

Rate-Distortion Function of the Stochastic Block Model

Martin Wachiye Wafula*, Praneeth Kumar Vippathalla[†], Justin Coon[‡] and Mihai-Alin Badiu[§]

Dept. of Engineering Science, University of Oxford

Oxford, United Kingdom

Email: *martin.wafula, [†]praneeth.vippathalla, [‡]justin.coon, [§]mihai.badiu@eng.ox.ac.uk

Abstract—The stochastic block model (SBM) is extensively used to model networks in which users belong to certain communities. In recent years, the study of information-theoretic compression of such networks has gained attention, with works primarily focusing on lossless compression. In this work, we address the lossy compression of SBM graphs by characterizing the rate-distortion function under a Hamming distortion constraint. Specifically, we derive the conditional rate-distortion function of the SBM with community membership as side information to both the encoder and decoder. We approach this problem as the classical Wyner-Ziv lossy problem by minimising mutual information of the graph and its reconstruction conditioned on community labels. Lastly, we also derive the rate-distortion function of the Erdős-Rényi (ER) random graph model.

Index Terms—Stochastic block model, lossy compression, rate-distortion function

I. INTRODUCTION

Over the past two decades, there has been a significant interest in analysing large graphs representing relationships between various data entities, including those in machine learning, medicine, communications, social and information network analysis, and transportation networks [1]–[4]. Compressing these large graphs has become a topic of interest for academia and industry due to its potential benefits, such as reducing the number of input-output operations, increasing the speed of graph analysis, reducing the amount of data communicated over the network [5] and for efficient storage and transmission of graphs themselves [6].

In this work, we examine the stochastic block model (SBM), which partitions the nodes into communities and assigns independent edge probabilities based on the community membership of the nodes. The SBM and its variants have been extensively studied and reviewed, particularly in community recovery problems, [7]–[13]. The SBM can be utilised to analyse the community structure of Wireless Sensor Networks (WSNs) as it can capture the natural clustering of nodes based on spatial proximity or functionality, allowing for the investigation of the performance of routing protocols, energy management schemes and other network services that rely on the community structure of the network. For instance, by utilising the community structure to minimise the number

of nodes needed to transmit data, energy-efficient routing protocols can be designed for WSNs. Clustered networks outperform non-clustered ones [14].

Motivated by their importance in practical applications, many works studied the information-theoretic compression of these graphs. In [15], Abbe addressed the problem of lossless compression of graphs with clusters, while Bhatt et al. dealt with universal compression of SBM [16]. Recently, Han et al. obtained the partitioned structural entropy of SBM similar to Choi’s and Szpankowski’s [17] structural entropy of the ER graph, which generalises the structural entropy of unlabelled graphs and encodes the partition information [18]. To consider data on SBM graphs, Asadi, Abbe, and Verdú [19] explored data compression limits on graphs using community dependencies, providing optimal compressor lengths when the community signal is strong. They termed the source model data block model (DBM).

There has also been research on lossy graph compression, primarily focusing on heuristics and algorithms with little emphasis on information-theoretic approaches. Preserving graph structure with fidelity is a significant research area [20]. Recent work explores compressing directed graphs by maintaining local structure [21], and there is a quantitative definition for compressibility based on network properties like transitivity and degree heterogeneity [22]. SLIMgraph is a recent lossy compression algorithm that has gained attention for its promising results on experimental data [5]. The technique leverages compression kernels and adopts a triangle reduction method, which involves the removal of an edge of the triangle with a specific probability. This removal is done while ensuring that the same edge is not selected for removal again and that edges belonging to multiple triangles, and hence more likely to be part of multiple shortest paths, are not removed more often than those belonging to only one triangle. Another example is shrinkage approximation, a graph summarization algorithm commonly used in call graphs that prunes edges, not nodes [23]. Both SLIMGraph and shrinkage approximation algorithms raise the question of determining the amount of distortion that can be tolerated. Specifically, in these cases, what is the maximum number of edges that can be pruned or discarded during lossy compression while still achieving acceptable performance levels? So far, to the best of our knowledge, there is no work on lossy compression of the SBM. Our work is a step towards addressing this problem.

This research was funded in whole or in part by EPSRC (EP/T02612X/1) and U.S. Army Research Laboratory and the U.S. Army Research Office (W911NF-22-1-0070). For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript (AAM) version arising from this submission.

