Article

Structural Entropy of the Stochastic Block Models

Jie Han [1], **Tao Guo** [1,*], **Qiaoqiao Zhou** [2], **Wei Han** [1], **Bo Bai** [1] and **Gong Zhang** [1]

[1] Theory Lab, Central Research Institute, 2012 Labs, Huawei Tech. Co., Ltd., Hong Kong SAR, China; han.jie@huawei.com (J.H.); harvey.hanwei@huawei.com (W.H.); baibo8@huawei.com (B.B.); nicholas.zhang@huawei.com (G.Z.)

[2] Department of Computer Science, School of Computing, National University of Singapore, Singapore 11741, Singapore; zhouqq@comp.nus.edu.sg

[*] Correspondence: guo.tao1@huawei.com

Abstract: With the rapid expansion of graphs and networks and the growing magnitude of data from all areas of science, effective treatment and compression schemes of context-dependent data is extremely desirable. A particularly interesting direction is to compress the data while keeping the "structural information" only and ignoring the concrete labelings. Under this direction, Choi and Szpankowski introduced the structures (unlabeled graphs) which allowed them to compute the structural entropy of the Erdős–Rényi random graph model. Moreover, they also provided an asymptotically optimal compression algorithm that (asymptotically) achieves this entropy limit and runs in expectation in linear time. In this paper, we consider the stochastic block models with an arbitrary number of parts. Indeed, we define a partitioned structural entropy for stochastic block models, which generalizes the structural entropy for unlabeled graphs and encodes the partition information as well. We then compute the partitioned structural entropy of the stochastic block models, and provide a compression scheme that asymptotically achieves this entropy limit.

Keywords: structural entropy; stochastic block model (SBM); network compression; optimal compression algorithm

1. Introduction

Shannon's metric of "Entropy" of information is a foundational concept of information theory [1,2]. Given a discrete random variable X with support set (that is, the possible outcomes) $x_1, x_2, \ldots, x_n$, which occurs with probability $p_1, p_2, \ldots, p_n$, the entropy of X is defined as

$$H(X) := -\sum_{i=1}^{n} p_i \log p_i,$$

where the logarithm here and throughout this paper is of base 2. Note that the entropy of X is a function of the probability distribution of X.

The entropy was originally created by Shannon in [3] as part of their theory of communication, where a data communication system consists of a data source X, a channel and a receiver. The fundamental problem of communication is for the receiver to reliably recover what data was generated by the source, based on the bits it receives through the channel. Shannon proved that the entropy of the source X plays a central role—in their source coding theorem it is shown that the entropy is the mathematical limit on how well the data can be losslessly compressed.

The question then arises: *How to compress data that has structures, e.g., data in social networks?* In Shannon's 1953 less known paper [4] he argued for an extension of information theory, where data is considered as observations of a source, to "non-conventional data" (that is, lattices). Indeed, nowadays data appears in various formats and structures (e.g., sequences, expressions, interactions) and in drastically increasing amounts. In many scenarios, data is highly context-dependent and in particular, the structural information

and the context information seem to be two conceptually different aspects. Therefore it is desirable to develop novel theory and efficient algorithms for extracting useful information from non-conventional data structures. Roughly speaking, such data consists of structural information, which, might be understood as the "shape" of the data, and context information which should be recognized as data labels.

It is well-known that complex networks (e.g., social networks) admit community structures [5]. That is, users within a group interact with each other more frequently than those outside the group. The stochastic block model (SBM) [6] is a celebrated random graph model that has been widely used to study the community structures in graphs and networks. It provides a good benchmark to evaluate the performance of community detection algorithms and inspires the design of many algorithms for community detection tasks. The theoretical underpinnings of the SBM have been extensively studied and sharp thresholds for exact recovery have been successively established [7–12]. We refer readers to [13] for a recent survey, where other interesting and important problems in SBM are also discussed.

In addition to the SBM discussed in [13], there are other angles to study compression of data with graph structures. Asadi et al. [14] investigated data compression on graphs with clusters. Zenil et al. [15] have surveyed information-theoretic methods, in particular Shannon entropy and algorithmic complexity, for characterizing graphs and networks.

1.1. Compression of Graphs

In recent years, graphical data and the network structures supporting them are becoming increasingly common and important in branches of engineering and sciences. To better represent and transmit graphical data, many works consider the problem of compressing the (random) graph up to isomorphism, i.e., compressing the structure of a graph. A graph G contains a finite set V of vertices and a set E of edges each of which connects two vertices. A graph can be represented by a binary matrix (the adjacency matrix) that further can be viewed as a binary sequence. Thus, encoding a labeled graph (that is, all vertices need to be distinguished) is equivalent to encoding the $\binom{|V|}{2}$-digit binary sequence, given certain probability distribution on all $\binom{|V|}{2}$ possible edges. However, such a string does not reflect internal symmetries that are conveyed by the graph automorphism, and sometimes we are only interested in the local or global structures in the graph, rather than the exact vertex labelings. The structural entropy is defined when the graphs are considered unlabeled, or simply called structures, where the vertices are viewed as undistinguishable. The goal of this natural definition is to capture the information of the structure, and thus provides a fundamental measure in graph/structure compression schemes.

