As a result, the edge densities are always rational numbers, which explains the corresponding definition used in $\mathcal{M}_{n/k}$.

The nontrivial part of maximum likelihood estimation within $\mathcal{M}_{n/k}$ is to find the optimal partition, but, as noted in the introduction, this has also been well studied in the literature.

Kraft’s inequality (e.g. [4]) implies that if letters are drawn from an alphabet with probabilities p_1, \dots, p_N , there exists a prefix coding with code lengths $\lceil -\log p_1 \rceil, \dots, \lceil -\log p_N \rceil$, and moreover, this coding scheme minimizes the mean code length. Thus, for any probability distribution P on the space of all graphs G with n vertices, there exists a prefix coding with code lengths $\lceil -\log_2 P(\{G\}) \rceil$. In particular, to every graph G_0 with $|V| = n$ we can associate an encoding with code lengths $\lceil -\log_2 P(G | \hat{\Theta}_k(G_0)) \rceil$ corresponding to the SBM $\hat{\Theta}_k(G_0) \in \mathcal{M}_{n/k}$. However, this is not all, since in order to be able to decode we must know what particular model was used. This means that $\hat{\Theta}_k(G_0)$ must also be prefix-encoded, with some code length $L(\hat{\Theta}_k(G_0))$. We end up with a description length called the *two-part MDL* [6]:

$$L(G) = \lceil -\log_2 P(G | \hat{\Theta}_k(G)) \rceil + L(\hat{\Theta}_k(G)). \tag{2.5}$$

A detailed implementation of this principle is proposed in the next section.

3. MDL analysis of stochastic block models

3.1. Block model codes

In the rest of this paper, we define information-theoretic functions in terms of natural logarithms, and certain notions (such as code lengths) should be divided by $\log 2$ to obtain their

values in bits. We denote by $H(\cdot)$ both Shannon’s entropy function for a partition and the entropy of a Bernoulli distribution:

$$H(\eta) = - \sum_{B \in \eta} \frac{|B|}{|V|} \log \frac{|B|}{|V|}, \quad H(p) = -p \log p - (1 - p) \log (1 - p).$$

Definition 3.1. A block model code of a graph $G = (V, E)$ with respect to a partition $\eta = \{B_1, \dots, B_m\}$ of V is a code with the structure described below. The model part is as follows:

1. The sizes of the blocks $B \in \eta$ are given as integers.
2. The edge density $d(B)$ inside each block $B \in \eta$ and the edge density $d(B, B')$ between each pair of distinct blocks $B, B' \in \eta$ are given as the numerators of the rational numbers presenting the exact (empirical) densities.
3. For $v \in V$, define $i_v = \sum_{i=1}^m i 1_{\{v \in B_i\}}$. The partition η is specified by encoding the sequence $(i_v)_{v \in V}$ by a prefix code corresponding to the membership distribution $P(i \in B_j) = |B_j|/n$.

The data part is as follows:

4. The edges inside each block $B_i \in \eta$ are specified by a prefix code corresponding to random distribution of edges with density d_i .
5. The edges between each pair of distinct blocks $B_i, B_j \in \eta$ are specified by a prefix code corresponding to random distribution of edges with density d_{ij} .

Note that a block model code can be given for any graph with respect to any partition of its vertices. Next, we shall specify the functions we use to estimate the lengths of each of the five code parts described in Definition 3.1. Parts 3–5 present long sequences from a finite ‘alphabet’. By classical information theory, basically Kraft’s inequality, the required code length per element is the entropy of the distribution of the ‘letters’.

Parts 1 and 2 encode certain integers. To obtain mathematically tractable and well-behaved estimates, we make some simplifications. Instead of Rissanen’s l^* function (2.1), we use the plain logarithm, and instead of the lengths of numerators of rational numbers, we use the lengths of their denominators. (Besides simplicity, this choice has the desirable effect that the respective code length estimates then depend only on the partition structure, not on the edge densities.) All length estimates should be nonzero. On the other hand, we do not need to care about ceiling functions. These considerations lead us to estimate the code length of a positive integer N as $\log(1 + N)$. With the additional simplification $\binom{m}{2} \approx m^2/2$, the estimate of the length of Part 2 of the code would then be

$$L'_2(G|\eta) = \sum_{B \in \eta} \log \frac{(1 + |B|)^2}{2} + \frac{1}{2} \sum_{B, B' \in \eta, B \neq B'} \log((1 + |B|)(1 + |B'|)). \quad (3.1)$$

So far we have followed the MDL methodology quite strictly, up to the above simplifications. Now, however, we introduce a twist that is motivated only by its usefulness in proving Theorem 3.3: we strengthen the impact of Part 2 by multiplying $L'_2(G|\eta)$ by $|\eta| + 1$.

Thus, we propose to base SBM selection on the following weighted estimate of the length of a block model code:

$$\begin{aligned}
 L(G|\eta) &= L_1(G|\eta) + L_2(G|\eta) + L_3(G|\eta) + L_4(G|\eta) + L_5(G|\eta), \quad \text{where} \\
 L_1(\eta) &= \sum_{i=1}^m \log(1 + |B_i|), \\
 L_2(\eta) &= (m + 1) \left(\sum_{i=1}^m \log \frac{(1 + |B_i|)^2}{2} + \sum_{i < j} \log((1 + |B_i|)(1 + |B_j|)) \right), \quad (3.2) \\
 L_3(G|\eta) &= |V|H(\eta), \\
 L_4(G|\eta) &= \sum_{i=1}^m \binom{|B_i|}{2} H(d(B_i)), \\
 L_5(G|\eta) &= \sum_{i < j} |B_i||B_j|H(d(B_i, B_j)).
 \end{aligned}$$

For sums of only some of the functions L_i , we define $L_{12}(G|n) := L_1(G|n) + L_2(G|n)$, etc.

Proposition 3.1. *For any graph $G = (V, E)$, the model part $L_{123}(G|\eta)$ is monotonically increasing and the data part $L_{45}(G|\eta)$ is monotonically decreasing in η with respect to refinement of partitions. That is, for any partitions η and ζ of V such that $\eta \leq \zeta$, we have*

$$L_{123}(G|\eta) \leq L_{123}(G|\zeta), \quad L_{45}(G|\eta) \geq L_{45}(G|\zeta).$$

Proof. The claim concerning the model part follows from the subadditivity of $\log(1 + x)$ for $x \geq 1$ and from the monotonicity of $\eta \mapsto H(\eta)$.

As regards the data part, note that the internal density of $B \in \eta$ can be written as a convex combination of densities related to the partition $\zeta \cap B$ of B :

$$\begin{aligned}
 d(B) &= \frac{|E(B)|}{\binom{|B|}{2}} \\
 &= \frac{1}{\binom{|B|}{2}} \left(\sum_{C \in \eta \cap B} |E(C)| + \sum_{\substack{C, C' \in \zeta \cap B \\ C \neq C'}} |E(C, C')| \right) \\
 &= \sum_{C \in \eta \cap B} \frac{\binom{|C|}{2}}{\binom{|B|}{2}} d(C) + \frac{1}{2} \sum_{\substack{C, C' \in \zeta \cap B \\ C \neq C'}} \frac{|C||C'|}{\binom{|B|}{2}} d(C, C').
 \end{aligned}$$

By the concavity of H , we then have

$$L_{45}(G|\eta) = \sum_{B \in \eta} \binom{|B|}{2} H(d(B)) + \frac{1}{2} \sum_{\substack{B, B' \in \eta \\ B \neq B'}} |B||B'|H(d(B, B')).$$

$$\begin{aligned}
 &\geq \sum_{B \in \eta} \sum_{C \in \zeta \cap B} \binom{|C|}{2} H(d(C)) \\
 &\quad + \sum_{B \in \eta} \frac{1}{2} \sum_{\substack{C, C' \in \zeta \cap B \\ C \neq C'}} |C||C'| H(d(C, C')) \\
 &\quad + \frac{1}{2} \sum_{\substack{B, B' \in \eta \\ B \neq B'}} \sum_{C \in \zeta \cap B} \sum_{C' \in \zeta \cap B'} |C||C'| H(d(C, C')) \\
 &= L_{45}(G|\zeta), \tag{3.3}
 \end{aligned}$$

where two first sums after the inequality sign come from $L_4(G|\eta)$ and the third from $L_5(G|\eta)$. □