Our study uses rate-distortion theory to determine limits for lossy compression algorithms, specifically focusing on the stochastic block model (SBM). By leveraging the Wyner-Ziv problem [24], [25], we derive a conditional rate with community labels as side information to both the encoder and decoder. We extend the findings to both homogeneous and inhomogeneous ER graph models. We apply a simple Hamming distortion measure to the graph edges, which makes practical sense for worst-case edge removal in algorithms such as SLIMGraph. The rate-distortion function serves as a boundary for compression algorithms, utilizing expected distortion as a metric to identify the maximum number of removable edges.

II. PRELIMINARIES

In this paper, we restrict our attention to an undirected and unweighted graph $G = (\mathcal{V}, \mathcal{E})$ constructed on a fixed vertex set $\mathcal{V} = [n] := \{1, 2, \dots, n\}$ where $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the set of edges. Moreover, we assume that there are no self-loops, i.e., $(v, v) \notin \mathcal{E}$ for all $v \in \mathcal{V}$. It is sometimes useful to represent a graph in terms of its adjacency matrix representation. For a graph G , we set $E_{ij} = 1$ if the vertices i and j share an edge, and $E_{ij} = 0$ otherwise. As there are no self-loops, which means that $E_{ii} = 0$ for $i \in [n]$, the values $\{E_{ij} : 1 \leq i < j \leq n\}$ specify the adjacency matrix uniquely, and hence the graph. Next, we present a few fundamental results related to the entropy of various random graph models.

A. Erdős-Rényi Graph Model

In the Erdős-Rényi (random) graph model $\mathcal{G}(n, p)$ model, we connect any two distinct vertices in $[n]$ by an edge independently with probability p . The ER model is the canonical example, to begin with, in most studies on graphs. One observes that under this model the probability of a graph G containing k edges is $P(G) = p^k(1-p)^{\binom{n}{2}-k}$.

In a generalisation of the ER model, which is commonly referred to as an inhomogeneous ER graph model or a generalised binomial graph, the probability of connecting two vertices i and j with $i < j$ by an edge is p_{ij} . This model was first defined by Kovalenko in [26]. Under this model, the probability of a graph G specified by $E_{ij} \in \{0, 1\}$ for $i < j$ is given by

$$P(G) = \prod_{i < j} p_{ij}^{E_{ij}} (1 - p_{ij})^{1 - E_{ij}}$$

Lemma 1. *The entropy of an inhomogeneous ER graph model with edge probabilities $\{p_{ij} : 1 \leq i < j \leq n\}$ is*

$$H(G) = \sum_{i < j} h(p_{ij}).$$

In particular, the entropy of the ER graph model $\mathcal{G}(n, p)$ is $H(G) = \binom{n}{2} h(p)$. Here $h(t) := -t \log_2 t - (1-t) \log_2 (1-t)$ is the binary entropy function.

B. Graph Entropy of the Stochastic Block Model (SBM)

Consider a graph with n nodes. Let $\mathbf{X} = (X_1, \dots, X_n)$ be a random vector whose components $X_i \in [k]$ are random variables that describes the community membership of node i and are identically and independently distributed (i.i.d.) according to $\mathbf{p} = (p_1, \dots, p_k)$ such that $\mathbb{P}[X_i = \ell] = p_\ell$ for all $i \in [n]$ and $\ell \in [k]$. Let \mathbf{W} be a $k \times k$ symmetric matrix, whose component $w_{\ell, m}$ is the probability that a node in community ℓ is connected to another node in community m . We refer to the pair (\mathbf{X}, G) as a stochastic block model if the generated graph G is such that an edge (i, j) exists with probability w_{X_i, X_j} .

Given the community label vector \mathbf{X} , the graph G (or equivalently, $\{E_{ij} : 1 \leq i < j \leq n\}$) is generated according to the inhomogeneous ER model with edge probabilities $\{w_{X_i, X_j} : 1 \leq i < j \leq n\}$.

Lemma 2 (Abbe, 2016 [15]). *Let $(\mathbf{X}, G) \sim \text{SBM}(n, \mathbf{p}, \mathbf{W})$. The entropy of the SBM graph is*

$$H(G) = \binom{n}{2} \mathbf{p}^T h_2(\mathbf{W}) \mathbf{p} + \mathcal{O}(n), \quad (1)$$

where $h_2(\mathbf{W})$ is an $n \times n$ matrix whose (i, j) -th entry is $h_2(w_{ij})$. The entropy of the SBM graph conditioned on the community labels is given by $H(G|\mathbf{X}) = \binom{n}{2} \mathbf{p}^T h_2(\mathbf{W}) \mathbf{p}$.

