Regular Decomposition: an information and graph theoretic approach to stochastic block models

Hannu Reittu† Fülöp Bazsó‡ Ilkka Norros§

Abstract

A method for compression of large graphs and matrices to a block structure is proposed. Szemerédi’s regularity lemma is used as a generic motivation of the significance of stochastic block models. Another ingredient of the method is Rissanen’s minimum description length principle (MDL). We propose practical algorithms and provide theoretical results on the accuracy and consistency of the method.

Keywords: Szemerédi’s Regularity Lemma, Minimum Description Length Principle, stochastic block model, big data

Contents

1 Introduction 2
2 Basics and definitions 3
  2.1 Szemerédi’s Regularity Lemma 3
  2.2 Stochastic block models 5
  2.3 The Minimum Description Length (MDL) principle 7
3 The family of stochastic block models in the MDL framework 8
  3.1 Two-part MDL for simple graphs 8
  3.2 Two-part MDL for matrices 10
  3.3 Refined MDL and asymptotic model complexity 11
4 The Regular Decomposition approach to stochastic block models 13
  4.1 Block model codes 13
  4.2 Algorithms for description length minimization 15
  4.3 Accuracy of block structure identification by MDL 20
  4.4 $\varepsilon$-regularity vs. stochastic block models 21
5 Proof of Theorem 4.13 22

*This work has been submitted to the IEEE for possible publication. Copyright may be transferred without notice, after which this version may no longer be accessible.
†VTT Technical Research Centre of Finland Ltd, P.O. Box 1000, 02044 VTT. Email hannu.reittu@vtt.fi
‡Department of Theory, Institute for Particle and Nuclear Physics, Wigner Research Centre for Physics, Hungarian Academy of Sciences, P.O. Box 49, H-1525 Budapest, Hungary. Email bazso.fulop@wigner.mta.hu.
§VTT Technical Research Centre of Finland Ltd, P.O. Box 1000, 02044 VTT. Email ilkka.norros@vtt.fi
1 Introduction

Szemerédi’s Regularity Lemma (SRL) is a fundamental result in graph theory. Roughly speaking, SRL states that any large enough graph can be approximated arbitrarily well by nearly regular, pseudo-random bipartite graphs, induced by a partition of the node set into a bounded number of equal-sized sets. For many graph problems it suffices to study the problem on a corresponding random structure, resulting in a much easier problem (for a review, see, e.g., [10]). SRL is fundamental also in theoretical computer science, for instance, in showing the existence of polynomial time approximations for solving dense graph problems, and in characterizing the class of so-called testable graph properties [2].

Despite the impressive theoretical applications of SRL, it has had only few applications to ‘real-life’ problems. The main reason might be that SRL has extremely bad worst cases in the sense that the lower bound of graph sizes, for which the partition claim holds with reasonable accuracy and without a single exception, is enormous. Thus, a real-world application of SRL in literal sense is impossible. However, this drawback does not mean that regular partitions could not appear in much smaller scales relevant to applications. On the contrary, one could conjecture that regular partitions or structures be commonplace and worth of revealing. The goal of our work is on such realistic yet preferably large networks that nowadays appear in almost all imaginable application areas. The regular structure granted by SRL is then replaced by a probabilistic model, substituting the regular bipartite components with truly random bipartite graphs. This model class is well known as stochastic block models [9], and it has recently gained much attention in research and popularity in practical applications like community detection [16, 7, 11]. However, the fundamental nature of SRL suggests the (heuristic) conjecture that stochastic block models in fact present a very generic form of the separation of structure and randomness in large real-world systems. These models have good algorithmic properties in maximum likelihood fitting, and one can use powerful tools like expectation maximization, simulated annealing and Monte Carlo Markov Chain algorithms.

There are also some examples of graph decomposition applications that have been explicitly inspired by SRL. The practical contexts are varied: brain cortex analysis [12], image processing [23], peer-to-peer network [15], analysing Functional Magnetic Resonance (fMRI) data to depict functional connectivity of the brain [14], a matrix of multiple time series [18]. In the last-mentioned work, the method was generalized from graphs to arbitrary positive matrices using a Poissonian construction as an intermediary step.

Interestingly, the authors of the recent work [21] define a ‘practical’ variant of SRL by relaxing algorithmic SRL in a certain way to make it more usable in machine-learning tasks, see also [17]. Bolla [3] has developed a spectral approach for finding regular structures of graphs and matrices. Our emphasis is more information-theoretical by nature and continues the works [12, 15, 18]. It would be very interesting to compare the methods of [21, 3] and those of this paper in depth.

The third main ingredient in this paper is 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 has the form of a graph or matrix, and we present the stochastic block models as a modeling space in the sense of the MDL theory. Within this modeling space, the model corresponding to minimum code length coding presents the optimally regular decomposition (partition) of the data. Therefore we call this Szemerédi-motivated and technically MDL-based approach Regular Decomposition.

By information theory, the optimal coding reveals as much redundancy in the data as is possible from the given modeling point of view. The regular structure has a high degree of redundancy: a regular pair is an almost structureless subgraph where almost all nodes have similar and uniform connectivity patterns. By definition, the MDL principle should be able to discover regular structures. Note also that this principle presents a case of 'Occam’s Razor’, a general rule of reasoning that has proved fruitful in all areas of science.

In non-hierarchic clustering tasks, it has been a major challenge to select the 'right' size of a partition ($k$). Intuitively, the optimal choice of $k$ will strike a balance between the simplest partition of the data using a single cluster and the maximal partition assigning each data point to its own cluster, and selecting something in between. A popular device has been the Akaike information criterion (AIC), which was applied also in [12]. The AIC simply adds the 'number of model parameters’ to the maximal log-likelihood and chooses the model that minimizes the sum. Our MDL-based approach solves the corresponding model selection task in a better founded way.

The contributions of this paper are the following: (i) the linkage of SRL, stochastic block models and the MDL principle, (ii) the unified handling of graphs and matrices, (iii) effective practical algorithms for revealing regular structures in data, and (iv) Theorem 4.13 that characterizes how accurately the MDL principle identifies a stochastic block model.

The paper is structured as follows. Section 2 presents the definitions of the central notions: SRL, stochastic block models and MDL. The last topic is expanded in Section 3. Section 4 presents our main technical results, in particular the algorithms and Theorem 4.13. The proof of Theorem 4.13 is given in Section 5. It is structured into several propositions and makes strong use of information-theoretic tools presented in Appendix A.

2 Basics and definitions

2.1 Szemerédi’s Regularity Lemma

Consider simple graphs $G(V, E)$, where $V$ is the set of nodes (vertices) and $E$ is the set of links (edges). The link density of a non-empty node set $X \subseteq V$ is defined as

$$d(X) = \frac{e(X)}{\binom{|X|}{2}},$$

where $e(X) = \{\{v, w\} \in E : v, w \in X\}$. 

3
and $|\cdot|$ denotes the cardinality of a set. Similarly, the link density between two disjoint non-empty node sets $X, Y \subseteq V$ is defined as

$$d(X, Y) := \frac{e(X, Y)}{|X||Y|},$$

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 $A, B \subseteq V$ is called $\epsilon$-regular, if for every $X \subseteq A$ and $Y \subseteq B$, such that $|X| > \epsilon |A|$ and $|Y| > \epsilon |B|$ we have

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

A partition $\xi = \{V_0, V_1, V_2, \cdots, V_k\}$ of $V$ into $k + 1$ sets, where all except $V_0$ have equal cardinalities, is called $\epsilon$-regular, iff all except at most $\epsilon k^2$ pairs are $\epsilon$-regular and $|V_0| < \epsilon |V|$.

**Theorem 2.2** (Szemerédi’s Regularity Lemma, [24]) For every $\epsilon > 0$ and for any positive integer $m$, there are positive integers $N(\epsilon, m)$ and $M(\epsilon, m)$, such that for any graph $G(E, V)$ with $|V| \geq N(\epsilon, m)$ there is an $\epsilon$-regular partition of $V$ into $k + 1$ classes with $m \leq k \leq M(\epsilon, m)$.

SRL states, roughly speaking, that nodes of any large enough graph can be partitioned into a bounded number, $k$, of equal-sized ‘clusters’ and into one small set in such a way that most pairs of clusters look like random bipartite graphs, whose link probability equals the link density between the pair.

The claim of SRL is significant for sufficiently dense graphs, i.e., when the link density is higher than $\epsilon$. However, the result can be modified to apply to sparse graphs by multiplying $\epsilon$ at the right-hand side of the regularity definition by the link density of the entire graph [22].

The amazing fact in SRL is that the regularity claim holds for all graphs starting from a lower bound for size that depends only on $\epsilon$. However, it is also well known that this dependence on $\epsilon$ is of extremely bad kind: the known lower bound for the graph size $N(\epsilon, m)$ behaves like a tower of powers of 2:

$$2^{2^{2^{\cdots}}},$$

where the height of the tower is upper-bounded by $1/\epsilon^5$. Such a number is too big to be considered in any applications. Thus, all real-world networks fall into a ‘gray area’ with respect to SRL.

The magnitude of $N(\epsilon, m)$ is a problem for using algorithmic versions of the SRL, like the one introduced by Alon et al. [1]. Although its time complexity is only polynomial $O(n^{2.376})$, corresponding to the time for multiplying two $n \times n$ binary matrices, it requires this enormous size ($n \geq N(\epsilon, m)$) of graph to be able to find a regular partition. A considerable improvement was found in the recent work [5], where the running time is only linear in the graph size. Perhaps even more importantly, this randomized algorithm works in a more realistic fashion: for any graph, it either finds an $\epsilon$-regular partition or concludes that such a partition does not exists. Thus, it does not need to work in the safe region of graph sizes specified in SRL. Another algorithm with the same feature was suggested by Tao [25]. In principle, such algorithms could possibly be applicable for real-world graphs, although some extra-big numbers are hidden in some of the constants of the algorithm.
2.2 Stochastic block models

The notion of an $\epsilon$-regular partition is purely combinatorial. The stochastic model closest to this notion is the following.

**Definition 2.3** Let $V$ be a finite set and $\xi = \{A_1, \ldots, A_k\}$ a partition of $V$. A stochastic block model is a random graph $G = (V, E)$ with the following structure:

- There is a symmetric $k \times k$ matrix $D = (d_{ij})_{i,j=1}^k$ of real numbers $d_{ij} \in [0, 1]$ satisfying the irreducibility condition that no two rows are equal, i.e.,

  \[ d_{ijq} \neq d_{jq}, \quad \text{for all } i, j, q, i < j, \]

- For every pair \{v, w\} of distinct nodes of $V$ such that $v \in A_i$, $w \in A_j$, let $e_{vw} = e_{wv}$ be a Bernoulli random variable with parameter $d_{ij}$, assuming that all $e_{vw}$'s are independent. The edges of $G$ are

  \[ E = \{\{v, w\} : v, w \in V, v \neq w, e_{vw} = 1\}. \]

Note that the case of the trivial partition $\xi = \{V\}$ yields the classical random graph with edge probability $d_{11}$.

A graph sequence $G_n = (V_n, E_n)$ presenting copies of the same stochastic block model in different sizes can, for definiteness, be constructed as follows.

**Construction 2.4** Let $\gamma_1, \ldots, \gamma_k$ be positive real numbers such that $\sum_{i=1}^k \gamma_i = 1$. Divide the interval $(0, 1]$ into $k$ segments

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

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

\[ V_n = \left\{\frac{i}{n} : i \in \{1, \ldots, n\}\right\}. \]

For each $n$, let $\xi_n$ be the partition of $V_n$ into the blocks

\[ A_i^{(n)} = I_i \cap V_n, \quad i = 1, \ldots, k. \]

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 non-empty and $\xi_n = \{A_1^{(n)}, \ldots, A_k^{(n)}\}$ is a genuine partition of $V_n$. We can then generate stochastic block models based on $(V_n, \xi_n, D)$ according to Definition 2.3.

**Remark 2.5** A slightly different kind of stochastic block model can be defined by drawing first the sizes of blocks $A_i^{(n)}$ as independent Poisson($\gamma_i n$) random variables and proceeding then with the matrix $D$ as before. The additional level of randomness, regarding the block sizes, is however of no interest in the present paper.

Next, we define the notion of a Poissonian block model in complete analogy with Definition 2.3.
Definition 2.6 Let $V$ be a finite set of vertices, $n = |V|$, and let $\xi = \{A_1, \ldots, A_k\}$ be a partition of $V$. The symmetric Poissonian block model is a symmetric random $n \times n$ matrix $E$ with the following structure:

- There is a symmetric $k \times k$ matrix $\Lambda = (\lambda_{ij})_{i,j=1}^k$ of non-negative real numbers satisfying the irreducibility condition that no two rows are equal, i.e.,
  
  for all $i$, $j$, $i < j$, there is $q_{ij} \in \{1, \ldots, k\}$ such that $\lambda_{iq_{ij}} \neq \lambda_{jq_{ij}}$; \hspace{1cm} (2)

- For every unordered pair \{v, w\} of distinct nodes of $V$ such that $v \in A_i$, $w \in A_j$, let $e_{vw} = e_{wv}$ be a Poisson random variable with parameter $\lambda_{ij}$, assuming that all $e_{vw}$’s are independent. The matrix elements of $E$ are $e_{vw}$ for $v \neq w$, and $e_{vv} = 0$ for the diagonal elements.