The problem actually has a strong theoretical background. Back to 1984, Turán [16] raised the question of finding an efficient coding method for general unlabeled graphs on n vertices, where a lower bound of $\binom{n}{2} - n \log n + O(n)$ bits is suggested. This lower bound can be seen by the number of unlabeled graphs [17]. The question was later answered by Naor [18] in 1990 who proposed such a representation that is optimal up to the first two leading terms when all unlabeled graphs are equally likely. In a recent paper Kieffer et al. [19] proved a structural complexity of a binary tree. There also have been some heuristic methods for real-world graph compression schemes, see [20–24]. Rather recently, Choi and Szpankowski [25] studied the structural entropy of the Erdős–Rényi random graph $\mathcal{G}(n,p)$. They computed the structural entropy given that p is not (very) close to 0 or 1 and also gave a compression scheme that matches their computation. Later, the structural entropy for other randomly generated graphs, e.g., the preferential attachment graphs and web graphs are also studied [26–29].

However, it is well-known that the Erdős–Rényi model is too simplistic to model real networks, in particular due to its strong homogeneity and absence of community structure. In this paper, we consider the compression of graphical structures of the SBM, which in general model real networks better and circumvent the issues of the ER-model. In summary, our contributions are as follows:

- We introduce the partitioned structural entropy which generalizes the structural entropy for unlabeled graphs and we show that it reflects the partition information of the SBM.
- We provide an explicit formula for the partitioned structural entropy of the SBM.
- We also propose a compression scheme that asymptotically achieves this entropy limit.

Semantic communications are considered as a key component of future generation networks, where a natural problem to consider is how to efficiently extract and transmit the "semantic information". In the case of graph data, one may view the (partitioned) structures as the information that needs to be abstracted while the concrete labeling information is considered redundant. From this point of view, our result is a step for the study of semantic compression/communication under appropriate contexts.

1.2. Related Works

Finally, we would like to point out that there are some other information metrics defined on graphs. The term "graph entropy" has been defined and used in the history. For example, graph entropy introduced by Kőrner in [30] denotes the number of bits one has to convey to resolve the ambiguity of a vertex in a graph. This notion also turns out to be useful in other areas, including combinatorics. Chromatic entropy introduced in [31] is the lowest entropy of any coloring of a graph. It finds application in zero-error source coding. We remark that the structural entropy we considered is quite different from the Kőrner graph entropy and chromatic entropy.

On the other hand, a concept of graph entropy (also called *topological information content of a graph*) was introduced by Rashevsky [32] and Trucco [33], and later by Mowshowitz [34–39], which is defined as a function of (the structure of) a graph and an equivalence relation defined on its vertices or edges. Such a concept is a measure of the graph itself and does not involve any probability distribution.

2. Preliminaries

2.1. Structural Entropy of Unlabeled Graphs

Now let us formally define the structural entropy given a probability distribution on unlabeled graphs. In this subsection, we use notations borrowed from [25].

Given an integer n, define $\mathcal{G}_n$ as the collection of all n-vertex labeled graphs.

Definition 1 (Entropy of Random Graph). *Given an integer n and a random graph $\mathcal{G}$ distributed over $\mathcal{G}_n$, the entropy of $\mathcal{G}$ is defined as:*

$$H_{\mathcal{G}} = \mathbb{E}[-\log P(G)] = -\sum_{G \in \mathcal{G}_n} P(G) \log P(G)$$

where $P(G) \triangleq P(\mathcal{G} = G)$ is the probability of a graph G in $\mathcal{G}_n$.

Then the random structure model $\mathcal{S}_n$ associated with the probability distribution $\mathcal{G}_n$, is defined as the unlabeled version of $\mathcal{G}_n$. For a given $S \in \mathcal{S}_n$, the probability of S can be computed as:

$$P(S) = \sum_{G \cong S, G \in \mathcal{G}_n} P(G).$$

Here $G \cong S$ means that G and S have the same structure, that is, S is isomorphic to G. Clearly if all isomorphic labeled graphs have the same probability, then for any labeled graph $G \cong S$, one has:

$$P(S) = N(S) \cdot P(G)$$

where $N(S)$ stands for the number of different labeled graphs that have the same structure as S.

Definition 2 (Structural Entropy)**.** *The structural entropy $H_\mathcal{S}$ of a random graph $\mathcal{G}$ is defined as the entropy of a random structure $\mathcal{S}$ associated with $\mathcal{G}_n$, that is,*

$$H_\mathcal{S} = \mathbb{E}[-\log P(S)] = -\sum_{S \in \mathcal{S}} P(S) \log P(S)$$

where the sum is over all distinct structures.

The Erdős–Rényi random graph $\mathcal{G}(n, p)$, also called the binomial random graph, is a fundamental random graph model, which has n vertices and each pair of vertices is connected with probability p, independent of other pairs. In 2012, Choi and Szpankowski [25] proved the following for the Erdős–Rényi random graphs.

Theorem 1 (Choi and Szpankowski, [25])**.** *For large n and all p satisfying $n^{-1} \ln n \ll p$ and $1 - p \gg n^{-1} \ln n$, the following holds:*

1. *The structural entropy $H_\mathcal{S}$ of $\mathcal{G}(n, p)$ is:*

$$H_\mathcal{S} = \binom{n}{2} h(p) - \log n! + O\left(\frac{\log n}{n^\alpha}\right)$$

   for some $\alpha > 0$.
2. *For a structure S of n vertices and $\varepsilon > 0$*

$$P\left(\left|-\frac{1}{\binom{n}{2}} \log P(S) - h(p) + \frac{\log n!}{\binom{n}{2}}\right| < \varepsilon\right) > 1 - 2\varepsilon$$

   where $h(p) = -p \log p - (1 - p) \log(1 - p)$ is the entropy rate of a binary memoryless source.

Furthermore, they [25] also presented a compression algorithm for unlabeled graphs that asymptotically achieves the structural entropy up to an $O(n)$ error term.

