Universal Graph Compression: Stochastic Block Models

by

Ziao Wang

B.Eng., Nanyang Technological University, 2019

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF APPLIED SCIENCE

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Electrical and Computer Engineering)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

August 2021

The following individuals certify that they have read, and recommend to the Faculty of Graduate Studies for acceptance, the thesis entitled:

Universal Graph Compression: Stochastic Block Models

submitted by Ziao Wang in partial fulfillment of the requirements for

the degree of Master of Applied Science

in Electrical and Computer Engineering

Examining Committee:

Dr. Lele Wang, Assistant Professor, Electrical and Computer Engineering, UBC

Supervisor

Dr. Lutz Lampe, Professor, Electrical and Computer Engineering, UBC

Supervisory Committee Member

Dr. Christos Thrampoulidis, Assistant Professor, Electrical and Computer Engineering, UBC

Supervisory Committee Member

Abstract

Motivated by the prevalent data science applications of processing and mining large-scale graph data such as social networks, web graphs, and biological networks, as well as the high I/O and communication costs of storing and transmitting such data, this thesis investigates lossless compression of data appearing in the form of a labeled graph. In particular, we consider a widely used random graph model, stochastic block model (SBM), which captures the clustering effects in social networks. An information-theoretic *universal compression* framework is applied, in which one aims to design a *single* compressor that achieves the asymptotically optimal compression rate, for every SBM distribution, without knowing the parameters of the SBM that generates the data. Such a graph compressor is proposed in this thesis, which universally achieves the optimal compression rate for a wide class of SBMs with edge probabilities ranging from $O(1)$ to $\Omega(1/n^{2-\epsilon})$ for any $0 < \epsilon < 1$.

Existing universal compression techniques are developed mostly for *stationary ergodic* one-dimensional sequences with *fixed* alphabet size and entropy *linear* in the number of variables. However, the adjacency matrix of SBM has complex two-dimensional correlations and sublinear entropy in the sparse regime. These challenges are alleviated through a carefully designed transform that converts two-dimensional correlated data into *almost* i.i.d. submatrices. The sequence of submatrices is then compressed by a Krichevsky–Trofimov compressor, whose length analysis is generalized from i.i.d. sequences to identically distributed but arbitrarily correlated sequences. In four benchmark graph datasets (protein-to-protein interaction, LiveJournal friendship, Flickr, and YouTube), the compressed files from competing algorithms (including CSR, Ligra+, PNG image compressor, and Lempel–Ziv compressor for two-dimensional data) take 2.4 to 27 times the space needed by the proposed scheme.

Lay Summary

Nowadays, graphical data is gaining popularity for its wide applications in social science, biology and so on. Many of these graphical data possess two features. Firstly, the size of the graphs are quite large. In many cases, there are millions or even billions of vertices in the graph. Secondly, the graphs are quite sparse. In a friendship network, the number of people in the graph can be huge while every person has very limited social circle. How to efficiently compress graphical data to facilitate data storage and transmission remains a big challenge. This thesis aims to find an efficient algorithm to compress such graphical data.

Comparing to traditional one-dimensional sequential data, graphical data possesses structural information which leads to complicated two-dimensional correlation, so the traditional compression techniques for sequential data may not be efficient for graphical data. In this thesis, we explore the correlation in stochastic block model, which is a random graph model widely used in social network, and design a graph compression algorithm that achieves the optimal compression rate of this model.

Preface

This thesis is the result of the joint work carried out by myself, in collaboration with Alankrita Bhatt, Dr. Chi Wang and my supervisor Dr. Lele Wang. All the content are based on the preprint [7]. A shorter version of this work is published in the *2021 IEEE International Symposium on Information Theory*.

Alankrita Bhatt and I contributed equally to this work. Alankrita Bhatt and Dr. Lele Wang are pioneers in this project. They together proposed the problem and came up with the graph compression algorithm. Alankrita Bhatt proposed an initial version of the universality proof using the Laplace probability assignment. My contribution to this work includes (1) deriving the upper bound of the redundancy of Krichevsky–Trofimov probability assignment (which is the final version of the graph compressor presented in the work), (2) refining the universality proof to provide a second-order analysis, (3) designing and conducting the experiments on real-world data, and (4) analyzing the proposed algorithm under the local weak convergence framework. Dr. Chi Wang provided precious advice on simulation experiment design and practical aspects of compression algorithm design. Dr. Lele Wang was responsible for providing feedback and overseeing the project.

Table of Contents

List of Tables

List of Figures

Acknowledgements

First and foremost, I would like to thank my supervisor Dr. Lele Wang for her invaluable advice and continuous support during my Master study. Thank you for leading me to the world of information theory and giving me the opportunity to explore my academic interests. She is the most patient supervisor I could ever imagine. Her support was not limited to high-level research ideas, whenever I got stuck on research, she was always there trying to understand the problem with me and providing precious advice. Also, her guidance on academic writing skills will benefit my entire research career.

Secondly, I would like to thank my collaborators Alankrita Bhatt and Dr. Chi Wang for your hard work and advice on this project. This project couldn't have been done without anyone's dedication.

Next, I am grateful to Dr. Lutz Lampe and Dr. Christos Thrampoulidis for dedicating their time to serve on my defense committee. They provided invaluable feedback and comments on this thesis.

I would also like to thank Natural Sciences and Engineering Research Council (NSERC) for funding this research.

Finally, I would like to express my deepest appreciation to my parents. It is your unconditional support, morally and financially, that helped me through this hard time of worldwide pandemic.

Chapter 1

Introduction

In many data science applications, data appears in the form of large-scale graphs. For example, in social networks, vertices represent users and an edge between vertices represents friendship; in the World Wide Web, vertices are websites and edges indicate the hyperlinks from one site to the other; in biological systems, vertices can be proteins and edges illustrate protein-to-protein interaction. Such graphs may contain billions of vertices. In addition, edges tend to be correlated with each other since, for example, two people sharing many common friends are likely to be friends as well. How to efficiently compress such large-scale structural information to reduce the I/O and communication costs in storing and transmitting such data is a persisting challenge in the era of big data.

In this thesis, we take an information theoretic approach to study lossless compression of graphs with vertex labels. We assume the graph is generated by some random graph model and investigate lossless compression schemes that achieve the theoretical limit, i.e., the entropy of the graph, asymptotically as the number of vertices goes to infinity. When the underlying distribution/statistics of the random graph model is known, optimal lossless compression can be achieved by methods like Huffman coding. However, in most real-world applications, the exact distribution is usually hard to obtain and the data we are given is a single realization of this distribution. This motivates us to consider the framework of *universal compression*, in which we assume the underlying distribution belongs to a known family of distributions and require that the encoder and the decoder should not be a function of the underlying distribution. The goal of universal compression is to design a single compression scheme that universally achieves the optimal theoretical limit, for every distribution in the family, without knowing which distribution generates the data. For this thesis, we focus on the family of *stochastic block models*, which are widely used random graph models that capture the clustering effect in social networks. Our goal is to develop a universal graph compression scheme for a family of stochastic block models with as wide range of parameters as possible.

How to design computationally efficient universal compression scheme is

a fundamental question in information theory. In the past several decades, a large number of universal compressors were proposed for one-dimensional sequences with fixed alphabet size, whose entropy is linear in the number of variables. Prominent results include the Laplace and Krichevsky–Trofimov (KT) compressors for i.i.d. processes [54, 55], Lempel–Ziv compressor [57, 58] and Burrows–Wheeler transform[22] for stationary ergodic processes, and context tree weighting [53] for finite memory processes. Many of these have been adopted in standard data compression applications such as `compress`, gzip, GIF, TIFF, and bzip2. More details on universal sequence compression will be introduced in Chapter 2. Despite these exciting developments, existing universal compression techniques fall short of establishing optimality results for graph data due to the following challenges. Firstly, graph data generated from a stochastic block model has non-stationary two-dimensional correlation, so existing techniques do not immediately apply here. Secondly, in many practical applications, where the graph is sparse, the entropy of the graph may be sublinear in the number of entries in the adjacency matrix.

For the first challenge, a natural question arising is: can we convert the two-dimensional adjacency matrix of the graph into a one-dimensional sequence in some order and apply a universal compressor for the sequence? For some simple graph model such as Erdős–Rényi graph, where each edge is generated i.i.d. with probability p, this would indeed work. For more complex graph models including stochastic block models, it is unclear whether there is an ordering of the entries that results in a stationary process. We will show in Chapter 8 several orders including row-by-row, column-by-column, and diagonal-by-diagonal fail to produce a stationary process. We alleviate this challenge by designing a decomposition of the adjacency matrix into blocks. We then show in Proposition 4 that with a carefully chosen parameter, the block decomposition converts two-dimensional correlated entries into a sequence of *almost* i.i.d. blocks with slowly growing alphabet size. To address the second challenge, we adjust the standard definition of universality, which normalizes the compression length by the number of variables. The new definition of universality accommodates data with unknown leading order in its entropy expression.

1.1 Problem Setup

1.1.1 Universal Graph Compressor

For simplicity, we focus on simple (undirected, unweighted, no self-loop) graphs with labeled vertices in this thesis. But our compression scheme and

the corresponding analysis can be extended to more general graphs. Let $\mathcal{A}_n$ be the set of all labeled simple graphs on n vertices. Let $\{0,1\}^i$ be the set of binary sequences of length i, and set $\{0,1\}^* = \cup_{i=0}^{\infty}\{0,1\}^i$. A lossless graph compressor $C\colon \mathcal{A}_n \to \{0,1\}^*$ is a one-to-one function that maps a graph to a binary sequence. Let $\ell(C(A_n))$ denote the length of the output sequence. When A_n is generated from a distribution, it is known that the entropy $H(A_n)$ is a fundamental lower bound on the expected length of any lossless compressor [43, Theorem 8.3]

$$H(A_n) - \log(e(H(A_n)+1)) \le \mathsf{E}[\ell(C(A_n))], \tag{1.1}$$

where $\log(\cdot) \coloneqq \log_2(\cdot)$, therefore

$$\liminf_{n\to\infty} \frac{\mathsf{E}[\ell(C(A_n))]}{H(A_n)} \ge 1.$$

Thus, a graph compressor is said to be *universal* for the family of distributions $\mathscr{P}$ if for all distribution $\mathsf{P} \in \mathscr{P}$ and $A_n \sim \mathsf{P}$, we have

$$\limsup_{n\to\infty} \frac{\mathsf{E}[\ell(C(A_n))]}{H(A_n)} = 1. \tag{1.2}$$

1.1.2 Stochastic Block Model

A stochastic block model $\mathrm{SBM}(n, L, \mathbf{p}, \mathbf{W})$ defines a probability distribution over $\mathcal{A}_n$. Here n is the number of vertices, L is the number of communities. Each vertex $i \in [n] \coloneqq \{1,\dots,n\}$ is associated with a community assignment $X_i \in [L]$. The length-L column vector $\mathbf{p} = (p_1, p_2, \dots, p_L)^T$ is a probability distribution over $[L]$, where p_i indicates the probability that any vertex is assigned community i. $\mathbf{W}$ is an $L \times L$ symmetric matrix, where W_{ij} represents the probability of having an edge between a vertex with community assignment i and a vertex with community assignment j. We say $A_n \sim \mathrm{SBM}(n, L, \mathbf{p}, \mathbf{W})$ if the community assignments $X_1, X_2, \dots, X_n$ are generated i.i.d. according to $\mathbf{p}$ and for every pair $1 \le i < j \le n$, an edge is generated between vertex i and vertex j with probability W_{X_i,X_j}. In other words, in the adjacency matrix A_n of the graph, $A_{ij} \sim \mathrm{Bern}(W_{X_i,X_j})$ for $i < j$; the diagonal entries $A_{ii} = 0$ for all $i \in [n]$; and $A_{ij} = A_{ji}$ for $i > j$. We write $\mathbf{W} = f(n)\mathbf{Q}$, where $\mathbf{Q}$ is an $L \times L$ symmetric matrix with $\max_{i,j} Q_{ij} = \Theta(1)$[1]. We assume all entries in $\mathbf{p}$ are $\Theta(1)$ and $L = \Theta(1)$.

[1] This thesis follows the standard order notation: $f(n) = O(g(n))$ if $\lim_{n\to\infty} \frac{|f(n)|}{g(n)} < \infty$; $f(n) = \Omega(g(n))$ if $\lim_{n\to\infty} \frac{f(n)}{g(n)} > 0$; $f(n) = \Theta(g(n))$ if $f(n) = O(g(n))$ and $f(n) = \Omega(g(n))$; $f(n) = o(g(n))$ if $\lim_{n\to\infty} \frac{f(n)}{g(n)} = 0$; $f(n) = \omega(g(n))$ if $\lim_{n\to\infty} \frac{|f(n)|}{|g(n)|} = \infty$; and $f(n) \sim g(n)$ if $\lim_{n\to\infty} \frac{f(n)}{g(n)} = 1$.

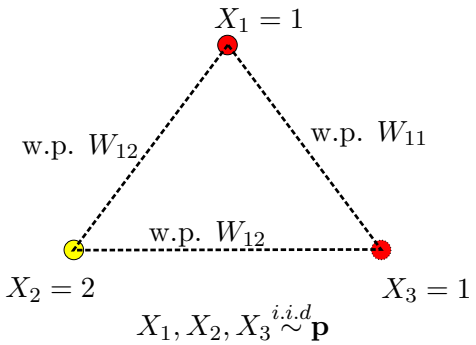

Figure 1.1: An illustration of edge generation process in stochastic block model. The community labels X_1, X_2 and X_3 are drawn i.i.d according to the community distribution $\mathbf{p}$. Vertices 1 and 3 got community label 1 (red) and vertex 2 got community label 2 (yellow) in this example. Conditioned on these community labels, the probability of generating an edge between two red vertices is W_{11} and the probability of generating an edge between a red vertex and a yellow vertex is W_{12}. The edge generation is independent of each other conditioning on the community labels.

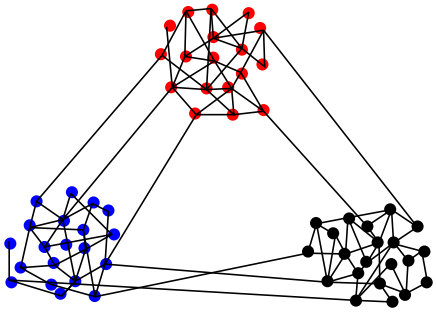

Figure 1.2: An example graph generated from stochastic block model. For illustration purpose, we use three different colors to indicator three different communities. In the actual graph generated from SBM, the community label is *hidden*. In the model generating the example graph, the probability of generating an edge within the community is set much higher than the probability of generating an edge between two different communities. The clustering effect of social network can be well characterized.

We will consider two families of stochastic block models: For $0 < \epsilon < 1$,

$$\mathscr{P}_1(\epsilon)\colon \mathrm{SBM}(n, L, \mathbf{p}, \mathbf{W}) \text{ with } f(n) = O(1), f(n) = \Omega\left(\frac{1}{n^{2-\epsilon}}\right), \tag{1.3}$$

$$\mathscr{P}_2(\epsilon)\colon \mathrm{SBM}(n, L, \mathbf{p}, \mathbf{W}) \text{ with } f(n) = o(1), f(n) = \Omega\left(\frac{1}{n^{2-\epsilon}}\right). \tag{1.4}$$

Note that the edge probability $\frac{1}{n^2}$ is the threshold for a random graph to contain an edge with high probability [24]. Thus, the family $\mathscr{P}_1(\epsilon)$ covers most non-trivial SBM graphs. Clearly, $\mathscr{P}_2(\epsilon)$ is a strict subset of $\mathscr{P}_1(\epsilon)$, as it does not contain the constant regime $f(n) = 1$.

1.1.3 Minimax Redundancy

Besides universality, which only concerns the first order terms in the expected length of the compressor and the entropy of the graph, the redundancy of a universal compressor is another important metric we wish to optimize in the design of universal graph compressors. Define the redundancy of a lossless compressor C for $A_n \sim \mathsf{P}$ as

$$R(C, A_n) = \mathsf{E}[\ell(C(A_n))] - H(A_n). \tag{1.5}$$

Then the minimax redundancy over a distribution family $\mathscr{P}$ is defined as

$$R_n^*(\mathscr{P}) = \inf_C \sup_{\mathsf{P}\in\mathscr{P}} R(C, A_n). \tag{1.6}$$

The minimax redundancy represents the lowest achievable redundancy of any lossless compressor for the family $\mathscr{P}$. Define the family of SBM in regime $f(n)$ with maximum entry $Q_{\max}$ in $\mathbf{Q}$ as

$$\mathscr{P}_3(f(n), Q_{\max}) : \mathrm{SBM}(n, L, \mathbf{p}, f(n)\mathbf{Q}) \text{ with } Q_{i,j} \le Q_{\max} \text{ for all } i, j \in [n]. \tag{1.7}$$

Notice that the $Q_{\max}$ and $f(n)$ are also parameters for SBMs, but we fix them here instead of taking supremum over them. That is because both the expected length of the compressor and the entropy scale with them. In this work, we are interested in analysing the minimax redundancy for the family $\mathscr{P}_3(f(n), Q_{\max})$ for each $f(n) = o(1)$ and $f(n) = \Omega(1/n^{2-\epsilon})$ for some $0 < \epsilon < 1$.

1.2 Literature Review

The literature on graph compression is vast. Existing compression schemes follow various different methodologies. Several methods exploited combinatorial properties such as cliques and cuts in the graph [31, 45]. Many works targeted at domain-specific graphs such as web graphs [9], biology networks [15, 27], and social network graphs [13]. Various representations of graphs were proposed, such as the text-based method, where the neighbour list of each vertex is treated as a "word" [38, 46], and the k^2-tree method, where the adjacency matrix is recursively partitioned into k^2 equal-size submatrices [12]. *Succinct* graph representations that enable certain types of fast computation, such as adjacency query or vertex degree query, were also widely studied [23]. While most compression schemes are for labeled graphs (graphs with vertex labels), there are also works considering lossless compression of unlabeled graphs [14, 37, 52], graphs with marks on its edges and vertices [19–21], or (correlated) data on the graph [1, 5]. We refer the readers to [6] for an exhaustive survey on lossless graph compression and space-efficient graph representations.

Lossless compression for graphs in an information-theoretic framework has been studied in [1, 5, 14, 18, 21, 25, 28, 30, 32, 39–42, 56]. In [14], universal compression of *unlabeled* graphs (isomorphism classes) generated from Erdős–Rényi models is investigated. In [1, 5], lossless compression of labeled graphs generated from SBMs is studied, but it does not consider universal compression schemes. In [30], a universal compression algorithm is proposed for two-dimensional arrays of data by first converting the two-dimensional data into a one-dimensional sequence using a Hilbert–Peano curve and then applying the Lempel–Ziv algorithm [58] for sequences. In [25, 28, 32, 56], universal compression on binary trees were studied. In [39–42], a minimum description length based method was proposed to infer network structures in stochastic block models. Recently, universal compression of graphs with marked edges and vertices is studied by Delgosha and Anantharam [18, 21]. They focus on the *sparse* graph regime, where the number of edges is in the same order as the number of vertices n. They employ the framework of local weak convergence, which provides a technique to view a sequence of graphs as a sequence of distributions on neighbourhood structures. Built on this framework, they propose an algorithm that compresses graphs by describing the local neighbourhood structures. Moreover, they introduce a universality/optimality criterion through a notion of entropy for graph sequences under the local weak convergence framework, known as the BC *entropy* [11]. This universality criterion is stronger than the one used in this

thesis. It requires the asymptotic length of the compressor to match the constants in both first and second order terms in Shannon entropy, whereas the universality criterion we use only requires to match the first order term. As a consequence of the stronger criterion, the compressor in [18] is universal over a smaller random graph family. In comparison, we expand the range of edge numbers from $\Theta(n)$ in the sparse regime to $\Theta(n^\alpha)$ for every $0 < \alpha \leq 2$ and propose a single universal compressor for the whole family under the weaker universality criterion. In Chapter 7, we evaluate the proposed compressor under the criterion in [18] for the family of SBMs. The proposed compressor achieves a similar performance in terms of BC entropy in the sparse regime.

1.3 Thesis Organization

In Section 1.1, we have defined universality over a family of graph distribution, the stochastic block models and the minimax redundancy for a family of graph distribution. The rest of the thesis is organized as follows. In Chapter 2, we introduce some preliminary on traditional universal sequence compression. We present our main result in Chapter 3, which is a graph compressor that is universal for a family containing most non-trivial stochastic block models and a bound for the minimax redundancy for this family of stochastic block models. We describe the proposed graph compressor in Section 3.1 and introduce the main theoretical results about our compressor in Section 3.2. In Section 3.3, we implement our compressor in four benchmark graph datasets and compare its empirical performance to four competing algorithms. We illustrate key steps in establishing universality in Chapter 4 and elaborate the proof of each step in Chapter 5. The result on minimax redundancy is proven in Chapter 6. In Chapter 7, we provide the second order analysis of the expected length of our compressor and compare it to the one in [18]. In Chapter 8, we explain why existing universal compressors developed for stationary processes may not be immediately applicable for some one-dimensional ordering of entries in the adjacency matrix. We also introduce some open problems and potential future works.

Chapter 2

Preliminary

To better explain our universal graph compressor, it is helpful to briefly go over traditional universal sequence compression. This chapter provides an overview for universal sequence compression. We will highlight Laplace compressor and Krichevsky–Trofimov compressor which are universal for i.i.d. sequences. These two compressors will later be used in our graph compressor.