C. Rate-Distortion Theory

In this section, we will review rate-distortion theory, primarily used to study lossy compression. Consider two measurable spaces $(\mathcal{U}, \mathcal{B}_{\mathcal{U}})$ and $(\hat{\mathcal{U}}, \mathcal{B}_{\hat{\mathcal{U}}})$ for the source and its compressed version, respectively, equipped with a nonnegative measurable mapping $d: \mathcal{U} \times \hat{\mathcal{U}} \rightarrow \mathbb{R} \cup \{+\infty\}$. This mapping is referred to as the distortion function. Suppose a distortion level D and probability measure P on $(\mathcal{U}, \mathcal{B}_{\mathcal{U}})$ are given. In lossy compression, one wishes to represent the elements u of \mathcal{U} using the elements \hat{u} of the $\hat{\mathcal{U}}$ such that the average of the distortions $d(u, \hat{u})$ is at most D . The aim is to achieve this in a parsimonious way, for example, by using as few elements of $\hat{\mathcal{U}}$ as possible or by associating a fixed or variable-length codeword to each element of $\hat{\mathcal{U}}$ and then minimizing the average length of the codewords corresponding to a (lossy) representation.

The focus of the classical rate-distortion theory is sequential (random) sources X^n taking values in \mathcal{X}^n using the sequences from $\hat{\mathcal{X}}^n$ for $n \geq 1$. Given a single-letter distortion measure $d' : \mathcal{X} \times \hat{\mathcal{X}} \rightarrow \mathbb{R}$, the distortion measure for sequences is defined as $d(x^n, \hat{x}^n) = \sum_{i=1}^n d'(x_i, \hat{x}_i)/n$. We say that a rate R is achievable at a distortion level D if there exists a sequence of codes $g_n : \mathcal{X}^n \rightarrow \hat{\mathcal{X}}^n$ such that $\limsup_{n \rightarrow \infty} \log |g_n(\mathcal{X}^n)|/n \leq R$ and $\limsup_{n \rightarrow \infty} \mathbb{E}[d(X^n, g(X^n))] \leq D$. The goal is to characterize the smallest achievable rate, denoted by $R_X(D)$, at a distortion level D . In his seminal work [27], Shannon showed that for an i.i.d. source X^n ,

$$R_X(D) = \min_{P_{\hat{\mathcal{X}}|\mathcal{X}} : \mathbb{E}[d(X, \hat{X})] \leq D} I(X; \hat{X}). \quad (2)$$

The rate-distortion function $R(D)$ is a non-increasing convex function of $D \in [0, \infty)$ [28]. In [24], Wyner and Ziv studied the lossy compression in the presence of side information $Y^n \in \mathcal{Y}^n$, which is correlated with the source X^n . In this case, the map g_n is a function of $\mathcal{X}^n \times \mathcal{Y}^n$, and $R_{X|Y}(D)$ is the smallest achievable rate at a distortion level D . It was shown in [24] that if (X^n, Y^n) is jointly i.i.d., then

$$R_{X|Y}(D) = \min_{p(\hat{x}|x,y): \mathbb{E}[d(X, \hat{X})] \leq D} I(X; \hat{X}|Y). \quad (3)$$

Interestingly, the rate-distortion functions (2) and (3) satisfy [29] the inequality

$$R_{X|Y}(D) \leq R_X(D) \leq R_{X|Y}(D) + I(X; Y), \quad (4)$$

which is analogous to the entropic inequality $H(X|Y) \leq H(X) = H(X|Y) + I(X; Y)$.

In general, sources of interest need not have the sequential structure; for example, graphical sources. In such cases, we can still study the mutual information between a source and its reconstructed version, minimized over all conditional distributions satisfying the distortion condition, as in (2). It can be argued that this quantity gives a converse bound on the compression rate (appropriately scaled) in a lossy compression problem.