Poissonian block models

Let us now consider Poissonian block models, where the entries of E are Poisson-distributed integers. For a pair of disjoint sets $A, B \subset V$, the set $\{e_{ij} : i \in A, j \in B\}$ is a sample from a distribution $R = (r_\ell)_{\ell \geq 0}$ that is a mixture of Poisson distributions. It would be hard to encode the sample by first estimating the unknown mixture distribution. Instead, we base the code simply on the sample mean

$$e_{AB} = \frac{1}{|A||B|} \sum_{i \in A, j \in B} e_{ij}$$

and encode $\{e_{ij} : i \in A, j \in B\}$ as if it came from a Poisson distribution $P = (p_\ell)$ with parameter e_{AB} . If the e_{ij} really came from a Poisson distribution, a value $e_{ij} = \ell$ would, by Kraft’s inequality, be well encoded by a code word with approximate length

$$-\log p_\ell = -\log \left(\frac{e_{AB}^\ell}{\ell!} e^{-e_{AB}} \right).$$

However, a mixture of different Poisson distributions is not a Poisson distribution. Denote by $R = (r_\ell)$ the empirical distribution of $\{e_{ij} : i \in A, j \in B\}$; i.e., r_ℓ is the number of e_{ij} with value ℓ , divided by $|A||B|$. The fundamental information inequality

$$\sum_{\ell} r_\ell (-\log p_\ell) \geq \sum_{\ell} r_\ell (-\log r_\ell) = H(R) \tag{3.4}$$

implies that this encoding is suboptimal for arbitrary disjoint subsets A and B , but it is optimal when A and B are blocks of the model partition ξ and R comes from a sample from a pure Poisson distribution. (The inequality (3.4) is a special case of the nonnegativity of the Kullback–Leibler divergence; it says that using a code recommended by Kraft’s inequality with wrong probabilities increases the expected code word length.) This suboptimality is no problem, because for our purposes it only improves the contrast between ξ and other partitions of V .

Using the above coding rule for an arbitrary partition $\eta = \{B_1, \dots, B_m\}$, the amount of code needed to encode all of the e_{ij} is

$$\begin{aligned} & \sum_{i=1}^m \binom{|B_i|}{2} \sum_{\ell \geq 0} r_\ell^{(B_i)} \left(-\log \left(\frac{e_{B_i}^\ell}{\ell!} e^{-e_{B_i}} \right) \right) \\ & + \sum_{i < j} |B_i||B_j| \sum_{\ell \geq 0} r_\ell^{(B_i B_j)} \left(-\log \left(\frac{e_{B_i B_j}^\ell}{\ell!} e^{-e_{B_i B_j}} \right) \right) \\ & = \sum_{i=1}^m \binom{|B_i|}{2} (-\log e_{B_i}) \sum_{\ell \geq 0} \ell r_\ell^{(B_i)} + \sum_{i < j} |B_i||B_j| (-\log e_{B_i B_j}) \sum_{\ell \geq 0} \ell r_\ell^{(B_i B_j)} \\ & + \sum_{i=1}^m \binom{|B_i|}{2} \sum_{\ell \geq 0} r_\ell^{(B_i)} \log \ell! + \sum_{i < j} |B_i||B_j| \sum_{\ell \geq 0} r_\ell^{(B_i B_j)} \log \ell! \\ & + \sum_{i=1}^m \binom{|B_i|}{2} e_{B_i} + \sum_{i < j} |B_i||B_j| e_{B_i B_j} \\ & = \sum_{i=1}^m \binom{|B_i|}{2} \phi(e_{B_i}) + \sum_{i < j} |B_i||B_j| \phi(e_{B_i B_j}) + \frac{1}{2} \sum_{v, w \in V, v \neq w} \log e_{vw}! + \sum_{v, w \in V, v \neq w} e_{vw}, \end{aligned}$$

where

$$e_{B_i} := \frac{\sum_{v, w \in B_i} e_{vw}}{|B_i|(|B_i| - 1)}, \quad e_{B_i B_j} := \frac{\sum_{v, w \in B_i} e_{vw}}{|B_i||B_j|},$$

the r -symbols refer to the empirical distribution of the e_{vw} , and $\phi(x) = -x \log x$. Now, we define our MDL-based objective function for a Poissonian block model E with respect to any partition $\eta = \{B_1, \dots, B_m\}$ as

$$\begin{aligned} L(E|\eta) &= L_0(E|\eta) + L_1(E|\eta) + L_2(E|\eta) + L_3(E|\eta) + L_4(E|\eta) + L_5(E|\eta), \\ L_0(E|\eta) &= \frac{1}{2} \sum_{v, w \in V, v \neq w} \log e_{vw}! + \sum_{v, w \in V, v \neq w} e_{vw}, \\ L_1(E|\eta) &= \sum_{i=1}^m \log(1 + |B_i|), \\ L_2(E|\eta) &= (m + 1) \left(\sum_{i=1}^m \log \frac{(1 + |B_i|)^2}{2} + \sum_{i < j} \log((1 + |B_i|)(1 + |B_j|)) \right), \tag{3.5} \\ L_3(E|\eta) &= |V|H(\eta), \\ L_4(E|\eta) &= \sum_{i=1}^m \binom{|B_i|}{2} \phi(e_{B_i}), \\ L_5(E|\eta) &= \sum_{i < j} |B_i||B_j| \phi(e_{B_i B_j}). \end{aligned}$$

Note that the term $L_0(E|\eta)$ is independent of the partition η and can be neglected when minimizing over η . The parts $L_1(E|\eta)$ and $L_2(E|\eta)$ are copied from the case of plain graphs.

One application of Poisson block models was found in [16], which used them to analyze graph distance matrices. This approach can help in finding communities in large and sparse networks. In [15], it was suggested that the Poisson block model objective function (3.5) could be used for any matrix with nonnegative entries, resulting in the same cost function as in the integer case.

3.2. Accuracy of model selection

Our results on MDL-based model selection are summarized in the following theorem. It is formulated in terms of the asymptotic behavior of a model sequence as specified by Construction 1. In this framework, an event is said to happen *with high probability* if its probability tends to 1 when $n \rightarrow \infty$.

Consider a sequence of irreducible SBMs (G_n, ξ_n) based on a vector $(\gamma_1, \dots, \gamma_k)$ of relative block sizes and a matrix $D = (d_{ij})_{i,j=1}^k$ of edge probabilities. Define

$$\vartheta(\eta, \xi_n) = \frac{1}{n} \max_{B \in \eta} \min_{A \in \xi_n} |B \setminus A|.$$

Thus, $\vartheta(\eta, \xi_n) = 0$ if and only if η is a refinement of ξ_n . For $\sigma \in (0, 1)$, denote by $\mathcal{P}_{n,\sigma}$ the set of partitions of V_n whose blocks each have at least σn vertices. Obviously, $|\eta| \leq 1/\sigma$ for $\eta \in \mathcal{P}_{n,\sigma}$. If $v \in V_n$ and $B \in \eta$, denote by $\eta_{v,B}$ the partition obtained from η by moving vertex v to block B (if $v \in B$, then $\eta_{v,B} = \eta$).

Theorem 3.1. Fix some minimal relative block size $\sigma \in (0, \min_i \gamma_i)$. There are two numbers $\epsilon_0 > 0$ and $\kappa_0 > 0$ such that the following holds with high probability: Let a partition $\eta \in \mathcal{P}_{n,\sigma}$ satisfy $|\eta| \geq k$ and $\vartheta(\eta, \xi_n) \leq \epsilon_0$. If $A \in \xi_n$ and $B \in \eta$ are such that $0 < |B \setminus A|/n \leq \epsilon_0$, choose any vertex $v \in B \setminus A$. Then there is a block $B' \in \eta$ such that

$$L(G_n|\eta_{v,B'}) < L(G_n|\eta) - \kappa_0 n.$$

Moreover, $L(G_n|\eta) \geq L(G_n|\tilde{\eta})$, where $\tilde{\eta}$ is a refinement of ξ_n .

Theorem 3.2. For any $\epsilon > 0$ and positive integer m , there is a constant $c_{\epsilon,m} > 0$ such that the following holds with high probability:

$$\text{For all } \eta \text{ such that } |\eta| \leq m \text{ and } \vartheta(\eta, \xi_n) > \epsilon, \quad L(G_n|\xi_n \vee \eta) < L(G_n|\eta) - c_{\epsilon,m} n^2.$$

Theorem 3.3. With high probability, minimizing $L(\eta)$ identifies ξ_n among all refinements η of ξ_n .

Corollary 3.1. For any $\sigma \in (0, \min_i \gamma_i)$, ξ_n is with high probability the unique minimizer of $L(G_n|\eta)$ for $\eta \in \mathcal{P}_{n,\sigma}$, and

$$L(G_n|\eta) \geq L(G_n|\eta \vee \xi_n) \geq L(G_n|\xi_n).$$

The corresponding results hold *mutatis mutandis* for Poissonian block models. The proofs are essentially identical. The differences, required by the replacement of binomial distributions by Poisson distributions, are indicated at the end of Appendix A.

