

# Type Size Code for Compressing Erdős-Rényi Graphs \*

Nematollah Iri<sup>†</sup>

September 4, 2020

## Abstract

We consider universal source coding of unlabeled graphs which are commonly referred to as graphical structures. We adopt an Erdős-Rényi model to generate the random graphical structures. We propose a variant of the previously introduced Type Size code, where type classes are characterized based on the number of edges of the graphical structures. The proposed scheme sorts the graphical structures based on the size of their type classes and assigns binary sequences to them in this order. The  $\epsilon$ -coding rate of the Type Size code (up to the third-order term) for compressing graphical structures is derived.

## 1 Introduction

Many problems in social networks, world wide web, recommendation systems, biology and etc., reduce to computations and processings over graphs. With the emergence of big data, such graphs may contain trillion edges. Moreover, next generation applications in the big data era, enforce more stringent I/O access and latency requirements. Therefore, new methods for compressing such massive graphs are essential for retrieval and processing over short time scales.

We aim at compressing the underlying graph up to isomorphism, i.e. to compress the structure of the graph. Many works, such as those in the area of graph summarization [1], consider the compressed version of the graphical structure to be a graph itself [2]. Other works, such as [3, 4, 5], follow an information theoretic paradigm, where the graph structure is mapped into binary strings. We focus on lossless graph compression to faithfully recover the original graph structure from its encoded bits. Numerous lossless graph compression algorithms have been proposed in the literature. See [6], for an exhaustive survey on the existing works, .

Prefix-free assumption has been traditionally imposed to unambiguously decode the block of codewords. However, for many applications in storage and retrieval, there are out-of-band markers to navigate the boundaries of the data to be compressed, hence, one may relax the prefix-free assumption for compression. Such a compression is referred to as one-to-one coding [7].

We follow a *universal* one-to-one source coding setup, where the underlying probability distribution generating the data is arduous to estimate or unknown, yet presumed to belong to a known class of distributions. A universal one-to-one compression scheme has been first introduced in [8]—the Type Size (TS) code— to optimally compress the class of all stationary memoryless sources.

---

\*This research was conducted independently by the author.

<sup>†</sup>nematollah.iri@asu.edu

TS code, universally *orders* sequences based on the size of their type classes, and subsequently map them to binary strings in a lexicographic order. The result in [8] shows a gain of logarithmic order in the input size by relaxing the prefix-free assumption.

There is an underlying flexibility in defining type classes from a TS code perspective. In fact, the one-to-one compression problem is equivalent to characterizing the type classes that lead to optimal performance [9]. Characterizing the type classes based on the empirical probability mass function of the sequences is shown to be optimal for compression of the class of *all* i.i.d. [8] and Markov [10] sources over a finite alphabet, while the quantized type classes [11] are shown to be optimal for compression of the parametric exponential family of distributions.

We adopt the TS approach for compressing unlabeled Erdős-Rényi graphs, which we refer to as graphical *structures*, with an unknown edge probability of  $p$ . We define two graphical structures to be in the same type class if and only if they have the same number of edges. Note that two graphical structures within a type class have the same probability regardless of the enforced Erdős-Rényi model. We order the graphical structures based on the size of the type class they belong to, from smallest to largest, and subsequently map them to binary strings lexicographically (See Figure 1 for an example). We show that with probability at least  $(1 - \epsilon)$ , the TS code requires at most

$$\binom{n}{2}H(p) + \sigma(p)\sqrt{\binom{n}{2}}Q^{-1}(\epsilon) - \log n! + \mathcal{O}(1) \quad (1)$$

bits, to compress a graphical structure with  $n$  nodes, where  $H(p)$  and  $\sigma^2(p)$  are the entropy and the varentropy of the underlying Erdős-Rényi model  $p$ , respectively. The first and third-order terms in (1) do match with an earlier result in [3]. However, we depart from an average case analysis in [3] to a probabilistic analysis.

The required memory to store the TS code ordering grows super-exponentially with the graph size <sup>1</sup>. Even though, this memory requirement may sound prohibitive, however, many applications may afford this storage requirement in order to achieve the rate optimality. A few examples include but are not limited to

- Long distance communications such as satellite communications,
- Transmission over VLF and ELF frequencies which offer a very low bandwidth. Optimal-rate compression comes into awareness in such scenarios, since one can transmit only a limited few (hundreds) bits in a minute.