III. RATE DISTORTION FUNCTION OF THE SBM

We focus on the Hamming distortion measure on the set of all graphs with n vertices. For two graphs G_n and \hat{G}_n , the Hamming distortion is defined as $d(G_n, \hat{G}_n) = \sum_{i < j} \mathbb{1}(E_{i,j} \neq \hat{E}_{i,j})$, where the edges $E_{i,j}$ and $\hat{E}_{i,j}$, respectively, correspond to the graphs G_n and \hat{G}_n , and $\mathbb{1}(\cdot)$ is the indicator function. Clearly, $0 \leq d(G_n, \hat{G}_n) \leq \binom{n}{2}$. For a stochastic block model $(\mathbf{X}, G_n) \sim \text{SBM}(n, \mathbf{p}, \mathbf{W})$, the rate-distortion function at the Hamming distortion D is defined as

$$R_{G_n}(D) \triangleq \min_{P_{\hat{G}_n|G_n}: \mathbb{E}[d(G_n, \hat{G}_n)] \leq D} I(G_n; \hat{G}_n), \quad (5)$$

where \hat{G}_n represents the reconstructed version of the original graph G_n . Notice that if $\binom{n}{2} \min\{\sum_{l,m} p_l p_m w_{l,m}, \sum_{l,m} p_l p_m (1 - w_{l,m})\} \leq D$, $R_{G_n}(D) = 0$ because \hat{G}_n can be made independent of G_n while satisfying the constraint $\mathbb{E}[d(G_n, \hat{G}_n)] \leq D$. So in this case, we restrict to the case $0 \leq D \leq \binom{n}{2} \min\{\sum_{l,m} p_l p_m w_{l,m}, \sum_{l,m} p_l p_m (1 - w_{l,m})\}$.

In the case when \mathbf{X} is available as side information to both the encoder and decoder, the rate-distortion function (with side information \mathbf{X}) at the Hamming distortion D is defined as

$$R_{G_n|\mathbf{X}}(D) \triangleq \min_{P_{\hat{G}_n|G_n, \mathbf{X}}: \mathbb{E}[d(G_n, \hat{G}_n)] \leq D} I(G_n; \hat{G}_n|\mathbf{X}). \quad (6)$$

As in previous case, if $\binom{n}{2} \sum_{l,m} p_l p_m \min\{w_{l,m}, 1 - w_{l,m}\} \leq D$, then $R_{G_n|\mathbf{X}}(D) = 0$. Therefore, the interval that is of interest here is $0 \leq D \leq \binom{n}{2} \sum_{l,m} p_l p_m \min\{w_{l,m}, 1 - w_{l,m}\}$.

In the next theorem, we give a characterization of the rate-distortion function of the stochastic block model.

Theorem 1 (RDF of SBM). *Let $(\mathbf{X}, G_n) \sim \text{SBM}(n, \mathbf{p}, \mathbf{W})$.*

- 1) *For $0 \leq D \leq \binom{n}{2} \sum_{l,m} p_l p_m \min\{w_{l,m}, 1 - w_{l,m}\}$, the rate-distortion function of the SBM model at the Hamming distortion D is*

$$R_{G_n}(D) = \binom{n}{2} \mathbf{p}^T [h_2(\mathbf{W}) - h_2(\mathbf{D}^*)] \mathbf{p} + \mathcal{O}(n), \quad (7)$$

where \mathbf{D}^* is a matrix with entries chosen according to $d_{l,m}^* = \min\{\min\{w_{l,m}, 1 - w_{l,m}\}, \mu\}$, for $l, m \in [k]$, and μ is chosen such that the constraint $\mathbf{p}^T \mathbf{D}^* \mathbf{p} = \frac{D}{\binom{n}{2}}$. The rate-distortion function of the SBM graph conditioned on the community labels is given by

$$R_{G_n|\mathbf{X}}(D) = \binom{n}{2} \mathbf{p}^T [h_2(\mathbf{W}) - h_2(\mathbf{D}^*)] \mathbf{p} \quad (8)$$

with \mathbf{D}^* defined above.

- 2) *For $\binom{n}{2} \sum_{l,m} p_l p_m \min\{w_{l,m}, 1 - w_{l,m}\} \leq D \leq \min\{\sum_{l,m} p_l p_m w_{l,m}, \sum_{l,m} p_l p_m (1 - w_{l,m})\}$, $R_{G_n|\mathbf{X}}(D) = 0$ and $R_{G_n}(D) = \mathcal{O}(n)$*

Proof. See Section IV. \square

The result in (8) is shown in Figure 1.