2.1 Sequence Compression When Distribution is Known

Let $\mathcal{X}$ denote a finite alphabet set and let $\mathscr{P}$ be a family of probability distributions on $\mathcal{X}^\infty$. Each $P \in \mathscr{P}$ specifies a random process $(X_n)_{n=1}^\infty$ with n-th order pmf $p(x^n)$. Suppose we have a random process $(X_n)_{n=1}^\infty \sim P$ following some underlying distribution $P \in \mathscr{P}$. A compressor C_n is defined as

$$C_n : \mathcal{X}^n \to \{0,1\}^*,$$

where, recall, $\{0,1\}^* = \cup_{i=0}^\infty \{0,1\}^i$. A compressor C_n is said to be *lossless* if the mapping is one-to-one, i.e, $C_n(x^n) \neq C_n(\tilde{x}^n)$ for all $x^n \neq \tilde{x}^n$. Moreover, C_n is said to be *uniquely decodable* if any concatenations of C_n is lossless. Let $\ell(C_n(x^n))$ denote the output length for some input sequences x^n. It is well-known that the Shannon entropy $H(X^n) \triangleq \sum_{x^n} -p(x^n) \log p(x^n)$ is the information theoretical limit for lossless compression, i.e, $\mathsf{E}[\ell(C_n(X^n))] \geq H(X^n)$ for any lossless compressor C_n. In the case that the underlying probability distribution P is known to us, we can design the compressor C_n such that we assign a code of length $\lceil \log 1/p(x^n) \rceil$ to each input sequence x^n using, say, Shannon code [48]. Compressor C_n is a well-defined uniquely decodable compressor and its expected length can be bounded as

$$H(X^n) \leq \mathsf{E}[\ell(C_n(X^n))] \leq H(X^n) + 1.$$

If we look at the per-symbol length of the compressor C_n, we have

$$\lim_{n\to\infty} \mathsf{E}\left[\frac{\ell(C_n(X^n))}{n}\right] = \lim_{n\to\infty} \frac{H(X^n)}{n} \triangleq \overline{H}((X_n)_{n=1}^{\infty}),$$

where $\overline{H}((X_n)_{n=1}^{\infty})$ denotes the entropy rate of the process. We can see that C_n performs quite well in the sense that its per-symbol length is converging to the entropy rate of the random process.

However, in most real life applications of data compression, only the data is seen to us and we have very limited knowledge about the underlying probability distribution. Therefore, people are motivated to consider universal compression.

2.2 Sequence Compression When Distribution is Unknown

In this section, we consider the compression problem in the case that the underlying distribution is unknown and we define the *universality* of a compressor.

Continuing current setup, we consider the random process $(X_n)_{n=1}^{\infty} \sim P$ for some underlying distribution $P \in \mathscr{P}$. The specific distribution P is *unknown* to us in this case. Again, we consider the per-symbol length of compressor C_n and say a compressor C_n is *universal* over the family $\mathscr{P}$ if

$$\lim_{n\to\infty} \frac{\mathsf{E}[\ell(C_n(X^n))]}{n} = \overline{H}((X_n)_{n=1}^{\infty}), \quad \forall P \in \mathscr{P},$$

i.e, the per-symbol length of the compressor converges to the entropy rate of the random process regardless of the specific distribution P in the family $\mathscr{P}$. In the next section, we will introduce one way of constructing a universal compressor.

2.3 Universality Compression via Probability Assignment

A common approach to design universal compressor is to first assign a probability distribution to all possible realizations of data and then apply adaptive arithmetic coding. We explain this in the sequence compression setting here. This will be the basis for the design of universal graph compressor later.

We construct our compressor C_n as follows. First, for each sequence $x^n \in \mathcal{X}^n$, we assign a probability $q(x^n)$ such that $\sum_{x^n} q(x^n) = 1$. Then, based on the probability we assigned to each sequence, we assign a code with length $\lceil \log 1/q(x^n) \rceil + 1$ to each sequence x^n using, say, adaptive arithmetic coding [33]. By Theorem 5.4.3 in [17], we can bound the expected length of this compressor as

$$H(X^n) + D(p(x^n)||q(x^n)) \leq \mathsf{E}[\ell(C_n(X^n))] \leq H(X^n) + D(p(x^n)||q(x^n)) + 2,$$

where $D(p(x^n)||q(x^n)) \triangleq \sum_{x^n} p(x^n) \log \frac{p(x^n)}{q(x^n)}$ is the Kullback–Leibler divergence between $p(x^n)$ and $q(x^n)$. Again, if we consider the per-symbol length, we have

$$\frac{H(X^n) + D(p(x^n)||q(x^n))}{n} \leq \frac{\mathsf{E}[\ell(C_n(X^n))]}{n} \leq \frac{H(X^n) + D(p(x^n)||q(x^n)) + 2}{n}.$$

Suppose our q satisfies that $D(p(x^n)||q(x^n)) = o(n)$ for any $P \in \mathscr{P}$, then our compressor C_n will be universal over the family $\mathscr{P}$. From above, we can see the problem of universal compression has been translated to the problem of finding the probability assignment that can always simulate the true underlying distribution. Formally, we say a sequence of pmfs $(q(x^n))_{n=1}^{\infty}$ is *mean universal* with respect to the family $\mathscr{P}$ if

$$\lim_{n\to\infty} \frac{1}{n} D(p(x^n)||q(x^n)) = 0, \quad \forall P \in \mathscr{P}.$$

In the following two sub-sections, we will introduce two probability assignments that are mean universal to the family of i.i.d. processes. These two probability assignments play important role in our graph compressor.

2.3.1 Laplace Probability Assignment

Let $\mathscr{P}$ be the family of i.i.d. processes, i.e, the symbols $X_1, X_2, \ldots$ are identically and independently distributed. Let $|\mathcal{X}| = m$. Notice that by chain rule, the probability assignment for a sequence can be written as the product of n conditional probabilities, i.e,

$$q(x^n) = q(x_1)q(x_2|x_1) \cdots q(x_n|x_1 \ldots x_{n-1}).$$

Therefore, we can construct the probability assignment $q(x^n)$ in a sequential manner. Given a m-ary sequence $x_1, \ldots, x_n$, *Laplace sequential probability*

assignment defines n conditional probability distributions over $[m]$ as follows. For $j = 0, 1, 2, \ldots, n-1$, we assign conditional probability

$$q_{\mathrm{L}}(X_{j+1} = i | X^j = x^j) := \frac{N_i(x^j) + 1}{j + m} \quad \text{for each } i \in [m], \tag{2.1}$$

where $X^j := (X_1, \ldots, X_j), x^j := (x_1, x_2, \ldots, x_j)$, and $N_i(x^j) := \sum_{k=1}^{j} \mathbb{1}\{x_k = i\}$ counts the number of symbol i in x^j. Intuitively, the conditional probability assigned to a symbol is proportional to the number of that symbol observed so far plus one. The universality of Laplace probability assignment is implied in the following Proposition.

Proposition 1. *Suppose $\mathscr{P}$ is the family of identical and independent distributions, i.e, $X_1, X_2, \ldots X_n$ are identically and independently distributed. Let $q_{\mathrm{L}}(\cdot)$ be the marginal distribution induced by Laplace sequential probability assignment defined in* (2.1)

$$q_{\mathrm{L}}(x^n) := \frac{N_1! N_2! \cdots N_m!}{n!} \cdot \frac{1}{\binom{n+m-1}{m-1}},$$

where $N_i := N_i(x^n)$ for all $i \in [m]$. We then have

$$D(p(x^n) || q_{\mathrm{L}}(x^n)) \leq (1 + o(1))(m-1) \log n, \quad P \in \mathscr{P}.$$

Proof. For some distribution $P \in \mathscr{P}$, we assume $\mathsf{P}(X_1 = i) = \theta_i$ for any $i \in [m]$. Since $X_1, \ldots, X_n$ are i.i.d, for any realization $x_1, \ldots, x_n$, we have $p(x^n) = \theta_1^{N_1} \cdots \theta_m^{N_m}$. Therefore, we can upper bound the KL divergence between p and q_{L}

$$\begin{aligned}
&D(p(x^n) || q_{\mathrm{L}}(x^n)) \\
&= \sum_{x^n} p(x^n) \log \frac{p(x^n)}{q_{\mathrm{L}}(x^n)} \\
&= \sum_{x^n} \theta_1^{N_1} \cdots \theta_m^{N_m} \log \frac{\theta_1^{N_1} \cdots \theta_m^{N_m} n! \binom{n+m-1}{m-1}}{N_1! N_2! \cdots N_m!} \\
&= \sum_{x^n} \theta_1^{N_1} \cdots \theta_m^{N_m} \left(\log \frac{\theta_1^{N_1} \cdots \theta_m^{N_m} n!}{N_1! N_2! \cdots N_m!} + \log \binom{n+m-1}{m-1} \right) \\
&\overset{(a)}{\leq} \sum_{x^n} \theta_1^{N_1} \cdots \theta_m^{N_m} \log \binom{n+m-1}{m-1} \\
&= \log \binom{n+m-1}{m-1}
\end{aligned}$$

$$
\begin{aligned}
&\leq \log(n+m-1)^{m-1} \\
&= (1+o(1))(m-1)\log n,
\end{aligned}
$$

where (a) follows since $\frac{\theta_1^{N_1}\cdots\theta_m^{N_m} n!}{N_1!N_2!\cdots N_m!}$ is a multinomial probability which is upper bounded by 1. The above bound holds for any distribution $P \in \mathscr{P}$; therefore the proof is completed. □

2.3.2 Krichevsky–Trofimov Probability Assignment

Similar to Laplace probability assignment, Krichevsky–Trofimov(KT) probability assignment is also defined in a sequential manner. Given an m-ary sequence $x_1, \ldots, x_n$, *KT sequential probability assignment* defines n conditional probability distributions over $[m]$ as follows. For $j = 0, 1, 2, \ldots, n-1$, assign conditional probability

$$
q_{\mathrm{KT}}(i|x^j) \coloneqq q_{\mathrm{KT}}(X_{j+1} = i | X^j = x^j) = \frac{N_i(x^j) + 1/2}{j + m/2} \quad \text{for each } i \in [m], \tag{2.2}
$$

where $X^j \coloneqq (X_1, \ldots, X_j)$, $x^j \coloneqq (x_1, x_2, \ldots, x_j)$, and $N_i(x^j) \coloneqq \sum_{k=1}^{j} \mathbb{1}\{x_k = i\}$ counts the number of symbol i in x^j. KT probability also assigns conditional probability to each symbols based on the number of that symbol observed so far. However, the difference is that the conditional probability is proportional to the number of observed symbol plus half. The universality of KT probability assignment is implied in the following Proposition.

Proposition 2. *Suppose $\mathscr{P}$ is the family of identical and independent distributions, i.e, $X_1, X_2, \ldots, X_n$ are identically and independently distributed. Let $q_{\mathrm{KT}}(\cdot)$ be the marginal distribution induced by KT sequential probability assignment defined in* (2.2)

$$
q_{\mathrm{KT}}(x^n) = \frac{(2N_1 - 1)!!(2N_2 - 1)!! \cdots (2N_m - 1)!!}{m(m+2)\cdots(m+2n-2)},
$$

where $(-1)!! \coloneqq 1$. We then have

$$
D(p(x^n)||q_{\mathrm{KT}}(x^n)) = (1+o(1))\frac{m-1}{2}\log n.
$$

Proof. See for example [55]. □

From above, we can see that both probability assignments are universal with respect to the family of i.i.d. processes. Moreover, KT probability

assignment has a lower redundancy than Laplace probability assignment. Both probability assignments are used in our graph compressor. Moreover, we extend the result of Proposition 1 and Proposition 2 in Proposition 5 and Proposition 6 respectively.

Chapter 3

Main Results

We present our compression scheme in Section 3.1 and state its performance guarantee in Theorems 1 and 2. The key idea in our design is a decomposition of the adjacency matrix into blocks, which, with a carefully chosen parameter, converts the two-dimensional correlated entries in the adjacency matrix into a sequence of *almost* i.i.d. blocks. We then compress the blocks using a Krichevsky–Trofimov or Laplace compressor and generalize their length analyses from i.i.d. processes to *arbitrarily correlated* but identically distributed processes.

3.1 Algorithm: Universal Graph Compressor

In this section, we describe our universal graph compression scheme. For each integer $k \leq n$, the graph compressor $C_k \colon \mathcal{A}_n \to \{0,1\}^*$ is defined as follows.

- **Block decomposition.** Let $n' = \lfloor n/k \rfloor$ and $\tilde{n} = \lfloor n/k \rfloor k$. For $1 \leq i, j \leq n'$, let $\mathbf{B}_{ij}$ be the submatrix of A_n formed by the rows $(i-1)k+1, (i-1)k+2, \ldots, ik$ and the columns $(j-1)k+1, (j-1)k+2, \ldots, jk$. For example, we have

$$\mathbf{B}_{12} = \begin{bmatrix} A_{1,k+1} & A_{1,k+2} & \cdots & A_{1,2k} \\ A_{2,k+1} & A_{2,k+2} & \cdots & A_{2,2k} \\ \vdots & \vdots & \ddots & \vdots \\ A_{k,k+1} & A_{k,k+2} & \cdots & A_{k,2k} \end{bmatrix}. \tag{3.1}$$

  We then write the top-left $\tilde{n} \times \tilde{n}$ submatrix of A_n in the block-matrix form as

$$\begin{bmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} & \cdots & \mathbf{B}_{1,n'} \\ \mathbf{B}_{21} & \mathbf{B}_{22} & \cdots & \mathbf{B}_{2,n'} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{B}_{n',1} & \mathbf{B}_{n',2} & \cdots & \mathbf{B}_{n',n'} \end{bmatrix}. \tag{3.2}$$

Denote

$$\mathbf{B}_{\text{ut}} \coloneqq \mathbf{B}_{12}, \mathbf{B}_{13}, \mathbf{B}_{23}, \mathbf{B}_{14}, \mathbf{B}_{24}, \mathbf{B}_{34}, \ldots, \mathbf{B}_{1,n'}, \cdots, \mathbf{B}_{n'-1,n'} \tag{3.3}$$

as the sequence of off-diagonal blocks in the upper triangle and

$$\mathbf{B}_{\text{d}} \coloneqq \mathbf{B}_{11}, \mathbf{B}_{22}, \ldots, \mathbf{B}_{n',n'} \tag{3.4}$$

as the sequence of diagonal blocks.

- **Binary to m-ary conversion.** Let $m := 2^{k^2}$. Each $k \times k$ block with binary entries in the two block sequences $\mathbf{B}_{\text{ut}}$ and $\mathbf{B}_{\text{d}}$ is converted into a symbol in $[m]$.

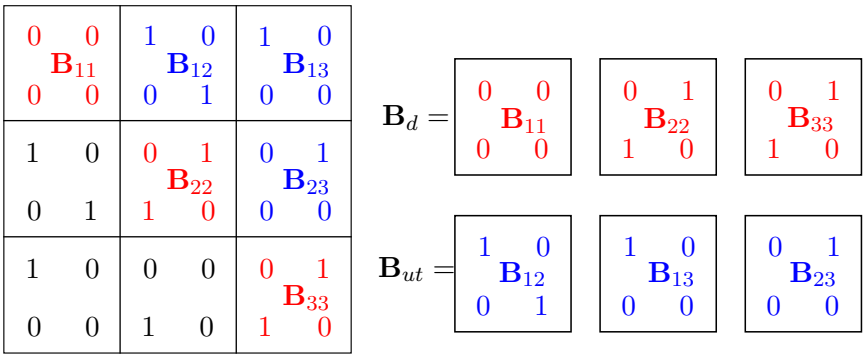

Figure 3.1: An illustration for the block decomposition of adjacency matrix. In this example, the block size k is set to 2. The adjacency matrix is cut into 2×2 sub-matrices. Each sub-matrix is viewed as a symbol from an alphabet set with size $2^{k^2} = 16$. The diagonal blocks $\mathbf{B}_{11}$, $\mathbf{B}_{22}$ and $\mathbf{B}_{33}$ are rearranged into sequence $\mathbf{B}_d$. The off-diagonal blocks $\mathbf{B}_{12}$, $\mathbf{B}_{13}$ and $\mathbf{B}_{23}$ are rearranged into sequence $\mathbf{B}_{ut}$.

- **KT probability assignment.** Apply KT sequential probability assignment defined in (2.2) for the two m-ary sequences $\mathbf{B}_{\text{ut}}$ and $\mathbf{B}_{\text{d}}$ respectively.

- **Adaptive arithmetic coding.** With the KT sequential probability assignments, compress the two sequences $\mathbf{B}_{\text{ut}}$ and $\mathbf{B}_{\text{d}}$ separately using adaptive arithmetic coding [33] (see description in Algorithm 1). In case $k = 1$, the diagonal sequence $\mathbf{B}_{\text{d}}$ becomes an all-zero sequence since we assume the graph is simple. So we will only compress the off-diagonal sequence $\mathbf{B}_{\text{ut}}$.

Algorithm 1: m-ary adaptive arithmetic encoding with KT probability assignment

Input : Data sequence x^N, alphabet size m

Initialize
$\texttt{lower} = 0, \texttt{upper} = 1, \texttt{logprob} = 0, N_1 = N_2 = \cdots = N_m = 0;$
for $j = 0, 1, \ldots, N-1$ **do**
  $\texttt{range} \leftarrow \texttt{upper} - \texttt{lower};$
  for $i = 1, 2, \ldots, x_{j+1}$ **do**
    Compute $q_{\mathrm{KT}}(i|x^j) = \frac{N_i + 1/2}{j+m/2}$;
  $\texttt{upper} \leftarrow \texttt{lower} + \texttt{range} \cdot \sum_{i=1}^{x_{j+1}} q_{\mathrm{KT}}(i|x^j);$
  $\texttt{lower} \leftarrow \texttt{upper} - \texttt{range} \cdot q_{\mathrm{KT}}(x_{j+1}|x^j);$
  $N_{x_{j+1}} \leftarrow N_{x_{j+1}} + 1;$
  $\texttt{logprob} \leftarrow \texttt{logprob} + \log(q_{\mathrm{KT}}(x_{j+1}|x^j));$

Output: the binary representation of $\frac{1}{2}(\texttt{lower} + \texttt{upper})$ with $\lceil -\texttt{logprob} \rceil + 1$ bits

- **Encoding the remaining bits.** The above process compressed the top-left $\tilde{n} \times \tilde{n}$ block of the adjacency matrix. For the remaining $(n - \tilde{n})\tilde{n} + \binom{n-\tilde{n}}{2}$ entries in the upper diagonal of the adjacency matrix, we simply use $2\lceil \log n \rceil$ bits to encode the row and column number of each ones.

Given the compressed graph sequence y^L, the number of vertices n and the block size k, the graph decompressor $D_k \colon \{0,1\}^* \to \mathcal{A}_n$ is defined as follows.

- **Adaptive arithmetic decoding.** With the KT sequential probability assignments defined in (2.2), decompress the two code sequences for $\mathbf{B}_{\mathrm{ut}}$ and $\mathbf{B}_{\mathrm{d}}$ separately using adaptive arithmetic decoding (see Algorithm 2). The length of data sequence $\mathbf{B}_{\mathrm{ut}}$ and $\mathbf{B}_{\mathrm{d}}$ are $\frac{n}{k}(\frac{n}{k} - 1)/2$ and $\frac{n}{k}$ respectively.

- **m-ary to binary conversion.** Each m-ary symbol in the sequence is converted to a k^2-bit binary number and further converted into a $k \times k$ block with binary entries.

- **Adjacency matrix recovery.** With the blocks in $\mathbf{B}_{\mathrm{ut}}$ and $\mathbf{B}_{\mathrm{d}}$, recover the top-left $\tilde{n} \times \tilde{n}$ submatrix of A_n in the order described in (3.2), (3.3), and (3.4).

Algorithm 2: m-ary adaptive arithmetic decoding with KT probability assignment

Input : Binary sequence y^L, alphabet size $m = 2^{k^2}$, length of data sequence N

Add '0.' before sequence y^L and convert it into a decimal real number Y. Initialize
$\texttt{lower} = 0, \texttt{upper} = 1, N_1 = N_2 = \cdots = N_m = 0;$
for $j = 0, 1, \ldots, N-1$ **do**
  $\texttt{range} \leftarrow \texttt{upper} - \texttt{lower};$
  for $i = 1, 2, \ldots, m$ **do**
    Compute $q_{\mathrm{KT}}(i|x^j) = \frac{N_i + 1/2}{j + m/2}$;
  Find minimum $z \in [m]$ such that
  $\texttt{lower} + \texttt{range} \cdot \sum_{i=1}^{z} q_{\mathrm{KT}}(i|x^j) > Y;$
  $\texttt{upper} \leftarrow \texttt{lower} + \texttt{range} \cdot \sum_{i=1}^{z} q_{\mathrm{KT}}(i|x^j);$
  $\texttt{lower} \leftarrow \texttt{upper} - \texttt{range} \cdot q_{\mathrm{KT}}(z|x^j);$
  $N_z \leftarrow N_z + 1;$
  $x_{j+1} \leftarrow z;$

Output: the m-ary data sequence $x_1, x_2, \cdots, x_N$

- **Decoding the remaining bits.** Recover the remaining $(n - \tilde{n})\tilde{n} + \binom{n-\tilde{n}}{2}$ entries in the A_n using the row and column numbers of the ones.

One can check that C_k is well-defined. The block decomposition and the binary to m-ary conversion are clearly one-to-one. It is also known that for any valid probability assignment, arithmetic coding produces a prefix code, which as also one-to-one.

The computational complexity of the proposed algorithm is $O(2^{k^2} n^2)$. For the choice of k that achieves universality over $\mathscr{P}_1(\epsilon)$ family in Theorem 1, $O(2^{k^2} n^2) = O(n^{2+\delta})$ for $\delta < \epsilon$. For the choice of k that achieves universality over $\mathscr{P}_2(\epsilon)$ family in Theorem 2, $O(2^{k^2} n^2) = O(n^2)$.