The rest of the paper is organized as follows. We introduce the lossless source coding of the graphical structures and related definitions in Section 2. In Section 3, we describe the Type Size code and provide tight bounds on the type class sizes. In Section 4, we present the main theorem of the paper which is proved in Section 5. We conclude in Section 6.

## 2 Preliminaries

We adopt an Erdős-Rényi graph generation mechanism [12], i.e., there is a simple undirected edge between pairs of nodes, independently with probability  $p$ . The structure of the graph is derived by removing the vertex labels. Let  $\mathcal{S}(n)$  (resp.  $\mathcal{G}(n)$ ) be the set of simple Erdős-Rényi graphical

---

<sup>1</sup>The size of a graph is the number of its nodes.

structures (resp. labeled graphs) on  $n$  vertices. For a structure  $S \in \mathcal{S}(n)$ , we define its size as the number of vertices and  $j(S)$  captures the number of edges in  $S$ . When it is clear from the context we simplify  $j(S)$  as  $j$ . For notational convenience, denote  $m = \binom{n}{2}$ . Let  $\pi = \pi_1, \pi_2, \dots, \pi_m$  be an arbitrary ordering of pairs of nodes in  $G$ . Construct a sequence  $X_i, i = 1, \dots, m$  as follows:  $X_i = 1$ , if there is an edge between pair of nodes corresponding to  $\pi_i$ , and 0 otherwise. Let  $x_i, i = 1, \dots, m$  denote a realization of  $X_i$  for an observed  $G$ . Since  $X_i$ 's are i.i.d., let  $X$  (resp.  $p(X)$ ) be a random variable (resp. probability distribution) representing the underlying model of  $X_i$ 's, i.e.  $p(X = 1) = p(X_i = 1) = p, i = 1 \dots m$ . Denote  $\mathcal{B}$  as the class of all Bernoulli distributions over  $\{0, 1\}$ . Let  $\mathbb{E}$  and  $\mathbb{V}$ , denote expectation and variance with respect to  $p(X)$ , respectively.

We consider a source coding scheme which maps graphical structures in  $\mathcal{S}(n)$  to variable length binary strings via a coding function

$$\phi : \mathcal{S}(n) \rightarrow \{0, 1\}^* = \{\emptyset, 0, 1, 00, 01, 10, 11, 000, \dots\}.$$

We do not impose the prefix-free condition on the coding scheme. Let  $l(\phi(S))$  be the length of the compressed binary string. For any  $S \in \mathcal{S}(n)$ , the underlying Erdős-Rényi model induces a probability distribution  $\mathbb{P}_S$  on the structures within  $\mathcal{S}$ . We evaluate the performance of coding schemes through the  $\epsilon$ -coding rate for graphs of size  $n$  given by

$$R_n(\epsilon, \phi, p) = \frac{1}{\binom{n}{2}} \min \left\{ k : \mathbb{P}_S(l(\phi(S)) \geq k) \leq \epsilon \right\}. \quad (2)$$

### 3 Type Size Code

#### 3.1 Type Size Code

For the class of all memoryless sources over a finite alphabet, the fixed-to-variable TS code is introduced in [8], which sorts sequences based on the size of their elementary type classes (from smallest to largest) and then encodes sequences to variable-length bit-strings in this order. We borrow the framework of the TS code, however, for the purpose of compressing graphical structures we define two graphical structures to be in the same type class if and only if they have the same number of edges, i.e.  $T_S = \{S' \in \mathcal{S}(n) : j(S) = j(S')\}$ , where  $T_S$  denotes the type class of  $S$ . We then sort graphical structures based on their type class sizes and map them to binary strings according to this type class size ordering. See Figure 1 for an example of the TS code for compressing graphical structures of size  $n = 4$ .

**Theorem 1.** [8] *For the TS code*

$$R_n(\epsilon, \phi, p) \leq \frac{1}{\binom{n}{2}} \lceil \log M(\epsilon) \rceil \quad (3)$$

where

$$M(\epsilon) = \inf_{\gamma : \mathbb{P}_S\left(\frac{1}{\binom{n}{2}} \log |T_S| > \gamma\right) \leq \epsilon} \sum_{S \in \mathcal{S}(n) : \frac{1}{\binom{n}{2}} \log |T_S| \leq \gamma} |T_S|. \quad (4)$$

### 3.2 Type Class Size

Let  $N(n, j)$  be the number of graphs with  $n$  unlabeled nodes and  $j$  simple undirected edges. Let

$$\Lambda(n, j) = \frac{\binom{n}{j}}{n!} \quad (5)$$

and

$$\mu = \frac{2j}{n} - \log n.$$

The following theorem by Wright [13], gives the number of graphical structures for a given number of nodes and edges.