4. Proofs of the theorems

Through this section, we consider a sequence $(G_n|\xi_n)$ of increasing versions of a fixed SBM based on a vector $(\gamma_1, \dots, \gamma_k)$ of relative block sizes and a matrix $D = (d_{ij})_{i,j=1}^k$ of edge probabilities, as specified in Construction 2.1.

Proof of Theorem 3.1. Let $\epsilon, \delta > 0$ be small numbers and m a positive integer to be specified. They can be chosen so that the following hold:

- The value of ϵ is small enough so that η is close to $\eta \vee \xi_n$ when $\mathfrak{d}(\eta, \xi_n) \leq \epsilon$:

$$m\epsilon < \delta \min_i \gamma_i. \tag{4.1}$$

- All the differing link probabilities are widely separated in δ units:

$$\delta \leq \frac{1}{m} \min \{ |d_{ij_1} - d_{ij_2}| : i, j_1, j_2 \in \{1, \dots, k\}, d_{ij_1} \neq d_{ij_2} \}. \tag{4.2}$$

For any $A_i, A_j \in \xi_n$ (possibly $i = j$), we then have

$$|d(A_i, A_j) - d_{ij}| + m\epsilon \leq \delta \tag{4.3}$$

with high probability.

Let η be a partition of V_n such that $\mathfrak{d}(\eta, \xi_n) \leq \epsilon$. For $i = 1, \dots, k$, denote by B_i a block $B \in \eta$ such that $|B \cap A_i|$ is maximal. Then $|B_i \setminus A_i|/n \leq \epsilon$, and the blocks B_1, \dots, B_k are distinct for small ϵ . Let us number arbitrarily any remaining blocks of $\eta = \{B_1, \dots, B_{|\eta|}\}$. We can further assume that if $|\eta| > k$, then for each $q > k$ there is a unique j_q such that $|B_q \setminus A_{j_q}|/n \leq \epsilon$ (for $q \leq k$, let $j_q = q$).

Assume now that $v \in B_s \setminus A_j, |B_s \setminus A_j|/n \leq \epsilon$, and $v \in A_i \cap B_s$ with $i \neq j$. We choose $B' = B_i$ and compare the partitions η and η_{v, B_i} . Let $b_q = |B_q|, q = 1, \dots, |\eta|$, and $\tilde{B}_i = B_i \cup \{v\}, \tilde{B}_s = B_s \setminus \{v\}$. Then

$$\begin{aligned} &L_{45}(G_n|\eta) - L_{45}(G_n|\eta_{v, B_i}) \\ &= \binom{b_i}{2} H(d(B_i)) + \binom{b_s}{2} H(d(B_s)) - \binom{b_i + 1}{2} H(d(\tilde{B}_i)) - \binom{b_s - 1}{2} H(d(\tilde{B}_s)) \\ &+ \sum_{q \neq i, s} \left[b_i b_q H(d(B_i, B_q)) + b_s b_q H(d(B_s, B_q)) \right. \\ &\quad \left. - (b_i + 1) b_q H(d(\tilde{B}_i, B_q)) + (b_s - 1) b_q H(d(\tilde{B}_s, B_q)) \right] \\ &+ b_i b_s H(d(B_i, B_j)) - (b_i + 1)(b_s - 1) H(d(\tilde{B}_i, \tilde{B}_s)). \end{aligned} \tag{4.4}$$

Consider first the sum over q . Leaving out the common factor b_q , each term of the sum can be written as

$$\begin{aligned} &b_s \left[H(d(B_s, B_q)) - \frac{b_s - 1}{b_s} H(d(\tilde{B}_j, B_q)) \right] \\ &- (b_i + 1) \left[H(d(\tilde{B}_i, B_q)) - \frac{b_i}{b_i + 1} H(d(B_i, B_q)) \right] \\ &= b_s \left[H \left(\frac{b_s - 1}{b_s} d(\tilde{B}_s, B_q) + \frac{1}{b_s} d(\{v\}, B_q) \right) \right. \\ &\quad \left. - \frac{b_s - 1}{b_s} H(d(\tilde{B}_j, B_q)) - \frac{b_i}{b_i + 1} H(d(\tilde{B}_i, B_q)) + \frac{b_i}{b_i + 1} H(d(B_i, B_q)) \right] \end{aligned}$$

$$\begin{aligned}
 & - \frac{b_s - 1}{b_s} H(d(\tilde{B}_s, B_q)) - \frac{1}{b_s} H(d(\{v\}, B_q)) \Big] \\
 & - (b_i + 1) \left[H \left(\frac{b_i}{b_i + 1} d(B_i, B_q) + \frac{1}{b_i + 1} d(\{v\}, B_q) \right) \right. \\
 & \left. - \frac{b_i}{b_i + 1} H(d(B_i, B_q)) - \frac{1}{b_i + 1} H(d(\{v\}, B_q)) \right]
 \end{aligned}$$

(note the addition and subtraction of the term $H(d(\{v\}, B_q))$). Using Lemmas A.2 and A.3 (ϵ -regularity) and the relations (4.1), (4.2), and (4.3), the last expression can be set to be, with high probability, arbitrarily close to the number

$$D_B(d_{iq} \| d_{jq}) - D_B(d_{ijq} \| d_{ijq}) = D_B(d_{ijq} \| d_{jq})$$

(the function $D_B(\cdot \| \cdot)$, the Kullback–Leibler divergence of Bernoulli distributions, is defined in (A.2)). Thus, the sum over q is, with high probability, close to

$$b_q \sum_{q \neq i, s} D_B(d_{ijq} \| d_{jq}).$$

Let us then turn to the remaining parts of (4.4), which refer to two codings of the internal links of $B_i \cup B_s$. Similarly as above, we can add and subtract terms to transform these parts into

$$\begin{aligned}
 & \binom{b_s}{2} \left[H \left(\frac{\binom{b_s-1}{2}}{\binom{b_s}{2}} d(\tilde{B}_s) + \frac{b_s - 1}{\binom{b_s}{2}} d(\{v\}, \tilde{B}_s) \right) \right. \\
 & \left. - \frac{\binom{b_s-1}{2}}{\binom{b_s}{2}} H(d(\tilde{B}_s)) - \frac{b_s - 1}{\binom{b_s}{2}} H(d(\{v\}, \tilde{B}_s)) \right] \\
 & - \binom{b_i + 1}{2} \left[H \left(\frac{\binom{b_i}{2}}{\binom{b_i+1}{2}} d(B_i) + \frac{b_i}{\binom{b_i+1}{2}} d(\{v\}, B_i) \right) \right. \\
 & \left. - \frac{\binom{b_i}{2}}{\binom{b_i+1}{2}} H(d(B_i)) - \frac{b_i}{\binom{b_i+1}{2}} H(d(\{v\}, B_i)) \right] \\
 & + b_i b_s \left[H \left(\frac{b_s - 1}{b_s} d(B_i, \tilde{B}_s) + \frac{1}{b_s} d(\{v\}, B_i) \right) \right. \\
 & \left. - \frac{b_s - 1}{b_s} H(d(B_i, \tilde{B}_s)) - \frac{1}{b_s} H(d(\{v\}, B_i)) \right] \\
 & - (b_i + 1)(b_s - 1) \left[H \left(\frac{b_i}{b_i + 1} d(B_i, \tilde{B}_s) + \frac{1}{b_i + 1} d(\{v\}, \tilde{B}_s) \right) \right. \\
 & \left. - \frac{b_i}{b_i + 1} H(d(B_i, \tilde{B}_s)) - \frac{1}{b_i + 1} H(d(\{v\}, \tilde{B}_s)) \right] \\
 & \approx (b_s - 1) D_B(d_{ij} \| d_{jj}) - (b_i + 1) D_B(d_{ii} \| d_{ii}) + b_i D_B(d_{ii} \| d_{ij}) - (b_s - 1) D_B(d_{ij} \| d_{ij}) \\
 & \approx b_s D_B(d_{ij} \| d_{jj}) + b_i D_B(d_{ii} \| d_{ij}).
 \end{aligned}$$

By the above analysis of (4.4), we have obtained

$$\begin{aligned}
 &L_{45}(G_n|\eta) - L_{45}(G_n|\eta_{v,B_i}) \\
 &\approx b_q \sum_{q \neq i,s} D_B(d_{jj_q} \| d_{ij_q}) + b_j D_B(d_{ij} \| d_{jj}) + b_i D_B(d_{ii} \| d_{ij}),
 \end{aligned} \tag{4.5}$$

where the approximation can be made arbitrarily accurate by the choice of ϵ , m , and δ . By the irreducibility assumption (1.1), there is a block A_{j_q} such that $d_{jq_i} \neq d_{j_qj}$, with the possibility that $q \in \{i, s\}$. It follows that at least one of the $D_B(x\|y)$ terms in (4.5) is positive. Let $\kappa^* = \min \{D_B(d_{ij_1} \| d_{ij_2}) : d_{ij_1} \neq d_{ij_2}\}$. Thus, with high probability,

$$L_{45}(G_n|\eta) - L_{45}(G_n|\eta_{v,B_i}) > \frac{1}{2}(\kappa^* \min_i \gamma_i)n.$$

Choose $\epsilon_0 = \epsilon$ and $\kappa_0 = \kappa^* \min_i \gamma_i/2$. As regards the other code parts, it is easy to compute that

$$L_3(G_n|\eta) - L_3(G_n|\eta_{v,B_i}) = n(H(\eta) - H(\eta_{v,B_i})) \rightarrow \log \frac{\gamma_j}{\gamma_i},$$

and the changes of L_1 and L_2 when moving from η to η_{v,B_i} are negligible. This concludes the proof of the claim concerning moving a single vertex from B_j to B_i . The second claim follows from noting that the first claim continues to hold *a fortiori* when moving similarly all remaining vertices that prevent η from being a refinement of ξ_n .