The orders in $\mathbf{B}_{\mathrm{ut}}$ and $\mathbf{B}_{\mathrm{d}}$ do not matter in terms of establishing universality. The current orders in (3.3) and (3.4) together with arithmetic coding enable a *horizon free* implementation. That is, the encoder does not need to know the *horizon* n to start processing the data and can output partial coded bits *on the fly* before receiving all the data. This leads to short encoding and decoding delay. For some real-world applications, for example, when the number of users increases in a large social network, this compressor has the advantage of not requiring to re-process existing data

and re-compress the whole graph from scratch.

Remark 1 (Laplace probability assignment). As an alternative to the KT sequential probability assignment, one can also use the Laplace sequential probability assignment as defined in (2.1).

Both methods can be shown to be universal, while Laplace probability assignment has a much cleaner derivation. However, KT probability assignment produces a better empirical performance. For this reason, we keep both in the thesis.

3.2 Theoretical Performance

We now state the main result of this thesis: the proposed compressor C_k, for a carefully chosen k, is universal over the classes $\mathscr{P}_1(\epsilon)$ and $\mathscr{P}_2(\epsilon)$ respectively for every $0 < \epsilon < 1$.

Theorem 1. *For every $0 < \epsilon < 1$, let $A_n \sim \mathrm{SBM}(n, L, \mathbf{p}, f(n)\mathbf{Q}) \in \mathscr{P}_1(\epsilon)$. Let $k = \omega(1)$ and $k \leq \sqrt{\delta \log n}$ for some $0 < \delta < \epsilon$.*

- *If $f(n) = o(1)$ and $f(n) = \Omega\left(\frac{1}{n^{2-\epsilon}}\right)$, then the expected length of compressor C_k defined in Section 3.1 is upper bounded as*

$$\begin{aligned}
&\mathsf{E}[\ell(C_k(A_n))] \\
&\leq H(A_n) + \binom{n}{2} f(n) \left(\mathbf{p}^T \mathbf{Q} \mathbf{p} \log \left(\frac{1}{\mathbf{p}^T \mathbf{Q} \mathbf{p}} \right) - \mathbf{p}^T \mathbf{Q}^* \mathbf{p} \right) + o(n^2 f(n)) \\
&= H(A_n) + o(H(A_n)),
\end{aligned} \tag{3.5}$$

  where $\mathbf{Q}^*$ denotes an $L \times L$ matrix whose (i, j) entry is $Q_{ij} \log(\frac{1}{Q_{ij}})$ when $Q_{ij} \neq 0$ and 0 when $Q_{ij} = 0$.

- *If $f(n) = \Theta(1)$, then the expected length of compressor C_k defined in Section 3.1 is upper bounded as*

$$\begin{aligned}
\mathsf{E}(\ell(C_k(A_n))) &\leq H(A_n) + H(\mathbf{p}) \frac{n^2}{k} + o\left(\frac{n^2}{k}\right) \\
&= H(A_n) + o(H(A_n)),
\end{aligned} \tag{3.6}$$

  where $H(\mathbf{p}) = \sum_{i=1}^{L} p_i \log\left(\frac{1}{p_i}\right)$.

Theorem 1 implies the existence of a universal graph compressor for the family $\mathscr{P}_1(\epsilon)$.

Corollary 1 (Universality over $\mathscr{P}_1$)**.** *For every $0 < \epsilon < 1$, the graph compressor C_k is universal over the family $\mathscr{P}_1(\epsilon)$ provided that*

$$0 < \delta < \epsilon, \quad k \leq \sqrt{\delta \log n}, \quad \text{and} \quad k = \omega(1).$$

Remark 2. Recall that $\mathscr{P}_1(\epsilon)$ is the family of SBMs with edge probability in the regime $\Omega(\frac{1}{n^{2-\epsilon}})$ and $O(1)$. Moreover, $\frac{1}{n^2}$ is the threshold for a random graph to contain an edge with high probability [24]. Thus, the family $\mathscr{P}_1(\epsilon)$ covers most non-trivial SBM graphs.

Theorem 2. *For every $0 < \epsilon < 1$, let $A_n \sim \mathrm{SBM}(n, L, \mathbf{p}, f(n)\mathbf{Q}) \in \mathscr{P}_2(\epsilon)$. Then the expected length of compressor C_1 defined in Section 3.1 is upper bounded as*

$$\begin{aligned} &\mathsf{E}[\ell(C_1(A_n))] \\ &\leq H(A_n) + \binom{n}{2} f(n) \left(\mathbf{p}^T \mathbf{Q} \mathbf{p} \log \left(\frac{1}{\mathbf{p}^T \mathbf{Q} \mathbf{p}} \right) - \mathbf{p}^T \mathbf{Q}^* \mathbf{p} \right) + o(n^2 f(n)) \\ &= H(A_n) + o(H(A_n)), \end{aligned} \tag{3.7}$$

where $\mathbf{Q}^*$ denotes an $L \times L$ matrix whose (i, j) entry is $Q_{ij} \log(\frac{1}{Q_{ij}})$ when $Q_{ij} \neq 0$ and 0 when $Q_{ij} = 0$.

Corollary 2 (Universality over $\mathscr{P}_2$)**.** *For every $0 < \epsilon < 1$, the graph compressor C_1 is universal over the family $\mathscr{P}_2(\epsilon)$.*

Remark 3. Any compressor universal over the class $\mathscr{P}_1(\epsilon)$ is also universal over the class $\mathscr{P}_2(\epsilon)$, but our compressor designed specifically for the class $\mathscr{P}_2(\epsilon)$ has a lower computational complexity.

The above results upper bound the redundancy and establish the universality of the proposed compressor over $\mathscr{P}_1$ and $\mathscr{P}_2$. The minimax redundancy represents the lowest achievable redundancy of any lossless compressor for the family of SBMs. We bound the minimax redundancy of the family of SBMs $\mathscr{P}_3$ with fixed regime $f(n)$ and fixed largest entry $Q_{\max}$ in $\mathbf{Q}$ in the following Theorem.

Theorem 3 (Minimax redundancy)**.** *Let $f(n) = o(1)$ and $f(n) = \Omega(1/n^{2-\epsilon})$ for some $0 < \epsilon < 1$. The minimax redundancy of the family $\mathscr{P}_3(f(n), Q_{\max})$ is bounded as*

$$\frac{1}{2} \log \left(\frac{\binom{n}{2} f(n) Q_{\max}}{\pi e} \right) \leq R^*(\mathscr{P}_3(f(n), Q_{\max})) \leq \frac{1}{e} (\log e) Q_{\max} \binom{n}{2} f(n).$$

In Chapter 7, we analyze the performance of the proposed compressor under the local weak convergence framework considered in [18]. It turns out our proposed compressor achieves the optimal compression rate under the stronger universality criterion considered in that framework. This result requires a lot more definitions to introduce and we defer it to Theorem 4 in Chapter 7.

3.3 Empirical Performance

We implement the proposed universal graph compressor (UGC) in four widely used benchmark graph datasets: protein-to-protein interaction network (PPI) [26], LiveJournal friendship network (Blogcatalog) [51], Flickr user network (Flickr) [51], and YouTube user network (YouTube) [36]. The block decomposition size k is chosen to be $1, 2, 3, 4$ and we present in Table 3.1 the compression ratios (the ratio between output length and input length of the encoder) of UGC for different choices of k.[2] We present in Table 3.2 the compression ratios of four competing algorithms. In Fig. 3.2, we combine the comparisons of the compression ratios in Tables 3.1 and 3.2 in the logarithmic scale.

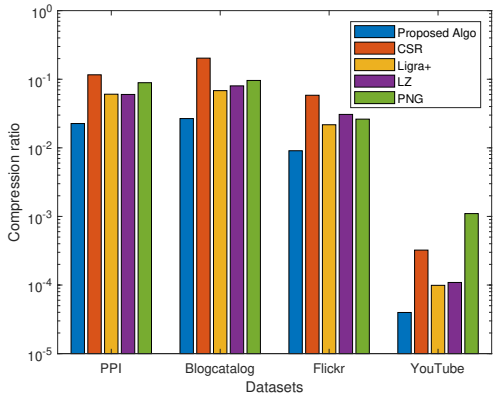

Figure 3.2: Log-scale comparisons of the compression ratios for the proposed universal graph compressor with other competing compressors. See Tables 3.1 and 3.2 for the exact compression ratios.

[2]Note that CSR and Ligra+ are designed to enable fast computation, such as adjacency query or vertex degree query, in addition to compressing the matrix. Our proposed compressor does not possess such functionality and is designed solely for compression purpose.

	$k=1$	$k=2$	$k=3$	$k=4$
PPI	0.0228	**0.0226**	0.0227	0.034
Blogcatalog	0.0275	0.0270	**0.0267**	0.0288
Flickr	0.00960	0.00935	0.00915	**0.00907**
YouTube	4.51×10^{-5}	4.11×10^{-5}	$\mathbf{3.98 \times 10^{-5}}$	4.00×10^{-5}

Table 3.1: Compression ratio of UGC under different k values.

	CSR	Ligra+	Lempel–Ziv	PNG
PPI	0.166	0.0605	0.06	0.089
Blogcatalog	0.203	0.0682	0.080	0.096
Flickr	0.0584	0.0217	0.0307	0.0262
YouTube	3.23×10^{-4}	9.90×10^{-5}	1.09×10^{-4}	1.10×10^{-3}

Table 3.2: Compression ratios of competing algorithms.

- CSR: Compressed sparse row is a widely used sparse matrix representation format. In the experiment, we further optimize its default compressor exploiting the fact that the graph is simple and its adjacency matrix is symmetric with binary entries.
- Ligra+: This is another powerful sparse matrix representation format [49, 50], which improves upon CSR using byte codes with run-length coding.
- LZ: This is an implementation of the algorithm proposed in [30], which first transforms the two-dimensional adjacency matrix into a one-dimensional sequence using the Peano–Hilbert space filling curve and then compresses the sequence using Lempel–Ziv 78 algorithm [58].
- PNG: The adjacency matrix of the graph is treated as a gray-scaled image and the PNG lossless image compressor is applied.

The compression ratios of the five algorithms implemented on four datasets are given as follows. The proposed UGC outperforms all competing algorithms in all datasets. The compression ratios from competing algorithms are 2.4 to 27 times that of the universal graph compressor.

Chapter 4

Main Ideas in Establishing Universality

In this section, we establish Theorems 1 and 2 in Section 3.1 based on four intermediate propositions. We defer the proofs of the propositions to Chapter 5.

4.1 Graph Entropy

We first calculate the entropy of the (random) graph A_n, which, recall, is the fundamental lower bound on the expected compression length for any compression scheme. Since we want to bound the redundancy of a universal compressor, we will be concerned with both the first and the second order term in $H(A_n)$.

Proposition 3 (Graph entropy)**.** *Let $A_n \sim \mathrm{SBM}(n, L, \mathbf{p}, f(n)\mathbf{Q})$ with $f(n) = O(1), f(n) = \Omega\left(\frac{1}{n^2}\right)$, and $L = \Theta(1)$. For $0 \le p \le 1$, let $h(p) \triangleq -p\log(p) - (1-p)\log(1-p)$ denote the binary entropy function. For a matrix W with entries in $[0,1]$, let $h(W)$ be a matrix of the same dimension whose (i,j) entry is $h(W_{ij})$. Then*

$$H(A_n) = \binom{n}{2} H(A_{12}|X_1, X_2)(1 + o(1)) \tag{4.1}$$

$$= \binom{n}{2} \mathbf{p}^T h\big(f(n)\mathbf{Q}\big)\mathbf{p} + o\left(n^2 h\big(f(n)\big)\right). \tag{4.2}$$

In particular, when $f(n) = \omega\left(\frac{1}{n}\right)$ and $f(n) = o(1)$, expression (4.2) can be simplified as

$$H(A_n) = \binom{n}{2} f(n) \left(\log\left(\frac{1}{f(n)}\right) \mathbf{p}^T\mathbf{Q}\mathbf{p} + \mathbf{p}^T\mathbf{Q}\mathbf{p}\log e + \mathbf{p}^T\mathbf{Q}^*\mathbf{p} + o(1) \right), \tag{4.3}$$

where $\mathbf{Q}^*$ denotes an $L \times L$ matrix whose (i,j) entry is $Q_{ij}\log(\frac{1}{Q_{ij}})$ when $Q_{ij} \neq 0$ and 0 when $Q_{ij} = 0$.

When $f(n) = \Omega\left(\frac{1}{n^2}\right)$ *and* $f(n) = O\left(\frac{1}{n}\right)$*, the entropy* $H(A_n)$ *can be upper bounded as*

$$H(A_n) \leq \binom{n}{2} f(n) \Bigg(\log\left(\frac{1}{f(n)}\right) \mathbf{p}^T \mathbf{Q}\mathbf{p} + \mathbf{p}^T \mathbf{Q}\mathbf{p} \log e + \mathbf{p}^T \mathbf{Q}\mathbf{p} \log \frac{1}{\mathbf{p}^T \mathbf{Q}\mathbf{p}} + o(1) \Bigg), \tag{4.4}$$

and lower bounded as

$$H(A_n) \geq \binom{n}{2} f(n) \left(\log\left(\frac{1}{f(n)}\right) \mathbf{p}^T \mathbf{Q}\mathbf{p} + \mathbf{p}^T \mathbf{Q}\mathbf{p} \log e + \mathbf{p}^T \mathbf{Q}^* \mathbf{p} + o(1) \right). \tag{4.5}$$

Remark 4. In the regime $f(n) = \Omega\left(\frac{1}{n}\right)$ and $f(n) = O(1)$, equation (4.2) has been established in [1]. We extend the analysis to the regime $f(n) = o\left(\frac{1}{n}\right)$ and $f(n) = \Omega(\frac{1}{n^2})$. Moreover, we establish the upper and lower bounds on the second order terms in the graph entropy.

Remark 5. Proposition 3 can be used to calculate the entropy of the graph for certain important regimes of $f(n)$, in which the SBM displays characteristic behavior. For $f(n) = 1$, we have $H(A_n) = \binom{n}{2}(\mathbf{p}^T h(\mathbf{Q})\mathbf{p} + o(1))$; for $f(n) = \frac{\log n}{n}$ (the regime where the phase transition for exact recovery of the community assignments occurs [3, 4]), we have $H(A_n) = \frac{n \log n}{2}(\mathbf{p}^T \mathbf{Q}\mathbf{p} \log n + \mathbf{p}^T \mathbf{Q}\mathbf{p} \log e + \mathbf{p}^T \mathbf{Q}^* \mathbf{p} + o(1))$; when $f(n) = \frac{1}{n}$ (the regime where the phase transition for detection between SBM and the Erdős–Rényi model occurs [35]), we have $H(A_n) = \frac{n \log n}{2}(\mathbf{p}^T \mathbf{Q}\mathbf{p} + o(1))$; when $f(n) = \frac{1}{n^2}$ (the regime where the phase transition for the existence of an edge occurs), we have $H(A_n) = \log n(\mathbf{p}^T \mathbf{Q}\mathbf{p} + o(1))$.

4.2 Asymptotic i.i.d. via Block Decomposition

To compress the matrix A_n, we wish to decompose it into a large number of components that have little correlation between them. This leads to the idea of block decomposition described previously. Since the sequence of blocks are used to compress A_n, the next proposition claims these blocks are identically distributed and asymptotically independent in a precise sense described as follows.

Proposition 4 (Block decomposition)**.** *Let* $A_n \sim \mathrm{SBM}(n, L, \mathbf{p}, f(n)\mathbf{Q})$ *with* $f(n) = \Omega\left(\frac{1}{n^{2-\epsilon}}\right)$ *for some* $0 < \epsilon < 1$*,* $f(n) = O(1)$*, and* $L = \Theta(1)$*. Let*

k be an integer that divides n and $n' = n/k$. Consider the $k \times k$ block decomposition in (3.2)*. We have all the off-diagonal blocks share the same joint distribution; all the diagonal blocks share the same joint distribution. In other words, for any $1 \le i_1, i_2, j_1, j_2 \le n'$ with $i_1 \ne j_1, i_2 \ne j_2$ and $1 \le l_1, l_2 \le n'$, we have*

$$\mathbf{B}_{i_1,j_1} \stackrel{d}{=} \mathbf{B}_{i_2,j_2},$$
$$\mathbf{B}_{l_1,l_1} \stackrel{d}{=} \mathbf{B}_{l_2,l_2}.$$

In addition, if $k = \omega(1)$ and $k = o(n)$, the off-diagonal blocks are asymptotically i.i.d. in the sense that

$$\lim_{n\to\infty} \frac{H(\mathbf{B}_{\mathrm{ut}})}{\binom{n'}{2} H(\mathbf{B}_{12})} = 1. \tag{4.6}$$

4.3 Length Analysis for Correlated Sequences

Thanks to this property of the block decomposition, we hope to compress these blocks as if they are independent using a Laplace probability assignment (which, recall, is universal for the class of all m-ary iid processes). However, since these blocks are still correlated (albeit weakly), we will need a result on the performance of Laplace probability assignment on correlated sequences with identical marginals, which we give next.

Proposition 5 (Laplace probability assignment for correlated sequence)**.** *Consider arbitrarily correlated $Z_1, Z_2, \ldots, Z_N$, where the marginal distribution of each Z_i is identically distributed over an alphabet of size $m \ge 2$. Let $\ell_{\mathrm{L}}(z^N) = \left\lceil \log \frac{1}{q_{\mathrm{L}}(z^N)} \right\rceil + 1$ where $q_{\mathrm{L}}(\cdot)$ is the marginal distribution induced by Laplace probability assignment in* (2.1)

$$q_{\mathrm{L}}(z^N) := \frac{N_1! N_2! \cdots N_m!}{N!} \cdot \frac{1}{\binom{N+m-1}{m-1}}. \tag{4.7}$$

We then have

$$\mathsf{E}[\ell_{\mathrm{L}}(Z^N)] \le m \log(2eN) + NH(Z_1) + 2. \tag{4.8}$$

We provide a similar result for the KT probability assignment.

Proposition 6 (KT probability assignment for correlated sequence)**.** *Consider arbitrarily correlated $Z_1, Z_2, \ldots, Z_N$, where the marginal distribution*

of each Z_i is identically distributed over an alphabet of size $m \geq 2$. Let $\ell_{\mathrm{KT}}(z^N) = \left\lceil \log \frac{1}{q_{\mathrm{KT}}(z^N)} \right\rceil + 1$ where $q_{\mathrm{KT}}(\cdot)$ is the marginal distribution induced by KT probability assignment in (2.2)

$$q_{\mathrm{KT}}(z^N) = \frac{(2N_1 - 1)!!(2N_2 - 1)!! \cdots (2N_m - 1)!!}{m(m+2)\cdots(m+2N-2)} \tag{4.9}$$

with $(-1)!! \triangleq 1$. *We then have*

$$\mathsf{E}[\ell_{\mathrm{KT}}(Z^N)] \leq \tfrac{m}{2} \log\left(e\left(1 + \tfrac{2N}{m}\right)\right) + \tfrac{1}{2}\log(\pi N) + NH(Z_1) + 2. \tag{4.10}$$

4.4 Proof of Theorem 1

We are now ready to prove Theorem 1.

Proof. We discuss the redundancy in two cases: (1) $f(n) = o(1)$ and $f(n) = \Omega(1/n^{2-\epsilon})$ and (2) $f(n) = \Theta(1)$.