2.2. Stochastic Block Model–Our Result

As the ER model is not appropriate to model real networks, the stochastic block model is introduced on the assumption that vertices in a network connect independently but with probability based on their profiles, or equivalently, on their community assignment. For example, in the SBM with two communities and symmetric parameters, also known as the planted bisection model, denoted by $\mathcal{G}(n, p, q)$, the vertex set is partitioned into two sets V_1 and V_2, any pair of vertices inside V_1 or V_2 are connected with probability p and any pair of vertices across the clusters are connected with probability q, and all these connections are independent.

As an illuminating example, consider a context G where there are $n/2$ users and $n/2$ devices, and each pair of users and each pair of devices are connected with probability p, a user and a device is connected with probability q and each of these connections is independent of all other connections. Suppose that we need to compress the information of G. However, in the context it is not appropriate to view G as an unlabeled graph, that is, in addition to the structure information, it is also important to keep the "community" information – the compression also needs to encode the information that who is a user and who is device.

Definition 3 (Partition-respecting isomorphism, Partitioned Unlabeled Graphs)**.** *Let $r \leq n$ be integers. Suppose V is a set of n vertices and $\mathcal{P} = \{V_1, V_2, \ldots, V_r\}$ is a partition of V into r parts. The partition-respecting isomorphism, denoted by "$\cong_\mathcal{P}$" is defined as follows. For any two labeled graphs G and G', we write $G \cong_\mathcal{P} G'$ if and only if $G \cong G'$ they are isomorphic via an isomorphism function $\phi : V \to V$ such that $\phi(V_i) = V_i$, for $1 \leq i \leq r$. Then $\Gamma_\mathcal{P}$ is defined as the*

collection of n-vertex graphs on V where we ignore the labels of vertices inside each V_i, $1 \leq i \leq r$, namely, the equivalence classes under partition-respecting isomorphism, with respect to $\mathcal{P}$.

Note that every labeled graph G corresponds to a unique structure $S \in \Gamma_{\mathcal{P}}$, and we use $G \cong_{\mathcal{P}} S$ to denote this relation. Furthermore, under the above definition, general unlabeled graphs correspond to the case $r = 1$.

Definition 4 (Partitioned Structural Entropy). *Let V be a set of n vertices where $n \in \mathbb{N}$. Suppose $\mathcal{P} = \{V_1, V_2, \ldots, V_r\}$ is a partition of V into r parts and $\mathcal{S}_n$ is a probability distribution over all partitioned unlabeled graphs on n vertices. Then the structural entropy $H_{\mathcal{S}}$ associated to $\mathcal{S}_n$ is defined by:*

$$H_{\mathcal{S}} = \mathbb{E}[-\log P(S)] = -\sum_{S \in \mathcal{S}_n} P(S) \log P(S).$$

In this paper, we extend Theorem 1 to the structural entropy of the stochastic block model with any given number of blocks, and provide a compression algorithm that asymptotically matches this structural entropy. For ease of comprehension, we first give the result for the balanced bipartition case $\mathcal{G}(n, p, q)$.

Theorem 2. *Let n be a positive even integer and let $V = V_1 \cup V_2$ be a set of n vertices with $|V_1| = |V_2| = n/2$. Suppose $\mathcal{G}(n, p, q)$ is a probability distribution of graphs on V where every edge inside V_1 or V_2 is present with probability p and every edge between V_1 and V_2 is present with probability q, and these edges are mutually independent. For large even n and all p satisfying $n^{-1} \ln n \ll p, q$ and $1 - p \gg n^{-1} \ln n$, the following holds:*

(i)   *The partitioned structural entropy $H_{\mathcal{S}}$ of $\mathcal{G}(n, p, q)$ is:*

$$H_{\mathcal{S}} = 2\binom{n/2}{2}h(p) + \frac{n^2}{4}h(q) - 2\log\left(\frac{n}{2}\right)! + O\left(\frac{\log n}{n^{\alpha}}\right) \tag{1}$$

for some $\alpha > 0$.

(ii)   *For a balanced bipartitioned structure S and $\varepsilon > 0$*

$$P\left(\left|-\frac{1}{\binom{n}{2}}\log P(S) - \frac{n-2}{2n-2}h(p) - \frac{n}{2n-2}h(q) + \frac{2\log(n/2)!}{\binom{n}{2}}\right| < 3\varepsilon\right) > 1 - 4\varepsilon$$

where $h(p) = -p\log p - (1-p)\log(1-p)$ is the entropy rate of a binary memoryless source.

Note that the structural entropy $H_{\mathcal{S}}$ here is larger than that in Theorem 1 (even if $p = q$), which reflects the fact that the SBM with "a planted (bi-)partition" contains *prefixed* structures, so has less symmetries than $\mathcal{G}(n, p)$, the pure random model (For $\mathcal{G}(n, p)$, when it is asymmetric, comparing with the completely labeled graphs, Theorem 1 saves a term as $\log n!$; this saving becomes $2\log(n/2)!$ for the planted balanced bipartition case in Theorem 2).

3. Proof of Theorem 2

One key ingredient in the proof of Theorem 1 in [25] is the following lemma on the symmetry of $\mathcal{G}(n, p)$. A graph is called asymmetric if its automorphism group does not contain any permutation other than identity; otherwise it is called symmetric.

Lemma 1 (Kim, Sudakov and Vu, 2002). *For all p satisfying $n^{-1} \ln n \ll p$ and $1 - p \gg n^{-1} \ln n$, a random graph $G \in \mathcal{G}(n, p)$ is symmetric with probability $O(n^{-w})$ for any positive constant w.*

Proof of Theorem 2. Note that every pair of vertices in V_1 or in V_2 should be considered as undistinguishable, but not the pairs of vertices in $V_1 \times V_2$. Recall that we write $G \cong_{\mathcal{P}} S$ for a graph G and a structure S if S represents the structure of G (with respect to the partition $\mathcal{P}$).