Thanks to the independence assumption, the sums $\sum_{u \in A} \sum_{v \in B} e_{uv}$ are Poisson distributed for any $A, B \in \xi$.

Remark 2.7 The rest of the technical contents of this paper focus on the simple binary and Poissonian models of Definitions 2.3 and 2.6. However, the following extensions are straightforward:

- bipartite graphs: this is just a subset of simple graphs;
- $m \times n$ matrices with independent Poissonian elements: a matrix can be seen as consisting of edge weights of a bipartite graph, where the parts are the index sets of the rows and columns of the matrix, respectively;
- directed graphs: a directed graph can be presented as a bipartite graph consisting of two parts of equal size, presenting the input and output ports of each node.

The following construction, inspired by our conditionally Poissonian graph model [13], is the key to extend the MDL methodology for data that has the most common form of a large non-negative matrix.

Construction 2.8 Let $C = (c_{ij})$ be a non-negative $m \times n$ matrix. Let $N$ be a (rather large) integer and denote $J = \{1, \ldots, N\}$. Let

$$V = J_1^{(1)} \cup \cdots \cup J_m^{(1)}, \quad W = J_1^{(2)} \cup \cdots \cup J_n^{(2)},$$

$J_i^{(1)} = \{(i, i)\} \times J$. Define a bipartite Poissonian block model $\mathcal{P}_N(C) = (V \cup W, \{J_i^{(1)}\}, C)$ with blocks $J_i^{(1)}$ and mean matrix $C$.

Although we have not studied this at the technical level, it is natural to expect that, with large $N$, a partition of $\mathcal{P}_N(C)$ with minimum description length would with high probability keep the blocks $J_i^{(1)}$ unbroken. Because the regular decomposition algorithm for Poissonian block models operates only on means over blocks (see subsection 12), Construction 2.8 is a strong heuristic argument that this algorithm applies as such to regular decomposition of non-negative matrices.
2.3 The Minimum Description Length (MDL) principle

The Minimum Description Length (MDL) Principle was introduced by Jorma Rissanen, inspired by Kolmogorov’s complexity theory, and an extensive presentation can be found in Grünwald’s monography [6], see also [20]. The basic idea is the following: a set $D$ of data is optimally explained by a model $M$, when the combined description of the (i) model and (ii) the data as interpreted in this model is as concise as possible. By description we mean here a code that specifies an object uniquely.

The principle is best illustrated by our actual case, 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 node 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. Our heuristic postulate is that in the case of graphs and similar objects a good a priori class of models should be inferred from SRL, which points to stochastic block models.

**Definition 2.9** Denote by $\mathcal{M}_{n/k}$ the set of irreducible stochastic block models $(V, \xi, D)$ with

- $|V| = n$,
- $|\xi| = k$, and, denoting $\xi = \{V_1, \ldots, V_k\}$,
- for $i, j \in \{1, \ldots, k\}$,

$$d_{ij} = \frac{h_{ij}}{|V_i||V_j|}, \ h_{ij} \in \mathbb{N}, \ d_{ii} = \frac{h_{ii}}{|V_i|^2}, \ h_{ii} \in \mathbb{N}.$$  

The condition in the last bullet entails that each modeling space $\mathcal{M}_{n/k}$ is finite.

**Remark 2.10** Without the irreducibility condition (1), there would not be a bijection between stochastic block models and their parameterizations.

The models in $\mathcal{M}_{n/k}$ are parameterized by $\Theta_k = (\xi, D)$. A good model for a graph $G$ is the one that gives maximal probability for $G$ and is called the maximum likelihood model. We denote the parameter of this model

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

where $P(G | \Theta_k)$ denotes the probability that the probabilistic model specified by $\Theta_k$ produces $G$.

One part of likelihood optimization is trivial: when a partition $\xi$ is selected for a given graph $G$, the optimal link probabilities are the empirical link densities:

$$d_{ij} = \frac{e(V_i, V_j)}{|V_i||V_j|}, \ i \neq j, \ d_{ii} = \frac{e(V_i)}{|V_i|^2}.$$  

Thus, the nontrivial part is to find the optimal partition for the given graph. This is the focus of the next sections.
3 The family of stochastic block models in the MDL framework

3.1 Two-part MDL for simple graphs

Let us denote the set of all simple graphs with \( n \) nodes as

\[ \Omega_n = \{ G : G = (V,E) \text{ is a graph, } |V| = n \}. \]

A prefix (binary) coding of a finite set \( \Omega \) is an injective mapping

\[ C : \Omega \rightarrow \cup_{s \geq 1} \{0,1\}^s \]

such that no code is a prefix of another code. Recall the following proposition from information theory (see, e.g., [4]):

Theorem 3.1 (Kraft’s Inequality) For an \( m \)-element alphabet there exists a binary prefix coding scheme with code lengths \( l_1, l_2, \cdots, l_m \) if and only if the code lengths satisfy:

\[ \sum_{i=1}^{m} 2^{-l_i} \leq 1. \]

An important application of Theorem 3.1 is the following: if letters are drawn from an alphabet with probabilities \( p_1, p_2, \cdots, p_m \), then there exists a prefix coding with code lengths \( \lceil -\log p_1 \rceil, \lceil -\log p_2 \rceil, \cdots, \lceil -\log p_m \rceil \), and such a coding scheme is optimal in the sense that it minimizes the expected code length (in this section, the logarithms are in base 2). In particular, any probability distribution \( P \) on the graph space \( \Omega_n \) indicates that there exists a prefix coding that assigns codes to elements of \( G \in \Omega_n \) with lengths equal to \( \lceil -\log P(\{G\}) \rceil \).

The code length \( l(\cdot) \) is the number of binary digits in the code of the corresponding graph. In case of a large set \( \Omega \), most such codes are long and as a result the ceiling function can be omitted, a case we assume in sequel. A good model results in good compression, meaning that a graph can be described by much less bits than there are elements in the adjacency matrix. An incompressible case corresponds to the uniform distribution on \( \Omega_n \) and results in code length \( -\log (1/|\Omega|) = (n^2)/2 \), equivalent to writing down all elements of the adjacency matrix.

For every graph \( G \) from \( \Omega_n \) and model \( P \) we can associate an encoding with code length distribution \( -\log P(\cdot | \hat{\Theta}_k(G)) \). However, this is not all, since in order to be able to decode we must know what particular probabilistic model \( P \) is used. This means that also \( \hat{\Theta}_k(G) \) must be prefix encoded, with some code-length \( L(\hat{\Theta}_k(G)) \). We end up with the following description length:

\[ l(G) = \lceil -\log P(G | \hat{\Theta}_k(G)) \rceil + L(\hat{\Theta}_k(G)). \]

Eq. (6) presents the so-called two-part MDL, [6]. In an asymptotic regime with \( n \to \infty \), we get an analytic expression of the refined MDL. A simple way of estimating \( L(\hat{\Theta}_k(G)) \) is just to map injectively every model in \( \mathcal{M}_{n/k} \) to an integer and then encode integers with \( l^*(|\mathcal{M}_{n/k}(G)|) \) as an upper bound of the code-length. Here

\[ l^*(m) = \max(0, \log(m)) + \max(0, \log \log(m)) + \cdots, \quad m \in \mathbb{N}, \]

gives, as shown by Rissanen, the shortest length prefix coding for integers (see [6, 19]). The size of the graph must also be encoded with \( l^*(n) \) bits (it is assumed that there is a way of defining an upper bound of the models with given \( n \)). In this point, it is necessary that the modeling space is finite. This results in
Proposition 3.2 For any graph $G \in \Omega_n$, there exists a prefix coding with code-length

$$l(G) = \lceil - \log P(G | \hat{\Theta}_k(G)) \rceil + m$$

$$\leq m_k := l^*(n) + l^* \left( S_2(n, k) \left( \binom{n-k+2}{2} + 1 \right) \right) + 1,$$

where $S_2(n, k)$ is the Stirling number of the second kind.

Proof The expression in (6) corresponds to a concatenation of two binary codes. The $L$-part is the length of a code for maximum likelihood parameters (in the case of a non-unique maximum, we take, say, the one with smallest number in the enumeration of all such models). The corresponding code is called the parametric code. The parametric code uniquely encodes the maximum-likelihood model. To create such an encoding, we just enumerate all possible models, given in Definition 2.9 and use the integer to fix the model. The length of a prefix code corresponding to an integer is the $l^*$-function computed for that integer, and we add 1 to handle the ceiling function.

To obtain an upper bound for the parametric code length $m_k$, we find an upper bound for the number of models in the modeling space. The number of models is upper-bounded by the product of the integers, each presenting the number of partitions of an $n$-element set into $k$ non-empty sets (blocks), which equals $S_2(n, k)$. We can view the blocks of a partition as the nodes of a ‘reduced multi-graph’ (in a multi-graph, there can be several links between a node pair, as well as self-loops). The range of multi-links is between zero and $\binom{n-k+2}{2}$: if we consider a pair of blocks (or one block internally), there can be at most $n-(k-2)$ nodes in such a pair (in one set, slightly less), since there must be at least $k-2$ nodes in the other blocks of the partition. Obviously, in such a subgraph of $n-(k-2)$ nodes there can be at most $\binom{n-k+2}{2}$ links. Thus, the number of values each multi-link can take is upper-bounded by $\binom{n-k+2}{2} + 1$. Since the number of node pairs in the reduced multi-graph is $\binom{k}{2} + k$, we obtain the second multiplier in the argument of $l^*$ in the proposition.

Finally, we should show that the coding of the graph is prefix. We concatenate both parts into one code that has the prescribed length and put first the prefix code of the integer that defines the parameters of the maximum likelihood model. When we start to decode from the beginning of the entire code, we first obtain a code of an integer, because we used a prefix coding for integers. At this stage we are able to define the probabilistic model that was used to create the other part of the code, corresponding to the probability distribution $P(\cdot | \hat{\Theta}_k(G))$. Using this information we can decode the graph $G$. It remains to show that the concatenated code itself is prefix. Assume the opposite: some prefix of such a code is prefix to some other similar code, say, the first code is a prefix to the second one. However, the parametric code was prefix, so both codes must correspond to the same model. Since the first two-part code is a prefix to the second, they both share the same parametric part, and the code for the graph of the first is a prefix of the second one. But this is impossible, since the encoding for graphs within the same model is prefix. This contradiction shows that the two-part coding is prefix. □
Finally, we call
\[ \mathcal{M}_n := \bigcup_{1 \leq k \leq n} \mathcal{M}_{n/k} \]  
the full regular decomposition modeling space of \( \Omega_n \).

### 3.2 Two-part MDL for matrices

In this section we consider input data in the form of a \( n \times m \) matrix \( A = (a_{ij}) \) with non-negative entries. With such a matrix we associate a random bipartite multi-graph. The set of rows and the set of columns form a bipartition. Between row \( i \) and column \( j \) there is a random number of links that are distributed according to Poisson distribution with mean \( a_{ij} \). Such a model was introduced in [13] and it has been used in various tasks in complex network analysis, see [8]. The aim of this model is to back up, heuristically, a corresponding practical decomposition model that minimizes the expected description length of such random matrices. We propose the following modeling spaces:

**Definition 3.3** For integers \( k_1, k_2 \) from ranges \( 1 \leq k_1 \leq n \) and \( 1 \leq k_2 \leq m \), the parameters of a model \( \Theta_{k_1,k_2} \) in the modeling space \( \mathcal{M}_{k_1,k_2} \) for an integer matrix \( X \) partition of rows into \( k_1 \) non-empty sets \( V = (V_1, \ldots, V_{k_1}) \) and partition of columns into \( k_2 \) non-empty sets \( U = (U_1, \ldots, U_{k_2}) \) and \( k_1 \times k_2 \) block average matrix \( P \), with elements \( (P)_{\alpha,\beta} := \sum_{i \in V_{\alpha}, j \in U_{\beta}} \frac{a_{ij}}{|V_{\alpha}||U_{\beta}|} \).