Remark 1. *The value of μ in the above solution can be chosen by “reverse water-filling,” but now the levels $d_{l,m}^*$ scaled by $p_l p_m$ should add up to $\frac{D}{\binom{n}{2}}$.*

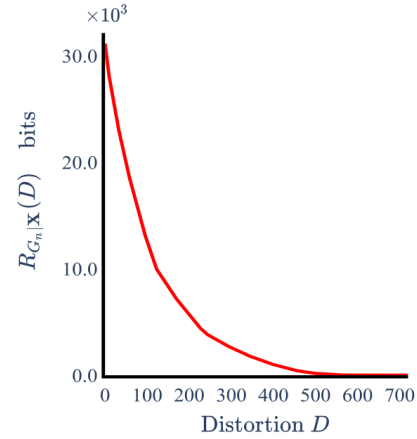


Fig. 1. Rate-distortion function (8) of an SBM over 100 nodes with $k=3$, $\mathbf{p} = (0.4, 0.3, 0.3)$ and $\mathbf{W} = [0.5, 0.2, 0.1; 0.2, 0.5, 0.1; 0.1, 0.1, 0.4]$.

A similar characterization can also be obtained for the rate-distortion function of the (inhomogeneous) ER model, defined as in (5). The next theorem gives a characterization of the rate-distortion function, whose proof is similar to that of Theorem 1

Theorem 2 (RDF of the inhomogeneous ER Model). *Given an inhomogeneous ER graph with an edge probability, $p_{i,j}$ between any two nodes i and j , $i, j \in \mathcal{V}$ and Hamming distortion $0 \leq D \leq \sum_{i < j} \min\{p_{i,j}, 1 - p_{i,j}\}$, the rate-distortion function of the graph is given by*

$$R(D) = \sum_{i < j} (h_2(p_{i,j}) - h_2(d_{i,j})) \quad (9)$$

where $d_{i,j} = \min\{\min\{p_{i,j}, 1 - p_{i,j}\}, \lambda\}$ for $i, j \in \mathcal{V}$ and λ is chosen such that $\sum_{i < j} d_{i,j} = D$.

From the above theorem, we can easily see that the rate-distortion function of a Erdős-Rényi(ER) graph model $\mathcal{G}(n, p)$ with the Hamming distortion D is

$$R(D) = \binom{n}{2} \left[h_2(p) - h_2\left(\frac{D}{\binom{n}{2}}\right) \right],$$

if $0 \leq D \leq \binom{n}{2} \min\{p, 1 - p\}$, and $R(D) = 0$ otherwise.

IV. PROOF OF THEOREM 1

By applying (4) to the pair (\mathbf{X}, G_n) , we obtain

$$R_{G_n|\mathbf{X}}(D) \leq R_G(D) \leq R_{G_n|\mathbf{X}}(D) + I(G; \mathbf{X}). \quad (10)$$

Since $I(G; \mathbf{X}) \leq nH(\mathbf{p})$, we have $R_{G_n|\mathbf{X}}(D) \leq R_G(D) \leq R_{G_n|\mathbf{X}}(D) + \mathcal{O}(n)$. The second part of the theorem immediately follows because $R_{G_n|\mathbf{X}}(D) = 0$ when $\binom{n}{2} \sum_{l,m} p_l p_m \min\{w_{l,m}, 1 - w_{l,m}\} \leq D \leq \min\{\sum_{l,m} p_l p_m w_{l,m}, \sum_{l,m} p_l p_m (1 - w_{l,m})\}$. For the first part, it is enough to show that $R_{G_n|\mathbf{X}}(D)$ satisfies (8).

To this end, consider a conditional joint distribution $P_{\hat{G}_n|G_n, \mathbf{X}}$ satisfying the constraint $\mathbb{E}[d(G_n, \hat{G}_n)] \leq D$. For this distribution, we have

$$\begin{aligned} I(G_n; \hat{G}_n|\mathbf{X}) &= H(G_n|\mathbf{X}) - H(G_n|\hat{G}_n, \mathbf{X}) \\ &\geq \sum_{i < j} H(E_{ij}|X_i, X_j) - H(E_{ij}|\hat{E}_{ij}, \mathbf{X}) \end{aligned} \quad (11)$$

$$\geq \sum_{i < j} \sum_{l,m} p_l p_m \max\{H(E_{ij}|X_i, X_j) - H(E_{ij} \oplus \hat{E}_{ij}|X_i = l, X_j = m), 0\} \quad (12)$$

$$= \sum_{i < j} \sum_{l,m} p_l p_m \max\{h_2(w_{l,m}) - h_2(d_{l,m}^{i,j}), 0\} \quad (13)$$

$$= \sum_{i < j} \sum_{l,m} p_l p_m h_2(w_{l,m}) - \min\{h_2(d_{l,m}^{i,j}), h_2(w_{l,m})\}$$

$$\geq \binom{n}{2} \mathbf{p}^T h_2(\mathbf{W}) \mathbf{p} - \sum_{l,m} p_l p_m \min\left\{ \sum_{i < j} h_2(d_{l,m}^{i,j}), \binom{n}{2} h_2(w_{l,m}) \right\} \quad (14)$$

$$\geq \binom{n}{2} \mathbf{p}^T h_2(\mathbf{W}) \mathbf{p} - \binom{n}{2} \sum_{l,m} p_l p_m \min\{h_2(d_{l,m}), h_2(w_{l,m})\}. \quad (15)$$