Let $\mathcal{G} := \mathcal{G}(n, p, q)$. We first compute $H_{\mathcal{G}}$. Note that there are $\binom{n}{2}$ possible edges in $G \in \mathcal{G}$, and we can view it as a binary sequence of length $\binom{n}{2}$, where each digit is a Bernoulli random variable. Moreover, for edges inside V_1 or V_2, the random variable, denoted by X_1, has expectation p and for edges in $V_1 \times V_2$ the random variable, denoted by X_2, has expectation q. Thus, we have:

$$
\begin{aligned}
H_{\mathcal{G}} &= -\mathbb{E}[\log X_1^{2\binom{n/2}{2}} X_2^{n^2/4}] \\
&= -2\binom{n/2}{2}\mathbb{E}[\log X_1] - \frac{n^2}{4}\mathbb{E}[\log X_2] \\
&= 2\binom{n/2}{2}h(p) + \frac{n^2}{4}h(q).
\end{aligned}
$$

Now write $\mathcal{S}_n$ for the probability distribution on V over all partitioned unlabeled graphs inherited from $\mathcal{G}$, namely, given $S \in \Gamma_{\mathcal{P}}$, $P(S) = \sum_{G \cong_{\mathcal{P}} S} P(G)$. Let $H_{\mathcal{S}}$ be the partitioned structural entropy of $\mathcal{S}_n$. Therefore, compared with our goal, it remains to show that:

$$
H_{\mathcal{S}} - H_{\mathcal{G}} = -2\log(n/2)! + O\left(\frac{\log n}{n^{\alpha}}\right). \tag{2}
$$

Note that in $\mathcal{G}(n, p, q)$, all labeled graphs $G \in \mathcal{G}$ such that $G \cong_{\mathcal{P}} S$ have the same probability $P(G)$. Thus, given a (labeled) graph $G \in \mathcal{G}$, we have $P(G) = P(S)/N(S)$, where $S \in \mathcal{S}_n$ is such that $G \cong_{\mathcal{P}} S$. So the graph entropy of $\mathcal{G} = \mathcal{G}(n, p, q)$ can be written as:

$$
\begin{aligned}
H_{\mathcal{G}} &= -\sum_{G \in \mathcal{G}} P(G)\log P(G) \\
&= -\sum_{S \in \mathcal{S}_n}\sum_{G \cong_{\mathcal{P}} S, G \in \mathcal{G}} P(G)\log P(G) \\
&= -\sum_{S \in \mathcal{S}_n}\sum_{G \cong_{\mathcal{P}} S, G \in \mathcal{G}} \frac{P(S)}{N(S)}\log\frac{P(S)}{N(S)} \\
&= -\sum_{S \in \mathcal{S}_n} P(S)\log\frac{P(S)}{N(S)} \\
&= H_{\mathcal{S}} + \sum_{S \in \mathcal{S}} P(S)\log N(S) \tag{3}
\end{aligned}
$$

Define $S[W]$ be be S restricted on W for $W \in V$. Now we split S into S_1 and S_2, i.e., $S_1 = S[V_1]$ and $S_2 = S[V_2]$. Write $Aut(S_i)$ for the automorphism group for S_i, and we naturally have:

$$
N(S) = \frac{(n/2)! \cdot (n/2)!}{|Aut(S_1)||Aut(S_2)|}.
$$

Combining this with (2) and (3), it remains to show that:

$$
\sum_{S \in \mathcal{S}} P(S)\log|Aut(S_1)||Aut(S_2)| = O\left(\frac{\log n}{n^{\alpha}}\right).
$$

In the summation above we only need to focus on S such that either S_1 or S_2 is symmetric, as otherwise $\log|Aut(S_1)||Aut(S_2)| = \log 1 = 0$. By Lemma 1, we conclude that the probability of S restricted on V_1 or V_2 is symmetric is $O(n^{-1-\alpha})$ for some $\alpha > 0$,

and for such S we use the trivial bound $\log|Aut(S_1)||Aut(S_2)| \leq 2\log(n/2)! \leq 2n\log n$. This gives us the desired estimate in (i)

$$\sum_{S \in \mathcal{S}} P(S)\log|Aut(S_1)||Aut(S_2)| \leq 2n\log n \cdot O(n^{-1-\alpha}) = O\left(\frac{\log n}{n^\alpha}\right).$$

To show (ii), for a set V of n vertices and a balanced bipartition $\mathcal{P} = (V_1, V_2)$ of V, we define the typical set T_ε^n as the set of structures S on n vertices satisfying:

(a) S is asymmetric on V_1 and V_2, respectively;

(b) $2^{-2\binom{n/2}{2}h(p) - \frac{n^2}{4}h(q) - \binom{n}{2}\varepsilon} \leq P(G) \leq 2^{-2\binom{n/2}{2}h(p) - \frac{n^2}{4}h(q) + \binom{n}{2}\varepsilon}$, for $G \cong_{\mathcal{P}} S$.