Thanks to the addition rule of Poisson distributions, the likelihood of \( X \) in a model \( \Theta_{k_1,k_2} \in \mathcal{M}_{k_1,k_2} \), corresponds to probabilistic models where the elements of \( X \) are independent and Poisson distributed with parameters \( x_{i,j} \sim \text{Poisson}(P_{\alpha(i),\beta(j)}) \), where \( i \in V_{\alpha(i)} \), \( j \in U_{\beta(j)} \) in the model \( \Theta_{k_1,k_2} \). The corresponding likelihood is denoted as \( P(X \mid \Theta_{k_1,k_2}) \), the actual probability of \( X \) is denoted as \( P(X \mid A) \). The maximum likelihood model is found from the program that maximizes the expected log-likelihood:

\[
\Theta_{k_1,k_2}^* = \arg \max_{\Theta_{k_1,k_2} \in \mathcal{M}_{k_1,k_2}} \sum_X P(X \mid A) \log P(X \mid \Theta_{k_1,k_2}) \\
= \arg \max_{\Theta_{k_1,k_2} \in \mathcal{M}_{k_1,k_2}} \left( \sum_X P(X \mid A) \log \frac{P(X \mid \Theta_{k_1,k_2})}{P(X \mid A)} + P(X \mid A) \log P(X \mid A) \right) \\
= \arg \max_{\Theta_{k_1,k_2} \in \mathcal{M}_{k_1,k_2}} \left( -D(P_A | P_{\Theta_{k_1,k_2}}) - H(P_A) \right)
\]

where \( D \) is the Kullback-Leibler divergence between distributions, \( H \) denotes entropy and \( P_A \) and \( P_{\Theta_{k_1,k_2}} \) are the two families of Poisson distributions for the matrix elements of \( X \). Since \( H(P_A) \) is independent on \( \Theta_{k_1,k_2} \), it does not affect the identification of the maximum likelihood model. Thus, the final program for finding the optimal model is

\[
\Theta_{k_1,k_2}^* = \arg \min_{\Theta_{k_1,k_2} \in \mathcal{M}} \left( D(P_A | P_{\Theta_{k_1,k_2}}) \right). \tag{9}
\]
The description length of a model \( l(\Theta_{k_1, k_2} \in \mathcal{M}_{k_1, k_2}) \) consist of the description length \( l(V) + l(U) \) of the two partitions and the description length of the block average matrix \( l(P(X)) \). For the latter we need to know only the integers presenting the block sums of \( X \), since the denominator is known for a fixed partition \((U, V)\). The code lengths of such integers are, for large matrices, simply the logarithms of the integers. For \( l(U) + l(V) \) we use the same entropy based formula as in (14). As a result we end up with the following expression for the description length of the random multi-graph model \( A \) using the modeling space \( \mathcal{M}_{k_1, k_2} \):

\[
l_{k_1, k_2}(A) = D(P_{\Theta_{k_1, k_2}} || P_{\Theta_{k_1, k_2}}) + l(V^*) + l(U^*) + \sum_{1 \leq \alpha \leq k_1; 1 \leq \beta \leq k_2} \log(a_{\alpha, \beta} + 1),
\]

where

\[
a_{\alpha, \beta} = \sum_{i \in V^*_\alpha, j \in U^*_\beta} a_{i, j}.
\]

The full two-part MDL would now be realized by finding the global minimum of this expression over various \((k_1, k_2)\). We return to this case in the algorithm section 4.2. Although a heuristic one, we believe that our method for matrices is both reasonable and easy to use and implement, see [18].

3.3 Refined MDL and asymptotic model complexity

Let us next consider Rissanen’s refined MDL variant (see [6]). The idea is to generate just one distribution on \( \Omega_n \), called the normalized maximum likelihood distribution \( P_{nml} \). Then a graph \( G \in \Omega_n \) has the description length \(-\log P_{nml}(G)\) which is at most as large as the one given by the two-part code in (9). The function \( P(\cdot \mid \hat{\Theta}_k(\cdot)) \) maps graphs of size \( n \) into \([0, 1]\), and it is not a probability distribution, because \( \sum_{G \in \Omega} P(G \mid \hat{\Theta}_k(G)) > 1 \). However, a related true probability distribution can be defined as

\[
P_{nml}(\cdot) = \frac{P(\cdot \mid \hat{\Theta}_k(\cdot))}{\sum_{G \in \Omega} P(G \mid \hat{\Theta}_k(G))}.
\]

The problem with this is that a computation of the normalization factor in (11) is far too involved: finding a maximum likelihood parametrization for a
single graph is a ‘macroscopic’ computational task by itself and it is not possible to solve such a problem explicitly for all graphs. Therefore the two-part variant is a more attractive choice in a practical context. However, the refined MDL approach is useful as an idealized target object for justifying various approximations implementations of the basic idea. It appears that in an asymptotic sense the problem is solvable for large simple graphs. The logarithm of the normalization factor in (11) is called the parametric complexity of the model space $M_{n/k}$:

$$COMP(M_{n/k}) := \log \left( \sum_{G \in \Omega_n} P(G \mid \hat{\Theta}_k(G)) \right).$$

(12)

In a finite modelling space case like in ours, this can be considered as a definition of model complexity. We have now the following simple bounds:

**Proposition 3.4**

$$\log(S_2(n, k)) \leq COMP(M_{n/k}) \leq m_k + 1,$$

where we use the same notation as in Proposition 3.2.

**Proof** The lower bound follows from the fact that we can have at least this number of graphs that have likelihood 1 in $M_{n/k}$. This corresponds to graphs for which the nodes can be partitioned into $k$ non-empty sets and inside each set we have a full graph and no links between the sets. Thus, for every partition there is at least one graph that has likelihood one and all such graphs are different from each other since there is a bijection between those graphs and partitions.

For the upper bound, we notice that according to Proposition 3.2, there is a prefix coding with code lengths that correspond to the two-part code. As a result, Kraft’s inequality yields that

$$1 \geq \sum_{G \in \Omega_n} 2^{-\log P(G \mid \hat{\Theta}_k(G)) - m_k} \geq \sum_{G \in \Omega_n} 2^{-\log P(G \mid \hat{\Theta}_k(G)) - m_k},$$

from which we get

$$\sum_{G \in \Omega_n} P(G \mid \hat{\Theta}_k(G)) \leq 2^{m_k + 1}.$$ 

Taking logarithms, we arrive at the claimed upper bound.  \[ \square \]

When considering large-scale structures corresponding to moderate $k$, the upper and lower bounds in Proposition 3.4 are asymptotically equivalent, and we have

**Corollary 3.5** Assume that $k > 1$ is fixed. Then

$$COMP(M_{n/k}) \sim n \log k, \quad n \to \infty.$$

**Proof** Denoting the lower and upper bound of parametric complexity in Proposition 3.4 respectively by $b_l$ and $b_u$, we argue that $b_u \sim b_l \sim n \log k$ asymptotically when $n \to \infty$. This follows from the fact that the dominant asymptotic component of both $b_u$ and $b_l$ is $\log S_2(n, k)$. Indeed, $S_2(n, k) \sim \frac{k^n}{n^k}$ for fixed $k$, the asymptotic of $\log S_2(n, k)$ is linear in $n$, and all other terms of the asymptotics of both bounds are additive and at most logarithmic in $n$.  \[ \square \]
Remark 3.6 The speed of convergence of the upper and lower bounds in Proposition 3.4 is of type $\frac{\log n}{n}$.

4 The Regular Decomposition approach to stochastic block models

4.1 Block model codes

The previous section developed both the two-part and refined variants of the MDL theory, as presented in [6], for the model space of stochastic block models. In the following, we formulate a variant of two-part MDL that allows both practical implementations and a proof of consistency, i.e., that the MDL principle identifies a block model. It was shown above that the most heavy task in the description of a block model is identifying the partition. The same is true for the model complexity, which is asymptotically just the logarithm of the number of partitions. It appears that in order to prove consistency, we need quite a delicate estimate for the description length of the partition. The asymptotic model complexity given in Corollary 3.5 seems to be too crude for consistency. A full resolution of this intriguing question is left for further investigations.

We call our two-part MDL construction a block model code of a graph with respect to a partition of its nodes that allows the computation of a tight upper bound of the code length. This upper bound is also consistent with a more generic information theoretic point of view with a semi-constructive coding scheme.

We denote by $H(\cdot)$ both Shannon’s entropy function of a partition and the entropy of a binomial distribution, i.e.

$$
H(\xi) = -\sum_{A \in \xi} \frac{|A|}{|V|} \log \frac{|A|}{|V|},
H(p) = -p \log p - (1 - p) \log (1 - p).
$$

Remark 4.1 In the rest of this paper, we define also information-theoretic functions in terms of natural logarithms, and certain notions like code lengths should be divided by $\log 2$ to obtain their values in bits.

Definition 4.2 A block model code of a graph $G = (V, E)$ with respect to a partition $\xi$ of $V$ is a code with the following structure:

The model part:

- first, the sizes of the blocks $A \in \xi$ are given as integers;
- second, the edge density $d(A)$ inside each block $A \in \xi$ and the edge density $d(A, B)$ between each pair of distinct blocks $A, B \in \xi$ are given as the denominators of the rational numbers presenting the exact densities.

The aim of these two codes is to describe the parameters of two probability distributions, one for the links and the other for the membership of nodes in the blocks of the partition.

The data part:
• third, the partition $\xi$ is specified by a prefix code corresponding to membership distribution $P(i \in A) = |A|/n$, where all nodes are independent of each other;

• fourth, the edges inside each block $A \in \xi$ are specified by a prefix code corresponding to a stochastic block model distribution of links inside each block of $\xi$;

• fifth, the edges between each pair of blocks $A, B \in \xi$ are specified by a prefix code corresponding to a block model distribution of links between pairs of blocks in $\xi$.

The choice of link densities as link probabilities (the second code) is natural, since conditionally to a partition the stochastic block model is just a collection of Bernoulli models, where the best choice is to use averages as parameters. Note that a block model code can be given for any graph with respect to any partition of its nodes.

From Kraft’s inequality and the above definitions, it follows that there exists a prefix code for a graph $G = (V, E)$ with respect to a partition $\xi = \{A_1, \ldots, A_k\}$ of $V$ with length at most (and, for large graphs, typically close to)

$$L(G|\xi) = L_1(G|\xi) + L_2(G|\xi) + L_3(G|\xi) + L_4(G|\xi) + L_5(G|\xi),$$

$$L_1(G|\xi) = \sum_{i=1}^{k} l^*(|A_i|),$$

$$L_2(G|\xi) = \sum_{i=1}^{k} l^* \left( \frac{|A_i|}{2} \right) d(A_i) + \sum_{i<j} l^* (|A_i||A_j|d(A_i, A_j)),$$

$$L_3(G|\xi) = |V|H(\xi),$$

$$L_4(G|\xi) = \sum_{i=1}^{k} \left( \frac{|A_i|}{2} \right) H(d(A_i)),$$

$$L_5(G|\xi) = \sum_{i<j} |A_i||A_j|H(d(A_i, A_j)),$$

where $l^*(m)$ was defined by (7). Below we shall approximate $l^*(m)$ by $\log m$ without further mentioning, because their difference is insignificant in our context. Similarly, we have dropped ceiling functions systematically. Also, recall Remark 4.1 on the use of natural logarithms.

Next, we shall extend the definition of block model codes to Poissonian block models. Because the entries of $E$ are integers, the matrix $E$ can be encoded in the same way as the simple graphs in Definition 4.2. Denote by $H_p(a)$ the entropy of the distribution Poisson$(a)$:

$$H_p(a) = -\sum_{k=0}^{\infty} \frac{a^k}{k!} e^{-k} (k \log a - \log k! - a).$$

Now, the Poissonian block model code length of $E$ with respect to any partition
\[ \eta = \{B_1, \ldots, B_m\} \] is defined as

\[ L(E|\eta) = L_1(E|\eta) + L_2(E|\eta) + L_3(E|\eta) + L_4(E|\eta) + L_5(E|\eta), \]

\[ L_1(E|\eta) = \sum_{i=1}^{m} \log(|B_i|), \]

\[ L_2(E|\eta) = \sum_{i=1}^{m} \log \left( \left( \frac{|B_i|}{2} \right) a(B_i) \right) + \sum_{i<j} \log (|B_i||B_j|a(B_i, B_j)), \] (14)

\[ L_3(E|\eta) = |V|H(\eta), \]

\[ L_4(E|\eta) = \sum_{i=1}^{m} \left( \frac{|B_i|}{2} \right) H_P(a(B_i)), \]

\[ L_5(E|\eta) = \sum_{i<j} |B_i||B_j|H_P(a(B_i, B_j)), \]

where

\[ a(B_i) := \frac{\sum_{v,w \in B_i} c_{vw}}{|B_i|(|B_i| - 1)}, \quad a(B_i, B_j) := \frac{\sum_{v,w \in B_i} c_{vw}}{|B_i||B_j|}. \]

Remark 4.3 Definition (14) parallels (13), and it works smoothly in Theorem 4.13 below. However, the expressions of \( L_4 \) and \( L_5 \) correspond to the case that each set of random variables is indeed Poisson distributed with the same parameter, which is not the case when \( \eta \) is not a refinement of \( \xi_n \). The definition given in Section 3.2 is based strictly on the notion of description length, and it is also applied in the practical algorithm in Section 4.2.

### 4.2 Algorithms for description length minimization

In this section we present algorithms that we have used in actual computations of regular decompositions of graph and matrix data. These are written for standard two-part MDL, where the code lengths \( L_4 \) and \( L_5 \) have a usual interpretation as a minus log-likelihood of a graph corresponding to a stochastic block model.