Proof of Theorem 3.2. Fix an $\epsilon \in (0, 1)$ and let η be a partition of V_n such that $\partial(\eta, \xi_n) > \epsilon$. Lemma 3.1 yields that $L_{45}(G_n|\eta) \geq L_{45}(G_n|\eta \vee \xi_n)$.

By assumption, there exists $B \in \eta$ such that $|B \setminus A| > \epsilon n$ for every $A \in \xi_n$. It is easy to see that there must be (at least) two distinct blocks, say A_i and A_j , such that

$$\min \{|A_i \cap B|, |A_j \cap B|\} \geq \frac{\epsilon}{k-1}n. \tag{4.6}$$

By the irreducibility assumption (1.1), there is a block A_q such that $d_{qi} \neq d_{qj}$, with the possibility that $q \in \{i, j\}$. Fix an arbitrary $\delta > 0$ to be specified later. By ϵ -regularity (Claim 2 of Lemma A.3), with high probability, every choice of a partition η satisfying $\exists B \in \eta \forall A \in \xi_n |B \setminus A| > \epsilon n$ results in some blocks A_i, A_j, A_q with the above characteristics plus the regularity properties

$$|d(A_i \cap B, A_q \cap B') - d_{iq}| \leq \delta, \quad |d(A_j \cap B, A_q \cap B') - d_{jq}| \leq \delta, \tag{4.7}$$

where B' denotes a block of η that maximizes $|A_q \cap B'|$ (note that because $|\eta| \leq m$, $|A_q \cap B'| \geq |A_q|/m$). By the concavity of H ,

$$\begin{aligned}
 &|A_i \cap B||A_q \cap B'|H(d(A_i \cap B, A_q \cap B')) \\
 &+ |A_j \cap B||A_q \cap B'|H(d(A_j \cap B, A_q \cap B')) \\
 &= |(A_i \cup A_j) \cap B||A_q \cap B'| \left[\frac{|A_i \cap B|}{|(A_i \cup A_j) \cap B|} H(d(A_i \cap B, A_q \cap B')) \right. \\
 &\quad \left. + \frac{|A_j \cap B|}{|(A_i \cup A_j) \cap B|} H(d(A_j \cap B, A_q \cap B')) \right] \\
 &< |(A_i \cup A_j) \cap B||A_q \cap B'|H(d((A_i \cup A_j) \cap B, A_q \cap B')).
 \end{aligned}$$

In the case that $q \in \{i, j\}$ and $B = B'$, we obtain a similar equation where $|A_q \cap B|$ is partly replaced by $|A_q \cap B| - 1$. Because of (4.7) and (4.6), the difference between the sides of the inequality has a positive lower bound that holds with high probability. On the other hand, this difference is part of the overall concavity inequality (3.3) in the proof of Lemma 3.1. Thus, we can choose $c_{\epsilon, m}$ so that

$$L_{45}(G_n|\xi_n \vee \eta) < L_{45}(G_n|\eta) - c_{\epsilon, m}n^2$$

with high probability. It remains to note that $L_{123}(G_n|\xi_n \vee \eta) - L_{123}(G_n|\eta)$ is proportional to n at most. □

Let us now turn to preparing the proof of Theorem 3.3 concerning refinements of ξ_n . For any partition $\eta \geq \xi_n$, write

$$\Delta L(\eta) = L(G_n|\eta) - L(G_n|\xi_n)$$

and, for individual components of the code,

$$\Delta L_{123}(\eta) := L_{123}(G_n|\eta) - L_{123}(G_n|\xi_n), \quad \text{etc.}$$

Lemma 4.1. *For any partition η of V_n ,*

$$L_2(\eta) = (|\eta| + 1)^2 \sum_{B \in \eta} \log(1 + |B|) - |\eta|(|\eta| + 1) \log 2.$$

Proof. The claim is proved by elaborating (3.1) as follows:

$$\begin{aligned} L'_2(\eta) &= 2 \sum_{B \in \eta} \log(1 + |B|) - |\eta| \log 2 + \frac{1}{2} \sum_{B \in \eta} \sum_{B' \in \eta \setminus \{B\}} (\log(1 + |B|) + \log(1 + |B'|)) \\ &= 2 \sum_{B \in \eta} \log(1 + |B|) - |\eta| \log 2 + \sum_{B \in \eta} \log(1 + |B|) \sum_{B' \in \eta \setminus \{B\}} 1 \\ &= 2 \sum_{B \in \eta} \log(1 + |B|) - |\eta| \log 2 + (|\eta| - 1) \sum_{B \in \eta} \log(1 + |B|) \\ &= (|\eta| + 1) \sum_{B \in \eta} \log(1 + |B|) - |\eta| \log 2, \end{aligned}$$

where the second line follows from symmetry. □

Lemma 4.2. *For $\eta \geq \xi_n$,*

$$\Delta L_2(\eta) \geq (|\eta| - k)(|\eta| + k + 1) (k \log n - c_n(\xi_n) - \log 2), \tag{4.8}$$

$$\log \Delta L_{23}(\eta) \leq \log n + 2 \log |\eta| + 2 \log 2, \tag{4.9}$$

where

$$c_n(\xi_n) = - \sum_{A \in \xi} \log \frac{1 + |A|}{n}. \tag{4.10}$$

Proof of (4.8). Note first that it follows from the subadditivity of $\log(1 + x)$ for $x \geq 1$ that for any $\eta \geq \xi$ we have

$$\sum_{B \in \eta} \log(1 + |B|) \geq \sum_{A \in \xi_n} \log(1 + |A|). \tag{4.11}$$

Using Lemma 4.1 and (4.11), we have

$$\begin{aligned}
 \Delta L_2(\eta) &= (|\eta| + 1)^2 \sum_{B \in \eta} \log(1 + |B|) - |\eta|(|\eta| + 1) \log 2 \\
 &\quad - (k + 1)^2 \sum_{A \in \xi_n} \log(1 + |A|) + k(k + 1) \log 2 \\
 &\geq (|\eta| - k)(|\eta| + k + 2) \sum_{A \in \xi_n} \log(1 + |A|) - (|\eta| - k)(|\eta| + k + 1) \log 2 \\
 &\geq (|\eta| - k)(|\eta| + k + 1) \left(\sum_{A \in \xi_n} \log(1 + |A|) - \log 2 \right) \\
 &= (|\eta| - k)(|\eta| + k + 1) \left(\sum_{A \in \xi_n} \left(\log n + \log \frac{1 + |A|}{n} \right) - \log 2 \right) \\
 &\geq (|\eta| - k)(|\eta| + k + 1)(k \log n - c_n(\xi_n) - \log 2).
 \end{aligned}$$

Proof of (4.9) We now use (4.11) in the other direction, and note that the concavity of $\log x$ implies

$$\begin{aligned}
 \sum_{B \in \eta} \log(1 + |B|) &= |\eta| \sum_{B \in \eta} \frac{1}{|\eta|} \log(1 + |B|) \\
 &\leq |\eta| \log \left(\sum_{B \in \eta} \frac{1 + |B|}{|\eta|} \right) \\
 &= |\eta| \log \left(1 + \frac{n}{|\eta|} \right) \\
 &\leq \log e^n = n.
 \end{aligned}$$

We then have

$$\begin{aligned}
 &\log \Delta L_{23}(\eta) \\
 &\leq \log \left((|\eta| - k)(|\eta| + k + 2) \sum_{B \in \eta} \log(1 + |B|) - (|\eta| - k)(|\eta| + 1) \log 2 + nH(\eta|\xi_n) \right) \\
 &\leq \log \left((|\eta| - k)(|\eta| + k + 2)n + n(H(\eta) - H(\xi_n)) \right) \\
 &\leq \log \left(n [(|\eta| - k)(|\eta| + k + 2) + \log |\eta|] \right). \quad \square
 \end{aligned}$$