Denote by T_1^n and T_2^n the sets of structures satisfying the properties (a) and (b), respectively, and thus we have $T_\varepsilon^n = T_1^n \cap T_2^n$. Firstly, by the asymmetry of $\mathcal{G}(n, p)$ (Lemma 1), we conclude that $P(T_1^n) > 1 - 2\varepsilon$ for large n. Secondly, we use a binary sequence of length $\binom{n}{2}$ to represent a (labeled) instance G of $\mathcal{G}(n, p, q)$, where the first $\binom{n/2}{2}$ bits $\mathcal{L}_1$ represent the induced subgraph on V_1, the next $\binom{n/2}{2}$ bits $\mathcal{L}_2$ represent the induced subgraph on V_2, and finally the rest $n^2/4$ bits $\mathcal{L}_{12}$ represent the bipartite graph on $V_1 \times V_2$. Since all edges of G are generated independently, both $\mathcal{L}_1$ and $\mathcal{L}_2$ have in expectation $\binom{n/2}{2}p$ 1's and the AEP property of the binary sequences implies that:

$$2^{-\binom{n/2}{2}h(p) - \binom{n}{2}\varepsilon} \leq P(G[V_1]), P(G[V_2]) \leq 2^{-\binom{n/2}{2}h(p) + \binom{n}{2}\varepsilon}$$

holds with probability at least $1 - 2\varepsilon$. Similarly, $\mathcal{L}_{12}$ has in expectation $(n^2/4)q$ 1's and the AEP property of the binary sequences gives that with probability at least $1 - \varepsilon$,

$$2^{-\frac{n^2}{4}h(q) - \binom{n}{2}\varepsilon} \leq P(G[V_1, V_2]) \leq 2^{-\frac{n^2}{4}h(q) + \binom{n}{2}\varepsilon}$$

Since these edges are independent, we finally conclude that (b) holds with probability at least $1 - 3\varepsilon$. Thus, $P(T_\varepsilon^n) \geq 1 - 4\varepsilon$. Now we can compute $P(S)$ for $S \in T_\varepsilon^n$. By (a), $P(S) = (n/2)!(n/2)!P(G)$ for any $G \cong S$. Together with (b) and straightforward computation, the assertion of (ii) follows. $\square$

4. SBM Compression Algorithm

Given the computation of the structural entropy, a natural next step is to design efficient compression schemes that are close to or even (asymptotically) achieve this entropy limit. Choi and Szpankowski [25] presented such an algorithm (which they named Szip) for (unlabeled) random graphs, which uses in expectation at most $\binom{n}{2}h(p) - n\log n + O(n)$ bits and asymptotically achieves the structural entropy given in Theorem 1. Roughly speaking, Szip greedily peels off vertices from the graph and (efficiently) store the neighborhood information. This procedure can be simply reversed but the labeling of the recovered graph may be different from the original graph, which is the reason on why a saving of the codeword length is achieved. Refinements and analysis [25] are also provided to achieve the proposed performance.

Here we give an algorithm that optimally compresses SBMs which uses the Szip algorithm as building blocks and matches the structural entropy computation in Theorem 2. The algorithm consists of two stages. It first compresses $S[V_1]$ and $S[V_2]$ using Szip and then compresses $S[V_1, V_2]$ using an arithmetic compression algorithm with the help of Szip decoding outputs.

To give a brief description of the compression algorithm, we again use the balanced bipartition $V_1 \cup V_2$ as an example. The encoding and decoding procedure of the algorithm is illustrated in Figure 1. The algorithm encodes the observed $\mathcal{S}(n, p, q)$ into a binary string as follows. It uses Szip as a subroutine to compress $S[V_1]$ and $S[V_2]$ into binary sequences $\mathcal{L}_1$ and $\mathcal{L}_2$. Then, as part of the encoder, we run the Szip decoder on $\mathcal{L}_1$ and $\mathcal{L}_2$ to obtain decoded structures $S'[V_1]$ and $S'[V_2]$, respectively. We then compress $S[V_1, V_2]$ as a labeled

bipartite graph under the vertex labeling of $S'[V_1]$ and $S'[V_2]$ into $\mathcal{L}_{12}$. This "Labeled Encoder" can be done by treating it as a binary sequence of length $n^2/4$ and compressing using a standard arithmetic encoder [40–42]. The concatenation of SZIP algorithms and the arithmetic encoder forms the cascade encoder of our algorithm and obtains the codeword $(\mathcal{L}_1, \mathcal{L}_2, \mathcal{L}_{12})$. Upon receiving the codeword, we decode them parallelly using SZIP decoder and the arithmetic decoder. This completes our algorithm.
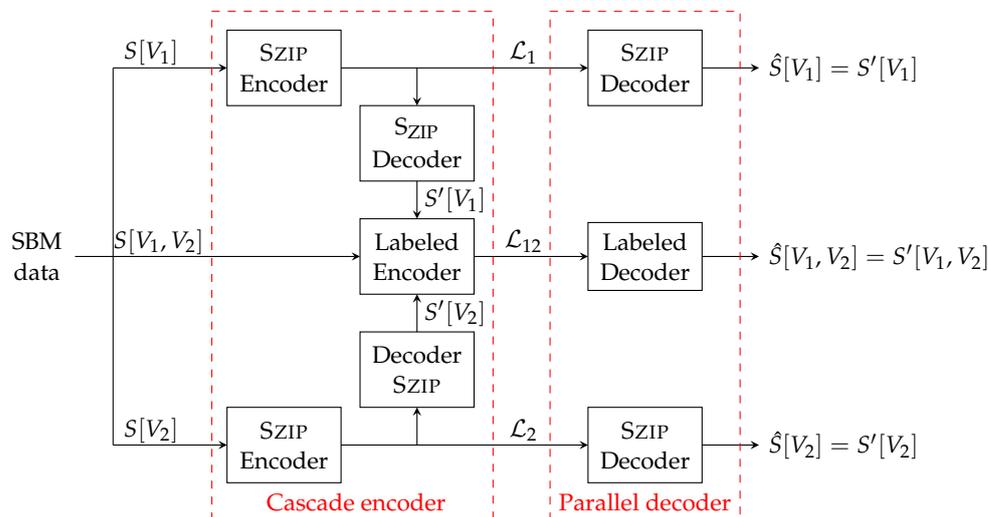

Figure 1. Illustration of compression algorithm.