Thus, we use link coding lengths found in the upper bound of Proposition 3.2. In many cases, this is all that can be computed realistically. We can obviously describe a partition into \( k \) nonempty sets using an \( n \times k \) binary matrix with all row sums equal to one and requiring that none of the column sums equals zero. The space of all such matrices we denote as \( \mathcal{R}_k \) and the members of this set as \( R \in \mathcal{R}_k \).

Definition 4.4 For a given graph \( G \in \Omega_n \) with adjacency matrix \( A \) and a partition matrix \( R \in \mathcal{R}_k \), denote

\[ P_1(R) := R^T AR, \]

where \(^T\) stands for matrix transpose, the column sums of \( R \) are denoted as

\[ n_\alpha := (R^T R)_{\alpha,\alpha}, \quad 1 \leq \alpha \leq k, \]

and the number of links within each block and between block pairs as

\[ e_{\alpha,\beta}(R) = (1 - \frac{1}{2} \delta_{\alpha,\beta})(P_1(R))_{\alpha,\beta}. \]
Then define

\[(P(R))_{\alpha,\alpha} := 1_{\{n_\alpha > 1\}} \frac{e_{\alpha,\beta}(R)}{n_\alpha n_\beta}, \quad (P(R))_{\alpha,\beta} := \frac{e_{\alpha,\beta}(R)}{n_\alpha n_\beta}, \quad \alpha \neq \beta.\]

Then the coding length of the graph corresponding to \(A\) using the model \(R\) is:

**Definition 4.5**

\[
l_k(G(A) \mid R \in \mathcal{R}_k) := \sum_{1 \leq i < j \leq k} n_i n_j H((P(R))_{i,j}) + \sum_{1 \leq i \leq k} \left(\frac{n_i}{2}\right) H((P(R))_{i,i}) + l_k(R),
\]

where

\[
l_k(R) = \sum_{1 \leq i \leq k} n_i H(n_i/n) + \sum_{1 \leq i < j \leq k} l^*(e_{i,j}(R))
\]

is the code length of the model, according to our theory and notation.

The **two-part MDL program** of finding the optimal model, denoted as \(R_{k^*}\), can now be written as:

\[
(k^*, R_{k^*}) := \arg \min_{1 \leq k \leq n} \min_{R \in \mathcal{R}_k} l_k(G(A) \mid R \in \mathcal{R}_k)
\]

To solve this program approximately, we can use the following greedy algorithm.

**Algorithm 4.6 Greedy Two-part MDL**

**Input:** \(G = G(A) \in \Omega_n\) a simple graph of size \(n\).

**Output:** \((k^*, R_{k^*} \in \mathcal{R}_{k^*})\), such that the two-part code for \(G\) is shortest possible for all models in \(\mathcal{M}_n\) by using this pair as a model.

**Start:** \(k = 1, l^* = \infty, R \in \mathcal{R}_n = \{I\}, k^* = 1\), where \(I\) denotes the \(n \times n\) unit matrix.

1. **Find** \(\hat{R}_k(G) := \arg \min_{R \in \mathcal{R}_k} l_k(G \mid R)\) using subroutine **ARGMAX k** (Algorithm 4.8).
2. **Compute** \(l_k(G) = [l_k(G \mid \hat{R}(G))] + l_k(R(G))\)
3. **If** \(l_k(G) < l^*\) **then** \(l^* = l_k(G), R_{k^*} = \hat{R}_k(G), k^* = k\)
4. \(k = k + 1\)
5. **If** \(k > n\), **Print** \((R_{k^*}, k^*)\) and **STOP** the program.
6. **GoTo** 1.

**Definition 4.7** A mapping \(\Phi : \mathcal{R}_k \to \mathcal{R}_k\) is defined as follows:

\[
(\log P(R))_{\alpha,\beta} := \log(P(R))_{\alpha,\beta},
(1 - \log P(R))_{\alpha,\beta} := 1 - \log(P(R))_{\alpha,\beta},
L(R) := -AR(\log P(R))^T - (1 - A)RL\log(1 - P(R)),
\]

where we set \(\log 0 = 0\) (all \(\log 0\) values will be later multiplied by 0),

\[
\beta(i, R) := \inf\{\beta : \beta = \arg \min_{1 \leq \alpha \leq k} (L(R))_{i,\alpha}\, 1 \leq i \leq n,
\]

and finally

\[
\Phi(R)_{i,\alpha} = \delta_{\alpha,\beta(i, R)}.
\]
The mapping \( \Phi(R) \) moves each node to a possibly different block such that the description length would be minimized if all other nodes stay in their current blocks.

**Algorithm 4.8 ARGMAX \( k \)**

*Algorithm for finding optimal regular decomposition for fixed \( k \)*

**Input:** \( A \): the adjacency matrix of a graph (an \( n \times n \) symmetric binary matrix with zero trace); \( N \): an integer (the number of iterations in the search of a global optimum); \( k \): a positive integer.

**Start:** \( m = 1 \).

1. \( i := 0 \); generate a uniformly random element \( R_i \in \mathcal{R}_k \).
2. If at least one of the column sums of \( R_i \) is zero, GoTo 1. Otherwise, set \( R_{i+1} := \Phi(R_i) \).
3. If \( R_{i+1} \neq R_i \), set \( i := i + 1 \) and GoTo 2.
4. \( R(m) := R_i \); \( m = m + 1 \); \( l(m) := \sum_{i=1}^{n} \min_{1 \leq \alpha \leq k} (L(R(m)))_{i,\alpha} \).
5. If \( m < N \), GoTo 1.
6. \( M := \{ m : l(m) \leq l(i), i = 1, 2, \ldots, N \} ; m^* := \inf M \).

**Output** optimal solution: \( R(m^*) \).

For very large graphs, the program may not be solvable in the sense that it is not possible and reasonable to go through all possible values of \( k \in \{1, 2, \ldots, n\} \). One option is to limit the range of \( k \). In case that no minimum is found, then use as an optimal choice the model found for the largest \( k \) within this range. Another option is to find the first minimum with smallest \( k \) and stop. When the graph is extremely large, it makes sense to use only a randomly sampled sub-graph as an input — indeed, when \( k^* \ll n \), a large-scale structure can be estimated from a sample.

Our algorithm for Poissonian block models and matrices is essentially similar, with certain differences in formulae as detailed below. Algorithms for other cases like directed graphs and non-quadratic matrices are written very similarly, although two partitions are needed, one for rows and one for columns. The logic of the solution remains the same however.

A semi-heuristic two-part MDL algorithm for finding a regular decomposition for an \( n \times m \) matrix \( A \) with non-negative entries works as follows. The decomposition takes a form of a bi-clustering: there are two partitions, one for rows and one for columns. Such partitions are described by binary matrices with row sums equal to one. The two-part MDL program to find an optimal regular decomposition is written as follows:

The row partition-matrices are denoted as \( R \in \mathcal{R}_{k_1} \) with dimensions \( n \times k_1, 1 \leq k_1 \leq n \), and the column partition matrices as \( C \in \mathcal{C}_{k_2} \) with dimensions \( m \times k_2, 1 \leq k_2 \leq m \).

Let us formulate the cost function for the matrix case that is derived from Eq (9). The number of matrix elements in row group \( \alpha \) and column group \( \beta \) can be written as a matrix element:

\[
(N)_{\alpha,\beta} = (R^T R)_{\alpha,\alpha} (C^T C)_{\beta,\beta} := n_{\alpha} m_{\beta}.
\]

Assuming that all blocks are non-empty, all \( N_{\alpha,\beta} > 0 \), we can define an average matrix element of block \( \alpha, \beta \). First compute the sum of all matrix elements of
A over such a block:

\[ e_{\alpha,\beta} = (RTAC)_{\alpha,\beta}. \]

The corresponding block averages form a \( k_1 \times k_2 \) \( P \)-matrix with elements

\[ (P)_{\alpha,\beta} = \frac{e_{\alpha,\beta}}{(N)_{\alpha,\beta}}. \]

The coding length of the matrix \( A \) using a two-part MDL code with partitions \((R,C)\) can be written as

\[
l_{k_1,k_2}(G(A) \mid R \in \mathcal{R}_{k_1}, C \in \mathcal{C}_{k_2}) = 
\sum_{1 \leq \alpha \leq k_1, 1 \leq \beta \leq k_2} \left\{ e_{\alpha,\beta}(1 - \log((P)_{\alpha,\beta}) + l^*([e_{\alpha,\beta}])) \right\} + 
\sum_{1 \leq \alpha \leq k_1} n_\alpha H\left(\frac{n_\alpha}{n}\right) + 
\sum_{1 \leq \beta \leq k_2} m_\beta H\left(\frac{m_\beta}{m}\right) + k_1 \times k_2 c.
\]

Here we assume a similar handling of log 0’s as in the binary case. \( n_\alpha \) and \( m_\beta \) are the sizes of row and column blocks, and \([e_{\alpha,\beta}]\) denotes the integer part of \( e_{\alpha,\beta} \); it is assumed that such block sums are large numbers with finite decimal precision \( c > 0 \). The description length of such decimals is the last term and it is small compared with other terms for large matrices and can be safely ignored.

Similarly to the binary case, the two-part MDL program is defined as:

**Definition 4.9**

\[
(k_{1}^*, k_{2}^*, R, C) := \arg \min_{(k_{1}, k_{2})} \left( \min_{R \in \mathcal{R}_{k_1}, C \in \mathcal{C}_{k_2}} l_{k_1,k_2}(G(A) \mid R \in \mathcal{R}_{k_1}, C \in \mathcal{C}_{k_2}) \right) \tag{16}
\]

where

\[ 1 \leq k_1 \leq n, \quad 1 \leq k_2 \leq m. \]

The greedy algorithm of solving this program is very similar to the case of a binary matrix. The difference is that two parametric sequences of partitions must be searched, \( \mathcal{R}_i \) and \( \mathcal{C}_j \), and in the subroutine that finds the optimal partitions for fixed \( i \) and \( j \). One may consider different strategies in the corresponding search of an optimal pair. For instance, moving first along the diagonal \( i = j \), and finding the value where the cost function (coding length of \( A \)) has a knee-point, and after that make an off-diagonal search near that value. Another option could be moving along the steepest descent direction of the cost function, or alternating the directions of increments in \( i \) and \( j \), until a saturation is reached in one direction, and then keeping that parameter fixed and finding the optimum on the second parameter. This question is a subject to further experimenting with real and artificial data. Therefore we write only the subroutine that finds the optimal partitioning with fixed \( k_1 \) and \( k_2 \) in a greedy fashion. First we need

**Definition 4.10** Define the mappings \( \Phi_R : \mathcal{R}_{k_1} \times \mathcal{C}_{k_2} \rightarrow \mathcal{R}_{k_1} \) and \( \Phi_C : \mathcal{R}_{k_1} \times \mathcal{C}_{k_2} \rightarrow \mathcal{C}_{k_2} \) as follows. Let \( E \) be an \( n \times m \) matrix with all elements equal to
1. Then, using the definition of $P$-matrix and LogP-matrix (related to $P$ as in binary case), define two matrices using a block-matrix notation:

$$L(R, C) = (E \ A) \left( \begin{array}{c} CP^T \\ -C(\log P)^T \end{array} \right)$$

and

$$M(R, C) = (E^T \ A^T) \left( \begin{array}{c} \log P \\ -R \log P \end{array} \right).$$

Define

$$\beta_1(i, R, C) = \inf \{ \beta : \beta = \arg\min_{1 \leq \alpha \leq k_1} (L(R, C))_{i, \alpha} \}, \quad 1 \leq i \leq n,$$

and

$$\beta_2(i, R, C) = \inf \{ \beta : \beta = \arg\min_{1 \leq \alpha \leq k_2} (M(R, C))_{i, \alpha} \}, \quad 1 \leq i \leq m.$$

Then,

$$\Phi_R(R, C)_{i, \alpha} = \delta_{\alpha, \beta_1(i, R, C)}, \quad 1 \leq \alpha \leq k_1, \quad 1 \leq i \leq n,$$

and

$$\Phi_C(R, C)_{i, \alpha} = \delta_{\alpha, \beta_2(i, R, C)}, \quad 1 \leq \alpha \leq k_2, \quad 1 \leq i \leq m.$$

The main greedy subroutine is:

**Algorithm 4.11 ARGMAX $(k_1, k_2)$**

*Algorithm for finding optimal regular decomposition for fixed $(k_1, k_2)$.*

**Input:** $A$: a real $n \times m$ matrix with non-negative entries; $N$: a positive integer (the number of iterations in the search of a global optimum); $(k_1, k_2)$: a pair of positive integers.

**Start:** $m = 1$.

1. $i = 0$; generate uniformly random elements $R_i \in \mathbb{R}_{k_1}$ and $C_i \in \mathbb{C}_{k_2}$.

2. If at least one of column sums of $R_i$ or $C_i$ is zero, GoTo 1. Otherwise, set $R_{i+1} := \Phi_R(R_i, C_i), \quad C_{i+1} := \Phi_C(R_i, C_i)$.