First assume $f(n) = o(1)$ and $f(n) = \Omega(1/n^{2-\epsilon})$. By Proposition 3, we can lower bound the graph entropy

$$H(A_n) \geq \binom{n}{2} f(n) \left(\log\left(\frac{1}{f(n)}\right) \mathbf{p}^T \mathbf{Q} \mathbf{p} + \mathbf{p}^T \mathbf{Q} \mathbf{p} \log e + \mathbf{p}^T \mathbf{Q}^* \mathbf{p} + o(1) \right). \tag{4.11}$$

Now, we will upper bound the redundancy of C_k for both KT probability assignment and Laplace probability assignment. Since the expected length of the KT probability assignment in (4.10) is upper bounded by the length of the Laplace probability assignment in (4.8), it suffices to upper bound the redundancy for the Laplace probability assignment. Recall that here we compress the diagonal blocks $\mathbf{B}_{\mathrm{d}}$ ($m = 2^{k^2}$-sized alphabet, $N = n'$ blocks), the off-diagonal blocks $\mathbf{B}_{\mathrm{ut}}$ ($m = 2^{k^2}$-sized alphabet, $N = \binom{n'}{2}$ blocks) and the remaining $(n-\tilde{n})\tilde{n} + \binom{n-\tilde{n}}{2}$ entries separately. Let N_r denote the number of ones in the remaining $(n-\tilde{n})\tilde{n} + \binom{n-\tilde{n}}{2}$ entries. We have,

$$\begin{aligned}
&\mathsf{E}(\ell(C_k(A_n))) = \mathsf{E}(\ell_{\mathrm{L}}(\mathbf{B}_{\mathrm{ut}})) + \mathsf{E}(\ell_{\mathrm{L}}(\mathbf{B}_{\mathrm{d}})) + \mathsf{E}(2N_r \lceil \log n \rceil) \\
&\leq \binom{n'}{2} H(\mathbf{B}_{12}) + 2^{k^2} \log\left(2e\binom{n'}{2}\right) + n' H(\mathbf{B}_{11}) + 2^{k^2}\log(2en') + 4 \\
&+ \mathsf{E}(2N_r \lceil \log n \rceil) \\
&\overset{(a)}{\leq} \binom{n'}{2} H(\mathbf{B}_{12}) + 2^{k^2} \log\left(en^2\right) + nH(\mathbf{B}_{11}) + 2^{k^2}\log(2en) + 4 \\
&+ \mathsf{E}(2N_r \lceil \log n \rceil)
\end{aligned}$$

$$\overset{(b)}{\leq} \binom{n'}{2} H(\mathbf{B}_{12}) + 2^{k^2} \log\left(2e^2 n^3\right) + (nk^2 H(A_{12}) + 4) + \mathsf{E}(2N_r \lceil \log n \rceil), \tag{4.12}$$

where in (a) we bound $2\binom{n'}{2} \leq n^2$ and $n' \leq n$, and in (b) we note that $H(\mathbf{B}_{11}) \leq k^2 H(A_{12})$ since there are $k^2 - k$ elements of the matrix (all apart from the diagonal elements) are distributed identically as A_{12}. We will now analyze each of these four terms separately. Firstly, we have

$$\begin{aligned}
&\binom{n'}{2} H(\mathbf{B}_{12}) \leq \binom{n'}{2} k^2 H(A_{1,k+1}) \\
&= \binom{n'}{2} k^2 h(f(n)\mathbf{p}^T\mathbf{Q}\mathbf{p}) \\
&\overset{(c)}{=} \binom{n}{2} f(n) \left(\left(\log \frac{1}{f(n)} \right) \mathbf{p}^T\mathbf{Q}\mathbf{p} + \mathbf{p}^T\mathbf{Q}\mathbf{p} \log e + \mathbf{p}^T\mathbf{Q}\mathbf{p} \log \frac{1}{\mathbf{p}^T\mathbf{Q}\mathbf{p}} + o(1) \right),
\end{aligned} \tag{4.13}$$

where (c) follows since $h(g(n)) = g(n) \log \frac{1}{g(n)} + g(n) \log e + o(g(n))$ (see, for example, [16]). Next, since $k \leq \sqrt{\delta \log n}$ and $\delta < \epsilon$, we have

$$2^{k^2} \log\left(2e^2 n^3\right) \leq n^{\delta} \log(2e^2 n^3) = o(n^2 f(n)). \tag{4.14}$$

Moreover, we have

$$\begin{aligned}
nk^2 H(A_{12}) + 4 &\leq (n\delta \log n) H(A_{12}) + 4 \\
&= (n\delta \log n) O\left(f(n) \log \frac{1}{f(n)} \right) + 4 \\
&= o(n^2 f(n)).
\end{aligned} \tag{4.15}$$

Finally, recall that N_r is the number of ones in the remaining $(n-\tilde{n})\tilde{n} + \binom{n-\tilde{n}}{2}$ entries, we have

$$\begin{aligned}
\mathsf{E}(2N_r \lceil \log n \rceil) &= f(n)\mathbf{p}^T\mathbf{Q}\mathbf{p} \left((n-\tilde{n})\tilde{n} + \binom{n-\tilde{n}}{2} \right) 2\lceil \log n \rceil \\
&= O(knf(n) \log n) \\
&= o(n^2 f(n)).
\end{aligned} \tag{4.16}$$

Substituting (4.13), (4.14), (4.15), (4.16) into the upper bound (4.12) and

combining with the lower bound for the entropy (4.11) gives

$$
\begin{aligned}
&\mathsf{E}(\ell(C_k(A_n))) - H(A_n) \\
&\le \binom{n}{2} f(n) \left(\left(\log \frac{1}{f(n)} \right) \mathbf{p}^T \mathbf{Q}\mathbf{p} + \mathbf{p}^T \mathbf{Q}\mathbf{p} \log e + \mathbf{p}^T \mathbf{Q}\mathbf{p} \log \frac{1}{\mathbf{p}^T \mathbf{Q}\mathbf{p}} + o(1) \right) \\
&\quad - \binom{n}{2} f(n) \left(\log \left(\frac{1}{f(n)} \right) \mathbf{p}^T \mathbf{Q}\mathbf{p} + \mathbf{p}^T \mathbf{Q}\mathbf{p} \log e + \mathbf{p}^T \mathbf{Q}^* \mathbf{p} + o(1) \right) \\
&\quad + o(n^2 f(n)) \\
&= \binom{n}{2} f(n) \left(\mathbf{p}^T \mathbf{Q}\mathbf{p} \log \left(\frac{1}{\mathbf{p}^T \mathbf{Q}\mathbf{p}} \right) - \mathbf{p}^T \mathbf{Q}^* \mathbf{p} \right) + o(n^2 f(n)) \\
&= o(H(A_n))
\end{aligned}
$$

as required.

Now assume $f(n) = \Theta(1)$. Then $H(A_n)$ can be lower bounded as

$$
H(A_n) \ge H(A_n | X^n) = \binom{n}{2} H(A_{12} | X_1, X_2) = \binom{n}{2} \mathbf{p}^T h(f(n)\mathbf{Q})\mathbf{p}. \quad (4.17)
$$

Still, we can upper bound the expected length of the compressor C_k using (4.12) and we will bound the four term separately. We have

$$
\begin{aligned}
\binom{n'}{2} H(\mathbf{B}_{12}) &= \binom{n'}{2} (H(\mathbf{B}_{12} | X_1^{2k}) + I(X_1^{2k}; \mathbf{B}_{12})) \\
&= \binom{n'}{2} (k^2 H(A_{1,k+1} | X_1, X_{k+1}) + I(X_1^{2k}; \mathbf{B}_{12})) \\
&\le \binom{n'}{2} (k^2 \mathbf{p}^T h(f(n)\mathbf{Q})\mathbf{p} + H(X_1^{2k})) \\
&= \binom{n'}{2} (k^2 \mathbf{p}^T h(f(n)\mathbf{Q})\mathbf{p} + 2kH(\mathbf{p})) \\
&= \frac{n(n-k)}{2} \mathbf{p}^T h(f(n)\mathbf{Q})\mathbf{p} + n(n'-1)H(\mathbf{p}), \qquad (4.18)
\end{aligned}
$$

$$
2^{k^2} \log(2e^2 n^3) \le n^\delta \log(2e^2 n^3) = o\left(\frac{n^2}{k} \right), \qquad (4.19)
$$

$$
nk^2 H(A_{12}) + 4 \le n\delta \log n H(A_{12}) + 4 = o\left(\frac{n^2}{k} \right). \qquad (4.20)
$$

and

$$\begin{aligned}\mathsf{E}(2N_r\lceil\log n\rceil) &= f(n)\mathbf{p}^T\mathbf{Q}\mathbf{p}\left((n-\tilde{n})\tilde{n}+\binom{n-\tilde{n}}{2}\right)2\lceil\log n\rceil \\ &= O(kn\log n) = o\left(\frac{n^2}{k}\right). \end{aligned} \tag{4.21}$$

Combining the bounds (4.17), (4.18), (4.19), (4.20), (4.21) gives

$$\begin{aligned}&\mathsf{E}(\ell(C_k(A_n))) - H(A_n) \\ &\leq \frac{n(n-k)}{2}\mathbf{p}^T h(f(n)\mathbf{Q})\mathbf{p} + n(n'-1)H(\mathbf{p}) \\ &- \frac{n(n-1)}{2}\mathbf{p}^T h(f(n)\mathbf{Q})\mathbf{p} + o\left(\frac{n^2}{k}\right) \\ &= n(n'-1)H(\mathbf{p}) + \frac{n(1-k)}{2}\mathbf{p}^T h(f(n)\mathbf{Q})\mathbf{p} + o\left(\frac{n^2}{k}\right) \\ &\leq H(\mathbf{p})\frac{n^2}{k} + o\left(\frac{n^2}{k}\right) \\ &= o(H(A_n)),\end{aligned}$$

where the last line follows since $k = \omega(1)$ and $H(A_n) = \Theta(n^2)$ when $f(n) = \Theta(1)$. □

4.5 Proof of Theorem 2

Proof. To upper bound the redundancy, we need to lower bound the entropy $H(A_n)$ and upper bound the expected length of the compressor C_1. We notice that the entropy lower bound (4.11) is still valid and it suffices to upper bound the expected length. Following a similar argument as in the proof of Theorem 1, it suffices to upper bound the expected length of Laplace probability assignment. In the case when $k = 1$, we do not need to compress the diagonal block sequence since they are all zeros and we do not have any remaining bits to compress separately. Proposition 5 yields

$$\mathsf{E}(\ell(C_1(A_n))) = \mathsf{E}(\ell_{\mathrm{L}}(\mathbf{B}_{\mathrm{ut}})) \tag{4.22}$$

$$\leq \binom{n}{2}H(A_{12}) + \left(2\log\left(2e\binom{n}{2}\right)+2\right), \tag{4.23}$$

We will now analyze each of these two terms separately. Firstly, we have

$$\begin{aligned}
&\binom{n}{2} H(A_{12}) = \binom{n}{2} h(f(n)\mathbf{p}^T\mathbf{Q}\mathbf{p}) \\
&\overset{(a)}{=} \binom{n}{2} f(n) \left(\left(\log \frac{1}{f(n)} \right) \mathbf{p}^T\mathbf{Q}\mathbf{p} + \mathbf{p}^T\mathbf{Q}\mathbf{p} \log e + \mathbf{p}^T\mathbf{Q}\mathbf{p} \log \frac{1}{\mathbf{p}^T\mathbf{Q}\mathbf{p}} + o(1) \right),
\end{aligned} \tag{4.24}$$

where (a) follows since $h(g(n)) = g(n) \log \frac{1}{g(n)} + g(n) \log e + o(g(n))$. Next, we have

$$2 \log \left(2e \binom{n}{2} \right) + 2 = O(\log n) = o(n^2 f(n)). \tag{4.25}$$

Substituting bounds (4.24), (4.25) into (4.23) and combining with bound (4.11) gives

$$\begin{aligned}
&\mathsf{E}(\ell(C_1(A_n))) - H(A_n) \\
&\leq \binom{n}{2} f(n) \left(\left(\log \frac{1}{f(n)} \right) \mathbf{p}^T\mathbf{Q}\mathbf{p} + \mathbf{p}^T\mathbf{Q}\mathbf{p} \log e + \mathbf{p}^T\mathbf{Q}\mathbf{p} \log \frac{1}{\mathbf{p}^T\mathbf{Q}\mathbf{p}} + o(1) \right) \\
&\quad - \binom{n}{2} f(n) \left(\log \left(\frac{1}{f(n)} \right) \mathbf{p}^T\mathbf{Q}\mathbf{p} + \mathbf{p}^T\mathbf{Q}\mathbf{p} \log e + \mathbf{p}^T\mathbf{Q}^*\mathbf{p} + o(1) \right) \\
&\quad + o(n^2 f(n)) \\
&= \binom{n}{2} f(n) \left(\mathbf{p}^T\mathbf{Q}\mathbf{p} \log \left(\frac{1}{\mathbf{p}^T\mathbf{Q}\mathbf{p}} \right) - \mathbf{p}^T\mathbf{Q}^*\mathbf{p} \right) + o(n^2 f(n)) \\
&= o(H(A_n))
\end{aligned}$$

as required. □

Remark 6. When $f(n) = 1$, the compressor C_1 is strictly suboptimal. This is because the length achieved by C_1 is $\binom{n}{2} h\left(f(n)\mathbf{p}^T\mathbf{Q}\mathbf{p}\right)(1 + o(1))$, whereas the first order term in the entropy is $\binom{n}{2} \mathbf{p}^T h\left(f(n)\mathbf{Q}\right)\mathbf{p}$. When $f(n)$ is $o(1)$, these two have the same first order term. However, when $f(n)$ is constant, $\mathbf{p}^T h\left(f(n)\mathbf{Q}\right)\mathbf{p}$ is strictly smaller than $h\left(f(n)\mathbf{p}^T\mathbf{Q}\mathbf{p}\right)$ by concavity of entropy.

Chapter 5

Proof of Intermediate Propositions

5.1 Graph Entropy

Proof of Proposition 3. Note that

$$\begin{aligned} H(A_n) &= H(A_n|X^n) + I(X^n; A_n) \\ &= \binom{n}{2} H(A_{12}|X_1, X_2) + I(X^n; A_n) \qquad (5.1) \\ &= \binom{n}{2} \mathbf{p}^T h\big(f(n)\mathbf{Q}\big)\mathbf{p} + I(X^n; A_n), \qquad (5.2) \end{aligned}$$

where (5.2) follows since all the $\binom{n}{2}$ edges are identically distributed and also independent given X^n and consequently

$$\begin{aligned} H(A_n|X^n) &= \binom{n}{2} H(A_{12}|X_1, X_2) \\ &= \binom{n}{2} \sum_{i,j} H(A_{12}|X_1 = i, X_2 = j) p_i p_j \\ &= \binom{n}{2} \mathbf{p}^T h\left(f(n)\mathbf{Q}\right)\mathbf{p}. \end{aligned}$$

In the following, we first establish statement (4.2) for $f(n) = \Theta(1)$ and then prove the simplified expression (4.3) for $f(n) = \omega\left(\frac{1}{n}\right)$ and $f(n) = o(1)$. Finally, we prove the refined upper and lower bounds on $H(A_n)$ in (4.4) and (4.5) for $f(n) = \Omega\left(\frac{1}{n^2}\right)$ and $f(n) = O\left(\frac{1}{n}\right)$, which implies statement (4.2) for $f(n)$ in the same regime.

For $f(n) = \Theta(1)$, we have that

$$\begin{aligned}\binom{n}{2}\mathbf{p}^T h\big(f(n)\mathbf{Q}\big)\mathbf{p} &= \binom{n}{2}\sum_{i,j=1}^{L} p_i h(f(n)Q_{ij})p_j \\ &\overset{(a)}{=} \binom{n}{2}(1+o(1))\sum_{i,j:Q_{ij}=\Theta(1)} p_i h(f(n)Q_{ij})p_j \qquad (5.3)\\ &= \Theta(n^2), \qquad (5.4)\end{aligned}$$

where (a) follows since $h(f(n)Q_{ij}) = o(h(f(n)))$ if $Q_{ij} = o(1)$. Moreover, we have

$$0 \le I(X^n; A_n) \le H(X^n) = nH(X_1) \le n\log L. \qquad (5.5)$$

Substituting (5.4) and (5.5) into (5.2) yields $H(A_n) = \binom{n}{2}\mathbf{p}^T h\big(f(n)\mathbf{Q}\big)\mathbf{p} + o\left(n^2 h(f(n))\right)$.

Next, consider the case when $f(n) = \omega\left(\frac{1}{n}\right)$ and $f(n) = o(1)$. Since $h(g(n)) = g(n)\log\frac{1}{g(n)} + g(n)\log e + o(g(n))$ and $f(n) = o(1)$, we can rewrite the term $\mathbf{p}^T h(f(n)\mathbf{Q})\mathbf{p}$ as

$$\begin{aligned}&\mathbf{p}^T h(f(n)\mathbf{Q})\mathbf{p}\\ &= \sum_{i,j\in[L]} p_i p_j h(f(n)Q_{ij})\\ &= \sum_{i,j\in[L]} p_i p_j \left(f(n)Q_{ij}\log\left(\frac{1}{f(n)Q_{ij}}\right) + f(n)Q_{ij}\log e + o(f(n)Q_{ij})\right)\\ &= \sum_{i,j\in[L]} p_i p_j \Bigg(f(n)Q_{ij}\log\left(\frac{1}{f(n)}\right) + f(n)Q_{ij}\log e\\ &\quad + f(n)Q_{ij}\log\left(\frac{1}{Q_{ij}}\right) + o(f(n)Q_{ij})\Bigg)\\ &= f(n)\left(\log\left(\frac{1}{f(n)}\right)\mathbf{p}^T\mathbf{Q}\mathbf{p} + \mathbf{p}^T\mathbf{Q}\mathbf{p}\log e + \mathbf{p}^T\mathbf{Q}^*\mathbf{p} + o(1)\right). \qquad (5.6)\end{aligned}$$

Since $f(n) = \omega\left(\frac{1}{n}\right)$, we have

$$I(X^n; A_n) \le H(X^n) = nH(X_1) \le n\log L = o\left(f(n)n^2\right). \qquad (5.7)$$

Substituting (5.6) and (5.7) into (5.2) yields

$$H(A_n) = \binom{n}{2} f(n)\left(\log\left(\frac{1}{f(n)}\right)\mathbf{p}^T\mathbf{Q}\mathbf{p} + \mathbf{p}^T\mathbf{Q}\mathbf{p}\log e + \mathbf{p}^T\mathbf{Q}^*\mathbf{p} + o(1)\right).$$

Finally, consider the case when $f(n) = \Omega\left(\frac{1}{n^2}\right)$ and $f(n) = O\left(\frac{1}{n}\right)$. By properties of the entropy, we have

$$H(A_n|X^n) \le H(A_n) \le \binom{n}{2} H(A_{12}). \tag{5.8}$$

Note that

$$\mathsf{P}(A_{12} = 1) = \sum_{i,j} \mathsf{P}(A_{12} = 1|X_1 = i, X_2 = j) p_i p_j = \mathbf{p}^T f(n) \mathbf{Q} \mathbf{p},$$

which yields that $H\left(A_{12}\right) = h\left(f(n)\mathbf{p}^T\mathbf{Q}\mathbf{p}\right)$. Substituting this into (5.8) gives

$$\binom{n}{2} \mathbf{p}^T h(f(n)\mathbf{Q})\mathbf{p} \le H(A_n) \le \binom{n}{2} h\left(f(n)\mathbf{p}^T\mathbf{Q}\mathbf{p}\right). \tag{5.9}$$

By (5.6), we can rewrite the lower bound as

$$\begin{aligned}
&\binom{n}{2} \mathbf{p}^{\mathbf{T}} h(f(n)\mathbf{Q})\mathbf{p} \\
&= \binom{n}{2} f(n) \left(\log\left(\frac{1}{f(n)}\right) \mathbf{p}^T\mathbf{Q}\mathbf{p} + \mathbf{p}^T\mathbf{Q}\mathbf{p} \log e + \mathbf{p}^T\mathbf{Q}^*\mathbf{p} + o(1) \right).
\end{aligned} \tag{5.10}$$

Note that $p_i = \Theta(1)$ for any $i \in [L]$ and $\max_{i,j} Q_{ij} = \Theta(1)$, we have

$$\mathbf{p}^T\mathbf{Q}\mathbf{p} = \sum_{i,j\in[L]} p_i Q_{ij} p_j = (1 + o(1)) \sum_{i,j:Q_{ij}=\Theta(1)} p_i Q_{ij} p_j. \tag{5.11}$$

Now, we can rewrite the upper bound of $H(A_n)$ in (5.9) as

$$\begin{aligned}
\binom{n}{2} h\left(f(n)\mathbf{p}^T\mathbf{Q}\mathbf{p}\right) =& \binom{n}{2} f(n) \Bigg(\log\left(\frac{1}{f(n)}\right) \mathbf{p}^T\mathbf{Q}\mathbf{p} + \mathbf{p}^T\mathbf{Q}\mathbf{p} \log e \\
&+ \mathbf{p}^T\mathbf{Q}\mathbf{p} \log \frac{1}{\mathbf{p}^T\mathbf{Q}\mathbf{p}} + o(1) \Bigg).
\end{aligned} \tag{5.12}$$

With bounds (5.10) and (5.12), we

$$\begin{aligned}
&\binom{n}{2} h\left(f(n)\mathbf{p}^T\mathbf{Q}\mathbf{p}\right) - \binom{n}{2} \mathbf{p}^T h(f(n)\mathbf{Q})\mathbf{p} \\
&= \binom{n}{2} f(n) \left(\mathbf{p}^T\mathbf{Q}^*\mathbf{p} - \mathbf{p}^T\mathbf{Q}\mathbf{p} \log \frac{1}{\mathbf{p}^T\mathbf{Q}\mathbf{p}} + o(1) \right) \\
&= o\left(n^2 h\big(f(n)\big)\right),
\end{aligned}$$

which proves $H(A_n) = \binom{n}{2}\mathbf{p}^T h(f(n)\mathbf{Q})\mathbf{p} + o\left(n^2 h\big(f(n)\big)\right)$ in this case. □

5.2 Asymptotic i.i.d. via Block Decomposition

We first invoke a known property of stochastic block models (see, for example, [29]). We include the proof here for completeness.