The main challenge in the design of our algorithm is how the decoder can retrieve the consistency between the bipartite graph $S[V_1, V_2]$ and the decoded version of $S[V_1]$ and $S[V_2]$. A key observation here is that since SZIP is a deterministic algorithm, although it may permute the vertex labelings, its output is an invariant given the same input. Given this, our solution here is to first run SZIP (both encoding and decoding) at the encoder, and obtain structures $S'[V_1]$ and $S'[V_2]$, respectively. We then compress $S[V_1, V_2]$ (as a labeled bipartite graph) under the vertex labeling of $S'[V_1]$ and $S'[V_2]$. This would guarantee that the decoded structures $\hat{S}[V_1]$, $\hat{S}[V_2]$ and $\hat{S}[V_1, V_2]$ share the same vertex labeling as $S'[V_1]$ and $S'[V_2]$, namely, S is recovered.

Before discussing the performance of the algorithm, we first describe some useful properties of the arithmetic compression algorithm in the following lemma. We omit the proof of the lemma, which follows from the analysis in [40–42] and AEP properties in [1,2].

Lemma 2. *Let L be the codeword length of the arithmetic compression algorithm when compressing a binary sequence with length m and entropy rate h. For large m, the following holds:*

(i)   *The expected codeword length asymptotically achieves the entropy of the message, i.e.,*

$$\mathbb{E}[L] = mh + O(\log m). \tag{4}$$

(ii)   *For any $\epsilon > 0$,*

$$P(|L - \mathbb{E}[L]| \leq \epsilon \log m) \geq 1 - o(1). \tag{5}$$

(iii)   *The arithmetic algorithm runs in time $O(m)$.*

The following theorem characterizes the performance of our algorithm. It is immediate from Theorem 2 in [25] (performance of SZIP) and Lemma 2, we omit the detailed proofs here.

Theorem 3. *Let $V = V_1 \cup V_2$ be a set of n vertices and $|V_1| = |V_2| = n/2$. Given a partitioned unlabeled graph S on V, let $L(S)$ be the codeword length given by our algorithm. For large n, our algorithm runs in time $O(n^2)$, and satisfies the following:*

(i) *The algorithm asymptotically achieves the structural entropy in (1) (Note that $(n/2)\log(n/2) = n\log n + O(n)$.), i.e.,*

$$\mathbb{E}[L(S)] \leq 2\binom{n/2}{2}h(p) + \frac{n^2}{4}h(q) - n\log n + O(n).$$

(ii) *For any $\epsilon > 0$,*

$$P(|L(S) - \mathbb{E}[L(S)]| \leq \epsilon n \log n) \geq 1 - o(1).$$

5. General SBM with $R \geq 2$ Blocks

In previous sections, we discussed the structural entropy of SBM and the compression algorithm that asymptotically achieves this structural entropy for the balanced bipartition case ($r = 2$). The corresponding results in Theorem 2 and 3 can be easily generalized to the general r-partition case. We briefly describe the generalizations below.

5.1. Structural Entropy

Our approach can deal with general SBMs similarly. In a general SBM with $r \geq 2$ parts, the transition matrix, an $r \times r$ symmetric matrix $P = (p_{ij})$ is used to describe the probabilities between and within the communities, where two vertices $u \in V_i$ and $v \in V_j$ are connected by an edge with probability $p_{ij} \in [0,1]$ (i and j are not necessarily distinct). We first give the result on the computation of the partitioned structural entropy of SBM.

Theorem 4. *Fix r reals $x_1, x_2, \ldots, x_r$ in $(0,1)$ whose sum is 1. Let $V = V_1 \cup V_2 \cup \cdots \cup V_r$ be a set of n vertices with a partition into r parts such that $|V_i| = x_i n$. Let S be a partitioned structure on V with transition matrix $P = (p_{ij})$. For large n and all $1 \leq i \leq r$ satisfying $n^{-1}\ln n \ll p_{i,i}$ and $1 - p_{i,i} \gg n^{-1}\ln n$, the following holds:*

(i) *The r-partitioned structural entropy H_S^r for $\mathcal{S}$ is*

$$H_S^r = \sum_{i=1}^{r}\binom{x_i n}{2}h(p_{i,i}) + \sum_{1 \leq i < j \leq r} x_i x_j n^2 h(p_{i,j}) - \sum_{i=1}^{r}\log(x_i n)! + O\left(\frac{\log n}{n^\alpha}\right) \quad (6)$$

for some $\alpha > 0$.

(ii) *For $\varepsilon > 0$,*

$$P\left(\frac{1}{\binom{n}{2}} \cdot \left| -\log P(S) - \sum_{i=1}^{r}\binom{x_i n}{2}h(p_{i,i}) - \sum_{1 \leq i < j \leq r} x_i x_j n^2 h(p_{i,j}) + \sum_{i=1}^{r}\log(x_i n)! \right| < 3\varepsilon\right) > 1 - 4\varepsilon.$$

5.2. Compression Algorithm

The compression algorithm for a general r with vertex partition $\{V_1, V_2, \ldots, V_r\}$ can be viewed as a union of the compression algorithms for $S[V_i]$ and $S[V_i, V_j]$ ($i < j \in \{1, 2, \ldots, r\}$). To be more precise, we describe the algorithm as follows. It first compresses all $S[V_i]$ into $\mathcal{L}_i$ using SZIP. Then run the SZIP decoder with input $\mathcal{L}_i$ to obtain the decoded structure $S'[V_i]$. With the indices of $S'[V_i]$, $i = 1, 2, \ldots, r$, we can compress $S[V_1, V_2, \ldots, V_r]$ as a labeled r-partite graph into $\mathcal{L}$ using an arithmetic encoder. This completes the encoding procedure and gives the codewords $\mathcal{L}_1, \ldots, \mathcal{L}_r, \mathcal{L}$, for which we concatenate together and get the final codeword. The decoding is to simply run the SZIP decoders and labeled (arithmetic) decoders parallelly. The correctness of the decoding output can also be argued accordingly.