3. If $R_{i+1} \neq R_i$ or $C_{i+1} \neq C_i$, set $i := i + 1$ and GoTo 2.

4. $R(m) := R_i; \quad C(m) := C_i; \quad m := m + 1$;

   $$e_{\alpha, \beta} := \left( T^T (m) A C(m) \right)_{\alpha, \beta}, \quad 1 \leq \alpha \leq k_1, \quad 1 \leq \beta \leq k_2;$$

   $$N_{\alpha, \beta} := \left( R^T (m) R(m) \right)_{\alpha, \alpha} \left( C^T (m) C \right)_{\beta, \beta}, \quad 1 \leq \alpha \leq k_1, \quad 1 \leq \beta \leq k_2;$$

   $$l(m) := \sum_{1 \leq \alpha \leq k_1, 1 \leq \beta \leq k_2} e_{\alpha, \beta} \left( 1 - \log \frac{e_{\alpha, \beta}}{N_{\alpha, \beta}} \right).$$

5. If $m < N$, GoTo 1.

6. $M := \{ m : l(m) \leq l(i), i = 1, 2, ..., N \}; \quad m^* := \inf M$.

**OUTPUT** optimal solution: $(R(m^*), C(m^*))$.

**Remark 4.12** It is also possible to find a regular decompositions in the case of partly missing matrix elements and also in the case of mixed positive and negative entries. In the latter case, we can use the idea of directed links already used in [12]. In the first case, we note that the main characteristic of the regular decomposition is the $P$-matrix with elements that are averages of the data matrix over large blocks that can be estimated, in many cases, despite a portion of data is missing.
4.3 Accuracy of block structure identification by MDL

The general idea of the Regular Decomposition method is to be a generic tool for separating structure and randomness in large data sets of graph or matrix form. A partition of a real-world data set that minimizes the (nominal) code length given in (13) resp. (14) can often not be compared with a ‘true solution’ for the simple reason that there may not be any objective notion of a ‘true structure’ of the data. However, it is important to analyse and understand how the method performs when the data really originates from a stochastic block model. This question is called the consistency of MDL. Our results on this question are summarized in the following theorem, formulated in terms of the asymptotic behavior of a model sequence as specified by Construction 2.4. In such a framework, an event is said to happen with high probability, if its probability tends to 1 when $n \to \infty$.

**Theorem 4.13** Consider a sequence of stochastic block models $(G_n, \xi_n)$ based on a vector $(\gamma_1, \ldots, \gamma_k)$ of relative block sizes and a matrix $D = (d_{ij})_{i,j=1}^k$ of link probabilities, as described in Definition 2.3 and Construction 2.4. With high probability, the following hold:

(i) Among all partitions $\eta$ of $V_n$ such that $|\eta| \leq k$, $\xi_n$ is the single minimizer of $L(G_n|\eta)$.

(ii) For any fixed $\epsilon \in (0, \min_i \gamma_i)$, $\xi_n$ is the single minimizer of $L(G_n|\eta)$ among partitions $\eta$ with minimal block size higher than $n\epsilon$.

(iii) No refinement $\eta$ of $\xi_n$ with $|\eta| \leq m$ improves $L(G_n|\xi_n)$ by more than $\text{const}(k,m) \log n$.

The corresponding claims hold for the Poissonian block model mutatis mutandis.

**Proof** The proof is given in the next section and structured into several propositions. Claim (i) follows from Proposition 5.1, Proposition 5.2 and Corollary 5.6. Claim (ii) is Proposition 5.7. Claim (iii) is Proposition 5.5.

The results of Section 5 offer a richer picture than what was distilled into Theorem 4.13. For example, Proposition 5.1 shows that if $|\eta| = k$ and $\eta$ differs from $\xi_n$ only a little, then the remaining misplaced nodes can be immediately identified by computing their effect to the value of $L(G_n|\eta)$. On the other hand, we have not been able to exclude the possibility that a refinement of $\eta$ could yield a slight $O(\log n)$ improvement of the code length.

**Remark 4.14** It is rather obvious that with large $n$, the identification of the block structure $\xi_n$ is robust against independent noise. The simplest case is that the Poissonian block model is disturbed by additive Poissonian noise to each matrix element:

$$\tilde{e}_{ij} = e_{ij} + \phi_{ij},$$

where the $\phi_{ij}$s are i.i.d. with $\phi_{ij} \sim \text{Poisson}(\nu)$, $\nu > 0$. Then $(\tilde{e}_{ij})$ is again an irreducible Poissonian block model with the same partition. More interesting cases are binary flips in the graph case and multiplicative noise with mean 1 in the case of non-negative matrices. We leave these for forthcoming work.
4.4 $\epsilon$-regularity vs. stochastic block models

Although the structure that a MDL-based algorithm finds typically looks like an $\epsilon$-regular structure, there is a principal difference. In particular cases, an $\epsilon$-regular graph can have a structure that allows much better compression that that provided by the $\epsilon$-regular partition. In this section we give an explicit example of such a case.

An important point in SRL is that for any $\epsilon > 0$, there is an upper bound for the size of regular partition, $M(\epsilon)$ so that for any graph with size above some finite threshold $N(\epsilon)$, all such graphs have a regular partition with at most $M(\epsilon)$ sets. Based on this, we show that the $\epsilon$-regular structure of SRL and the structure induced by the MDL need not coincide. Let us fix an order of graph $2n$, large enough so that SRL holds for some $\epsilon > 0$, and that $M(\epsilon) < n^{1-\alpha}$ for some fixed $0 < \alpha < 1/2$.

**Proposition 4.15** There is a graph of order $2n$ such that it has a MDL structure with code length $o(n^2)$ and an $\epsilon$-regular structure that allows only $O(n^2)$ code length.

**Proof** Take $n$ large enough as prescribed above. Then construct a bipartite graph $(X, Y)$ with $|X| = |Y| = n$ such that $n$ is divisible by $n^\alpha$ with some rational $\alpha \in (0, 1/2)$. Assume that both parts of the bipartition are further partitioned into equal size blocks: $X = \sum X_i, Y = \sum Y_i$, $|X_i| = |Y_i| = n^\alpha$.

Define then a random graph $G_p = (X, Y, E)$ as follows. For each pair $(X_i, Y_j)$, take $e(X_i, Y_j) = \xi_{i,j}|X_i||Y_j|$, where $\xi_{i,j} \sim Ber(p)$ is a Bernoulli random variable with parameter $0 < p < 1$, and the variables for different pairs are independent. Assume that there are no other edges.

We show that, with high probability, $G_p$ is $\epsilon$-regular with regular partition $(X, Y)$. In the MDL-approach, such a structure has a coding length at least $\binom{n}{2} H(p + o(1)) = O(n^2)$. This comes from the log-likelihood part, and the $o(1)$ corresponds to very small deviations of link densities from the expected value $p$ that can be made arbitrarily small by increasing $n$. Now we check that the $\epsilon$-regularity of graph $G_p$ has a positive probability, which implies that such an $\epsilon$-regular pair exists (actually, it appears that this happens with high probability).

$\epsilon$-regularity means that for any $X' \subseteq X, Y' \subseteq Y$, $|X'|, |Y'| > \epsilon n$, the link density $d(X', Y')$ deviates from the link density of the pair, $d(X, Y)$, no more than by $\epsilon$. By definition,

$$d(X', Y') = \frac{\sum_{i,j} \xi_{i,j} |X' \cap X_i || Y' \cap Y_j|}{|X'| |Y'|},$$

and as a result the expectation is

$$Ed(X', Y') = \frac{\sum_{i,j} E\xi_{i,j} |X' \cap X_i || Y' \cap Y_j|}{|X'| |Y'|} = p.$$

Denote

$$x_{i,j} := \frac{\xi_{i,j} |X' \cap X_i || Y' \cap Y_j|}{|X'| |Y'|}$$

The range of $x_{i,j}$ is interval $[0, 1]$ of unit length. Hoeffding’s inequality yields for $S = \sum x_{i,j}$ that

$$P(|S - ES| > t) \leq 2e^{-\frac{t^2}{\sum x_{i,j}^2}}.$$

21
where \(b_{i,j} = \frac{|X' \cap X_i| |Y' \cap Y_j|}{|X||Y'|} \) is the range of variable \(x_{i,j} \). The denominator of the exponent in the right-hand side of the Hoeffding inequality can be bounded as
\[
\sum b_{i,j}^2 \leq \frac{1}{(en)^2} \sum |X' \cap X_i|^2 |Y' \cap Y_j|^2 \leq \frac{1}{(en)^2} n^{2\alpha} n^{2\alpha} \left( \frac{n}{en} \right)^2 = \frac{1}{e^2} n^{2\alpha}.
\]
By taking \(t = \epsilon |X'| |Y'| \geq \epsilon (en)^2 \) we get for the link density:
\[
P(\| d(X', Y') - p \| > \epsilon) \leq 2e^{-2\epsilon^8 n^2 (1-\alpha)}.
\]
Finally, since there are at most \(4^n \) pairs of subsets \((X', Y')\) from which to choose, the probability that none of them violates regularity is lower bounded by
\[
1 - 2 \times 4^n e^{-2\epsilon^8 n^{2(1-\alpha)}} \to 1,
\]
if the exponent has a positive power of \(n\), and this happens when \(\alpha < 1/2\).

Thus, all large subsets have densities that deviate from expectation less than \(\epsilon\) with a probability tending to one. Thus, we have shown the \(\epsilon\)-regularity of the partition \((X, Y)\).

On the other hand, using MDL, we could reach the level of small sets \(X_i\) and \(Y_i\), and the corresponding log-likelihood is zero. The model complexity is \(o(n^2)\), as can be easily seen from asymptotic formulas for the upper bound for \(m_k\), with \(k = n^{1-\alpha}\).

\[\Box\]

## 5 Proof of Theorem 4.13

Through this section, we consider a sequence \((G_n|\xi_n)\) of increasing versions of a fixed stochastic block model based on a vector \((\gamma_1, \ldots, \gamma_k)\) of relative block sizes and a matrix \(D = (d_{ij})_{i,j=1}^k\) of link probabilities, as specified in Construction 2.4.

Consider partitions \(\eta\) of \(V_n\). Denote
\[
\mathcal{D}(\eta, \xi_n) = \frac{1}{n} \max_{B \in \eta} \min_{A \in \xi_n} |B \setminus A|.
\]
Thus, \(\mathcal{D}(\eta, \xi_n) = 0\) if and only if \(\eta\) is a refinement of \(\xi_n\). If \(v \in V_n\) and \(B \in \eta\), denote by \(\eta_{v,B}\) the partition obtained from \(\eta\) by moving node \(v\) to block \(B\) (if \(v \in B\), then \(\eta_{v,B} = \eta\)).

**Proposition 5.1** There is a number \(\epsilon_0 > 0\) such that the following holds with high probability: if \(|\eta| = k\) and \(\mathcal{D}(\eta, \xi_n) \leq \epsilon_0\), then

if \(A \in \xi_n\), \(B \in \eta\), \(\frac{1}{n} |B \setminus A| \leq \epsilon_0\) and \(v \in A \setminus B\), then \(L(G_n|\eta_{v,B}) < L(G_n|\eta)\).