Lemma 1 (Exchangeability of SBM)**.** *Let $A_n \sim \mathrm{SBM}(n, L, \mathbf{p}, \mathbf{W})$. For a permutation $\pi : [n] \to [n]$, let $\pi(A_n)$ be an $n \times n$ matrix whose (i, j) entry is given by $A_{\pi(i),\pi(j)}$. Then, for any permutation $\pi : [n] \to [n]$, the joint distribution of A_n is the same as the joint distribution of $\pi(A_n)$, i.e.,*

$$A_n \stackrel{d}{=} \pi(A_n). \tag{5.13}$$

Proof. Let a_n be a realization of the random matrix A_n and $\pi(X^n)$ be the permuted vector $(X_{\pi(1)}, \ldots, X_{\pi(n)})$. For any symmetric binary matrix a_n with zero diagonal entries, we have

$$\begin{aligned}
&\mathsf{P}(A_n = a_n) \\
&= \sum_{x^n \in [L]^n} \mathsf{P}(A_n = a_n, X^n = x^n) \\
&= \sum_{x^n \in [L]^n} \mathsf{P}(A_n = a_n | X^n = x^n) \prod_{i=1}^{n} \mathsf{P}(X_i = x_i) \\
&\stackrel{(a)}{=} \sum_{x^n \in [L]^n} \prod_{\substack{i,j \\ 1 \le i < j \le n}} \mathsf{P}(A_{ij} = a_{ij} | X_i = x_i, X_j = x_j) \prod_{i=1}^{n} \mathsf{P}(X_{\pi(i)} = x_i) \\
&\stackrel{(b)}{=} \sum_{x^n \in [L]^n} \prod_{\substack{i,j \\ 1 \le i < j \le n}} (W_{x_i,x_j})^{a_{ij}} (1 - W_{x_i,x_j})^{1-a_{ij}} \prod_{i=1}^{n} \mathsf{P}(X_{\pi(i)} = x_i) \\
&\stackrel{(c)}{=} \sum_{x^n \in [L]^n} \prod_{\substack{i,j \\ 1 \le i < j \le n}} \mathsf{P}(A_{\pi(i),\pi(j)} = a_{ij} | X_{\pi(i)} = x_i, X_{\pi(j)} = x_j) \prod_{i=1}^{n} \mathsf{P}(X_{\pi(i)} = x_i) \\
&= \sum_{x^n \in [L]^n} \mathsf{P}(\pi(A_n) = a_n, \pi(X^n) = x^n) \\
&= \mathsf{P}(\pi(A_n) = a_n),
\end{aligned}$$

where (a) follows since X^n are i.i.d. and thus $\mathsf{P}(X_i = x_i) = \mathsf{P}(X_{\pi(i)} = x_i)$

and (b) follows since $A_{ij} \sim \text{Bern}(W_{X_i,X_j})$, and thus

$$\mathsf{P}(A_{ij} = a_{ij} | X_i = x_i, X_j = x_j) = \begin{cases} W_{x_i,x_j} & \text{if } a_{ij} = 1 \\ 1 - W_{x_i,x_j} & \text{if } a_{ij} = 0 \end{cases} \tag{5.14}$$

$$= (W_{x_i,x_j})^{a_{ij}} (1 - W_{x_i,x_j})^{1-a_{ij}}. \tag{5.15}$$

The step in (c) follows since $A_{\pi(i),\pi(j)} \sim \text{Bern}(W_{X_{\pi(i)},X_{\pi(j)}})$ and the conditional probability has the same expression as in (5.15). □

Now we are ready to establish Proposition 4.

Proof of Proposition 4. For any $i_1 \neq j_1$ and $i_2 \neq j_2$, $1 \leq i_1, j_1, i_2, j_2 \leq n'$, consider a permutation $\pi_1 : [n] \to [n]$ that has

$$\pi_1(x) = \begin{cases} x + (i_2 - i_1)k & \text{for } (i_1 - 1)k + 1 \leq x \leq i_1 k \\ x + (j_2 - j_1)k & \text{for } (j_1 - 1)k + 1 \leq x \leq j_1 k \end{cases}$$

and the remaining $n - 2k$ arguments are mapped to the $n - 2k$ values in $[n] \setminus \{(i_2-1)k+1, \ldots, i_2 k, (j_2-1)k, \ldots, j_2 k\}$ in any order. Lemma 1 implies that $\mathbf{B}_{i_1,j_1}$, which is the submatrix formed by the rows $(i_1 - 1)k + 1, \ldots, i_1 k$ and the columns $(j_1 - 1)k + 1, \ldots, j_1 k$ has the same distribution as the submatrix formed by the rows $\pi_1((i_1-1)k+1), \ldots, \pi_1(i_1 k)$ and the columns $\pi_1((j_1 - 1)k + 1), \ldots, \pi_1(j_1 k)$. From the definition of π_1, we see that the latter submatrix is $\mathbf{B}_{i_2,j_2}$ and we establish that $\mathbf{B}_{i_1,j_1} \stackrel{d}{=} \mathbf{B}_{i_2,j_2}$. Similarly, defining a permutation $\pi_2 : [n] \to [n]$ which has

$$\pi_2(x) = \; x + (l_2 - l_1)k \quad \text{for } (l_1 - 1)k + 1 \leq x \leq l_1 k$$

and invoking Lemma 1 establishes $\mathbf{B}_{l_1,l_1} \stackrel{d}{=} \mathbf{B}_{l_2,l_2}$.

Now, clearly $H(\mathbf{B}_{\text{ut}}) \leq \binom{n'}{2} H(\mathbf{B}_{12})$, and therefore we have

$$\limsup_{n\to\infty} \frac{H(\mathbf{B}_{\text{ut}})}{\binom{n'}{2} H(\mathbf{B}_{12})} \leq 1. \tag{5.16}$$

Moreover we have $H(A_n) = H(\mathbf{B}_{\text{ut}}, \mathbf{B}_{\text{d}}) \leq H(\mathbf{B}_{\text{ut}}) + H(\mathbf{B}_{\text{d}}) \leq H(\mathbf{B}_{\text{ut}}) + n' H(\mathbf{B}_{11}) \leq H(\mathbf{B}_{\text{ut}}) + n' k^2 h(A_{12})$ where the last inequality follows by noting that except for the diagonal elements of $\mathbf{B}_{\text{d}}$ (which are zero and thus have zero entropy), all other elements have the same distribution as A_{12}. We therefore obtain $H(\mathbf{B}_{\text{ut}}) \geq H(A_n) - n' k^2 h(A_{12}) = H(A_n) - nkh(A_{12}) \geq H(A_n | X_1^n) - nkh(A_{12}) = \binom{n}{2} \mathbf{p}^T h(f(n)\mathbf{Q})\mathbf{p} - nkh(f(n)\mathbf{p}^T\mathbf{Q}\mathbf{p})$. Consequently,

$$\frac{H(\mathbf{B}_{\text{ut}})}{\binom{n'}{2} H(\mathbf{B}_{12})} \geq \frac{\binom{n}{2} \left(\mathbf{p}^T h(f(n)\mathbf{Q})\mathbf{p} - \frac{2kh(f(n)\mathbf{p}^T\mathbf{Q}\mathbf{p})}{n-1} \right)}{\binom{n'}{2} H(\mathbf{B}_{12})}. \tag{5.17}$$

We will now analyze the right hand side of (5.17) in two parameter regimes.

- $\underline{f(n) = 1}$: We have

$$\begin{aligned} H(\mathbf{B}_{12}) &\overset{(a)}{\leq} H(\mathbf{B}_{12}|X_1^{2k}) + H(X_1^{2k}) \\ &\leq H(\mathbf{B}_{12}|X_1^{2k}) + 2kH(\mathbf{p}) \\ &\overset{(b)}{=} k^2 H(A_{1,k}|X_1, X_k) + 2kH(\mathbf{p}) \\ &\leq k^2 \left(\mathbf{p}^T h(\mathbf{Q})\mathbf{p} + 2\frac{\log L}{k} \right), \end{aligned} \tag{5.18}$$

  where (a) follows from the chain rule and (b) follows since all elements of the matrix $\mathbf{B}_{12}$ are independent given $X_1, \cdots, X_{2k}$. Plugging this into the right hand side of (5.17) we obtain

$$\frac{H(\mathbf{B}_{\text{ut}})}{\binom{n'}{2} H(\mathbf{B}_{12})} \geq \frac{\binom{n}{2}\left(\mathbf{p}^T h(\mathbf{Q})\mathbf{p} - \frac{2kh(\mathbf{p}^T\mathbf{Q}\mathbf{p})}{n-1} \right)}{\binom{n'}{2} k^2 \left(\mathbf{p}^T h(\mathbf{Q})\mathbf{p} + 2\frac{\log L}{k} \right)}. \tag{5.19}$$

  Since $k = o(n), k = \omega(1)$ and $\binom{n'}{2}k^2 \sim \binom{n}{2}$, we have from (5.19)

$$\liminf_{n\to\infty} \frac{H(\mathbf{B}_{\text{ut}})}{\binom{n'}{2} H(\mathbf{B}_{12})} \geq 1, \tag{5.20}$$

  which together with (5.16) yields the required result.

- $\underline{f(n) = \Omega\left(\frac{1}{n^2}\right), f(n) = o(1)}$: Since $\mathbf{B}_{12}$ is a matrix of k^2 identically distributed Bernoulli random variables, we have

$$H(\mathbf{B}_{12}) \leq k^2 h(A_{1,k}) = k^2 h\left(f(n)\mathbf{p}^T\mathbf{Q}\mathbf{p} \right). \tag{5.21}$$

  Plugging this into the RHS of (5.17) then yields

$$\frac{H(\mathbf{B}_{\text{ut}})}{\binom{n'}{2} H(\mathbf{B}_{12})} \geq \frac{\binom{n}{2}\left(\mathbf{p}^T h(f(n)\mathbf{Q})\mathbf{p} - \frac{2kh(f(n)\mathbf{p}^T\mathbf{Q}\mathbf{p})}{n-1} \right)}{\binom{n'}{2} k^2 h\left(f(n)\mathbf{p}^T\mathbf{Q}\mathbf{p} \right)}. \tag{5.22}$$

  We first observe that in this parameter range, since $f(n) = o(1)$, we have by equations (5.10) and (5.12)

$$\mathbf{p}^T h(f(n)\mathbf{Q})\mathbf{p} \sim h\left(f(n)\mathbf{p}^T\mathbf{Q}\mathbf{p} \right). \tag{5.23}$$

Finally using that $k = o(n)$ and $\binom{n'}{2}k^2 \sim \binom{n}{2}$ establishes

$$\liminf_{n\to\infty} \frac{H(\mathbf{B}_{\mathrm{ut}})}{\binom{n'}{2}H(\mathbf{B}_{12})} \geq 1, \tag{5.24}$$

which together with (5.16) yields the required result.

□

5.3 Length of the Laplace Probability Assignment

Proof of Proposition 5. Let us first elaborate the relation between probability assignment and compression length. In Algorithm 1, the terms $\log(q(x_{j+1}|x^j))$ are added up, which lead to the marginal probability implied by the sequential probability assignment

$$\sum_{j=0}^{N-1} \log(q(x_{j+1}|x^j)) = \log\left(\prod_{j=0}^{N-1} q(x_{j+1}|x^j)\right) = \log(q(x^N)). \tag{5.25}$$

The compression output length of Algorithm 1 is $\left\lceil \log \frac{1}{q(x^N)} \right\rceil + 1$.

Now we analyze the compression length of Laplace compressor for the sequence $Z_1, Z_2, \ldots, Z_N$. Define $\theta_i \coloneqq \mathsf{P}(Z_1 = i), N_i \coloneqq \sum_{k=1}^{N} \mathbb{1}\{Z_k = i\}, i \in [m]$. We have

$$\begin{aligned}
&\log \frac{1}{q_{\mathrm{L}}(z^N)} \\
&= \log \frac{\theta_1^{N_1}\theta_2^{N_2}\cdots\theta_m^{N_m}}{q_{\mathrm{L}}(z^N)} + \log \frac{1}{\theta_1^{N_1}\theta_2^{N_2}\cdots\theta_m^{N_m}} \\
&= \log \binom{N+m-1}{m-1} + \log\left(\frac{N!}{N_1!N_2!\cdots N_m!}\theta_1^{N_1}\theta_2^{N_2}\cdots\theta_m^{N_m}\right) \\
&+ \log \frac{1}{\theta_1^{N_1}\theta_2^{N_2}\cdots\theta_m^{N_m}} \\
&\overset{(a)}{\leq} \log \binom{N+m-1}{m-1} + \log \frac{1}{\theta_1^{N_1}\theta_2^{N_2}\cdots\theta_m^{N_m}} \\
&\overset{(b)}{\leq} (m-1)\log\left(e\left(\frac{N}{m-1}+1\right)\right) + \log \frac{1}{\theta_1^{N_1}\theta_2^{N_2}\cdots\theta_m^{N_m}}
\end{aligned}$$

$$\leq m \log(2eN) + \sum_{i=1}^{m} N_i \log \frac{1}{\theta_i}, \tag{5.26}$$

where (a) follows since $\frac{N!}{N_1!N_2!\cdots N_m!}\theta_1^{N_1}\theta_2^{N_2}\cdots\theta_m^{N_m}$ is a multinomial probability which is always upper bounded by 1, and (b) follows since $\binom{n}{k} \leq \left(\frac{en}{k}\right)^k$. Taking expectation on both sides of (5.26), we obtain

$$\begin{aligned}
\mathsf{E}\left[\log \frac{1}{q_{\mathrm{L}}(Z^N)}\right] &\leq m \log(2eN) + \sum_{i=1}^{m} \mathsf{E}[N_i] \log \frac{1}{\theta_i} \\
&\overset{(a)}{=} m \log(2eN) + \sum_{i=1}^{m} N\theta_i \log \frac{1}{\theta_i} \\
&= m \log(2eN) + NH(Z_1),
\end{aligned}$$

where (a) follows since $\mathsf{E}[N_i] = \sum_{k=1}^{N} \mathsf{E}[\mathbb{1}\{Z_k = i\}] = N\mathsf{P}(Z_1 = i)$ since the Z_i are identically distributed. Finally, we have

$$\mathsf{E}[\ell_{\mathrm{L}}(Z^N)] \leq \mathsf{E}\left[\log \frac{1}{q_{\mathrm{L}}(Z^N)}\right] + 2 \leq m \log(2eN) + NH(Z_1) + 2$$

as required. □

5.4 Length of the KT Probability Assignment

Lemma 2. *For any integer $m > 0$, $N_1, N_2, \cdots N_m \in \mathbb{N}$ and probability distribution $(\theta_1, \cdots \theta_m)$,*

$$\frac{\binom{N}{N_1, N_2 \cdots N_m}\theta_1^{N_1}\cdots\theta_m^{N_m}}{\binom{2N}{2N_1, 2N_2 \cdots 2N_m}\theta_1^{2N_1}\cdots\theta_m^{2N_m}} \geq 1,$$

where $N = \sum_{i=1}^{m} N_i$.

Remark 7. Equivalently, consider an urn containing known number of balls with m different colours. The lemma claims that the probability of getting N_1 balls of colour 1, N_2 of balls of colour 2, $\cdots$ N_m balls of colour m out of N draws with replacement is always greater than the probability of getting $2N_1$ balls of colour 1, $2N_2$ of balls of colour 2, $\cdots$ $2N_m$ balls of colour m out of $2N$ draws with replacement.

Proof. Let $p_1 = N_1/N, p_2 = N_2/N, \cdots, p_m = N_m/N$. Notice that $\sum_{i=1}^m p_i = 1$, so $(p_1, \cdots p_m)$ can be viewed as a probability distribution. And the entropy of this distribution is $H(p_1, \cdots p_m) = \sum_{i=1}^m -p_i \log p_i$. Firstly we consider the case when $N_1, N_2 \cdots N_m$ are all positive and none of them equal to N. By Stirling's approximation for factorial $\sqrt{2\pi n}(\frac{n}{e})^n e^{1/(12n+1)} \leq n! \leq \sqrt{2\pi n}(\frac{n}{e})^n e^{1/12n}$, we can bound

$$\begin{aligned}\binom{N}{N_1, N_2 \cdots N_m} &\geq \frac{\sqrt{2\pi N} N^N \exp\left(\frac{1}{12N+1} - \frac{1}{12N_1} - \frac{1}{12N_2} - \cdots - \frac{1}{12N_m}\right)}{(2\pi)^{m/2}(N_1 N_2 \cdots N_m)^{1/2} N_1^{N_1} N_2^{N_2} \cdots N_m^{N_m}} \\ &= \frac{\exp\left(\frac{1}{12N+1} - \frac{1}{12N_1} - \frac{1}{12N_2} - \cdots - \frac{1}{12N_m}\right)}{(2\pi)^{\frac{m-1}{2}}(p_1 p_2 \cdots p_m)^{1/2} N^{\frac{m-1}{2}} 2^{-NH(p_1, p_2, \cdots, p_m)}}.\end{aligned}$$

Similarly, we have

$$\binom{2N}{2N_1, 2N_2 \cdots 2N_m} \leq \frac{\exp\left(\frac{1}{24N} - \frac{1}{24N_1+1} - \frac{1}{24N_2+1} - \cdots - \frac{1}{24N_m+1}\right)}{(2\pi)^{\frac{m-1}{2}} 2^{\frac{m-1}{2}}(p_1 p_2 \cdots p_m)^{1/2} N^{\frac{m-1}{2}} 2^{-2NH(p_1 \cdots p_m)}}.$$

Consider the function

$$\begin{aligned}f(N_1, N_2, \cdots, N_m) =& \tfrac{1}{12N+1} - \tfrac{1}{24N} + (\tfrac{1}{24N_1+1} - \tfrac{1}{12N_1}) + (\tfrac{1}{24N_2+1} - \tfrac{1}{12N_2}) \\ &+ \cdots + (\tfrac{1}{24N_m+1} - \tfrac{1}{12N_m})\end{aligned}$$

and the function

$$g(n) = \frac{1}{24n+1} - \frac{1}{12n},$$

where n is a positive integer. Function $g(n)$ is minimized with $n = 1$ and $\min g(n) = 1/25 - 1/12$ and we can bound function $f(N_1, N_2, \cdots, N_m) \geq \frac{1}{12N+1} - \frac{1}{24N} + (1/25 - 1/12)m$. Finally we are ready to prove the lemma.

$$\begin{aligned}\frac{\binom{N}{N_1, N_2 \cdots N_m}\theta_1^{N_1} \cdots \theta_m^{N_m}}{\binom{2N}{2N_1, 2N_2 \cdots 2N_m}\theta_1^{2N_1} \cdots \theta_m^{2N_m}} &\geq \frac{2^{\frac{m-1}{2}} \exp(f(N_1, N_2, \cdots, N_m))}{2^{NH(p_1 \cdots p_m)}\theta_1^{N_1} \cdots \theta_m^{N_m}} \\ &\geq \frac{2^{\frac{m-1}{2}} \exp\left(\frac{1}{12N+1} - \frac{1}{24N} + (1/25 - 1/12)m\right)}{2^{-ND_{\mathrm{KL}}(p||\theta)}} \\ &= 2^{\frac{m-1}{2}} 2^{ND_{\mathrm{KL}}(p||\theta)} 2^{\log e(\frac{1}{12N+1} - \frac{1}{24N} + (1/25-1/12)m)}.\end{aligned}$$

Notice that $\frac{1}{12N+1} - \frac{1}{24N}$ goes to zero when $N \to \infty$, $\frac{m-1}{2} > (1/25 - 1/12)m$ and $D_{\mathrm{KL}}(P||\theta) \geq 0$. Therefore in this case,

$$\frac{\binom{N}{N_1, N_2 \cdots N_m}\theta_1^{N_1} \cdots \theta_m^{N_m}}{\binom{2N}{2N_1, 2N_2 \cdots 2N_m}\theta_1^{2N_1} \cdots \theta_m^{2N_m}} \geq 1.$$

When one of $\{N_i\}_{i=1}^N$ equals to N, without loss of generality, we assume that $N_1 = N$. We have

$$\frac{\binom{N}{N_1,N_2\cdots N_m}\theta_1^{N_1}\cdots\theta_m^{N_m}}{\binom{2N}{2N_1,2N_2\cdots 2N_m}\theta_1^{2N_1}\cdots\theta_m^{2N_m}} = \frac{1}{\theta_1^{N_1}\cdots\theta_m^{N_m}} > 1.$$

When there are k numbers out of $N_1, N_2, \cdots, N_m$ that equal to zero, we can simply remove these values and consider the case with alphabet size $m-k$. And this will yield the same result. □

Proof of Proposition 6. In this proof, we define a generalized form of factorial function. Let x be a positive integer, $(x+\frac{1}{2})! = \frac{1}{2}\frac{3}{2}\cdots(x+\frac{1}{2})$. Since $(2N_1-1)!! = \frac{(2N_1)!}{2^{N_1}(N_1)!}$, we have

$$\begin{aligned}
&m(m+2)\cdots(m+2N-2)\\
&= 2^N\left(\frac{m}{2}\right)\left(\frac{m+2}{2}\right)\cdots\left(\frac{m+2N-2}{2}\right) = 2^N\frac{\left(\frac{m}{2}+N-1\right)!}{\left(\frac{m}{2}-1\right)!}.
\end{aligned}$$