Proposition 4.1. For any refinement η of ξ_n , we have

$$-\Delta L_{45}(\eta) \stackrel{(st)}{\leq} \sum_{j=1}^{M(|\eta|) - M(k)} (\log 2 + Y_j), \quad (4.12)$$

where (st) refers to stochastic order, the Y_i are independent $\text{Exp}(1)$ random variables, and

$$M(x) = \frac{x(x+1)}{2}. \quad (4.13)$$

Proof. Here we apply results presented in Appendix A. Denote by $\eta \cap A_i$ the subset of η whose members are subsets of the block A_i of ξ_n . Writing the edge code lengths of the coarser and finer partition similarly as in (3.3), taking the difference, and using (A.5), we obtain

$$\begin{aligned}
 & -L_{45}(\eta) \\
 = & \sum_{i=1}^k \left(\sum_{B \in \eta \cap A_i} \binom{|B|}{2} D_B(d(B) \| d_{ii}) + \frac{1}{2} \sum_{\substack{B, B' \in \eta \cap A_i \\ B \neq B'}} |B||B'| D_B(d(B, B') \| d_{ii}) \right. \\
 & \left. - \binom{|A_i|}{2} D_B(d(A_i) \| d_{ii}) \right) \\
 & + \sum_{i < j} \left(\sum_{B \in \eta \cap A_i} \sum_{B' \in \eta \cap A_j} |B||B'| D_B(d(B, B') \| d_{ij}) - |A_i||A_j| D_B(d(A_i, A_j) \| d_{ij}) \right).
 \end{aligned} \tag{4.14}$$

Applying Proposition A.4 to each term of both outer sums now yields the claim, because

$$\sum_{i=1}^k (M(|\eta \cap A_i|) - 1) + \sum_{i < j} (|\eta \cap A_i| |\eta \cap A_j| - 1) = M(|\eta|) - M(k). \quad \square$$

It is rather surprising that the stochastic bound (4.12) depends only on the number of blocks in η —not on their relative sizes, nor on the overall model size n .

Proof of Theorem 3.3. Let η be a refinement of ξ_n and recall Lemma 3.1: refining the partition with respect to ξ_n yields a gain, based on the concavity of H , in the data code part L_{45} , but additional costs in the model code part L_{123} . We have to relate these to each other. Since $L_1(G_n | \eta)$ is just a small addition to $L_2(G_n | \eta)$ (see Lemma 4.1), we can ignore it and focus on $L_2(G_n | \eta)$ and $L_3(G_n | \eta)$ as regards the model part.

The refinement gain in code part L_{45} was bounded in Proposition 4.1 stochastically by $Exp(1)$ random variables. The rate function (see the beginning of Appendix A) of the distribution $Exp(1)$ is

$$I_E(x) = x - 1 - \log x.$$

Proposition 4.1 yields, using (A.1) and Proposition A.1 that for $y > \log 2$

$$\begin{aligned}
 & \mathbb{P}(-\Delta L_{45}(\eta) > y) \\
 & \leq \mathbb{P} \left(\sum_{j=1}^{M(|\eta|) - M(k)} Y_j > y - (M(|\eta|) - M(k)) \log 2 \right) \\
 & \leq \exp \left(-(M(|\eta|) - M(k)) I_E \left(\frac{y - (M(|\eta|) - M(k)) \log 2}{M(|\eta|) - M(k)} \right) \right) \\
 & \leq \exp \left(-y + (M(|\eta|) - M(k)) \log y \right) \left(\frac{2e}{M(|\eta|) - M(k)} \right)^{M(|\eta|) - M(k)}.
 \end{aligned}$$

For two refinements of ξ_n , write $\eta' \sim \eta$ if the block sizes of η' in each A_i are identical to those of η . The number of refinements η' of ξ_n with $\eta' \sim \eta$ is bounded above by

$$\exp \left(\sum_{A \in \xi_n} |A| H(\eta \cap A) \right) = e^{nH(\eta|\xi_n)},$$

where $H(\eta \cap A)$ denotes the entropy of the partition of A induced by η and $H(\eta|\xi_n)$ the conditional entropy of η given ξ_n . On the other hand, we have

$$\Delta L_3(\eta) = nH(\eta) - nH(\xi_n) = nH(\eta|\xi_n).$$

Let $m = |\eta|$. With $y = \Delta L'_{23}$, the union bound and Lemma 2 yield

$$\begin{aligned} & \mathbb{P} \left(\inf_{\eta' \sim \eta} \Delta L(\eta) < 0 \right) \\ & \leq \exp \left(nH(\eta|\xi_n) - \Delta L_3(\eta) - \Delta L'_2(\eta) + \frac{1}{2}(m-k)(m+k+1) \log \Delta L'_{23}(\eta) \right) \\ & \quad \cdot \left(\frac{2e}{M(m) - M(k)} \right)^{M(m) - M(k)} \\ & \leq \exp \left(-(m-k)(m+k+1) (k \log n - c_n(\xi_n) - \log 2) \right. \\ & \quad \left. + \frac{1}{2}(m-k)(m+k+1) \log (n [(m-k)(m+k+2) + \log m]) \right. \\ & \quad \left. - \frac{1}{2}(m-k)(m+k+1) (\log ((m-k)(m+k+1)) - 2 \log 2 - 1) \right) \\ & = \exp \left\{ -(m-k)(m+k+1) \left(k - \frac{1}{2} \right) \log n \right. \\ & \quad \left. + \frac{1}{2}(m-k)(m+k+1) \rho_1(m) \right\}, \end{aligned} \tag{4.15}$$

where

$$\rho_1(m) = \log \left(\frac{(m-k)(m+k+2) + \log m}{(m-k)(m+k+1) - 2c_n(\xi) - 4 \log 2 - 1} \right), \tag{4.16}$$

and $\rho_1(m) = o(m)$ as $m \rightarrow \infty$.

Consider all different block size sequences of partitions $\eta > \xi$ such that $|\eta| = m$:

$$\begin{aligned} & \kappa_{11} \geq \kappa_{12} \geq \dots \geq \kappa_{1\sigma_1}; \dots; \kappa_{k1} \geq \dots \geq \kappa_{k\sigma_k}, \\ & \sum_{j=1}^{\sigma_i} \kappa_{ij} = |A_i|, \quad i = 1, \dots, k; \quad \sum_{i=1}^k \sigma_i = m. \end{aligned}$$

A rough upper bound of their number is

$$k^{m-k} n^{m-k} = \exp \left((m-k) \log n + (m-k) \log k \right). \tag{4.17}$$

Indeed, choose first the number of subblocks in each block A_i of ξ_n , then the sizes of the subblocks, each in decreasing order. Note that within each block of ξ_n , the size of its smallest subblock is determined by the others; i.e., k of the m block sizes ‘come for free’.

Using (4.15), (4.17), and the union bound, we obtain for any integer $m > k$

$$\begin{aligned} & \mathbb{P}\left(\inf_{\eta>\xi, |\eta|=m} \Delta L(\eta) < 0\right) \\ & \leq \exp\left\{(m-k)\log n + (m-k)\log k - (m-k)(m+k+1)\left(k - \frac{1}{2}\right)\log n\right. \\ & \quad \left. + \frac{1}{2}(m-k)(m+k+1)\rho_1(m)\right\} \\ & = \exp\left\{-(m-k)\left((m+k+1)\left(k - \frac{1}{2}\right) - 1\right)\log n + m^2\rho_2(m)\right\}, \end{aligned} \tag{4.18}$$

where

$$\rho_2(m) = \frac{1}{m^2}\left(\frac{1}{2}(m-k)(m+k+1)\rho_1(m) + (m-k)\log k\right),$$

and $\rho_2(m) = o(m)$. Adding and subtracting $m \log n$ and using the facts $1 \leq k < m$ and $k \leq m$, (4.18) can be further bounded by

$$\begin{aligned} & \exp\left(-m \log n - \left[(m-k)\left(\frac{m}{2} - \frac{1}{2}\right) - m\right]\log m + m^2\rho_2(m)\right) \\ & = \exp\left(-m \log n - \left(\frac{1}{4}\log m - \rho_2(m)\right)m^2 - \frac{m}{2}\left(\frac{m}{2} - k - 3 + \frac{k}{m}\right)\right). \end{aligned}$$

Now define

$$m^* := \inf\left\{m \geq k + 1 : \rho_2(m') \leq \frac{1}{4}\log m' \quad \forall m' \geq k + 1\right\} \vee (2k + 6).$$

Let $n \geq m^*$. Then we have

$$\mathbb{P}\left(\inf_{\eta>\xi, |\eta|=m} \Delta L(\eta) < 0\right) \leq \begin{cases} n^{-k} \exp\left(m^2\rho_2(m) + \frac{m}{2} + 4\right), & m < m^*, \\ e^{-m \log n}, & m \geq m^*. \end{cases}$$