**Proof** Let \(\epsilon, \delta > 0\) be small numbers and \(m\) a positive integer to be specified. They can be chosen so that the following holds:

- \(\epsilon\) is small so that \(\eta\) and \(\xi_n\) nearly overlap when \(\mathcal{D}(\eta, \xi_n) \leq \epsilon\):

  \[m \epsilon \leq \delta \min_{A \in \xi_n} |A|; \quad (17)\]
• all the differing link probabilities are widely separated in $\delta$ units:

$$\delta \leq \frac{1}{m} \min \{|d_{ij} - d_{j\bar{i}}| : i, j, j_2 \in \{1, \ldots, k\}, \ d_{ij} \neq d_{j\bar{i}}\}; \quad (18)$$

• the empirical densities are close to their mean values: for any $A_i, A_j \in \xi_n$ (possibly $i = j$), we have, with high probability,

$$|d(A_i, A_j) - d_{ij}| + m\epsilon \leq \delta. \quad (19)$$

Let $\eta$ be a partition of $V_n$ such that $d(\eta, \xi_n) \leq \epsilon$. Condition $[17]$ entails that for each block $A_j \in \xi_n$ there is a unique block $B_j \in \eta$ such that $|B_j \setminus A_j| \leq \epsilon$. Let us now assume that $v \in A_i \cap B_j$ and $i \neq j$, and compare the partitions $\eta$ and $\eta_{v, B_i}$. Denote $b_i = |B_i|$, $i = 1, \ldots, k$, and $\tilde{B}_i = B_i \cup \{v\}$, $\tilde{B}_j = B_j \setminus \{v\}$. Then

$$L_4(G_n|\eta) + L_5(G_n|\eta) - (L_4(G_n|\eta_{v, B_i}) + L_5(G_n|\eta_{v, B_j}))$$

$$= \left(\frac{b_i}{2}\right)H(d(B_i)) + \left(\frac{b_j}{2}\right)H(d(B_j)) - \left(\frac{b_i + 1}{2}\right)H(d(\tilde{B}_i)) - \left(\frac{b_j - 1}{2}\right)H(d(\tilde{B}_j))$$

$$+ \sum_{q \neq i,j} [b_i b_q H(d(B_i, B_q)) + b_j b_q H(d(B_j, B_q))$$

$$- (b_i + 1)b_q H(d(\tilde{B}_i, B_q)) + (b_j - 1)b_q H(d(\tilde{B}_j, B_q))]$$

$$+ b_i b_j H(d(B_i, B_j)) - (b_i + 1)(b_j - 1)H(d(\tilde{B}_i, \tilde{B}_j)). \quad (20)$$

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

$$b_j[H(d(B_j, B_q)) - \frac{b_j - 1}{b_j}H(d(\tilde{B}_j, B_q))]$$

$$- (b_i + 1)[H(d(\tilde{B}_i, B_q)) - \frac{b_i}{b_j + 1}H(d(B_i, B_q))]$$

$$= b_j \left[H \left(\frac{b_j - 1}{b_j}d(B_j, B_q) + \frac{1}{b_j}d(\{v\}, B_q)\right)$$

$$- \frac{b_j - 1}{b_j}H(d(\tilde{B}_j, B_q)) - \frac{1}{b_j}H(d(\{v\}, B_q))\right]$$

$$- (b_i + 1) \left[H \left(\frac{b_i}{b_i + 1}d(B_i, B_q) + \frac{1}{b_i + 1}d(\{v\}, B_q)\right)$$

$$- \frac{b_i}{b_i + 1}H(d(B_i, B_q)) - \frac{1}{b_i + 1}H(d(\{v\}, B_q))\right)$$

(note the addition and subtraction of the term $H(d(\{v\}, B_q))$). Using Lemma $[A, 7]$ and the assumptions on $\epsilon$, $\delta$ and $m$, the last expression can be set to be, with high probability, arbitrarily close to the number

$$I(d_{jq} : d_{iq}) + I(d_{iq} : d_{iq}) = I(d_{jq} : d_{iq})$$

(the function $I(\cdot, \cdot)$, the Kullback-Leibler divergence of Bernoulli distributions, is defined in $(22)$). Thus, the sum over $q$ is, with high probability, close to

$$\sum_{q \neq i,j} b_q I(d_{jq}, d_{iq}).$$

23
Let us then turn to the remaining parts of (20) that refer to two codings of the internal links of $B_i \cup B_j$. Similarly as above, we can add and subtract terms to transform these parts into

\[
\binom{b_j}{2} \left[ H \left( \frac{b_j-1}{b_j} d(\tilde{B}_j) + \frac{b_j-1}{b_j} d(\{v\}, \tilde{B}_j) \right) \right.
- \left. \frac{b_j-1}{b_j} H(d(\tilde{B}_j)) \right]
\]

\[
- \binom{b_i+1}{2} \left[ H \left( \frac{b_i}{b_i+1} d(B_i) + \frac{b_i}{b_i+1} d(\{v\}, B_i) \right) \right.
- \left. \frac{b_i}{b_i+1} H(d(B_i)) \right]
\]

\[
+ b_ib_j \left[ H \left( \frac{b_j-1}{b_j} d(B_j, \tilde{B}_j) + \frac{1}{b_j} d(\{v\}, B_i) \right) \right.
- \left. \frac{b_j-1}{b_j} H(d(B_i, \tilde{B}_j)) \right]
\]

\[
- \binom{b_i+1}{2} \left[ H \left( \frac{b_i}{b_i+1} d(B_i, \tilde{B}_j) + \frac{1}{b_i+1} d(\{v\}, \tilde{B}_j) \right) \right.
- \left. \frac{b_i}{b_i+1} H(d(\{v\}, \tilde{B}_j)) \right]
\]

\[
\approx (b_j - 1) I(d_{ij} : d_{jj}) - (b_i + 1) I(d_{ii} : d_{ij}) + b_i I(d_{ii} : d_{jj}) - (b_j - 1) I(d_{ij} : d_{ij})
\]

\[
= b_j I(d_{ij} : d_{jj}) + b_i I(d_{ii} : d_{ij}).
\]

By the above analysis of (20), we have obtained

\[
L_4(G_n|\eta) + L_5(G_n|\eta) - (L_4(G_n|\eta_{e,B_i}) + L_5(G_n|\eta_{e,B_i})) 
\approx b_q \sum_{q \neq i,j} I(d_{jq} : d_{ij}) + b_j I(d_{ij} : d_{jj}) + b_i I(d_{ii} : d_{ij}). \tag{21}
\]

By the irreducibility assumption (1), there is a block $A_q$ such that $d_{qj} \neq d_{jj}$, with the possibility that $q \in \{i,j\}$. It follows that at least one of the $I(x : y)$'s in (21) is positive. Denote

\[
\kappa^* = \min \{ I(d_{ij1} : d_{ij2}) : d_{ij1} \neq d_{ij2} \}.
\]

Thus, with high probability,

\[
L_4(G_n|\eta) + L_5(G_n|\eta) - (L_4(G_n|\eta_{e,B_i}) + L_5(G_n|\eta_{e,B_i})) > \frac{1}{2} (\kappa^* \min \gamma_i) n.
\]

On the other hand, it is easy to compute that

\[
L_3(G_n|\eta) - L_3(G_n|\eta_{e,B_i}) = n(H(\eta) - H(\eta_{e,B_i})) \rightarrow \log \frac{2j}{\gamma_i}.
\]

The changes of $L_1$ and $L_2$ when moving from $\eta$ to $\eta_{e,B_i}$ are negligible. This concludes the proof. \qed
The proof of Proposition 5.1 showed that when \( d(\eta, \xi_n) \leq \epsilon_0 \), moving any node to its correct block decreases \( L(G_n|\eta) \) at least by \( (1/2) \left( \kappa^\star \min_i \gamma_i \right) n \). In particular, with high probability, \( \xi_n \) is the unique minimizer of \( L(G_n|\eta) \) among \( k \)-partitions \( \eta \) satisfying \( d(\eta, \xi_n) \leq \epsilon_0 \).

**Proposition 5.2** For any \( \epsilon \in (0, 1) \) and positive integer \( m \), there is a constant \( c_\epsilon \) such that the following holds with high probability:

\[
\text{if } |\eta| \leq m \text{ and } d(\eta, \xi_n) > \epsilon, \text{ then } \frac{1}{n^2} (L(G_n|\eta) - L(G_n|\xi_n \lor \eta)) \geq c_\epsilon.
\]

**Proof** Fix an \( \epsilon \in (0, 1) \) and let \( \eta \) be a partition of \( V_n \) such that \( d(\eta, \xi_n) > \epsilon \).

By the concavity of \( H \), we have

\[
L_4(G_n|\eta) + L_5(G_n|\eta) = \sum_{B \in \eta} \left( \begin{array}{c} |B| \\ 2 \end{array} \right) H(d(B)) + \frac{1}{2} \sum_{B, B' \in \eta} |B| |B'| H(d(B, B'))
\]

\[
\geq \sum_{B \in \eta} \sum_{A \in \xi_n} \left( \begin{array}{c} |A \cap B| \\ 2 \end{array} \right) H(d(A \cap B))
\]

\[
+ \frac{1}{2} \sum_{B \in \eta} \sum_{A, A' \in \xi_n} |A \cap B| |A' \cap B| H(d(A \cap B, A' \cap B)) \quad (22)
\]

\[
+ \frac{1}{2} \sum_{B, B' \in \eta} \sum_{A, A' \in \xi_n} |A \cap B| |A' \cap B'| H(d(A \cap B, A' \cap B'))
\]

\[
= L_4(G_n|\eta \lor \xi_n) + L_5(G_n|\eta \lor \xi_n).
\]

By assumption, there is a block \( B \in \eta \) 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

\[
\min \{|A_i \cap B|, |A_j \cap B|\} \geq \frac{\epsilon}{k-1} n. \quad (23)
\]

By the irreducibility assumption (ii), there is a block \( A_q \) such that \( d_{q_i} \neq d_{q_j} \), with the possibility that \( q \in \{i, j\} \). Fix an arbitrary \( \delta > 0 \) to be specified later.

By \( \epsilon \)-regularity (claim [ii]) of Lemma A.7, with high probability, every choice of a partition \( \eta \) with \( |B \setminus A| > \epsilon n \) results in some blocks \( A_i, A_j, A_q \) with the above characteristics plus the regularity properties

\[
|d(A_i \cap B, A_q \cap B') - d_{q_i}| \leq \delta, \quad |d(A_j \cap B, A_q \cap B') - d_{q_j}| \leq \delta \quad (24)
\]

where \( B' \) denotes a block of \( \eta \) that maximizes \( |A_q \cap B'| \) (note that because
because the difference between the sides of the equality has a positive lower bound that holds with high probability. On the other hand, this difference is part of the overall concavity inequality (22).

**Proposition 5.3** For any refinement \( \eta \) of \( \xi_n \), we have

\[
L_4(G_n|\xi_n) + L_5(G_n|\xi_n) - (L_4(G_n|\eta) + L_5(G_n|\eta)) \leq \sum_{j=1}^{M(|\eta|)-M(k)} \log 2 + Y_i,
\]

where (st) refers to stochastic order, the \( Y_i \)'s are i.i.d. \( \text{Exp}(1) \) random variables, and

\[
M(x) = \frac{x(x+1)}{2}.
\]

**Proof** Here we apply results presented in Appendix A. Denote by \( \eta \cap A_i \) the subset of \( \eta \) whose members are subsets of the block \( A_i \) of \( \xi_n \). Writing the edge code lengths of the coarser and finer partition similarly as in (22), taking the difference and using (35), we obtain

\[
L_4(G_n|\xi_n) + L_5(G_n|\xi_n) - (L_4(G_n|\eta) + L_5(G_n|\eta)) = \sum_{i=1}^{k} \left( \sum_{B \in \eta \cap A_i} \left( \frac{|B|}{2} \right) I(d(B):d_{ii}) + \frac{1}{2} \sum_{B,B' \in \eta \cap A_i, B \neq B'} |B||B'|I(d(B,B'):d_{ii}) \right.
\]

\[
- \left( \frac{|A_i|}{2} \right) I(d(A_i):d_{ii})
\]

\[
+ \sum_{i<j} \left( \frac{|A_i|}{2} \sum_{B \in \eta \cap A_i} \sum_{B' \in \eta \cap A_j} |B||B'|I(d(B,B'):d_{ij}) - |A_i||A_j|I(d(A_i,A_j):d_{ij}) \right).
\]

(27)

Applying Proposition A.6 to each term of both outer sums now yields the claim, because

\[
\sum_{i=1}^{k} (M(|\eta \cap A_i|) - 1) + \sum_{i<j} (|\eta \cap A_i||\eta \cap A_j| - 1) = M(|\eta|) - M(k).
\]
Remark 5.4 It is rather surprising that the stochastic bound (25) depends only on the number of blocks in $\eta$ — not on their relative sizes, nor on the overall model size $n$!

Proposition 5.5 For any positive integer $m > k$, the following holds with high probability:

$$L(G_n|\xi_n) - \min_{\eta \geq \xi_n, |\eta| \leq m} L(G_n|\eta) \leq (m + M(m - k)) \log n,$$

where the relation $\eta \geq \xi_n$ means that $\eta$ is a refinement of $\xi_n$, and $M(\cdot)$ was defined in (26).

Proof Let $\eta$ be a refinement of $\xi_n$. Refining the partition w.r.t. $\xi_n$ yields a gain, based on the concavity of $H$, in the code part $L_4 + L_5$, but costs in the parts $L_1, L_2$ and $L_3$. We have to relate these to each other.