In (11), we use conditional independence of the $E_{i,j}$'s. The inequality (12) follows from the fact that mutual information is nonnegative. In (13), we define

$$\begin{aligned} d_{l,m}^{i,j} &\triangleq \min\left\{ \mathbb{P}(E_{ij} \oplus \hat{E}_{ij} = 1 | X_i = l, X_j = m), \right. \\ &\quad \left. \mathbb{P}(E_{ij} \oplus \hat{E}_{ij} = 0 | X_i = l, X_j = m) \right\}. \end{aligned}$$

For (14), we rely on a simple fact that for two real number sequences (a_1, a_2, \dots, a_n) and (b_1, b_2, \dots, b_n) , $\sum_i \min\{a_i, b_i\} \leq$

$\min\{\sum_i a_i, \sum_i b_i\}$. The inequality (15), where we set $d_{l,m} \triangleq \sum_{i < j} \frac{d_{l,m}^{i,j}}{u}$, follows immediately from Jensen's inequality and concavity of the entropy. Notice that $d_{l,m}$'s satisfy the condition

$$D \geq \mathbb{E}[d(E^u, \hat{E}^u)] \geq \sum_{i < j} \sum_{l,m} d_{l,m}^{i,j} p_l p_m = u \sum_{l,m} p_l p_m d_{l,m}.$$

From (15), we have

$$\begin{aligned} R_{G|\mathbf{X}}(D) &\geq \binom{n}{2} \left[\mathbf{p}^T h_2(\mathbf{W}) \mathbf{p} \right. \\ &\quad \left. - \max\left\{ \sum_{l,m} p_l p_m \min\{h_2(d_{l,m}), h_2(w_{l,m})\} \right\} \right], \end{aligned} \quad (16)$$

where the maximization in (16) is with respect to the constraint $\binom{n}{2} \sum_{l,m} d_{l,m} p_l p_m \leq D$. Notice that the optimization problem in (16) is equivalent to the following optimization problem:

$$\begin{aligned} \text{minimize} \quad & - \sum_{l,m} p_l p_m h_2(d_{l,m}) \\ \text{subject to} \quad & \binom{n}{2} \sum_{l,m} p_l p_m d_{l,m} \leq D, \quad l, m \in [k], \\ & d_{l,m} \leq \min\{w_{l,m}, 1 - w_{l,m}\}, \quad l, m \in [k]. \end{aligned} \quad (17)$$

This is simply because we can map an optimizer of (16) to a point in the constraint of (17) such that the objective functions of both these problems evaluated at these points remain the same, and vice versa. The Lagrangian for the above optimization problem is

$$\begin{aligned} L &= - \sum_{l,m} p_l p_m H(d_{l,m}) + \nu \left(\sum_{l,m} p_l p_m d_{l,m} - \frac{D}{u} \right) \\ &\quad + \sum_{l,m} \lambda_{l,m} (d_{l,m} - \min\{w_{l,m}, 1 - w_{l,m}\}). \end{aligned} \quad (18)$$

Since (17) is a convex optimization problem function satisfying Slater's condition, the KKT conditions give sufficient and necessary conditions for an optimizer. By solving the corresponding KKT conditions, we can show that an optimizer of (17) is

$$d_{l,m}^* = \min\{\min\{w_{l,m}, 1 - w_{l,m}\}, \mu\}, \quad \forall l, m \in [k],$$

where μ is chosen such that $\binom{n}{2} \sum_{l,m} p_l p_m d_{l,m}^* = D$. Hence, we obtain from (16) that

$$R_{G|\mathbf{X}}(D) \geq \binom{n}{2} \mathbf{p}^T [h_2(\mathbf{W}) - h_2(\mathbf{D}^*)] \mathbf{p}, \quad (19)$$

where \mathbf{D}^* is a matrix with entries $d_{l,m}^*$.