Therefore we can rewrite the KT probability assignment in (4.9) as

$$\begin{aligned}
q_{\mathrm{KT}}(z^N) &= \frac{(\frac{m}{2}-1)!}{2^N(\frac{m}{2}+N-1)!}\frac{\binom{2N}{N}}{\binom{2N}{N}}\prod_{i=1}^m\frac{(2N_i)!}{N_i!2^{N_i}}\\
&= \frac{(\frac{m}{2}-1)!}{2^N(\frac{m}{2}+N-1)!}\binom{2N}{N}N!\frac{N!}{(2N)!}\prod_{i=1}^m\frac{(2N_i)!}{N_i!2^{N_i}}\\
&\overset{(a)}{\geq} \frac{\left(\frac{m}{2}-1\right)!\binom{2N}{N}}{4^N(N+\frac{m}{2}-\frac{1}{2})^{\frac{m-1}{2}}}\frac{N!}{(2N)!}\prod_{i=1}^m\frac{(2N_i)!}{N_i!}\\
&\overset{(b)}{=} \frac{\theta_1^{N_1}\cdots\theta_m^{N_m}(\frac{m}{2}-1)!\binom{2N}{N}}{4^N(N+\frac{m}{2}-\frac{1}{2})^{\frac{m-1}{2}}}\frac{\binom{N}{N_1,N_2\cdots N_m}\theta_1^{N_1}\cdots\theta_m^{N_m}}{\binom{2N}{2N_1,2N_2\cdots 2N_m}\theta_1^{2N_1}\cdots\theta_m^{2N_m}},
\end{aligned}$$

where (a) follows that when m is even, $\frac{N!}{(\frac{m}{2}+N-1)!} = \frac{1}{(N+1)\cdots(\frac{m}{2}+N-1)} \geq \frac{1}{(N+\frac{m}{2}-\frac{1}{2})^{\frac{m-1}{2}}}$ and when m is odd, $\frac{N!}{(\frac{m}{2}+N-1)!} \geq \frac{N!}{(\frac{m}{2}+N-\frac{1}{2})!} = \frac{1}{(N+1)\cdots(\frac{m}{2}+N-\frac{1}{2})} \geq \frac{1}{(N+\frac{m}{2}-\frac{1}{2})^{\frac{m-1}{2}}}$, (b) follows that $\binom{N}{N_1,N_2\cdots N_m} = \frac{N!}{\prod_{i=1}^m N_i!}$ and $\theta_i \triangleq \mathbb{P}(Z_1 = i)$. By lemma 2, we have $q_{\mathrm{KT}}(z^N) \geq \frac{\theta_1^{N_1}\cdots\theta_m^{N_m}(\frac{m}{2}-1)!\binom{2N}{N}}{4^N(N+\frac{m}{2}-\frac{1}{2})^{\frac{m-1}{2}}}$. Thus,

$$\log\frac{1}{q_{\mathrm{KT}}(z^N)}$$

$$
\begin{aligned}
&\le \log \frac{1}{\theta_1^{N_1}\cdots\theta_m^{N_m}} + \log \frac{4^N (N+\frac{m}{2}-\frac{1}{2})^{\frac{m-1}{2}}}{(\frac{m}{2}-1)!\binom{2N}{N}} \\
&= \log \frac{1}{\theta_1^{N_1}\cdots\theta_m^{N_m}} + \frac{m-1}{2}\log\left(N+\frac{m-1}{2}\right) + \log\frac{4^N}{\binom{2N}{N}} - \log\Big(\frac{m}{2}-1\Big)! \\
&\overset{(a)}{\le} \log \frac{1}{\theta_1^{N_1}\cdots\theta_m^{N_m}} + \frac{m-1}{2}\log\left(N+\frac{m-1}{2}\right) + \log\frac{4^N}{\binom{2N}{N}} \\
&\quad - \Big(\frac{m}{2}-1\Big)\log(\frac{\frac{m}{2}-1}{e}) \\
&\overset{(b)}{\sim} \log \frac{1}{\theta_1^{N_1}\cdots\theta_m^{N_m}} + \frac{m-1}{2}\log\left(N+\frac{m-1}{2}\right) + \log\sqrt{\pi N} \\
&\quad - \Big(\frac{m}{2}-1\Big)\log\left(\frac{\frac{m}{2}-1}{e}\right) \\
&\sim \frac{m}{2}\log\frac{e(\frac{m}{2}+N)}{m/2} + \log\sqrt{\pi N} + \log\frac{1}{\theta_1^{N_1}\cdots\theta_m^{N_m}} \\
&= \frac{m}{2}\log\big(e\big(1+\tfrac{2N}{m}\big)\big) + \frac{1}{2}\log(\pi N) + \sum_{i=1}^{m} N_i \log\frac{1}{\theta_i},
\end{aligned}
$$

where (a) follows Stirling's approximation $k! \ge \sqrt{2\pi k}(\frac{k}{e})^k e^{\frac{1}{12k+1}}$ and (b) follows Stirling's approximation for binomial coefficient, i.e., $\binom{2N}{N} \sim \frac{4^N}{\sqrt{\pi N}}$. Therefore, we have

$$
\mathsf{E}\left[\log\frac{1}{q_{\mathrm{KT}}(z^N)}\right] \le \frac{1}{2}m\log\big(e\big(1+\tfrac{2N}{m}\big)\big) + \frac{1}{2}\log(\pi N) + NH(Z_1).
$$

Finally, it follows that

$$
\begin{aligned}
\mathsf{E}[\ell_{\mathrm{KT}}(Z^N)] &\le \mathsf{E}\left[\log\frac{1}{q_{\mathrm{KT}}(z^N)}\right] + 2 \\
&\le \tfrac{m}{2}\log\big(e\big(1+\tfrac{2N}{m}\big)\big) + \tfrac{1}{2}\log(\pi N) + NH(Z_1) + 2.
\end{aligned}
$$

□

Chapter 6

Minimax Redundancy Analysis

Proof of Theorem 3. Firstly, we prove the lower bound for the minimax redundancy. The proof essentially follows from the *redundancy-capacity theorem* [34]. To start with, we will define a vector $\mathbf{\Theta}$ of parameters $L, \mathbf{p}$ and $\mathbf{Q}$ for SBM. Consider stochastic block model $\mathrm{SBM}(n, L, \mathbf{p}, f(n)\mathbf{Q})$. Recall that L is a scalar. Vector $\mathbf{p} = (p_1, \ldots, p_L)^T$ represents a distribution over $[L]$, so it can be written as $\mathbf{p} = (p_1, \ldots, p_{L-1}, 1 - \sum_{i=1}^{L-1} p_i)^T$. For the L-by-L symmetric matrix $\mathbf{Q}$, we can write the upper-triangle entries into a vector $(Q_{11}, \ldots, Q_{1L}, Q_{22}, \ldots, Q_{2L}, \ldots, Q_{LL})$. We define parameter vector

$$\begin{aligned}&\mathbf{\Theta}(\mathrm{SBM}(n, L, \mathbf{p}, f(n)\mathbf{Q}))\\ &:= (L, p_1, \ldots, p_{L-1}, Q_{11}, \ldots, Q_{1L}, Q_{22}, \ldots, Q_{2L}, \ldots, Q_{LL}).\end{aligned}$$

Notice that by our definition, $\mathbf{\Theta}$ is a variable-length vector. Its length $L + \frac{L(L+1)}{2}$ is a function of its first entry L. For some fixed $f(n)$ and $Q_{\max}$, let

$$\mathcal{W} := \{\mathbf{\Theta}(\mathrm{SBM}(n, L, \mathbf{p}, f(n)\mathbf{Q})) : \mathrm{SBM}(n, L, \mathbf{p}, f(n)\mathbf{Q}) \in \mathscr{P}_3(f(n), Q_{\max})\}$$

denote the set of all parameter vectors for the family of SBMs $\mathscr{P}_3(f(n), Q_{\max})$. Let $\mu(\cdot)$ denote a probability distribution supported on $\mathcal{W}$. Suppose $A_n \sim \mathrm{SBM}(n, L, \mathbf{p}, f(n)\mathbf{Q})$ with random SBM parameters following prior μ, by redundancy-capacity theorem (see, for example, [34]), we have

$$R_n^*(\mathscr{P}_3(f(n), Q_{\max})) = \sup_{\mu} I_\mu(\mathbf{\Theta}; A_n),$$

where $I_\mu(\mathbf{\Theta}; A_n)$ denote the mutual information between $\mathbf{\Theta}$ and A_n under the prior μ and the supremum is taken over all possible probability distributions supported on $\mathcal{W}$. To lower bound the quantity $\sup_\mu I_\mu(\mathbf{\Theta}; A_n)$, we consider a particular distribution $\nu \in \mathcal{W}$ such that L is fixed to 1 and $Q_{11} \sim \mathrm{Unif}(0, Q_{\max})$, i.e, the Erdős–Rényi model with edge probability

being uniform from $(0, f(n)Q_{\max}]$. Since ν is supported on $\mathcal{W}$, we have $\sup_\mu I_\mu(\mathbf{\Theta}; A_n) \geq I_\nu(\mathbf{\Theta}; A_n)$.

We now move on to evaluate the mutual information $I_\nu(\mathbf{\Theta}; A_n)$. For an adjacency matrix A_n, we define $\hat{Q}_{11} := \frac{K}{\binom{n}{2} f(n)}$ where $K = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} A_{ij}$ denotes the number of ones in upper-triangle of A_n. We have

$$\begin{aligned}
I_\nu(\mathbf{\Theta}; A_n) &= I_\nu(Q_{11}; A_n) \\
&= h(Q_{11}) - h(Q_{11}|A_n) \\
&\overset{(a)}{\geq} \log Q_{\max} - h(Q_{11}|\hat{Q}_{11}) \\
&= \log Q_{\max} - h(Q_{11} - \hat{Q}_{11}|\hat{Q}_{11}) \\
&\geq \log Q_{\max} - h(Q_{11} - \hat{Q}_{11}) \\
&\overset{(b)}{\geq} \log Q_{\max} - \frac{1}{2} \log \left(2\pi e \mathrm{Var}(Q_{11} - \hat{Q}_{11}) \right),
\end{aligned}$$

where (1) follows since $\hat{Q}_{11}$ is a function of A_n and (b) follows since Gaussian distribution maximizes differential entropy for a given variance. To evaluate the variance $\mathrm{Var}(Q_{11} - \hat{Q}_{11})$, we consider the conditional variance

$$\begin{aligned}
\mathrm{Var}(Q_{11} - \hat{Q}_{11}|Q_{11} = q) &= \mathsf{E}\left[\left(\frac{K}{\binom{n}{2} f(n)} - Q_{11} \right)^2 \middle| Q_{11} = q \right] \\
&= \frac{1}{\binom{n}{2}^2 (f(n))^2} \mathsf{E}\left[\left(K - \binom{n}{2} f(n) q \right)^2 \middle| Q_{11} = q \right] \\
&\overset{(c)}{=} \frac{\binom{n}{2} f(n) q (1 - f(n)q)}{\binom{n}{2}^2 (f(n))^2} \\
&= \frac{q(1 - f(n)q)}{\binom{n}{2} f(n)},
\end{aligned}$$

where (c) follows since $K|Q_{11} = q \sim \mathrm{Binom}(\binom{n}{2}, f(n)q)$. By law of total variance and since $\mathsf{E}[\hat{Q}_{11} - Q_{11}|Q_{11} = q] = 0$, we have

$$\mathrm{Var}(Q_{11} - \hat{Q}_{11}) = \mathsf{E}\left[\frac{Q_{11}(1 - f(n)Q_{11})}{\binom{n}{2} f(n)} \right] \leq \frac{\mathsf{E}[Q_{11}]}{\binom{n}{2} f(n)} = \frac{Q_{\max}}{2\binom{n}{2} f(n)}.$$

Finally, we have

$$\begin{aligned} I_\nu(\mathbf{\Theta}; A_n) &\geq \log Q_{\max} - \frac{1}{2}\log\left(2\pi e \mathrm{Var}(Q_{11} - \hat{Q}_{11})\right) \\ &\geq \log Q_{\max} - \frac{1}{2}\log\left(\frac{Q_{\max}\pi e}{\binom{n}{2} f(n)}\right) \\ &= \frac{1}{2}\log\left(\frac{\binom{n}{2} f(n) Q_{\max}}{\pi e}\right), \end{aligned}$$

which proves the lower bound.

Now, we prove the upper bound for the minimax redundancy. Recall that the minimax redundancy is defined as $R_n^*(\mathscr{P}_3(f(n), Q_{\max})) = \inf_C \sup_{L,\mathbf{p},\mathbf{Q}} (\mathsf{E}[\ell(C(A_n))] - H(A_n))$, where $L, \mathbf{p}, \mathbf{Q}$ are parameters of SBM in the family $\mathscr{P}_3(f(n), Q_{\max})$. Since our proposed compressor C_k is a valid lossless compressor for $\mathscr{P}_3$, we can upper bound the minimax redundancy with the maximum redundancy of our compressor. By Theorem 1, we have

$$\begin{aligned} &R_n^*(\mathscr{P}_3(f(n), Q_{\max})) \\ &= \inf_C \sup_{L,\mathbf{p},\mathbf{Q}} (\mathsf{E}[\ell(C(A_n))] - H(A_n)) \\ &\leq \sup_{L,\mathbf{p},\mathbf{Q}} (\mathsf{E}[\ell(C_k(A_n))] - H(A_n)) \\ &\overset{(a)}{\leq} \sup_{L,\mathbf{p},\mathbf{Q}} \binom{n}{2} f(n) \left(\mathbf{p}^T\mathbf{Q}\mathbf{p} \log \frac{1}{\mathbf{p}^T\mathbf{Q}\mathbf{p}} - \mathbf{p}^T\mathbf{Q}^*\mathbf{p}\right) + o(n^2 f(n)), \end{aligned}$$

where (a) follows since $\mathscr{P}_3$ is a sub-family of $\mathscr{P}_2$, so $f(n) = o(1)$. Therefore, it suffices to maximize the constant term $\left(\mathbf{p}^T\mathbf{Q}\mathbf{p} \log \frac{1}{\mathbf{p}^T\mathbf{Q}\mathbf{p}} - \mathbf{p}^T\mathbf{Q}^*\mathbf{p}\right)$ over all valid choices of $L, \mathbf{p}, \mathbf{Q}$. We can rewrite the term as

$$\begin{aligned} &\mathbf{p}^T\mathbf{Q}\mathbf{p} \log \frac{1}{\mathbf{p}^T\mathbf{Q}\mathbf{p}} - \mathbf{p}^T\mathbf{Q}^*\mathbf{p} \\ &= \sum_{i,j\in[L]} p_i p_j Q_{ij} \log \frac{1}{\sum_{k,l\in[L]} p_k p_l Q_{kl}} - \sum_{i,j\in[L]} p_i p_j Q_{ij} \log \frac{1}{Q_{ij}} \\ &= Q_{\max}\left(\sum_{i,j\in[L]} p_i p_j \frac{Q_{ij}}{Q_{\max}} \log \frac{1}{\sum_{k,l\in[L]} p_k p_l Q_{kl}} - \sum_{i,j\in[L]} p_i p_j \frac{Q_{ij}}{Q_{\max}} \log \frac{1}{Q_{ij}}\right) \\ &= Q_{\max}\left(\sum_{i,j\in[L]} p_i p_j \frac{Q_{ij}}{Q_{\max}} \log \left(\frac{1}{\sum_{k,l\in[L]} p_k p_l \frac{Q_{kl}}{Q_{\max}}} \frac{1}{Q_{\max}}\right)\right. \end{aligned}$$

$$
\begin{aligned}
&\quad - \sum_{i,j\in[L]} p_i p_j \frac{Q_{ij}}{Q_{\max}} \log\left(\frac{Q_{\max}}{Q_{ij}} \frac{1}{Q_{\max}}\right)\Bigg) \\
&= Q_{\max}\Bigg(\sum_{i,j\in[L]} p_i p_j \frac{Q_{ij}}{Q_{\max}} \log \frac{1}{\sum_{k,l\in[L]} p_k p_l \frac{Q_{kl}}{Q_{\max}}} \\
&\quad - \sum_{i,j\in[L]} p_i p_j \frac{Q_{ij}}{Q_{\max}} \log \frac{Q_{\max}}{Q_{ij}}\Bigg) \\
&= Q_{\max}\left(\mathbf{p}^T \mathbf{Q}'\mathbf{p} \log \frac{1}{\mathbf{p}^T\mathbf{Q}'\mathbf{p}} - \mathbf{p}^T (\mathbf{Q}')^* \mathbf{p}\right),
\end{aligned}
$$

where $\mathbf{Q}'$ denotes the matrix $\mathbf{Q}/Q_{\max}$ and $(\mathbf{Q}')^*$ denotes an $L \times L$ matrix whose (i,j) entry is $Q'_{ij} \log(\frac{1}{Q'_{ij}})$. By our assumptions, the entries in $\mathbf{Q}'$ are all non-negative and the largest entry is 1. Since $\mathbf{p}$ is a probability distribution, we have $0 < \mathbf{p}^T\mathbf{Q}'\mathbf{p} \leq 1$. Now, let us consider the function $g(x) = x \log \frac{1}{x}$. When $0 \leq x \leq 1$, its supremum is obtained at $x = 1/e$ and its infimum is obtained at $x = 0$ or $x = 1$. With these properties, we can bound our constant term

$$
Q_{\max}\left(\mathbf{p}^T \mathbf{Q}'\mathbf{p} \log \frac{1}{\mathbf{p}^T\mathbf{Q}'\mathbf{p}} - \mathbf{p}^T (\mathbf{Q}')^* \mathbf{p}\right) \leq Q_{\max} \frac{1}{e} \log e.
$$

In particular, the equality can be achieved when all the entries in $\mathbf{Q}'$ are either zero or one and $\mathbf{p}$ is carefully chosen such that $\mathbf{p}^T\mathbf{Q}'\mathbf{p} = 1/e$. This completes the proof for the upper bound. □

Chapter 7

Performance under the Local Weak Convergence Framework

In this Chapter, we analyze the performance of the proposed compressor under the local weak convergence framework introduced in [11, 18]. We focus on a subfamily of SBM in the $f(n) = 1/n$ regime, whose expected degree is identical for vertices in all communities, i.e., $\mathbf{Qp} = (\lambda, \ldots, \lambda)^T$. We will show that the proposed compressor achieves the same performance in terms of BC entropy as the compressor proposed in [19].

We first introduce some basic definitions on rooted graphs in Section 7.1. Then, we define the local weak convergence of graphs and derive the local weak convergence limit of the subfamily of stochastic block model in Section 7.2. Finally, we review the definition of BC entropy in Section 7.3 and state the performance guarantee of our compression algorithm in Section 7.4.

7.1 Basic Definitions on Rooted Graphs

Let $G = (V, E)$ be a simple graph (undirected, unweighted, no self-loop), with V a countable set of vertices and E a countable set of edges. Let $u \overset{G}{\sim} v$ denote the connectivity of vertices u and v in G. G is said to be *locally finite* if, for all $v \in V$, the degree of v in G is finite. A rooted graph (G, o) is a locally finite and connected graph $G = (V, E, o)$ with a distinguished vertex $o \in V$, called the root. Two rooted graphs $(G_1, o_1) = (V_1, E_1, o_1)$ and $(G_2, o_2) = (V_2, E_2, o_2)$ are *isomorphic*, denoted as $(G_1, o_1) \simeq (G_2, o_2)$, if there exists a bijection $\pi : V_1 \to V_2$ such that $\pi(o_1) = o_2$ and $u \overset{G_1}{\sim} v$ if and only if $\pi(u) \overset{G_2}{\sim} \pi(v)$ for all $u, v \in V_1$. One can verify that this notion of isomorphism defines an equivalence relation on rooted graphs. Let $[G, o]$ denote the equivalence class corresponding to (G, o). Let $\mathcal{G}^*$ denote the set of all locally finite and connected rooted graphs. For $(G, o) \in \mathcal{G}^*$ and $h \in \mathbb{N}$,

we write $(G, o)_h$ for the truncated graph at depth h of the graph (G, o), in other words, the induced subgraph on the vertices such that their distance from the root is less than or equal to h. The equivalence classes $[G, o]_h$ follows the similar definition. Let $\mathcal{G}_h^*$ denote the set of all $[G, o]_h$. Now, we define the metric d^* on $\mathcal{G}^*$. For any $[G_1, o_1]$ and $[G_2, o_2]$, let

$$\begin{aligned}\hat{h} := \sup\{h \in \mathbb{Z}^+ : (G_1, o_1)_h \simeq (G_2, o_2)_h \\ \text{for some } (G_1, o_1) \in [G_1, o_1], (G_2, o_2) \in [G_2, o_2]\}\end{aligned}$$

and define the metric d^* as

$$d^*([G_1, o_1], [G_2, o_2]) := \frac{1}{1 + \hat{h}}.$$

As shown in [19], equipped with the metric defined above, $\mathcal{G}^*$ is a Polish space, i.e, a complete separable metric space. For this Polish space, let $\mathcal{P}(\mathcal{G}^*)$ denote the Borel probability measures on it. We say that a sequence of measures $\mu_n \in \mathcal{P}(\mathcal{G}^*)$ *converges weakly* to $\mu \in \mathcal{P}(\mathcal{G}^*)$, written as $\mu_n \rightsquigarrow \mu$, if for any bounded continuous function f on $\mathcal{G}^*$, we have $\int f d\mu_n \to \int f d\mu$. It was shown in [8] that $\mu_n \rightsquigarrow \mu$ if for any uniformly continuous and bounded functions f, we have $\int f d\mu_n \to \int f d\mu$. For $\mu \in \mathcal{P}(\mathcal{G}^*)$, $h \in \{0, 1, 2, \ldots\}$, and $[G, o] \in \mathcal{G}^*$, let μ_h denote the h-neighbourhood marginal of μ

$$\mu_h([G, o]) = \sum_{[G', o] \in \mathcal{G}^* : [G', o]_h = [G, o]} \mu([G', o]).$$

For a locally finite graph $G = (V, E)$ and a vertex $v \in V$, let $G(v)$ denote the graph component in G that is connected to v. By our previous definitions, $(G(v), v)$ denotes the rooted graph of the connected component of v and the root is located at v and $[G(v), v]$ denotes the equivalence class corresponding to $(G(v), v)$. Now, the *rooted neighbourhood distribution* of G is defined as the distribution of the rooted graph when the root is chosen uniformly at random over V

$$U(G) := \frac{1}{|V|} \sum_{v \in V} \delta_{[G(v), v]}, \tag{7.1}$$

where δ is the Dirac delta function.

7.2 Local Weak Convergence

For our study of stochastic block model, which is a sequence of *random* graphs $\{A_n\}_{n=1}^{\infty}$, $U(A_n)$ as defined in (7.1) becomes a random distribution.

In the section, we establish the asymptotic behavior of the expected neighbourhood distribution $\mathsf{E}U(A_n)$, averaged over the randomness of the graph A_n, for a family of stochastic block models in the sparse regime. We will show in Proposition 8 that the neighbourhood distribution converges, in the local weak sense, to a Galton–Watson tree distribution.

Let $B \subseteq \mathcal{G}^*$ be a measurable event in $\mathcal{G}^*$. By the exchangability of stochastic block model, we have $\mathsf{E}U(A_n)(B) = \frac{1}{n}\sum_{i=1}^{n} \mathsf{P}([A_n(i), i] \in B) = \mathsf{P}([A_n(1), 1] \in B)$, in other words, $\mathsf{E}U(A_n)$ is simply the neighbourhood distribution at any vertex ρ in the graph A_n, denoted $\mathrm{Nbr}(\rho)$. Now, we study the R-neighbourhood marginal of $\mathrm{Nbr}(\rho)$, i.e. $\mathrm{Nbr}_R(\rho)$, under certain assumptions on the parameters of our stochastic block model. Since this distribution is identical for every vertex in $V(A_n)$, we will simply write Nbr_R for the R-neighbourhood distribution of any vertex.