Consider first the value of $L_2(G_n|\eta)$. In our analysis, it is important to distinguish between ‘large’ and ‘tiny’ blocks, where the relative sizes of large blocks exceed some pre-defined number $\epsilon$ and the rest can be arbitrarily small, even singletons. Now, each block $A_i \in \xi_n$ must contain at least one block $B_i \in \eta \cap A_i$ such that $|B_i| \geq |A_i|/m$. Define

$$\epsilon := \min_j \frac{|A_j|}{2mn}.$$ 

Because no concavity gain can be obtained with an index pair $\{i, j\}$ such that $d_{ij} \in \{0, 1\}$, it does not restrict generality to assume that $d_{ij} \in (0, 1)$ for all $i, j$. Then

$$L_2(G_n|\eta) = \sum_{i=1}^{k} \left( \sum_{B \in \eta \cap A_i} \log \left( \frac{|B|}{2} \right) d_{ii} + \frac{1}{2} \sum_{B, B' \in \eta \cap A_i, B \neq B'} \log(|B| |B'| d_{ii}) \right)$$

$$+ \sum_{i<j} \left( \sum_{B \in \eta \cap A_i, B' \in \eta \cap A_j} \log(|B| |B'| d_{ij}) \right)$$

$$\geq \sum_{i=1}^{k} \left( \log \left( \frac{\epsilon^2 n^2 d_{ii}}{2} \right) + (|\eta \cap A_i| - 1) \log(\epsilon n d_{ii}) \right)$$

$$+ \sum_{i<j} \left( \log \left( \epsilon^2 n^2 d_{ij} \right) + (|\eta \cap A_i| + |\eta \cap A_j| - 2) \log(\epsilon n d_{ij}) \right)$$

$$= (k|\eta| + k) \log n + c_1(D, \epsilon).$$

On the other hand, the bound $|A_i| \leq n$ yields

$$L_2(G_n|\xi_n) \leq (k^2 + k) \log n + c_2(D).$$

Thus,

$$L_2(G_n|\eta) - L_2(G_n|\xi_n) \geq k(|\eta| - k) + c_1(D, \epsilon) - c_2(D).$$

(28)

We obviously have also $L_1(G_n|\eta) \geq L_1(G_n|\xi_n)$, but this difference is insignificant in the present context.
The refinement gain in code parts $L_4$ and $L_5$ was bounded in Proposition 5.3 stochastically by $\Exp(1)$ random variables. The rate function (see the beginning of Appendix A) of the distribution $\Exp(1)$ is

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

Denoting

$$\text{Gain}_n(\eta) := L_4(G_n|\xi_n) + L_5(G_n|\xi_n) - (L_4(G_n|\eta) + L_5(G_n|\eta)),$$

Proposition 5.3 yields, for $y > \log 2$,

$$P(\text{Gain}_n(\eta) > y) \leq P\left(\sum_{j=1}^{M(|\eta|) - M(k)} Y_j > y - (M(|\eta|) - M(k)) \log 2\right)$$

$$\leq \exp\left(- (M(|\eta|) - M(k)) \frac{y - (M(|\eta|) - M(k)) \log 2}{M(|\eta|) - M(k)}\right)$$

$$\leq \exp\left(- y + (M(|\eta|) - M(k)) \log y \frac{2e}{M(|\eta|) - M(k)}\right)^{(M(|\eta|) - M(k))},$$

where the second factor is bounded and will be henceforth neglected.

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

$$\exp\left(\sum_{A \in \xi_n} |A| \log 2\right) = e^{n H(\eta|\xi_n)}.$$  

On the other hand, we have

$$L_3(G_n|\eta) - L_3(G_n|\xi_n) = n H(\eta) - n H(\xi_n) = n H(\eta|\xi_n).$$  

Write

$$\Delta L_{123}(\eta) = L_1(G_n|\eta) + L_2(G_n|\eta) + L_3(G_n|\eta)$$

$$- (L_1(G_n|\xi_n) + L_2(G_n|\xi_n) + L_3(G_n|\xi_n)), $$

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

Denote $z(\eta) := (|\eta| + M(|\eta| - k)) \log n$. Recalling (28), the union bound yields

$$P\left(\sup_{\eta' \sim \eta} \text{Gain}_n(\eta') > \Delta L_{123}(\eta) + z(\eta)\right)$$

$$\leq \exp\left(n H(\eta|\xi_n) - \Delta L_{123}(\eta) - z(\eta) + (M(|\eta|) - M(k)) \log(\Delta L_{123}(\eta) + z(\eta))\right)$$

$$\leq \exp\left(- z(\eta) + (|\eta| - k + M(|\eta|) - M(k)) \log n + c_3(D, \epsilon)\right).$$

The number of different block size sequences $\ell_1 \geq \cdots \geq \ell_{|\eta|}$ is upper bounded
by $n^{1-\|\eta\|} = e^{(1-\|\eta\|)\log n}$. Thus, a second application of the union bound yields

$$P\left(\sup_{\eta \geq \xi_n, \|\eta\| \leq m} \Delta L(\eta) > z(\eta)\right) \leq me^{-\left(\left[\frac{k(m-k)}{2}\right] + \left[M(m) - M(k)\right] - \left[m + M(m-k)\right] + \left[m-1\right]\right) \log n + \text{const}}.$$

Thus, $\sup_{\eta \geq \xi_n, \|\eta\| \leq m} \Delta L(\eta) > z(\eta)$ implies $m e^{-\left(\left[\frac{k(m-k)}{2}\right] + \left[M(m) - M(k)\right] - \left[m + M(m-k)\right] + \left[m-1\right]\right) \log n + \text{const}}$.

$$\lim_{n \to \infty} P = 0,$$

$$\lim_{n \to \infty} P = 0,$$

$\xi_n$ is the unique minimizer of $L(G_n|\eta)$ among partitions $\eta$ with $|\eta| = k$.

**Proof**

By Proposition 5.1,

$$\min_{|\eta| = k, \mathfrak{c}(\eta, \xi_n) < \epsilon} L(G_n|\eta) > L(G_n|\xi_n)$$

with high probability. On the other hand, Proposition 5.2 yields that, with high probability,

$$\min_{|\eta| = k, \mathfrak{c}(\eta, \xi_n) > \epsilon} L(G_n|\eta) > L(G_n|\xi_n \vee \eta) + c_\epsilon n^2.$$

By Proposition 5.5,

$$\min_{|\eta| = k} L(G_n|\xi_n \vee \eta) > L(G_n|\xi_n) - (k + M(k^2 - k)) \log n$$

with high probability. It remains to note that $n^2$ grows faster than $\log n$. $\square$

**Proposition 5.7** Let $\epsilon \in (0, \min_i \gamma_i)$. Consider refinements $\eta$ of $\xi_n$ with relative minimal block size $\epsilon$, i.e. the set

$$\mathcal{B}_\epsilon^{(n)} = \{\eta \geq \xi_n : |B| \geq n \epsilon \forall B \in \eta\}.$$  \hspace{1cm} (29)

With high probability,

$$\min_{\eta \in \mathcal{B}_{(n)}^{(n)} \setminus \{\xi_n\}} L(G_n|\eta) > L(G_n|\xi_n).$$  \hspace{1cm} (30)

**Proof** The restriction $\eta \in \mathcal{B}_{(n)}^{(n)}$ implies $|\eta| \leq \lfloor 1/\epsilon \rfloor = m$. The difference to the conditions of Proposition 5.5 is now just the magnitude of $L_2(\eta)$. When $\eta \in \mathcal{B}^{(n)}$,

$$L_2(G_n|\eta) \geq M(|\eta|) \log n + c(D, \epsilon, |\eta|),$$

so that

$$L_2(G_n|\eta) - L_2(G_n|\xi_n) \geq 2(M(|\eta|) - M(k)) \log n + c(D, \epsilon, |\eta|).$$

29
By a corresponding computation as in the proof of Proposition 5.5 we obtain for any fixed $\eta \in B(n)$ that

$$P \left( \sup_{\eta' \sim \eta} \text{Gain}_n(\eta') > \Delta_{L_{123}}(\eta) - \frac{1}{2} \log n \right) \leq \exp \left( - (M(|\eta|) - M(k)) \log n + \text{const} \right).$$

Proceeding with a second union bound like in the proof of Proposition 5.5 yields

$$P \left( \min_{\eta \in B(n) \setminus \{\xi_n\}} L(G_n|\eta) \leq L(G_n|\xi_n) + 1 + \frac{1}{2} \log n + \text{const} \right)$$

$$\leq \max_{q \leq m} \exp \left( \left( - (M(q) - M(k)) \right) + q - 1 \right) + \frac{1}{2} \log n + \text{const}$$

$$= \exp \left( - \frac{1}{2} \log n + \text{const} \right)$$

$$= \frac{\text{const}}{\sqrt{n}} \to 0, \text{ as } n \to \infty,$$

where the maximum over $q$ was obtained with the smallest value $q = k + 1$. □

### A Chernoff bounds and other information-theoretic preliminaries

Consider a random variable $X$ with moment generating function

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

and denote $D_X = \{\beta : \phi_X(\beta) < \infty\}$. We restrict to distributions of $X$ for which $D_X$ is an open (finite or infinite) interval. The corresponding rate function is

$$I_X(x) = - \inf_{\beta \in D_X} (\log \phi_X(\beta) - \beta x).$$

$I_X(x)$ is a strictly convex function with minimum at $\mathbb{E} X$ and value $+\infty$ outside the range of $X$. As a further restriction on the distribution of $X$, we assume that $I_X(x)$ approaches smoothly $+\infty$ both when $x$ decreases or increases from the point $\mathbb{E} X$. For the mean $\bar{X}_n = \frac{1}{n} \sum_i X_i$ of i.i.d. copies of $X$, we have

$$I_{\bar{X}_n}(x) = n I_X(x). \quad (31)$$

The Chernoff bound (also known as Cramér-Lundberg bound):

**Proposition A.1**

$$P(X < x) \leq e^{-I_X(x)} \quad \text{for } x < \mathbb{E} X,$$

$$P(X > x) \geq e^{-I_X(x)} \quad \text{for } x > \mathbb{E} X.$$
This has the following simple consequence. Let the support of \( X \) be the closure of \((x^-, x^+)\) and write

\[
I_X(x) = I_X^-(x)1_{(x^-, E X)}(x) + I_X^+(x)1_{(E X, x^+)}(x) + \infty \cdot 1_{(\mathbb{R} \setminus (x^-, x^+))}(x).
\]

With the assumptions made above, the functions \( I_X^-(x) \) and \( I_X^+(x) \) are, respectively, bijections from \((x^-, E X)\) and \([E X, x^+\)) to \([0, \infty)\).

**Lemma A.2** For any \( z \geq 0 \) we have

\[
I_X(X) \overset{\text{(st)}}{\leq} \log 2 + Y,
\]

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

**Proof**

\[
P(I_X(X) > z) = P(1_{\{X<0\}}I_X(X) > z \text{ or } 1_{\{X>0\}}I_X(X) > z)
= P(1_{\{X<0\}}I_X(X) > z) + P(1_{\{X>0\}}I_X(X) > z)
= P(X < I_X^-(z)) + P(X > I_X^+(z))
\leq \exp(-I_X(I_X^-(z))) + \exp(-I_X(I_X^+(z)))
= e^{-z} + e^{-z}
= e^{-(z-\log 2)^+},
\]

where the inequality comes from Proposition A.1. \( \square \)

In the case that \( X \) has the \( \text{Bernoulli}(p) \) distribution, we have

\[
I_X(x) = I(x : p) := q \log \frac{q}{p} + (1 - q) \log \frac{1 - q}{1 - p}.
\]  

(32)

**Lemma A.3** The first and second derivatives of the functions \( H(x) \) and \( I(x : p) \) are

\[
H'(x) = \log \frac{1-x}{x}, \quad H''(x) = \frac{1}{x(1-x)},
\]

(33)

\[
I'(x : p) = H'(p) - H'(x), \quad I''(x : p) = \frac{1}{x(1-x)}.
\]

(34)

In particular,

\[
H(q) - (H(p) + H'(p)(q - p)) = I(q : p),
\]

(35)

and

\[
\lim_{n \to \infty} n \left[ H \left( (1 - \frac{1}{n})y + \frac{1}{n}x \right) - \left( (1 - \frac{1}{n})H(y) + \frac{1}{n}H(x) \right) \right]
= (x - y)H'(y) - (H(x) - H(y))
= I(x : y).
\]

31
Proposition A.4 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. Denote \( X_{12} = X_1 + X_2 \) and \( \bar{X}_1 = X_1/m, \bar{X}_2 = X_1/(n-m), \bar{X}_{12} = X_{12}/n \). Then the following identities hold:

\[
\begin{align*}
& mI(\bar{X}_1 : p) + (n-m)I(\bar{X}_2 : p) - nI(\bar{X}_{12} : p) \quad \text{(36)} \\
& = X_n I\left( \frac{X_1}{X_{12}} : \frac{m}{n} \right) + (n-X_n)I\left( \frac{m-X_1}{n-X_{12}} : \frac{m}{n} \right) \quad \text{(37)} \\
& = mI(\bar{X}_1 : \bar{X}_{12}) + (n-m)I\left( \frac{X_{12}-X_1}{n-m} : \bar{X}_{12} \right). \quad \text{(38)}
\end{align*}
\]

The identities in Proposition A.4 are obtained by writing the full expression of (36) and re-arranging the log terms in two other ways. Formulae (37) and (38) are written without \( X_2 \), expressing the fact that any two of the three random variables \( X_1, X_2 \) and \( X_{12} \) contain same information as the full triple. Note that (37) and (38) 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 (36) and (38) can be interpreted so that the two positive terms of (36) measure exactly same amount of information about \( p \) as what is subtracted by the negative term. Moreover, (38) has the additional interpretation of presenting the rate function of the hypergeometric distribution:

Proposition A.5 Let \( X \) have the distribution \( \text{Hypergeometric}(n, m, z) \), i.e. the conditional distribution of \( X_1 \) of Proposition A.4 given that \( X_{12} = z \). The rate function of \( X \) is

\[
I_X(x) = mI\left( \frac{x}{m} : \frac{z}{n} \right) + (n-m)I\left( \frac{z-x}{n-m} : \frac{z}{n} \right). \quad \text{(39)}
\]