The performance of the algorithm can be obtained similar to Theorem 3 as follows.

Theorem 5. *Fix r reals $x_1, x_2, \ldots, x_r$ in $(0,1)$ whose sum is 1. Let $V = V_1 \cup V_2 \cup \cdots \cup V_r$ be a set of n vertices with a partition into r parts such that $|V_i| = x_i n$. Given a partitioned unlabeled graph S on V with transition matrix $P = (p_{ij})$, let $L(S)$ be the codeword length given by our algorithm. For large n, our algorithm runs in time $O(n^2)$, and satisfies the following:*

(*i*)　　　*The algorithm asymptotically achieves the structural entropy in* (6), *i.e.*,

$$\mathbb{E}[L(S)] \leq \sum_{i=1}^{r} \binom{x_i n}{2} h(p_{i,i}) + \sum_{1 \leq i < j \leq r} x_i x_j n^2 h(p_{i,j}) - n \log n + O(n).$$

(*ii*)　　　*For any* $\epsilon > 0$,

$$P(|L(S) - \mathbb{E}[L(S)]| \leq \epsilon n \log n) \geq 1 - o(1).$$

6. Conclusions

In this paper, we defined the partitioned unlabeled graphs and partitioned structural entropy, which generalize the structural entropy for unlabeled graphs introduced by Choi and Szpankowski [25]. We then computed the partitioned structural entropy for stochastic block models and gave a compression algorithm that asymptotically achieves this structural entropy limit. As mentioned earlier, we believe that in appropriate contexts the structural information of a graph or network can be interpreted as a kind of semantic information, in which case, the communication schemes may benefit from structural compressions which considerably reduce the cost.

Author Contributions: Conceptualization, J.H., W.H. and B.B.; methodology, J.H., T.G. and Q.Z.; software, J.H. and T.G.; validation, J.H., T.G., Q.Z. and B.B.; formal analysis, J.H. and T.G.; investigation, J.H., W.H. and B.B.; resources, W.H.; data curation, T.G.; writing—original draft preparation, J.H., T.G. and Q.Z.; writing—review and editing, J.H., T.G., Q.Z., W.H. and B.B.; visualization, J.H., T.G., W.H. and B.B.; supervision, W.H., B.B. and G.Z.; project administration, W.H., B.B. and G.Z.; funding acquisition, W.H., B.B. and G.Z. All authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

SBM　　Stochastic Block Models
Szip　　Compression algorithm in [25]
AEP　　Asymptotic Equipartition

References

1.　Yeung, R.W. *Information Theory and Network Coding*; Springer: Berlin/Heidelberg, Germany, 2008.
2.　Cover, T.M.; Thomas, J.A. *Elements of Information Theory*, 2nd ed.; Wiley-Interscience: Hoboken, NJ, USA, 2006.
3.　Shannon, C.E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [CrossRef]
4.　Shannon, C.E. The lattice theory of information. *Trans. Ire Prof. Group Inf. Theory* **1953**, *1*, 105–107. [CrossRef]
5.　Palla, G.; Derényi, I.; Farkas, I.; Vicsek, T. Uncovering the overlapping community structure of complex networks in nature and society. *Nature* **2005**, *435*, 814–818. [CrossRef]
6.　Holland, P.W.; Laskey, K.B.; Leinhardt, S. Stochastic blockmodels: First steps. *Social Netw.* **1983**, *5*, 109–137. [CrossRef]
7.　Abbe, E.; Bandeira, A.S.; Hall, G. Exact recovery in the stochastic block model. *IEEE Trans. Inf. Theory* **2015**, *62*, 471–487. [CrossRef]
8.　Mossel, E.; Neeman, J.; Sly, A. Consistency thresholds for the planted bisection model. In Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing, Portland, OR, USA, 14–17 June 2015; pp. 69–75.
9.　Abbe, E.; Sandon, C. Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery. In Proceedings of the 2015 IEEE 56th Annual Symposium on Foundations of Computer Science, Berkeley, CA, USA, 17–20 October 2015; pp. 670–688.
10.　Hajek, B.; Wu, Y.; Xu, J. Information limits for recovering a hidden community. *IEEE Trans. Inf. Theory* **2017**, *63*, 4729–4745. [CrossRef]
11.　Bianconi, G. Entropy of network ensembles. *Phys. Rev. E* **2009**, *79*, 036114. [CrossRef]

12. Peixoto, T.P. Entropy of stochastic blockmodel ensembles. *Phys. Rev. E* **2012**, *85*, 056122. [CrossRef]

13. Abbe, E. Community detection and stochastic block models: Recent developments. *J. Mach. Learn. Res.* **2017**, *18*, 6446–6531.