To state the limiting distribution, we need to define the *Galton–Watson tree* probability distribution on rooted trees $\mathrm{GWT}(\mathrm{Poi}(\lambda))$. Let $\mathrm{Poi}(\lambda)$ denote the Poisson distribution with mean λ. We take a vertex as the root and generate $Z^{(1)} \sim \mathrm{Poi}(\lambda)$ as the number of children of the first generation. For the first generation, independent of $Z^{(1)}$, we generate $\xi_1^{(1)}, \ldots, \xi_{Z^{(1)}}^{(1)}$ i.i.d. according to $\mathrm{Poi}(\lambda)$ as the number of children of each vertex in the first generation. Let $Z^{(2)} = \sum_{i=1}^{Z^{(1)}} \xi_i^{(1)}$ denote the total number of vertices in the first generation. In general, for the jth generation, $j = 1, 2, \ldots$, generate the number of children for each vertex in the jth generation $\xi_1^{(j)}, \ldots, \xi_{Z^{(j)}}^{(j)}$ i.i.d. according to $\mathrm{Poi}(\lambda)$, independent of all previous variables $\{\xi_1^{(i-1)}, \ldots, \xi_{Z^{(i-1)}}^{(i-1)}, Z^{(i)}, \text{ for all } i \le j\}$. Let $Z^{(j+1)} = \sum_{k=1}^{Z^{(j)}} \xi_k^{(j)}$ denote the total number of vertices in the jth generation. In this way, we iteratively defined a measure on rooted trees.

Remember that the total variation distance between two probability measures μ_1 and μ_2 is defined as $d_{\mathrm{TV}}(\mu_1, \mu_2) := \sup_{g:\mathcal{G}^* \to [-1,1]} \left| \int g d\mu_1 - \int g d\mu_2 \right|$. With the definitions above, we are ready to state the proposition that upper bounds the total variation distance between Nbr_R and a Galton–Watson tree.

Proposition 7. *Let* $A_n \sim \mathrm{SBM}(n, L, \mathbf{p}, \frac{1}{n}\mathbf{Q})$ *with* $\mathbf{Qp} = \lambda \mathbf{1}$ *for some positive constant* λ *and an all-*1 *vector* $\mathbf{1}$ *of dimension* $L \times 1$. *Let* $R = \left\lfloor \frac{1}{10 \log(2Q_{\max})} \log n \right\rfloor$, *where* $Q_{\max} \triangleq \max_{i,j} Q_{ij}$ *denotes the largest entry in* $\mathbf{Q}$. *Then, the total variation distance satisfies*

$$d_{\mathrm{TV}}(\mathrm{Nbr}_R, \mathrm{GWT}(\mathrm{Poi}(\lambda))_R) \to 0$$

as $n \to \infty$.

The proof of Proposition 7 follows the similar idea as the proof of Proposition 2 in [35]. The key idea essentially follows from the fact that $\mathrm{GWT}(\mathrm{Poi}(\lambda))$ can be constructed from a sequence of Poisson random variables with mean λ, while Nbr_R can be constructed from a sequence of binomial random variables with approximately the same mean λ. The case when $R = 1$ follows essentially from the Poisson approximation of Binomial in the $1/n$ probability regime. This Proposition generalizes it to $R = O(\log n)$.

To upper bound the total variation distance between these two distributions, we first review the definition of coupling between two probability distributions. Let μ and ν be probability measures on the same measurable space $(S, \mathcal{S})$. A coupling of μ and ν is a probability measure γ on the product space $(S \times S, \mathcal{S} \times \mathcal{S})$ such that the marginal of γ coincide with μ and ν, i.e.,

$$\gamma(A \times S) = \mu(A) \; and \; \gamma(S \times A) = \nu(A), \; \forall A \in \mathcal{S}.$$

It's known that for two probability distributions μ and ν on $(S, \mathcal{S})$. For any coupling (X, Y) for μ and ν,

$$d_{\mathrm{TV}}(\mu, \nu) \leq \mathsf{P}(X \neq Y),$$

(see, e.g., [44]). Therefore, it suffices to construct a coupling (G_R, T_R) such that the marginal distribution of G_R is Nbr_R, the marginal distribution of T_R is $\mathrm{GWT}(\mathrm{Poi}(\lambda))_R$ and $\mathsf{P}(G_R \neq T_R) \to 0$. To construct such a coupling, we will present an induction process in which each step holds with high probability.

To formally state the proof of Proposition 7, we need a few more definitions. Let $V = V(A_n)$ denote the set of vertices in the graph A_n. For a vertex $\rho \in V$ and an integer $r \in [R]$, let $G_r(\rho)$ denote the (random) r-neighbourhood of ρ, i.e, $G_r(\rho)$ is the induced subgraph on the vertex set $\{v \in V : d(v, \rho) \leq r\}$. We will simply write G_r for $G_r(\rho)$. Let $V_r \triangleq V \setminus V(G_r)$ denote the set of vertices in V that are not in $V(G_r)$. Let $\partial G_r \triangleq \{v \in V : d(v, \rho) = r\}$ denote the depth-r neighbourhood of ρ. For a vertex $v \in \partial G_r$, let Y_v denote the number of neighbours v has in V_r. Let T_R be a random tree generated from $\mathrm{GWT}(\mathrm{Poi}(\lambda))_R$, i.e., $T_R \sim \mathrm{GWT}(\mathrm{Poi}(\lambda))_R$. For a vertex $u \in V(T_R)$, let Z_u denote the number of children of u in T_R. In order to couple G_R with T_R such that $\mathsf{P}(G_R \neq T_R) \to 0$, a necessary condition is that G_R is a tree with high probability. Now, we define two sets of events to guarantee that G_R is a tree. For any $r \in [R]$, let C_r denote the event that no vertex in V_{r-1} has more than one neighbour in G_{r-1} and let D_r denote the event that there are no edges within ∂G_r. Clearly, if C_r and D_r holds for every $r \in [R]$, then G_R is indeed a tree. With these definitions, we state a lemma that introduces the induction step.

Lemma 3. *If*

1. $G_{r-1} = T_{r-1}$;
2. $Y_u = Z_u$ *for every* $u \in \partial G_{r-1}$;
3. C_r *and* D_r *hold*

then $G_r = T_r$.

Proof of Lemma 3. First of all, events C_r and D_r guarantee that G_r is a tree. Moreover, $G_{r-1} = T_{r-1}$ and $Y_u = Z_u$ for every vertex in the last generation of G_{r-1} and T_{r-1}, so we have $G_r = T_r$. □

Next, we define a set of auxiliary events in order to complete the proof of our induction step. For any $0 \le r \le R$, we define E_r to be the event

$$E_r = \{|\partial G_s| \le 2^s Q_{\max}^s \log n,\ \forall s \le r+1\}.$$

The utility of this auxiliary event is shown in the following two lemmas.

Lemma 4. *For all* $r \le R$ *and some constant* $c > 0$,

$$\mathsf{P}(E_r | E_{r-1}) \ge 1 - n^{-c}.$$

Moreover, $|G_r| = O(n^{1/8})$ *when* E_{r-1} *holds true.*

Proof of Lemma 4. First of all, Y_v is stochastically dominated by $\mathrm{Binom}(n, Q_{\max}/n)$ for any v. On E_{r-1}, $|\partial G_r| \le 2^r Q_{\max}^r \log n$ and so $|\partial G_{r+1}|$ is stochastically dominated by

$$Z \sim \mathrm{Binom}(2^r Q_{\max}^r n \log n, \frac{Q_{\max}}{n}).$$

Thus,

$$\begin{aligned}\mathsf{P}(E_r^c | E_{r-1}) &= \mathsf{P}(|\partial G_{r+1}| > 2^{r+1} Q_{\max}^{r+1} \log n | E_{r-1}) \\ &\le \mathsf{P}(Z \ge 2\mathsf{E}Z) \overset{(1)}{\le} \left(\frac{e}{4}\right)^{\mathsf{E}Z},\end{aligned}$$

where (1) follows by a multiplicative version of Chernoff's inequality

$$\mathsf{P}(X > (1+\delta)\mathsf{E}X) \le \left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)^{\mathsf{E}X}$$

for binomial random variable X and any $\delta > 0$. We have

$$\mathsf{E}Z = 2^r Q_{\max}^{r+1} \log n = \Theta(\log n),$$

which proves the first part of the Lemma.
For the second part, on E_{r-1}

$$|G_r| = \sum_{r=1}^{R} |\partial G_r| \leq \sum_{r=1}^{R} 2^r Q_{\max}^r \log n \leq (2Q_{\max})^{R+1} \log n = O(n^{1/8}),$$

since $R = \left\lfloor \frac{1}{10 \log(2Q_{\max})} \log n \right\rfloor$. □

Lemma 5. *For any r,*

$$\mathsf{P}(C_r | E_{r-1}) \geq 1 - O(n^{-3/4})$$

$$\mathsf{P}(D_r | E_{r-1}) \geq 1 - O(n^{-3/4}).$$

Proof. For the first claim, fix $u, v \in \partial G_r$. For any $w \in V_r$, the probability that (u, w) and (v, w) both appear is $O(n^{-2})$. Now, $|V_r| \leq n$ and Lemma 4 implies that $|\partial G_r|^2 = O(n^{1/4})$. Hence the result follows from the union bound over all triples u, v, w. For the second claim, the probability of an edge between any particular pair of vertices $u, v \in \partial G_r$ is $O(n^{-1})$. Lemma 4 implies that $|\partial G_r|^2 = O(n^{1/4})$ and the result follows from a union bound over all the pairs u, v. □

Moreover, we state a lemma to bound the total variation distance between Binomial distribution and Poisson distribution.

Lemma 6 (Lemma 5 in [35])**.** *If m and n are positive integers and $c > 0$ is a positive constant, then*

$$d_{\mathrm{TV}}(\mathrm{Binom}(m, c/n), \mathrm{Poi}(c)) = O\left(\frac{\max\{1, |m-n|\}}{n}\right).$$

Now, we are ready to prove Proposition 7.

Proof of Proposition 7. For any random variable X, let $\mathrm{dist}(X)$ denote the probability distribution of X. For a vertex $v \in V$, let X_v denote its community label. Fix r and suppose E_{r-1} holds and $T_{r-1} = G_{r-1}$. Let $V^{(1)}, \ldots, V^{(L)}$ denote the set of vertices in V with community label $1, \ldots, L$ respectively and similarly, let $V_{r-1}^{(1)}, \ldots, V_{r-1}^{(L)}$ denote the set of vertices in V_{r-1} with community label $1, \ldots, L$ respectively. Let us first provide a high

probability bound for $|V_{r-1}^{(1)}|, \ldots, |V_{r-1}^{(L)}|$. For all $i \in [L]$, by Hoeffding's inequality, we have

$$\mathsf{P}\left(||V^{(i)}| - np_i| \geq n^{3/4}\right) \leq 2e^{-\Omega(n^{1/2})}.$$

It follows from the union bound that

$$\mathsf{P}\left(\forall i \in [L], ||V^{(i)}| - np_i| \leq n^{3/4}\right) \geq 1 - 2Le^{-\Omega(n^{1/2})}.$$

Combining with Lemma 4, conditioned on event E_{r-1} and the event that $\{\forall i \in [L], ||V^{(i)}| - np_i| \leq n^{3/4}\}$, the number $|V_{r-1}^{(i)}|$ of vertices in V_{r-1} with community label i can be upper and lower bounded as

$$np_i + n^{3/4} \geq |V^{(i)}| \geq |V_{r-1}^{(i)}| \geq |V^{(i)}| - |G_{r-1}| \geq np_i - n^{3/4} - O(n^{1/8})$$

for any $i \in [L]$.

Next, let us analyze the distribution of Y_v, i.e., the number of neighbours in V_{r-1} for a vertex $v \in \partial G_{r-1}$. Let $Y_v^{(1)}, \ldots, Y_v^{(L)}$ denote the number of neighbours v has in V_{r-1} with community label $1, \ldots, L$ respectively. Then we have $Y_v = \sum_{i=1}^{L} Y_v^{(i)}$ and $Y_v^{(1)}, \ldots, Y_v^{(L)}$ are independent of each other given the community label X_v. We know that conditional distribution of $Y_v^{(i)}$ given $X_v = j$ is $\mathrm{Binom}(|V_{r-1}^{(i)}|, Q_{ji}/n)$. By Lemma 6, we have $d_{\mathrm{TV}}(\mathrm{Binom}(|V_{r-1}^{(i)}|, Q_{ji}/n), \mathrm{Poi}(p_i Q_{ji})) = O(n^{-1/4})$. For any $i \in [L]$, let $Z^{(i)}$ denote a Poisson random variable with mean $p_i Q_{ji}$, i.e., $Z^{(i)} \sim \mathrm{Poi}(p_i Q_{ji})$. Therefore, conditioning on any $X_v = j$ for any $i \in [L]$ we can couple $Y_v^{(i)}$ with $Z^{(i)}$ such that $\mathsf{P}(Y_v^{(i)} \neq Z^{(i)}) = O(n^{-1/4})$. By a union bound over L communities, we can couple the L-tuples $(Y_v^{(1)}, \ldots, Y_v^{(L)})$ and $(Z^{(1)}, \ldots, Z^{(L)})$ such that $\mathsf{P}((Y_v^{(1)}, \ldots, Y_v^{(L)}) \neq (Z^{(1)}, \ldots, Z^{(L)})) = O(n^{-1/4})$. By Poisson superposition, it follows that $d_{\mathrm{TV}}(\mathrm{dist}(Y_v | X_v = j), \mathrm{Poi}(\sum_{i=1}^{L} p_i Q_{ji})) = O(n^{-1/4})$. Recall our assumption that $\sum_{i=1}^{L} p_i Q_{ji} = \lambda$ for any $j \in [L]$ and that $Z_v \sim \mathrm{Poi}(\lambda)$. Moreover Y_v is conditionally independent of G_{r-1} given its community label X_v. Therefore, we have

$$\begin{aligned}
&d_{\mathrm{TV}}(\mathrm{dist}(Y_v | G_{r-1} = g_{r-1}), \mathrm{dist}(Z_v)) \\
&= d_{\mathrm{TV}}\left(\mathrm{dist}(Y_v | X_v = j), \mathrm{Poi}\left(\sum_{i=1}^{L} p_i Q_{ji}\right)\right) = O(n^{-1/4})
\end{aligned}$$

for any possible realization g_{r-1}. Thus, we can couple Y_v with Z_v such that $\mathsf{P}(Y_v \neq Z_v) = O(n^{-1/4})$. Then by a union bound over all vertices in ∂G_{r-1},

we have

$$\begin{aligned}\mathsf{P}(\exists v \in \partial G_{r-1} : Y_v \neq Z_v) &\leq |\partial G_{r-1}| \mathsf{P}(Y_v \neq Z_v) \\ &= O(n^{1/8}) O(n^{-1/4}) \\ &= O(n^{-1/8}).\end{aligned}$$

The argument above shows that we can find a coupling such that with probability at least $1 - O(n^{-1/8})$, $Y_u = Z_u$ for any $u \in \partial G_{r-1}$. Moreover, Lemmas 4 and 5 imply that C_r, D_r and E_r hold simultaneously with probability at least $1 - n^{-c} - O(n^{-3/4})$. Putting these all together, we see that the hypothesis of Lemma 3 holds with probability at least $1 - O(n^{-\min(c,1/8)})$. Thus,

$$\mathsf{P}(G_r = T_r, E_r | G_{r-1} = T_{r-1}, E_{r-1}) \geq 1 - O(n^{-\min(c,1/8)}).$$

But $\mathsf{P}(E_0) \to 1$ as $n \to \infty$ by the multiplicative Chernoff's inequality, and we can certainly couple G_0 with T_0 since they are both a single vertex. Therefore, we can apply a union bound over $r = 1, \ldots, R$,

$$\begin{aligned}\mathsf{P}(G_R \neq T_R) &= \mathsf{P}(\exists r \in [R] \text{ s.t. } G_r \neq T_r \text{ or } E_r^c | G_{r-1} = T_{r-1}, E_{r-1}) \\ &\leq R \cdot O(n^{-\min(c,1/8)}) \\ &= \Theta(\log n) O(n^{-\min(c,1/8)}) \\ &\to 0\end{aligned}$$

as $n \to \infty$, i.e., we can couple G_R and T_R such that $G_R = T_R$ with high probability. Therefore, we have $d_{\mathrm{TV}}(\mathrm{Nbr}_R, \mathrm{GWT}(\mathrm{Poi}(\lambda))_R) \to 0$ as $n \to \infty$. □

With Proposition 7, we are ready to establish the local weak convergence of the stochastic block model.

Proposition 8 (Local weak convergence of sparse SBMs)**.** *Let A_n denote a graph generated from a stochastic block model* $\mathrm{SBM}(n, L, \mathbf{p}, \frac{1}{n}\mathbf{Q})$ *with* $\mathbf{Qp} = \lambda \mathbf{1}$ *for some positive constant λ and an all-1 vector $\mathbf{1}$ of dimension $L \times 1$. Let $U(A_n)$, defined as in* (7.1)*, be the random rooted neighbourhood distribution of A_n. Then, the average neighbourhood distribution $\mathsf{E}U(A_n)$ converges weakly to a Poisson Galton–Watson tree*

$$\mathsf{E}U(A_n) \rightsquigarrow \mathrm{GWT}(\mathrm{Poi}(\lambda)).$$

Remark 8. When $Q_{i,j} = c$ for all $i, j \in [L]$, the stochastic block model recovers the well-known local weak convergence result on the Erdős–Rényi model (see, e.g., [10, Theorem 3.12]).

Proof of Proposition 8. We want to show that for any uniformly continuous and bounded function f,

$$\left| \int f d\mathsf{E}U(A_n) - \int f d\mathrm{GWT}(\mathrm{Poi}(\lambda)) \right| \to 0$$

as $n \to \infty$. Since f is a uniformly continuous function on $\mathcal{G}^*$, for every $\epsilon > 0$ there exists $\delta > 0$ such that, for any pair of rooted graphs $[G_1, o_1]$ and $[G_2, o_2] \in \mathcal{G}^*$ with $d^*([G_1, o_1], [G_2, o_2]) < \delta$ we have $|f(G_1, o_1) - f(G_2, o_2)| < \epsilon$. Recall that $d^*([G_1, o_1], [G_2, o_2]) := \frac{1}{1+\hat{h}}$, where $\hat{h}$ denotes the maximum layers of matching between $[G_1, o_1]$ and $[G_2, o_2]$. Therefore, as long as $h > \frac{1}{\delta} - 1$, we have $|f((G, o)_h) - f(G, o)| < \epsilon$. It follows that $|f([i, o]) - f([g, o])| < \epsilon$, if $[i, o]_h = [g, o]$. Let $\mu \in \mathcal{P}(\mathcal{G}^*)$ and assume $h > \frac{1}{\delta} - 1$. We have

$$\left| \int f d\mu_h - \int f d\mu \right| \tag{7.2}$$

$$= \left| \sum_{[g,o] \in \mathcal{G}_h^*} f([g, o]) \mu_h([g, o]) - \sum_{[i,o] \in \mathcal{G}^*} f([i, o]) \mu([i, o]) \right| \tag{7.3}$$

$$\leq \sum_{[g,o] \in \mathcal{G}_h^*} \left| f([g, o]) \mu_h([g, o]) - \sum_{[i,o] \in \mathcal{G}^* : [i,o]_h = [g,o]} f([i, o]) \mu([i, o]) \right| \tag{7.4}$$

$$\overset{(a)}{=} \sum_{[g,o] \in \mathcal{G}_h^*} \left| \sum_{[i,o] \in \mathcal{G}^* : [i,o]_h = [g,o]} (f([g, o]) - f([i, o])) \mu([i, o]) \right| \tag{7.5}$$

$$\leq \sum_{[g,o] \in \mathcal{G}_h^*} \sum_{[i,o] \in \mathcal{G}^* : [i,o]_h = [g,o]} |f([g, o]) - f([i, o])| \, \mu([i, o]) \tag{7.6}$$

$$\leq \sum_{[g,o] \in \mathcal{G}_h^*} \sum_{[i,o] \in \mathcal{G}^* : [i,o]_h = [g,o]} \epsilon \mu([i, o]) = \epsilon, \tag{7.7}$$

where (a) follows since $\mu_h([g, o]) = \sum_{[i,o] \in \mathcal{G}^* : [i,o]_h = [g,o]} \mu([i, o])$. Therefore, $|\int f d\mathsf{E}U(A_n)_h - \int f d\mathsf{E}U(A_n)| < \epsilon$ and $|\int f d\mathrm{GWT}(\mathrm{Poi}(\lambda))_h - \int f d\mathrm{GWT}(\mathrm{Poi}(\lambda))| < \epsilon$. By Proposition 7, there exists n_0 such that if $n \geq n_0$ and $\left\lfloor \frac{1}{10 \log(2Q_{\max})} \log n \right\rfloor \geq R$, we have $d_{\mathrm{TV}}(\mathrm{GWT}(\mathrm{Poi}(\lambda))_R, \mathsf{E}U(A_n)_R) < \epsilon$. Since f is a bounded function, we have $|\int f d\mathrm{GWT}(\mathrm{Poi}(\lambda))_R - \int f d\mathsf{E}U(A_n)_R| < \epsilon$, as long as n is large

enough. Therefore, if we take n large enough such that $\left\lfloor \frac{1}{10\log(2Q_{\max})} \log n \right\rfloor > \frac{1}{\delta} - 1$ and $|\int f d\mathrm{GWT}(\mathrm{Poi}(\lambda))_h - \int f d\mathsf{E}U(A_n)_h| < \epsilon$, we have

$$\begin{aligned}
&\left| \int f d\mathsf{E}U(A_n) - \int f d\mathrm{GWT}(\mathrm{Poi}(\lambda)) \right| \\
&\leq \left| \int f d\mathsf{E}U(A_n)_h - \int f d\mathsf{E}U(A_n) \right| \\
&\quad + \left| \int f d\mathrm{GWT}(\mathrm{Poi}(\lambda))_h - \int f d\mathrm{GWT}(\mathrm{Poi}(\lambda)) \right| \\
&\quad + \left| \int f d\mathrm{GWT}(\mathrm{Poi}(\lambda))_h - \int f d\mathsf{E}U(A_n)_h \right| \\
&< 3\epsilon,
\end{aligned}$$

which completes the proof. □

7.3 BC Entropy

In this section, we review the notion of BC entropy introduced in [11], which is shown to be the fundamental limit of universal lossless compression for certain graph family [19].