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

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

Write

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

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

Proposition A.6 Let \( k \geq 2 \) and let \( X_i, i \in \{1, \ldots, k\} \), be independent random variables with distributions \( \text{Bin}(n_i, p) \), respectively. Denote \( X_{1\ldots j} = \sum_{i=1}^j X_i, \bar{X}_i = X_i/n_i \) and \( \bar{X}_{1\ldots j} = X_{1\ldots j}/(\sum_{i=1}^j n_i) \). Then

\[
\sum_{i=1}^k n_i I(\bar{X}_i : p) - nI(\bar{X}_{1\ldots j} : p) \leq \sum_{i=1}^{k-1} (\log 2 + Y_i), \quad \text{(40)}
\]

(41)

where \( Y_1, \ldots, Y_{k-1} \) are independent \( \text{Exp}(1) \) random variables.
Proof For \( k = 2 \), the left hand side of (40) equals
\[
n_1 I(\bar{X}_1 : \bar{X}_{12}) + n_2 I\left(\frac{X_{12} - X_1}{n_2} : \bar{X}_{12}\right)
\]
by Proposition A.4. For any \( N \in \{0, \ldots, n\} \), consider the conditional distribution of (42), given that \( X_{12} = N \). By Proposition A.5, 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.2, 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
\[
\sum_{i=1}^{k} n_i I(\bar{X}_i : p) - n I(\bar{X}_{12} : p) = \sum_{i=1}^{k-1} n_i I(\bar{X}_i : p) - (n - n_k) I(\bar{X}_{1...(k-1)} : p) + n_k I(\bar{X}_k : p) + (n - n_k) I(\bar{X}_{1...(k-1)} : p) - n I(\bar{X}_{1...k} : p).
\]

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 \( X_{1...(k-1)} \). Similarly, the second row is stochastically bounded by \( \log 2 + Y_k \), where \( Y_k \sim \text{Exp}(1) \), irrespective of the value of \( X_{1...k} \). It remains to note that \( Y_k \) can be chosen to be independent of \( (Y_1, \ldots, Y_{k-2}) \), because \( X_k \) is independent of \( (X_1, \ldots, X_{k-1}) \), and of \( \bar{X}_{1...(k-1)} \) in particular.

Lemma A.7 Consider the sequence \((G_n, \xi_n)\) of stochastic block models as in Theorem 4.13. Then the following holds.

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

(ii) For any \( \epsilon > 0 \), the partition \( \xi_n \) is \( \epsilon \)-regular with high probability.

Proof Claim (i) By Proposition A.1 and (43),
\[
P\left( \max_{v \in A_i} \frac{e(\{v\}, A_j)}{|A_j|} > d_{ij} + h \right)
\leq \sum_{v \in A_i} P\left( \frac{e(\{v\}, A_j)}{|A_j|} > d_{ij} + h \right)
\leq |A_i| \exp\left( -|A_j| (d_{ij} + h) : d_{ij} \right)
\leq |A_i| \exp\left( -|A_j| \left( \frac{h^2}{2d_{ij}(1 - d_{ij})} + \frac{h^3}{6} I'''(z : d_{ij}) \right) \right).
\]

33
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 (ii). Fix \( \epsilon > 0 \) and consider any \( i, j \). Let \( U_1 \subseteq A_i \) and \( U_2 \subseteq A_j \) such that \(|U_1| \geq \epsilon |A_i|\) and \(|U_2| \geq \epsilon |A_j|\). By Proposition A.1

\[
P(|d(U_1, U_2) - d_{ij}| > \epsilon) \leq e^{-\alpha U_2} + e^{-\alpha U_1} + e^{-\alpha U_2} + e^{-\alpha U_1}.
\]

Let \( \iota(\epsilon) = \min \{ I(d_{ij} + \epsilon : d_{ij}) \} \). The union bound yields

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

because \( \iota(\epsilon) > 0 \).

\( \square \)

**Remark A.8** Because \( \iota(\epsilon) \propto \epsilon^2 \) as \( \epsilon \to 0 \), the proof of claim (ii) indicates that with a fixed \( \epsilon \), \( \epsilon \)-regularity starts to hold when \( n \gg \epsilon^{-4} \).

### Poissonian block model

Denote by \( H_P(a) \) the entropy of the distribution \( \text{Poisson}(a) \)

\[
H_P(a) = -\sum_{k=0}^{\infty} \frac{a^k}{k!} e^{-a} (k \log a - \log k! - a) \quad (43)
\]

\[
= a - a \log a + e^{-a} \sum_{k=0}^{\infty} \frac{a^k}{k!} \log k!, \quad (44)
\]

and by \( I_P(b : a) \) the Kullback-Leibler divergence between distributions \( \text{Poisson}(a) \) and \( \text{Poisson}(b) \)

\[
I_P(b : a) = a - b + b \log b - b \log a. \quad (45)
\]

We note the following facts:

**Lemma A.9** The entropy function \( H_P(a) \) is increasing and concave, and its two first derivatives have the expressions

\[
\frac{d}{da} H_P(a) = \mathbb{E} \log \frac{X + 1}{a}, \quad (46)
\]

\[
\frac{d^2}{da^2} H_P(a) = -\frac{1}{a} + \mathbb{E} \log \left( \frac{1}{X + 1} \right), \quad (47)
\]

where \( X \) denotes a random variable with distribution \( \text{Poisson}(a) \).

**Lemma A.10** For any \( a \in [0, 1] \) and \( x, y > 0 \), denote \( z = ax + (1-a)y \). Then we have

\[
H_P(z) - (aH_P(x) + (1-a)H_P(y)) \leq aI_P(x : z) + (1-a)I_P(y : z) \quad (48)
\]

\[
zI_{\text{Ber}(a)}\left( \frac{ax}{z} \right) \quad (49)
\]

\[
= I_{\text{Bin}(z, a)}(ax), \quad (50)
\]

where the form (50) is applicable when \( z \) is a positive integer.
Proof By (44) and (45),
\[
H_P(z) - (\alpha H_P(x) + (1 - \alpha) H_P(y)) = f(z) - (\alpha f(x) + (1 - \alpha) f(y)) \leq 0,
\]

because the function
\[
f(u) = e^{-u} \sum_{k=2}^{\infty} \frac{u^k}{k!} \log k!
\]
is convex:
\[
f''(u) = e^{-u} \left( \sum_{k=1}^{\infty} \frac{u^k}{k!} \log k + 2 \frac{k+1}{k} + \log 2 \right) > 0.
\]
The expression (50) is obtained by writing the right hand side of (48) with the substitution \(y = (z - \alpha x)/(1 - \alpha)\) and re-combining the log terms.

The Poissonian counterpart of Proposition A.6 is the following.

**Proposition A.11** Let \(a > 0\), \(k \geq 2\), \(n_i \geq 1\), \(i = 1, \ldots, k\), and \(n = \sum_i n_i\).

Let \(X_i, i \in \{1, \ldots, k\}\), be independent random variables with distributions Poisson\((n_i a)\), respectively. Denote \(X_{1\ldots j} = \sum_{i=1}^{j} X_i\), \(\bar{X}_i = X_i/n_i\) and \(\bar{X}_{1\ldots j} = X_{1\ldots j}/\sum_{i=1}^{j} n_i\). Then
\[
nH_P(\bar{X}_{12}) - \sum_{i=1}^{k} n_i H_P(\bar{X}_i) \leq \sum_{i=1}^{k-1} (\log 2 + Y_i),
\]
where \(Y_1, \ldots, Y_{k-1}\) are independent \(\text{Exp}(1)\) random variables.

Proof The proof of Proposition A.6 can be imitated, using the following observations:

- Using induction, it suffices to consider the case \(k = 2\).

- By (48), the left hand side of (10) can be upperbounded in terms of Kullback informations.

- The latter expression can be transformed to the form (50), i.e. the value of a the rate function of a binomial distribution taken at random variable \(X_1\).

- Now, the conditional distribution of \(X_1\) given \(X_{12}\) is the above binomial distribution. Thus, we can apply Lemma A.2 in a similar way as in the proof of Proposition A.6.

Acknowledgment. This work was partly supported by Academy of Finland project 294763 (Stomograph).
References

[1] N. Alon, R. Duke, H. Lefmann, V. Rödl, and R. Yuster. The algorithmic aspects of the regularity lemma. *Journal of Algorithms*, 16:80–109, 1994.

[2] N. Alon, E. Fischer, I. Newman, and A. Shapira. A combinatorial characterization of the testable graph properties: It’s all about regularity. In *STOC’06*, Seattle, U.S., May 2006.

[3] M. Bolla. *Spectral Clustering and Biclustering*. Wiley, 2013.

[4] T. Cover and J. Thomas. *Elements of Information Theory*. John Wiley and Sons, U.S.A., 1991.

[5] E. Fischer, A. Matsliah, and A. Shapira. Approximate hypergraph partitioning and applications. *Foundations of Computer Science*, pages 579–589, 2007.

[6] P. Grünwald. *Minimum Description Length Principle*. The MIT Press, 2007.

[7] S. Heimlicher, M. Lelarge, and L. Massoulié. Community detection in the labelled stochastic block model. In *NIPS Workshop on Algorithmic and Statistical Approaches for Large Social Networks*, 2012.

[8] R. v. d. Hofstad. *Random graphs and complex networks*. Cambridge University Press, 2016.

[9] P. Holland, K. Laskey, and S. Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, 1983.

[10] J. Komlós and M. Simonovits. Szemerédi’s regularity lemma and its applications in graph theory. In D. Miklós, V. Sós, and T. Szonyi, editors, *Combinatorics, Paul Erdős is Eighty*, pages 295–352. János Bolyai Mathematical Society, Budapest, 1996.

[11] L. Massoulié. Community detection thresholds and the weak Ramanujan property. In *Proceedings of the Forty-sixth Annual ACM Symposium on Theory of Computing*, STOC ’14, pages 694–703, New York, NY, USA, 2014. ACM.

[12] T. Nepusz, L. Négyessy, G. Tusnády, and F. Bazsó. Reconstructing cortical networks: case of directed graphs with high level of reciprocity. In B. Bollobás, R. Kozma, and D. Miklós, editors, *Handbook of Large-Scale Random Networks*, number 18 in Bolyai Society of Mathematical Studies, pages 325–368. Springer, 2008.

[13] I. Norros and H. Reittu. On a conditionally Poissonian graph process. *Adv. Appl. Prob.*, 38:59–75, 2006.

[14] I. Pappas and P. Pardalos. Identifying cognitive states using regularity partitions. *PLoS ONE*, 10(8): e0137012, 2015. https://doi.org/10.1371/journal.pone.0137012.
[15] V. Pehkonen and H. Reittu. Szemerédi-type clustering of peer-to-peer streaming system. In *Proc. Cnet 2011*, San Francisco, U.S.A., 2011.

[16] T. P. Peixoto. Entropy of stochastic blockmodel ensembles. *Physical Review E*, 85(056122), 2012.

[17] M. Pelillo, I. Elezi, and M. Fiorucci. Revealing structure in large graphs: Szemerédi’s regularity lemma and its use in pattern recognition. *Pattern Recognition Letters*, 87:4–11, 2017.

[18] H. Reittu, F. Bazsó, and R. Weiss. Regular decomposition of multivariate time series and other matrices. In P. Franti, G. Brown, M. Loog, F. Escolano, and M. Pelillo, editors, *Proc. S+SSPR 2014*, number 8621 in LNCS, pages 424–433. Springer-Verlag, 2014.

[19] J. Rissanen. A universal prior for integers and estimation by minimum description length. *The Annals of Statistics*, 11(2):416–431, 1983.

[20] J. Rissanen. *Stochastic Complexity in Statistical Inquiry*, volume 15. World scientific, 1998.

[21] G. Sárközy, F. Song, E. Szemerédi, and S. Trivedi. A practical regularity partitioning algorithm and its application in clustering. 2012. arXiv:1209.6540v1 [math.CO] 28 Sep 2012.

[22] A. Scott. Szemerédi’s regularity lemma for matrices and sparse graphs. *Combinatorics, Probability and Computing*, 20(3):455–466, 2011.

[23] A. Sperotto and M. Pelillo. Szemerédi’s regularity lemma and its applications to pairwise clustering and segmentation. In *Proc. EMMCVPR 2007*, Italy, 2007.

[24] E. Szemerédi. Regular partitions of graphs. In *Problèmes Combinatoires et Théorie des Graphes*, number 260 in Colloq. Intern. C.N.R.S., pages 399–401, Orsay, 1976.

[25] T. Tao. Szemerédi’s regularity lemma via random partitions, 2009. Blog entry, [https://terrytao.wordpress.com/2009/04/26/szemeredis-regularity-lemma-via-random-parti](https://terrytao.wordpress.com/2009/04/26/szemeredis-regularity-lemma-via-random-parti).