14. Asadi, A.R.; Abbe, E.; Verdú, S. Compressing data on graphs with clusters. In Proceedings of the 2017 IEEE International Symposium on Information Theory (ISIT), Aachen, Germany, 25–30 June 2017; pp. 1583–1587. [CrossRef]

15. Zenil, H.; Kiani, N.A.; Tegnér, J. A Review of Graph and Network Complexity from an Algorithmic Information Perspective. *Entropy* **2018**, *20*. [CrossRef]

16. Turán, G. On the succinct representation of graphs. *Discr. Appl. Math.* **1984**, *8*, 289–294. [CrossRef]

17. Harary, F.; Palmer, E.M. *Graphical Enumeration*; Academic Press: Cambridge, MA, USA, 1973.

18. Naor, M. Succinct representation of general unlabeled graphs. *Discr. Appl. Math.* **1990**, *28*, 303–307. [CrossRef]

19. Kieffer, J.C.; Yang, E.H.; Szpankowski, W. Structural complexity of random binary trees. In Proceedings of the 2009 IEEE International Symposium on Information Theory (ISIT), Seoul, Korea, 28 June–3 July 2009; pp. 635–639.

20. Adler, M.; Mitzenmacher, M. Towards compressing Web graphs. In Proceedings of the DCC 2001. Data Compression Conference, Snowbird, Utah, 27–29 March 2001; pp. 203–212.

21. Chierichetti, F.; Kumar, R.; Lattanzi, S.; Mitzenmacher, M.; Panconesi, A.; Raghavan, P. On Compressing Social Networks. In Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; Association for Computing Machinery: New York, NY, USA, 28 June–1 July 2009; pp. 219–228.

22. Peshkin, L. Structure induction by lossless graph compression. In Proceedings of the 2007 Data Compression Conference (DCC'07), Snowbird, Utah, 27–29 March 2007; pp. 53–62.

23. Savari, S.A. Compression of words over a partially commutative alphabet. *IEEE Trans. Inf. Theory* **2004**, *50*, 1425–1441. [CrossRef]

24. Sun, J.; Bollt, E.; Ben-Avraham, D. Graph Compression-Save Information by Exploiting Redundancy. *J. Statist. Mechan. Theory Exper.* **2008**, *2008*, 06001. [CrossRef]

25. Choi, Y.; Szpankowski, W. Compression of Graphical Structures: Fundamental Limits, Algorithms, and Experiments. *IEEE Trans. Inf. Theory* **2012**, *58*, 620–638. [CrossRef]

26. Łuczak, T.; Magner, A.; Szpankowski, W. Compression of Preferential Attachment Graphs. In Proceedings of the 2019 IEEE International Symposium on Information Theory (ISIT), Paris, France, 7–12 July 2019; pp. 1697–1701.

27. Łuczak, T.; Magner, A.; Szpankowski, W. Asymmetry and structural information in preferential attachment graphs. *Random Struct. Algorithms* **2019**, *55*, 696–718. [CrossRef]

28. Sauerhoff, M. On the Entropy of Models for the Web Graph. 2016. Available online: https://www.researchgate.net/publication/255589046_On_the_Entropy_of_Models_for_the_Web_Graph (accessed on 1 November, 2021).

29. Kontoyiannis, I.; Lim, Y.H.; Papakonstantinopoulou, K.; Szpankowski, W. Symmetry and the Entropy of Small-World Structures and Graphs. In Proceedings of the 2021 IEEE International Symposium on Information Theory (ISIT), Victoria, Australia, 11–16 July 2021; pp. 3026–3031.

30. Körner, J. Coding of an information source having ambiguous alphabet and the entropy of graphs. In *6th Prague Conference on Information Theory*; Walter de Gruyter: Berlin, Germany, 1973; pp. 411–425.

31. Alon, N.; Orlitsky, A. Source coding and graph entropies. *IEEE Trans. Inf. Theory* **1996**, *42*, 1329–1339. [CrossRef]

32. Rashevsky, N. Life information theory and topology. *Bull. Math. Biophys.* **1955**, *17*, 229–235. [CrossRef]

33. Trucco, E. A note on the information content of graphs. *Bull. Math. Biophys.* **1956**, *18*, 129–135. [CrossRef]

34. Mowshowitz, A. Entropy and the complexity of the graphs I: An index of the relative complexity of a graph. *Bull. Math. Biophys.* **1968**, *30* 175–204. [CrossRef] [PubMed]

35. Mowshowitz, A. Entropy and the complexity of graphs II: The information content of digraphs and infinite graphs. *Bull. Math. Biophys.* **1968**, *30*, 225–240. [CrossRef]

36. Mowshowitz, A. Entropy and the complexity of graphs III: Graphs with prescribed information content. *Bull. Math. Biophys.* **1968**, *30*, 387–414. [CrossRef]

37. Mowshowitz, A. Entropy and the complexity of graphs IV: Entropy measures and graphical structure. *Bull. Math. Biophys.* **1968**, *30*, 533–546. [CrossRef]

38. Mowshowitz, A.; Dehmer, M. Entropy and the Complexity of Graphs Revisited. *Entropy* **2012**, *14*, 559–570. [CrossRef]

39. Dehmer, M.; Mowshowitz, A. A History of Graph Entropy Measures. *Inf. Sci.* **2011**, *181*, 57–78. [CrossRef]

40. Pasco, R.C. Source Coding Algorithms for Fast Data Compression. Ph.D. Thesis, Stanford University, Stanford, CA, USA, 1976.

41. Rissanen, J.J. Generalized Kraft Inequality and Arithmetic Coding. *IBM J. Res. Dev.* **1976**, *20*, 198–203. [CrossRef]

42. Willems, F.; Shtarkov, Y.; Tjalkens, T. The context-tree weighting method: basic properties. *IEEE Trans. Inf. Theory* **1995**, *41*, 653–664. [CrossRef]