Finally,

$$\begin{aligned} \mathbb{P}\left(\inf_{\eta>\xi} \Delta L(\eta) < 0\right) & \leq \sum_{m=k+1}^n \mathbb{P}\left(\inf_{\eta>\xi, |\eta|=m} \Delta L(\eta) < 0\right) \\ & \leq n^{-k} \sum_{m=k+1}^{m^*-1} e^{m^2\rho_2(m)+m/2+4} + \sum_{m=m^*}^n e^{-m \log n} \\ & \leq \text{const}(k) \cdot n^{-k} + \frac{n^{-m^*}}{1 - n^{-1}} \rightarrow 0, \quad \text{as } n \rightarrow \infty. \end{aligned}$$

5. Discussion

The results of this paper provide rather clear insight into the description length optimization landscape for SBMs. When a partition η deviates only a little from the true partition ξ_n and $|\eta| = k = |\xi_n|$, minimization of $L_{45}(\eta)$ (likelihood maximization) in a single round by

moving each vertex to its most appropriate block leads to exact identification of ξ_n (Theorem 3.1). L_{45} is essentially the log-likelihood function. With $|\eta| > k$, however, minimization of L_{45} may lead instead to some refinement of ξ_n , with a value $L_{45}(\eta)$ that is strictly smaller than $L_{45}(\xi_n)$. We have quantified this additional gain, which is of order n , and shown that it is outweighed by additional model complexity, measured by functions suggested by the MDL principle. MDL seems outstanding as a theoretically solid way to weigh likelihood gain against model complexity.

In the earlier paper [18], we and our coauthors analyzed several questions on regular decomposition, i.e., our MDL-based methodology, focusing on the case in which k is known. It was found that not a single vertex can be misplaced without a substantial penalty in L_{45} , proportional to n . In this sense, the global minimum is well separated from all local minima. However, there is no easy way of testing whether a minimum found by a greedy algorithm is the global one. This problem is acute with real-life data, in which there is no SBM-type ground truth. It is an interesting theoretical challenge to determine the probability of finding the global optimum of MDL using a greedy minimization algorithm.

The paper [18] proposed a way to speed up the L_{45} minimization when k is known and $n = |V|$ arbitrarily large. The idea is that a moderately sized random sample from V reveals the block structure, which can be identified by some brute force method, in our case by repeating the greedy algorithm many times with random initial partitions. The majority of the vertices can then be placed into their blocks with very high accuracy.

In [18], we also compared regular decomposition by simulations with a spectral clustering method. The results suggested that the spectral clustering method is sensitive to the relative block sizes and requires them to be well balanced, whereas regular decomposition is not sensitive to the relative block sizes.

The additional insights provided by the present paper may be useful in designing algorithms where k is unknown and likelihood maximization is combined with some different kinds of operations. One simple operation worth studying is that of merging blocks according to the MDL criterion.

Practical implementation of MDL often requires a bit of art in addition to science. In the present work, we could not prove Theorem 3.3 completely without using the slightly bigger function L_2 instead of the better motivated L'_2 , although MDL is theoretically able to identify SBM-like models exactly. Another remaining technical question is whether the restriction to partitions with nonnegligible relative block sizes is really necessary for $L(\eta) \geq L(\eta \vee \xi_n)$.

Appendix A. Information-theoretic preliminaries

Consider a random variable X with moment-generating function

$$\phi_X(\beta) = \mathbb{E} e^{\beta X},$$

and let $\mathcal{D}_X = \{\beta : \phi_X(\beta) < \infty\}$. We restrict our attention to distributions of X for which \mathcal{D}_X is an open (finite or infinite) interval. The corresponding *rate function* is

$$I_X(x) = - \inf_{\beta \in \mathcal{D}_X} (\log \phi_X(\beta) - \beta x);$$

this is a strictly convex function with minimum 0 at $\mathbb{E} X$ and value $+\infty$ outside the range of X . For the mean $\bar{X}_n = \frac{1}{n} \sum_1^n X_i$ of independent and identically distributed copies of X , we have

$$I_{\bar{X}_n}(x) = nI_X(x). \quad (\text{A.1})$$

The following inequalities are known as the Chernoff bounds for the random variable X .

Proposition A.1 *We have*

$$\begin{aligned} \mathbb{P}(X \leq x) &\leq e^{-I_X(x)} \quad \text{for } x \leq \mathbb{E}X, \\ \mathbb{P}(X \geq x) &\geq e^{-I_X(x)} \quad \text{for } x \geq \mathbb{E}X. \end{aligned}$$

We make frequent use of the following consequence of the Chernoff bounds. Let the convex hull of the support of X be the closure of (x^-, x^+) , and define

$$a^- := \lim_{x \downarrow x^-} I_X(x) = -\log \mathbb{P}(X = x^-), \quad a^+ := \lim_{x \uparrow x^+} I_X(x) = -\log \mathbb{P}(X = x^+)$$

(note that $a^- < \infty$ if and only if the distribution of X has an atom at x^- , and similarly for a^+). We can now write

$$\begin{aligned} I_X(x) &= I_X^-(x)1_{\{(x^-, \mathbb{E}X]\}}(x) + I_X^+(x)1_{\{[\mathbb{E}X, x^+)\}}(x) \\ &\quad + a^- \cdot 1_{\{x^-\}}(x) + a^+ \cdot 1_{\{x^+\}}(x) + \infty \cdot 1_{\{\mathbb{R} \setminus [x^-, x^+]\}}(x). \end{aligned}$$

With the assumptions made above, the functions $I_X^-(x)$ and $I_X^+(x)$ are, respectively, bijections from $(x^-, \mathbb{E}X]$ and $[\mathbb{E}X, x^+)$ to $[0, a^-)$ and $[0, a^+)$.

Lemma A.1. *We have*

$$I_X(X) \stackrel{(st)}{\leq} \log 2 + Y,$$

where $\stackrel{(st)}{\leq}$ denotes stochastic order and Y is a random variable with distribution $\text{Exp}(1)$.

Proof. For any $z \geq 0$, we have

$$\begin{aligned} \mathbb{P}(I_X(X) > z) &= \mathbb{P}(1_{\{X < \mathbb{E}X\}}I_X(X) > z \text{ or } 1_{\{X > \mathbb{E}X\}}I_X(X) > z) \\ &= \mathbb{P}(1_{\{X < \mathbb{E}X\}}I_X^-(X) > z) + \mathbb{P}(1_{\{X > \mathbb{E}X\}}I_X^+(X) > z) \\ &= 1_{\{z < a^-\}}\mathbb{P}(X < I_X^{-(1)}(z)) + 1_{\{z < a^+\}}\mathbb{P}(X > I_X^{+(1)}(z)) \\ &\leq 1_{\{z < a^-\}} \exp(-I_X(I_X^{-(1)}(z))) + 1_{\{z < a^+\}} \exp(-I_X(I_X^{+(1)}(z))) \\ &\leq 2e^{-z} \\ &= e^{-(z - \log 2)}, \end{aligned}$$

where the first inequality comes from Proposition A.1. Thus,

$$\mathbb{P}(I_X(X) > z) \leq \min \left\{ 1, e^{-(z - \log 2)} \right\} = e^{-(z - \log 2)^+} = \mathbb{P}(\log 2 + Y > z). \quad \square$$

In the case that X has the *Bernoulli*(p) distribution, we have

$$I_X(x) = D_B(x||p) := x \log \frac{x}{p} + (1 - x) \log \frac{1 - x}{1 - p}. \tag{A.2}$$

Lemma A.2. *The first and second derivatives of the functions $H(x)$ and $x \mapsto D_B(x||p)$ are*

$$H'(x) = \log \frac{1-x}{x}, \quad H''(x) = -\frac{1}{x(1-x)}, \quad (\text{A.3})$$

$$D'_B(x||p) = H'(p) - H'(x), \quad D''_B(x||p) = \frac{1}{x(1-x)}. \quad (\text{A.4})$$

Since $D_B(p||p) = D'_B(p||p) = 0$ and $D''_B(p||p) = -H''(p)$, we also have

$$H(q) - (H(p) + H'(p)(q-p)) = -D_B(q||p), \quad (\text{A.5})$$

and

$$\begin{aligned} & \lim_{n \rightarrow \infty} n \left[H \left(\left(1 - \frac{1}{n}\right)p + \frac{1}{n}q \right) - \left(\left(1 - \frac{1}{n}\right)H(p) + \frac{1}{n}H(q) \right) \right] \\ &= (q-p)H'(p) - (H(q) - H(p)) \\ &= D_B(q||p). \end{aligned}$$