Achievability of the lower bound: Now it remains to show that inequality in (19) holds with equality. To this end, we have to produce a conditional probability distribution $P_{\hat{G}_n|G_n, \mathbf{X}}$ (or equivalently, $P_{\{\hat{E}_{i,j}: i < j\}|\{E_{i,j}: i < j\}, \mathbf{X}}$) that satisfies the distortion constraint and has $I(G_n; \hat{G}_n|\mathbf{X})$ equal to the expression on the right-hand side of (19). Let $d_{l,m}^*$ be chosen

according to the constraint in (19). Consider the conditional probability distribution of the form

$$P_{\{\hat{E}_{i,j}:i<j\}|\{E_{i,j}:i<j\},\mathbf{x}} = \prod_{i<j} \frac{P_{E_{ij}|\hat{E}_{ij},X_i,X_j} P_{\hat{E}_{ij}|X_i,X_j}}{P_{E_{ij}|X_i,X_j}},$$

where for $l, m \in [k]$, $P_{E_{ij}|\hat{E}_{ij},X_i,X_j}(\cdot | \cdot, l, m)$ is a binary symmetric channel with the crossover probability $d_{l,m}^*$ and $P_{\hat{E}_{ij}|X_i,X_j}(1 | l, m) = \frac{w_{l,m}-d_{l,m}^*}{1-2d_{l,m}^*}$ and $P_{\hat{E}_{ij}|X_i,X_j}(0 | l, m) = \frac{1-w_{l,m}-d_{l,m}^*}{1-2d_{l,m}^*}$, which are non-negative by virtue of the way $d_{l,m}^*$'s are defined. (Note that these distributions are the same for all $i < j$.) It is easy to see that this distribution satisfies the distortion criterion:

$$\mathbb{E}[d(G_n, \hat{G}_n)] = \sum_{i<j} \mathbb{P}(E_{ij} \oplus \hat{E}_{ij} = 1) = \binom{n}{2} \mathbf{p}^T \mathbf{D}^* \mathbf{p} = D.$$

This implies that

$$R_{G|\mathbf{X}}(D) \leq I(G_n; \hat{G}_n | \mathbf{X}) = \binom{n}{2} \mathbf{p}^T [h_2(\mathbf{W}) - h_2(\mathbf{D}^*)] \mathbf{p}.$$

By combining this inequality and (19), we get the desired result.

V. CONCLUSION AND FUTURE WORK

In this paper, we addressed the lossy compression of SBM by deriving its rate-distortion function under the Hamming distortion. Our approach involved considering the conditional rate-distortion function with community memberships as side information for the encoder and decoder. In addition, we characterized the rate-distortion function of the Erdős-Rényi random graph models. In both these models, the rate-distortion expressions contain a matrix \mathbf{D}^* , whose entries are chosen using a version of the “reverse-water filling” solution.

In future work, we plan to explore the case where the distortion measure takes into account the community labels. This could lead to the problem of partial recovery in the stochastic block model [15]. We also aim to investigate distortion measures that effectively preserve the graph's structure during reconstruction, such as the measure considered in [20], [21]. These results could also be incorporated in applications such as graph anomaly detection through the use of coding algorithms and in reducing overhead during topology inference in networks.