For a Polish space Ω, let $\mathcal{P}(\Omega)$ denote the set of all Borel probability measures on Ω. Let A be a Borel set in Ω, we define the *ϵ-extension of A*, denoted A^ϵ, as the union of the open balls with radius ϵ centered around the points in A. For two probability measures μ and ν in $\mathcal{P}(\Omega)$, we define the *Lévy–Prokhorov distance* $d_{\mathrm{LP}}(\mu,\nu) := \inf\{\epsilon > 0 : \mu(A) \leq \nu(A^\epsilon) + \epsilon \text{ and } \nu(A) \leq \mu(A^\epsilon) + \epsilon, \forall A \in \mathcal{B}(\Omega)\}$, where $\mathcal{B}(\Omega)$ denotes the Borel sigma algebra of Ω. Let $\rho \in \mathcal{P}(\mathcal{G}^*)$. Let d be the expected number of neighbours of root under the law ρ and let a sequence $m = m(n)$ such that $m/n \to d/2$, as $n \to \infty$. Define $\mathcal{G}_{n,m}$ to be the set of graphs with n vertices and m edges. For $\epsilon > 0$, define

$$\mathcal{G}_{n,m}(\rho,\epsilon) = \{G \in \mathcal{G}_{n,m} : U(G) \in B(\rho,\epsilon)\},$$

where $B(\rho,\epsilon)$ denotes the open ball with radius ϵ around ρ with respect to Lévy–Prokhorov metric. Now, we define the *ϵ-upper BC entropy* of ρ as

$$\overline{\Sigma}(\rho,\epsilon) = \limsup_{n\to\infty} \frac{\log|\mathcal{G}_{n,m}(\rho,\epsilon)| - m\log n}{n}$$

and define the *upper BC entropy* of ρ as

$$\overline{\Sigma}(\rho) = \lim_{\epsilon\to 0} \overline{\Sigma}(\rho,\epsilon).$$

Similarly we define the *ϵ-lower BC entropy* $\underline{\Sigma}(\rho, \epsilon)$ and *lower BC entropy* $\underline{\Sigma}(\rho)$ with lim sup replaced by lim inf in above definitions. If ρ is such that $\overline{\Sigma}(\rho) = \underline{\Sigma}(\rho)$, then this common limit is called the *BC entropy* of ρ

$$\Sigma(\rho) := \overline{\Sigma}(\rho) = \underline{\Sigma}(\rho).$$

The following lemma states the BC entropy of the Galton–Waston tree distribution.

Lemma 7 (Corollary 1.4 of [11])**.** *The BC entropy of the Galton–Watson tree distribution* $\mathrm{GWT}(\mathrm{Poi}(\lambda))$ *is given by*

$$\Sigma\left(\mathrm{GWT}(\mathrm{Poi}(\lambda))\right) = \frac{\lambda}{2} \log \frac{e}{\lambda} \quad \text{bits.}$$

7.4 Achieving BC Entropy in the Sparse Regime

With the Lemma above, we can give a performance guarantee of our algorithm corresponding to the BC entropy. It is a Theorem analog to Proposition 1 in[19].

Theorem 4. *Let* $A_n \sim \mathrm{SBM}\left(n, L, \mathbf{p}, \frac{1}{n}\mathbf{Q}\right)$ *with* $\mathbf{Qp} = \lambda\mathbf{1}$ *for some positive constant* λ *and an all-*1 *vector* $\mathbf{1}$ *of dimension* $L \times 1$*. Let* $m = \binom{n}{2}\frac{\lambda}{n}$ *be the expected number of edges in the model. Then, our compression algorithm achieves the BC entropy of the local weak limit of stochastic block models in the sense that*

$$\limsup_{n\to\infty} \frac{\mathsf{E}[\ell(C_k(A_n))] - m \log n}{n} \leq \Sigma\left(\mathrm{GWT}(\mathrm{Poi}(\lambda))\right).$$

Proof. By our proof of Theorem 1, we have

$$\mathsf{E}[\ell(C_k(A_n))] \leq \binom{n'}{2} H(\mathbf{B}_{12}) + 2^{k^2} \log(2en^3) + nk_n^2 H(A_{12}).$$

Notice that

$$\begin{aligned}
\binom{n'}{2} H(\mathbf{B}_{12}) &\le \binom{n'}{2} k^2 H(A_{12}) \\
&= \binom{n'}{2} k^2 h(\lambda/n) \\
&\overset{(1)}{=} \binom{n'}{2} k^2 \left(\frac{1}{n} \lambda \log \frac{ne}{\lambda} + o\left(\frac{1}{n}\right) \right) \\
&\overset{(2)}{\sim} \binom{n}{2} \left(\frac{1}{n} \lambda \log n + \frac{1}{n} \lambda \log \frac{e}{\lambda} + o\left(\frac{1}{n}\right) \right) \\
&= \binom{n}{2} \frac{1}{n} \lambda \log n + \frac{\lambda \log e - \lambda \log \lambda}{2} n + o(n) \\
&\overset{(3)}{=} m \log n + n\Sigma \left(\mathrm{GWT}(\mathrm{Poi}(\lambda)) \right) + o(n)
\end{aligned}$$

where (1) follows since $h(p) = p \log \frac{e}{p} - \frac{\log e}{2} p^2 + o(p^2)$, (2) follows since $n'k = n$ and (3) follows from Lemma 7. Then it suffices to that the remaining terms in the upper bound of $\mathsf{E}[\ell(C_k(A_n))]$ are all $o(n)$. Indeed we have

$$2^{k^2} \log(2en^3) \le 2^{\delta \log n} \log(2en^3) = n^{\delta} \log(2en^3) = o(n)$$

since $\delta < 1$ and

$$\begin{aligned}
nk_n^2 H(A_{12}) &= nk_n^2 h(\lambda/n) \\
&= nk_n^2 \left(\frac{1}{n} \lambda \log \frac{ne}{\lambda} + o\left(\frac{1}{n}\right) \right) \\
&\le n\delta \log n \left(\frac{1}{n} \lambda \log \frac{ne}{\lambda} + o\left(\frac{1}{n}\right) \right) \\
&= \delta \log n \left(\lambda \log \frac{ne}{\lambda} \right) + o(\log n) \\
&= o(n).
\end{aligned}$$

□

Theorem 4 shows that in the sparse regime $f(n) = \frac{1}{n}$, our compressor achieves the BC entropy of the Galton–Watson tree that is the local weak convergence limit of the underlying sequence of graphs.

Chapter 8

Concluding Remarks

In this chapter, we consider three alternative compression methods, and demonstrate why they are not applicable in our problem or not universal in the same sense as the proposed algorithm.

We first consider We first take a closer look at the correlation among entries in the adjacency matrix and explain why existing universal compressors developed for stationary processes may not be immediately applicable for certain orderings of the entries. Compressing A_n entails compressing

$$A_{12}, \ldots, A_{1,n}, A_{23}, \ldots, A_{n-1,n},$$

i.e. the bits in the upper triangle of A_n. Clearly, these are not independent (because of the dependency through X_1^n) so one cannot use any of the compressors universal for the class of iid processes to compress A_n. So, one hopes that it is possible to list the $\binom{n}{2}$ random variables $A_{12}, \ldots, A_{1,n}, A_{23}, \ldots, A_{n-1,n}$ in an order that makes the resulting sequence stationary, so that the Lempel–Ziv compressor (which, recall, is universal for the class of stationary processes) may be used. However, we can show that that some of the most natural orders of listing these $\binom{n}{2}$ bits result in a sequence that is nonstationary. First, consider listing the bits in the upper triangle row-wise (i.e. first listing the bits in the first row, followed by the bits in the second and so on, ending with $A_{n-1,n}$) we get the following sequence

$$A_{12}, \ldots, A_{1,n}, A_{23}, \ldots, A_{2,n}, \ldots, A_{n-1,n},$$

which can be seen to be nonstationary. Consider the case when $n = 4, L = 2, Q_{11} = Q_{12} = 1, Q_{12} = 0$. In this case the horizontal ordering is

$$A_{12}, A_{13}, A_{14}, A_{23}, A_{24}, A_{34}$$

and this is seen to be nonstationary by observing $P(A_{12} = 1, A_{13} = 0, A_{14} = 1) > 0$ but $P(A_{23} = 1, A_{24} = 0, A_{34} = 1) = 0$. Similar counterexamples can be given for establishing nonstationarity of a column-by-column ordering and a diagonal-by-diagonal ordering. Therefore, one cannot plug-in these

1-dimensional sequences into a compressor optimal for stationary sequences (e.g. the Lempel–Ziv compressor).

Another naive approach to compression is to first recover the community labels, since conditioned on the community labels all edges become i.i.d. However, it is known that exact recovery of the community labels is information-theoretically impossible when the edge probabilities are $o\left(\frac{\log n}{n}\right)$ but our compressor is universal even in this regime. This implies that universal compression is a fundamentally easier task than community detection for stochastic block models. Even in the regime where exact recovery is possible, all algorithms for community detection require knowledge of the graph parameters such as the number of communities [2].

Another approach is based on run-length coding for the adjacency matrix A_n, which, stores just the lengths of runs of one of the symbol for the entire sequence. For example, the sequence `AAAABAABBAA`, when encoded using the run length of `A`s, would be stored as `(4,B,2,B,0,B,2)`. When applied to our problem, the expected length of this method can be shown to be upperbounded by $\mathbf{p}^T Q \mathbf{p} \binom{n}{2} f(n)[\log(1/p_{\min}) + \log\log n]$ (by suitably modifying [47]) and thus it fails to achieve the leading term in entropy when $f(n) = \Omega(1/\log n)$, since $p_{\min} = \Theta(f(n))$ and the term with $\log\log n$ contributes to the leading term. The analysis also fails in the case when there are zero entries in W (since $p_{\min} = 0$), which is a reasonable choice of parameters for the SBM. In contrast, the analysis for C_k is valid even when there are zero entries in W as long as $\max_{ij} Q_{ij} = \Theta(1)$. Furthermore, in the regime $f(n) = 1/n$, C_k is shown to achieve the BC entropy for the family of symmetric SBMs, and it is not clear whether a similar result can be shown for the run-length encoder.

In this thesis, we constructed a compressor that is universal for the class of stochastic block models with connection probabilities ranging from the regime of $1/n^{2-\epsilon}$ to constant. There are many intriguing open problems that remain. Firstly, even though our algorithm takes polynomial time, $O(n^2)$ may be restrictive when n is too large, and a lower complexity method is desirable. Secondly, a tight characterization of the the minimax redundancy in this problem is of interest. Finally, we have considered the problem of universal compression in the stochastic setting, where we assumed that our graph is generated from a random graph model. This leads naturally to the question of the individual graph setting, where the redundancy must be established relative to a class of compressors. Each of these are promising avenues for further study.

Bibliography

[1] Emmanuel Abbe. Graph compression: The effect of clusters. In *Proc. 54th Ann. Allerton Conf. Commun. Control Comput.*, pages 1–8, 2016.

[2] Emmanuel Abbe. Community detection and stochastic block models: recent developments. *The Journal of Machine Learning Research*, 18(1):6446–6531, 2017.

[3] Emmanuel Abbe, Afonso S Bandeira, and Georgina Hall. Exact recovery in the stochastic block model. *IEEE Trans. Inf. Theory*, 62(1):471–487, 2015.

[4] Emmanuel Abbe and Colin Sandon. Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery. In *IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 670–688, 2015.

[5] Amir R. Asadi, Emmanuel Abbe, and Sergio Verdú. Compressing data on graphs with clusters. In *Proc. IEEE Internat. Symp. Inf. Theory*, pages 1583–1587, August 2017.

[6] Maciej Besta and Torsten Hoefler. Survey and taxonomy of lossless graph compression and space-efficient graph representations. 2018.

[7] Alankrita Bhatt, Ziao Wang, Chi Wang, and Lele Wang. Universal graph compression: Stochastic block models. 2020.

[8] Patrick Billingsley. *Convergence of probability measures*. Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons Inc., New York, second edition, 1999. A Wiley-Interscience Publication.

[9] P. Boldi and S. Vigna. The webgraph framework i: Compression techniques. In *Proceedings of the 13th International Conference on World Wide Web*, WWW '04, pages 595–602, New York, NY, USA, 2004. Association for Computing Machinery.

[10] Charles Bordenave. Lecture notes on random graphs and probabilistic combinatorial optimization. `https://www.math.univ-toulouse.fr/$\sim$bordenave/coursRG.pdf`, 2016.

[11] Charles Bordenave and Pietro Caputo. Large deviations of empirical neighborhood distribution in sparse random graphs. *Probability Theory and Related Fields*, 163(1-2):149–222, Nov 2014.

[12] Nieves R. Brisaboa, Susana Ladra, and Gonzalo Navarro. K2-trees for compact web graph representation. In *Proceedings of the 16th International Symposium on String Processing and Information Retrieval*, SPIRE '09, pages 18–30, Berlin, Heidelberg, 2009. Springer-Verlag.

[13] Flavio Chierichetti, Ravi Kumar, Silvio Lattanzi, Michael Mitzenmacher, Alessandro Panconesi, and Prabhakar Raghavan. On compressing social networks. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 219–228, New York, NY, USA, 2009. Association for Computing Machinery.

[14] Y. Choi and W. Szpankowski. Compression of graphical structures: Fundamental limits, algorithms, and experiments. *IEEE Trans. Inf. Theory*, 58(2):620–638, Feb 2012.

[15] Thomas C. Conway and Andrew J. Bromage. Succinct data structures for assembling large genomes. *Bioinformatics*, 27(4):479–486, 01 2011.

[16] Thomas Courtade. Properties of the binary entropy function, 2012. `https://blogs.princeton.edu/blogit/2012/10/26/properties-of-the-binary-entropy-function/`.

[17] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, USA, 2006.

[18] P. Delgosha and V. Anantharam. Universal lossless compression of graphical data. *IEEE Trans. Inf. Theory*, 66(11):6962–6976, 2020.

[19] Payam Delgosha and Venkat Anantharam. Universal lossless compression of graphical data. In *Proc. IEEE Internat. Symp. Inf. Theory*, June 2017.

[20] Payam Delgosha and Venkat Anantharam. Universal lossless compression of graphical data. 2019.

[21] Payam Delgosha and Venkat Anantharam. A universal low complexity compression algorithm for sparse marked graphs. In *Proc. IEEE Internat. Symp. Inf. Theory*, June 2020.

[22] Michelle Effros, Karthik Visweswariah, Sanjeev R Kulkarni, and Sergio Verdú. Universal lossless source coding with the burrows wheeler transform. *IEEE Trans. Inf. Theory*, 48(5):1061–1081, 2002.

[23] Arash Farzan and J. Ian Munro. Succinct encoding of arbitrary graphs. *Theoretical Computer Science*, 513:38 – 52, 2013.

[24] Alan Frieze and Michał Karoński. *Introduction to Random Graphs*. Cambridge University Press, 2015.

[25] M. Ganardi, D. Hucke, M. Lohrey, and L. Seelbach Benkner. Universal tree source coding using grammar-based compression. *IEEE Trans. Inf. Theory*, 65(10):6399–6413, 2019.

[26] Aditya Grover and Jure Leskovec. Node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 855–864, New York, NY, USA, 2016. Association for Computing Machinery.

[27] M. Hayashida and T. Akutsu. Comparing biological networks via graph compression. *BMC systems biology*, 4 Suppl 2(Suppl 2), 2010.

[28] J. C. Kieffer, E.-H. Yang, and W. Szpankowski. Structural complexity of random binary trees. In *Proc. IEEE Internat. Symp. Inf. Theory*, pages 635–639, 2009.

[29] Steffen Lauritzen, Alessandro Rinaldo, and Kayvan Sadeghi. Random networks, graphical models, and exchangeability. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80, 01 2017.

[30] Abraham Lempel and Jacob Ziv. Compression of two-dimensional data. *IEEE Trans. Inf. Theory*, 32(1):2–8, 1986.

[31] Y. Lim, U. Kang, and C. Faloutsos. Slashburn: Graph compression and mining beyond caveman communities. *IEEE Transactions on Knowledge and Data Engineering*, 26(12):3077–3089, 2014.

[32] A. Magner, K. Turowski, and W. Szpankowski. Lossless compression of binary trees with correlated vertex names. *IEEE Trans. Inf. Theory*, 64(9):6070–6080, 2018.

[33] D. Marpe, H. Schwarz, and T. Wiegand. Context-based adaptive binary arithmetic coding in the h.264/avc video compression standard. *IEEE Trans. Circuits Syst. Video Technol.*, 13(7), 2003.

[34] N. Merhav and M. Feder. A strong version of the redundancy-capacity theorem of universal coding. *IEEE Transactions on Information Theory*, 41(3):714–722, 1995.

[35] Elchanan Mossel, Joe Neeman, and Allan Sly. Reconstruction and estimation in the planted partition model. *Probability Theory and Related Fields*, 162(3-4):431–461, 2015.

[36] Sharad Nandanwar and M. N. Murty. Structural neighborhood based classification of nodes in a network. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1085–1094, New York, NY, USA, 2016. Association for Computing Machinery.

[37] Moni Naor. Succinct representation of general unlabeled graphs. *Discrete Applied Mathematics*, 28(3):303 – 307, 1990.

[38] Gonzalo Navarro. Compressing web graphs like texts. Technical report, Dept. of Computer Science, University of Chile, 2007.

[39] Tiago P. Peixoto. Parsimonious module inference in large networks. *Phys. Rev. Lett.*, 110:148701, Apr 2013.

[40] Tiago P. Peixoto. Hierarchical block structures and high-resolution model selection in large networks. *Phys. Rev. X*, 4:011047, Mar 2014.

[41] Tiago P. Peixoto. Model selection and hypothesis testing for large-scale network models with overlapping groups. *Phys. Rev. X*, 5:011033, Mar 2015.

[42] Tiago P. Peixoto. Nonparametric bayesian inference of the microcanonical stochastic block model. *Phys. Rev. E*, 95:012317, Jan 2017.

[43] Yury Polyanskiy and Yihong Wu. Lecture notes on information theory. 2014.

[44] Sebastien Roch. Modern discrete probability: An essential toolkit. `http://www.math.wisc.edu/~roch/mdp/`, 2020.

[45] R.A. Rossi and R. Zhou. GraphZIP: a clique-based sparse graph compression method. *Journal of Big Data*, 5(10), 2018.

[46] Kunihiko Sadakane. New text indexing functionalities of the compressed suffix arrays. *Journal of Algorithms*, 48(2):294 – 313, 2003.

[47] Mark F Schilling. The longest run of heads. *The College Mathematics Journal*, 21(3):196–207, 1990.

[48] Claude E. Shannon. A mathematical theory of communication. *Bell Syst. Tech. J.*, 27(3):379–423, 1948.

[49] J. Shun, L. Dhulipala, and G. E. Blelloch. Smaller and faster: Parallel processing of compressed graphs with ligra+. In *2015 Data Compression Conference*, pages 403–412.

[50] Julian Shun and Guy E. Blelloch. Ligra: A lightweight graph processing framework for shared memory. In *Proceedings of the 18th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming*, PPoPP '13, pages 135–146, New York, NY, USA, 2013. Association for Computing Machinery.

[51] Lei Tang and Huan Liu. Relational learning via latent social dimensions. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 817–826, New York, NY, USA, 2009. Association for Computing Machinery.

[52] György Turán. On the succinct representation of graphs. *Discrete Applied Mathematics*, 8(3):289 – 294, 1984.

[53] Frans MJ Willems, Yuri M Shtarkov, and Tjalling J Tjalkens. The context-tree weighting method: basic properties. *IEEE Trans. Inf. Theory*, 41(3):653–664, 1995.

[54] Qun Xie and Andrew R Barron. Minimax redundancy for the class of memoryless sources. *IEEE Trans. Inf. Theory*, 43(2):646–657, 1997.

[55] Qun Xie and Andrew R Barron. Asymptotic minimax regret for data compression, gambling, and prediction. *IEEE Trans. Inf. Theory*, 46(2):431–445, 2000.

[56] J. Zhang, E.-H. Yang, and J. C. Kieffer. A universal grammar-based code for lossless compression of binary trees. *IEEE Trans. Inf. Theory*, 60(3):1373–1386, 2014.

[57] Jacob Ziv and Abraham Lempel. A universal algorithm for sequential data compression. *IEEE Trans. Inf. Theory*, 23(3):337–343, 1977.

[58] Jacob Ziv and Abraham Lempel. Compression of individual sequences via variable-rate coding. *IEEE Trans. Inf. Theory*, 24(5):530–536, 1978.