Proposition A.2. *Let $n \geq 2$, and let X_1 and X_2 be independent random variables with distributions $\text{Bin}(m,p)$ and $\text{Bin}(n-m,p)$, respectively. Let $X_{12} = X_1 + X_2$ and $\bar{X}_1 = X_1/m$, $\bar{X}_2 = X_2/(n-m)$, $\bar{X}_{12} = X_{12}/n$. Then the following identities hold:*

$$mD_B(\bar{X}_1||p) + (n-m)D_B(\bar{X}_2||p) - nD_B(\bar{X}_{12}||p) \quad (\text{A.6})$$

$$= X_{12}D_B \left(\frac{X_1}{X_{12}} \parallel \frac{m}{n} \right) + (n - X_{12})D_B \left(\frac{m - X_1}{n - X_{12}} \parallel \frac{m}{n} \right) \quad (\text{A.7})$$

$$= mD_B(\bar{X}_1||\bar{X}_{12}) + (n-m)D_B \left(\frac{X_{12} - X_1}{n-m} \parallel \bar{X}_{12} \right). \quad (\text{A.8})$$

The identities in Proposition A.2 are obtained by writing out the full expression for (A.6) and rearranging the log terms in two other ways. The formulae (A.7) and (A.8) are written without X_2 , expressing the fact that any two of the three random variables X_1 , X_2 , and X_{12} contain the same information as the full triple. Note that (A.7) and (A.8) do not contain p . This reflects the fact that the conditional distribution of X_1 given X_{12} , known as the hypergeometric distribution, does not depend on p . The identity of (A.6) and (A.8) can be interpreted so that the two positive terms of (A.6) measure exactly the same amount of information about p as what is subtracted by the negative term. Moreover, (A.8) has the additional interpretation of presenting the rate function of the hypergeometric distribution, as stated in the following proposition.

Proposition A.3. *Let X have the distribution $\text{Hypergeometric}(n,m,z)$, i.e. the conditional distribution of X_1 of Proposition A.2 given that $X_{12} = z$. The rate function of X is*

$$I_X(x) = mD_B \left(\frac{x}{m} \parallel \frac{z}{n} \right) + (n-m)D_B \left(\frac{z-x}{n-m} \parallel \frac{z}{n} \right). \quad (\text{A.9})$$

Proof. Define the bivariate moment-generating function of (X_1, X_2) as

$$\phi(\alpha, \beta) = \mathbb{E} e^{\alpha X_1 + \beta X_2}.$$

Write

$$\mathbb{P} [X_1 = m \mid X_{12} = z] = \frac{\mathbb{P} (X_1 = m, X_2 = z - m)}{\mathbb{P} (X_{12} = z)},$$

and note that we can assume $p = z/n$. We can now derive the claim using $\phi(\alpha, \beta)$ in a similar manner as in the well-known proof of the one-dimensional Chernoff bound. \square

Proposition A.4. *Let $k \geq 2$, and let $X_i, i \in \{1, \dots, k\}$, be independent random variables with distributions $\text{Bin}(n_i, p)$, respectively. Let $n = \sum_{i=1}^k n_i, X_{1\dots j} = \sum_{i=1}^j X_i, \bar{X}_i = X_i/n_i$, and $\bar{X}_{1\dots j} = X_{1\dots j}/\sum_{i=1}^j n_i$. Then*

$$\sum_{i=1}^k n_i D_B(\bar{X}_i \| p) - n D_B(\bar{X}_{1\dots k} \| p) \stackrel{(st)}{\leq} \sum_{i=1}^{k-1} (\log 2 + Y_i), \tag{A.10}$$

where Y_1, \dots, Y_{k-1} are independent $\text{Exp}(1)$ random variables.

Proof. For $k = 2$, the left-hand side of (A.10) equals

$$n_1 D_B(\bar{X}_1 \| \bar{X}_{12}) + n_2 D_B\left(\frac{X_{12} - X_1}{n_2} \parallel \bar{X}_{12}\right) \tag{A.11}$$

by Proposition A.2 For any $N \in \{0, \dots, n\}$, consider the conditional distribution of (A.11), given that $X_{12} = N$. By Proposition A.3, this is the distribution of the *Hypergeometric*(n, n_1, N) rate function taken at the random variable X_1 with the same distribution. The claim now follows by Lemma A.1, because the stochastic upper bound does not depend on N , i.e. on the value of X_{12} .

For $k > 2$ we proceed by induction. Assume that the claim holds for $k - 1$ and write

$$\begin{aligned} & \sum_{i=1}^k n_i D_B(\bar{X}_i \| p) - n D_B(\bar{X}_{12} \| p) \\ &= \sum_{i=1}^{k-1} n_i D_B(\bar{X}_i \| p) - (n - n_k) D_B(\bar{X}_{1\dots(k-1)} \| p) \\ & \quad + n_k D_B(\bar{X}_k \| p) + (n - n_k) D_B(\bar{X}_{1\dots(k-1)} \| p) - n D_B(\bar{X}_{1\dots k} \| p). \end{aligned}$$

By the induction hypothesis, the first row of the second expression is stochastically bounded by $\sum_{i=1}^{k-2} (\log 2 + Y_i)$, irrespective of the value of $X_{1\dots(k-1)}$. Similarly, the second row is stochastically bounded by $\log 2 + Y_k$, where $Y_k \sim \text{Exp}(1)$, irrespective of the value of $X_{1\dots k}$. It remains to note that Y_k can be chosen to be independent of (Y_1, \dots, Y_{k-2}) , because X_k is independent of (X_1, \dots, X_{k-1}) , and of $\bar{X}_{1\dots(k-1)}$ in particular. \square

Lemma A.3. *Consider the sequence (G_n, ξ_n) of SBMs as in Subsection 3.2. Then the following hold:*

1. *For any blocks A_i and A_j such that $d_{ij} \notin \{0, 1\}$, it holds for an arbitrary $\epsilon > 0$ with high probability that*

$$\min_{v \in A_i} \frac{|e(\{v\}, A_j)|}{|A_j|} \geq d_{ij} - n^{-\frac{1}{2} + \epsilon}, \quad \max_{v \in A_i} \frac{|e(\{v\}, A_j)|}{|A_j|} \leq d_{ij} + n^{-\frac{1}{2} + \epsilon}.$$

2. *For any $\epsilon > 0$, all block pairs of the partition ξ_n are ϵ -regular (Definition 2.1) with high probability.*

Proof. Claim 1: By Proposition A.1 and (A.4),

$$\begin{aligned} & \mathbb{P} \left(\max_{v \in A_i} \frac{|e(\{v\}, A_j)|}{|A_j|} > d_{ij} + h \right) \\ & \leq \sum_{v \in A_i} \mathbb{P} \left(\frac{|e(\{v\}, A_j)|}{|A_j|} > d_{ij} + h \right) \\ & \leq |A_i| \exp \left(-|A_j| D_B(d_{ij} + h \parallel d_{ij}) \right) \\ & = |A_i| \exp \left(-|A_j| \left(\frac{h^2}{2d_{ij}(1 - d_{ij})} + \frac{h^3}{6} D_B'''(z \parallel d_{ij}) \right) \right). \end{aligned}$$

The last expression converges to zero with the choice $h = n^{-\frac{1}{2} + \epsilon}$ (recall that $|A_i| \sim n\gamma_i$ and $|A_j| \sim n\gamma_j$), which proves the claim on the maximum. The case of the minimum is symmetric.

Claim 2: Fix $\epsilon > 0$ and consider any i, j . Let $U_1 \subseteq A_i$ and $U_2 \subseteq A_j$ be such that $|U_1| \geq \epsilon|A_i|$ and $|U_2| \geq \epsilon|A_j|$. By Proposition A.1,

$$\mathbb{P} \left(|d(U_1, U_2) - d_{ij}| > \epsilon \right) \leq e^{-|U_1||U_2|D_B(d_{ij} + \epsilon \parallel d_{ij})} + e^{-|U_1||U_2|D_B(d_{ij} - \epsilon \parallel d_{ij})}.$$

Let $\iota(\epsilon) = \min \{D_B(d_{ij} + \epsilon \parallel d_{ij}), D_B(d_{ij} - \epsilon \parallel d_{ij})\}$. The union bound yields

$$\begin{aligned} & \mathbb{P} \left(\exists U_1 \subseteq A_i, U_2 \subseteq A_j : |U_1| \geq \epsilon|A_i|, |U_2| \geq \epsilon|A_j|, |d(U_1, U_2) - d_{ij}| > \epsilon \right) \\ & \leq 2|A_i||A_j| \exp \left((|A_i| + |A_j|) \log 2 - \epsilon^2|A_i||A_j|\iota(\epsilon) \right) \\ & \leq 2n^2 \exp \left(n^2 \left(\frac{(\gamma_i + \gamma_j) \log 2}{n} - \gamma_i\gamma_j\epsilon^2\iota(\epsilon) \right) \right) \rightarrow 0 \quad \text{as } n \rightarrow \infty, \end{aligned}$$

because $\iota(\epsilon) > 0$. □

Preliminaries for the Poissonian block model

For the Poissonian block model, the function $\phi(x) = -x \log x$ replaces binomial entropy in the counterparts of code lengths $L_4 + L_5$. We indicate below how the crucial steps of the proofs would change.