REFERENCES

- [1] A. Eisenman, L. Cherkasova, G. Magalhaes, Q. Cai, and S. Katti, “Parallel graph processing on modern multi-core servers: New findings and remaining challenges,” in *2016 IEEE 24th International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems (MASCOTS)*, pp. 49–58, 2016.
- [2] O. Batarfi, R. E. Shawi, A. G. Fayoumi, R. Nouri, S.-M.-R. Beheshti, A. Barnawi, and S. Sakr, “Large scale graph processing systems: Survey and an experimental evaluation,” *Cluster Computing*, vol. 18, no. 3, pp. 1189–1213, 2015.
- [3] L. V. S. Lakshmanan, “Social network analytics: Beyond the obvious,” in *Biomedical Data Management and Graph Online Querying*, (Cham), pp. 149–154, Springer International Publishing, 2016.
- [4] M. Besta and T. Hoefer, “Survey and taxonomy of lossless graph compression and space-efficient graph representations,” *arXiv preprint arXiv:1806.01799*, 2018.
- [5] M. Besta, S. Weber, L. Gianinazzi, R. Gerstenberger, A. Ivanov, Y. Oltchik, and T. Hoefer, “Slim graph: Practical lossy graph compression for approximate graph processing, storage, and analytics,” in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, SC '19*, (New York, NY, USA), Association for Computing Machinery, 2019.
- [6] M.-A. Badiu and J. P. Coon, “Structural complexity of one-dimensional random geometric graphs,” *IEEE Transactions on Information Theory*, vol. 69, no. 2, pp. 794–812, 2023.
- [7] P. W. Holland, K. B. Laskey, and S. Leinhardt, “Stochastic block models: First steps,” *Social Networks*, vol. 5, no. 2, pp. 109–137, 1983.
- [8] S. Fortunato and D. Hric, “Community detection in networks: A user guide,” *arXiv*, vol. abs/1608.00163, 2016.
- [9] C. Lee and D. J. Wilkinson, “A review of stochastic block models and extensions for graph clustering,” *Applied Network Science*, vol. 4, no. 1, p. 122, 2019.
- [10] H. Saad and A. Nosratinia, “Community detection with side information: Exact recovery under the stochastic block model,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 12, no. 5, pp. 944–958, 2018.
- [11] E. Abbe, “Community detection and stochastic block models: Recent developments,” *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 6446–6531, 2017.
- [12] F. Zhao, J. Sima, and S.-L. Huang, “On the optimal error rate of stochastic block model with symmetric side information,” in *2021 IEEE Information Theory Workshop (ITW)*, pp. 1–6, 2021.
- [13] J. Sima, F. Zhao, and S.-L. Huang, “Exact recovery in the balanced stochastic block model with side information,” in *2021 IEEE Information Theory Workshop (ITW)*, pp. 1–6, 2021.
- [14] D.-H. Jung, G. Im, J.-G. Ryu, S. Park, H. Yu, and J. Choi, “Satellite clustering for non-terrestrial networks: Concept, architectures, and applications,” *arXiv preprint arXiv:2301.08386*, 2023.
- [15] E. Abbe, “Graph compression: The effect of clusters,” in *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 1–8, 2016.
- [16] A. Bhatt, Z. Wang, C. Wang, and L. Wang, “Universal graph compression: Stochastic block models,” in *2021 IEEE International Symposium on Information Theory (ISIT)*, pp. 3038–3043, 2021.
- [17] Y. Choi and W. Szpankowski, “Compression of graphical structures: Fundamental limits, algorithms, and experiments,” *IEEE Transactions on Information Theory*, vol. 58, no. 2, pp. 620–638, 2012.
- [18] J. Han, T. Guo, Q. Zhou, W. Han, B. Bai, and G. Zhang, “Structural entropy of the stochastic block models,” *Entropy*, vol. 24, no. 1, 2022.
- [19] A. R. Asadi, E. Abbe, and S. Verdú, “Compressing data on graphs with clusters,” in *2017 IEEE International Symposium on Information Theory (ISIT)*, pp. 1583–1587, 2017.
- [20] R. Bustin and O. Shayevitz, “On lossy compression of binary matrices,” in *2017 IEEE International Symposium on Information Theory (ISIT)*, pp. 1573–1577, 2017.
- [21] R. Bustin and O. Shayevitz, “On lossy compression of directed graphs,” *IEEE Transactions on Information Theory*, vol. 68, no. 4, pp. 2101–2122, 2022.
- [22] C. W. Lynn and D. S. Bassett, “Quantifying the compressibility of complex networks,” *Proceedings of the National Academy of Sciences*, vol. 118, no. 32, p. e2023473118, 2021.
- [23] W. Henecka and M. Roughan, “Lossy compression of dynamic, weighted graphs,” in *2015 3rd International Conference on Future Internet of Things and Cloud*, pp. 427–434, 2015.
- [24] A. Wyner and J. Ziv, “The rate-distortion function for source coding with side information at the decoder,” *IEEE Transactions on Information Theory*, vol. 22, no. 1, pp. 1–10, 1976.
- [25] H. Yamamoto, “Wyner - Ziv theory for a general function of the correlated sources (Corresp.),” *IEEE Transactions on Information Theory*, vol. 28, no. 5, pp. 803–807, 1982.
- [26] I. N. Kovalenko, “Theory of random graphs,” *Cybernetics*, vol. 7, no. 4, pp. 575–579, 1971.
- [27] C. E. Shannon, “Coding theorems for a discrete source with a fidelity criterion,” *IRE Nat. Conv. Rec.*, vol. 4, no. 142-163, p. 1, 1959.
- [28] T. M. Cover and J. A. Thomas, *Elements of information theory* (Wiley Series in Telecommunications and Signal Processing). USA: Wiley-Interscience, 2006.
- [29] R. Gray, “A new class of lower bounds to information rates of stationary sources via conditional rate-distortion functions,” *IEEE Transactions on Information Theory*, vol. 19, no. 4, pp. 480–489, 1973.