Denote by $D_P(b \parallel a)$ the Kullback–Leibler divergence between the distributions *Poisson*(a) and *Poisson*(b),

$$D_P(b \parallel a) = a - b + b \log b - b \log a. \tag{A.12}$$

For a counterpart to Lemma A.2, note that, for any $z > 0$,

$$\phi''(x) = -\frac{1}{x} = -\frac{d^2}{dx^2} D_P(x \parallel z). \tag{A.13}$$

Lemma A.4. *For any $\alpha \in [0, 1]$ and $x, y > 0$, let $z = \alpha x + (1 - \alpha)y$. Then we have*

$$\phi(z) - (\alpha\phi(x) + (1 - \alpha)\phi(y)) = \alpha D_P(x \parallel z) + (1 - \alpha) D_P(y \parallel z) \tag{A.14}$$

$$= z I_{Ber(\alpha)} \left(\frac{\alpha x}{z} \right). \tag{A.15}$$

Proof. The equality (A.14) follows from (A.13) by an argument similar to the derivation of (A.5). The expression (A.15) is obtained by writing the right-hand side of (A.14) with the substitution $y = (z - \alpha x)/(1 - \alpha)$ and recombining the log terms.

The Poissonian counterpart of Proposition A.4 is the following.

Proposition A.5. *Let $a > 0$, $k \geq 2$, $n_i \geq 1$, $i = 1, \dots, k$, and $n = \sum_i n_i$. Let X_i , $i \in \{1, \dots, k\}$, be independent random variables with distributions $\text{Poisson}(n_i a)$, respectively. Let $X_{1\dots j} = \sum_{i=1}^j X_i$, $\bar{X}_i = X_i/n_i$, and $\bar{X}_{1\dots j} = X_{1\dots j} / \sum_{i=1}^j n_i$. Then*

$$n\phi(\bar{X}_{1\dots k}) - \sum_{i=1}^k n_i\phi(\bar{X}_i) \stackrel{(st)}{\leq} \sum_{i=1}^{k-1} (\log 2 + Y_i), \tag{A.16}$$

where Y_1, \dots, Y_{k-1} are independent $\text{Exp}(1)$ random variables.

Proof. The proof of Proposition A.4 can be imitated as follows:

- Using induction, it suffices to consider the case $k = 2$.
- Apply Lemma A.4 to the left-hand side of (A.16) with $x = \bar{X}_1$, $y = \bar{X}_2$, $z = \bar{X}_{12}$, and $\alpha = n_1/n$. This yields

$$X_{12} I_{\text{Ber}(\frac{n_1}{n})} \left(\frac{X_1}{X_{12}} \right) = I_{\text{Bin}(X_{12}, \frac{n_1}{n})}(X_1).$$

- Now, the conditional distribution of X_1 given X_{12} is the above binomial distribution. Thus, we can apply Lemma A.1 in a similar way as in the proof of Proposition A.4 \square

Funding

This work was supported by the Academy of Finland project 294763 (Stomograph).

Competing Information

There were no competing interests to declare which arose during the preparation or publication process for this article.

References

- [1] ALON, N., FISCHER, E., NEWMAN, I. AND SHAPIRA, A. (2006). A combinatorial characterization of the testable graph properties: it’s all about regularity. In *Proc. 38th Annual ACM Symposium on Theory of Computing (STOC ’06)*, Association for Computing Machinery, New York, pp. 251–260.
- [2] BOLLA, M. (2016). Relating multiway discrepancy and singular values of nonnegative rectangular matrices. *Discrete Appl. Math.* **203**, 26–34.
- [3] BOLLA, M. AND ELBANNA, A. (2015). Estimating parameters of a probabilistic heterogeneous block model via the EM algorithm. *J. Prob. Statist.* **2015**, 657965.
- [4] COVER, T. AND THOMAS, J. (1991). *Elements of Information Theory*. John Wiley, New York.
- [5] GAO, C., MA, Z., ZHANG, A. AND ZHOU, H. (2017). Achieving optimal misclassification proportion in stochastic block model. *J. Mach. Learning Res.* **18**, 1–45.
- [6] GRÜNWARD, P. (2007). *Minimum Description Length Principle*. MIT Press.
- [7] HEIMLICH, S., LELARGE, M. AND MASSOULIÉ, L. (2012). Community detection in the labelled stochastic block model. In *NIPS Workshop on Algorithmic and Statistical Approaches for Large Social Networks*, Lake Tahoe, NV. Available at <https://arxiv.org/abs/1209.2910>.
- [8] HOLLAND, P., LASKEY, K. AND LEINHARDT, S. (1983). Stochastic blockmodels: first steps. *Social Networks* **5**, 109–137.
- [9] KOMLÓS, J. AND SIMONOVITS, M. (1996). Szemerédi’s regularity lemma and its applications in graph theory. In *Combinatorics, Paul Erdős Is Eighty*, eds D. MIKLÓS, V. SÓS AND T. SZONYI, János Bolyai Mathematical Society, Budapest, pp. 295–352.
- [10] MASSOULIÉ, L. (2014). Community detection thresholds and the weak Ramanujan property. In *Proc. 46th Annual ACM Symposium on Theory of Computing (STOC ’14)*, Association for Computing Machinery, New York, pp. 694–703.

- [11] NEPUSZ, T., NÉGYESSY, L., TUSNÁDY, G. AND BAZSÓ, F. (2008). Reconstructing cortical networks: case of directed graphs with high level of reciprocity. In *Handbook of Large-Scale Random Networks* (Bolyai Society Mathematical Studies 18), eds B. BOLLOBÁS, R. KOZMA AND D. MIKLÓS, SPRINGER, BERLIN, Heidelberg, pp. 325–368.
- [12] PEHKONEN, V. AND REITTU, H. (2011). Szemerédi-type clustering of peer-to-peer streaming system. In *Proc. 2011 International Workshop on Modeling, Analysis, and Control of Complex Networks (Cnet '11)*, Association for Computing Machinery, New York, pp. 23–30.
- [13] PEIXOTO, T. P. (2012). Entropy of stochastic blockmodel ensembles. *Phys. Rev. E* **85**, 056122.
- [14] PEIXOTO, T. P. (2013). Parsimonious module inference in large networks. *Phys. Rev. Lett.* **110**, 148701.
- [15] REITTU, H., BAZSÓ, F. AND WEISS, R. (2014). Regular decomposition of multivariate time series and other matrices. In *Structural, Syntactic, and Statistical Pattern Recognition (S+SSPR 2014)*, eds P. FRÁNTI, G. BROWN, M. LOOG, F. ESCOLANO AND M. PELILLO, SPRINGER, BERLIN, Heidelberg, pp. 424–433.
- [16] REITTU, H., LESKELÄ, L., RÄTY, T. AND FIORUCCI, M. (2018). Analysis of large sparse graphs using regular decomposition of graph distance matrices. In *Proc. 2018 IEEE International Conference on Big Data*, Institute of Electrical and Electronics Engineers, New York, pp. 3784–3792.
- [17] REITTU, H., NORROS, I. AND BAZSÓ, F. (2017). Regular decomposition of large graphs and other structures: scalability and robustness towards missing data. In *Proc. 2017 IEEE International Conference on Big Data*, Institute of Electrical and Electronics Engineers, New York, pp. 3352–3357.
- [18] REITTU, H. *et al.* (2019). Regular decomposition of large graphs: foundation of a sampling approach to stochastic block model fitting. *Data Sci. Eng.* **4**, 44–60.
- [19] RISSANEN, J. (1983). A universal prior for integers and estimation by minimum description length. *Ann. Statist.* **11**, 416–431.
- [20] RISSANEN, J. (1998). *Stochastic Complexity in Statistical Inquiry*. World Scientific, Singapore.
- [21] ROSVALL, M. AND BERGSTROM, C. T. (2007). An information-theoretic framework for resolving community structure in complex networks. *Proc. Nat. Acad. Sci. USA* **104**, 7327–7331.
- [22] SZEMERÉDI, E. (1976). Regular partitions of graphs. In *Problèmes Combinatoires et Théorie des Graphes* (Colloq. Int. CNRS 260), CNRS, Orsay, pp. 399–401.
- [23] WANG, Y. X. R. AND BICKEL, P. J. (2017). Likelihood-based model selection for stochastic block models. *Ann. Statist.* **45**, 500–528.
- [24] YAGLOM, A. AND YAGLOM, I. (1983). *Probability and Information*. D. Reidel, Dordrecht.