**Theorem 2.** [13] *For a constant  $C_0$  independent of  $n$ ,*

$$N(n, j) = \Lambda(n, j) \left(1 + \mathcal{O}\left(e^{-C_0 \mu}\right)\right),$$

*if and only if  $\mu \rightarrow \infty$  as  $n \rightarrow \infty$ .*

Let  $\mathcal{E}$  be the event where the condition of Theorem 2 is not satisfied, i.e.  $\lim_{n \rightarrow \infty} \mu < \infty$ . In the following lemma, we show that the necessary and sufficient condition of Theorem 2 is satisfied with high probability. The proof is an straightforward application of Chernoff bound [14] and is provided in Appendix A.

**Lemma 3.** *There exist constants  $0 < \delta_1, \delta_2 < 1$  which are independent of  $n$ , such that*

$$\mathbb{P}_S \left( j(S) \leq (1 - \delta_1) \binom{n}{2} p \right) \leq e^{-\binom{n}{2} \delta_2}.$$

Size of the type class of  $S$ ,  $|T_S|$  is then given by

$$\begin{aligned} \log |T_S| &= \log N(n, j) \\ &= \log \left( \binom{n}{j} \right) - \log n! + \mathcal{O}(1) \end{aligned} \quad (6)$$

where  $\mathcal{O}(1)$  term is (with an abuse of notation) bounded between two positive constants independent of  $n$ .

Let  $j^c = m - j$ . Define *empirical entropy* of the graphical structure  $S$  as

$$H_{\text{empirical}}(S) = -\frac{j}{m} \log \frac{j}{m} - \frac{j^c}{m} \log \frac{j^c}{m}.$$

The following lemma provides upper and lower bounds on the size of the graphical type class.

**Lemma 4.** *With probability at least  $1 - e^{-\binom{n}{2} \delta_2}$ , we have the following upper and lower bounds for the size of the type class of a graphical structure  $S \in \mathcal{S}(n)$ :*

$$\binom{n}{2} H_{\text{empirical}}(S) - \log n! + C_L \leq \log |T_S| \leq \binom{n}{2} H_{\text{empirical}}(S) - \log n! + C_U$$

where  $C_L, C_U$  are constants independent of  $n$  and  $\delta_2$  is the constant in Lemma 3.

*Proof.* See Appendix B. □

## 4 Main Result

Let  $H(p) = \mathbb{E} \left( \log \frac{1}{p(X)} \right)$  be the entropy of the underlying source generating the Erdős-Rényi graph, and  $\sigma^2(p) = \mathbb{V} \left( \log \frac{1}{p(X)} \right)$  be the varentropy of it. The following theorem provides an achievability bound for the rate of the TS code in compressing graphical structures.

**Theorem 5.** *For the TS code and any Bernoulli distribution  $p \in \mathcal{B}$ ,*

$$R_n(\epsilon, \phi, p) \leq H(p) + \frac{\sigma(p)}{\sqrt{\binom{n}{2}}} Q^{-1}(\epsilon) - \frac{\log n!}{\binom{n}{2}} + \mathcal{O}\left(\frac{1}{n^2}\right).$$

## 5 Proof of Theorem

When it is clear from the context, we omit the underlying distribution and denote  $H := H(p)$  and  $\sigma := \sigma(p)$ . For a constant  $A > 0$  defined in the Berry-Esséen Lemma 8 (See Appendix C) which is independent of  $n$ , let

$$\gamma = H + \frac{\sigma}{\sqrt{m}} Q^{-1} \left( \epsilon - \frac{A}{\sqrt{m}} - e^{-m\delta_2} \right) - \frac{\log n!}{m} + \frac{C_U}{m} \quad (7)$$

where  $\delta_2, C_U$  are the constants in Lemmas 3 and 4, respectively. Denote

$$p_\gamma := \mathbb{P}_S (\log |T_S| > m\gamma).$$

Let  $q_S(X)$  be a derived Bernoulli distribution from the structure  $S$ , such that  $q_S(0) = \frac{1^c}{m}$  and  $q_S(1) = \frac{1}{m}$ . It is clear that

$$H_{\text{empirical}}(S) = \frac{1}{m} \sum_{i=1}^m -\log q_S(x_i). \quad (8)$$

Let

$$S_m = \frac{1}{\sigma(p)\sqrt{m}} \sum_{i=1}^m (-\log q_S(x_i) - H(p)). \quad (9)$$

Recall  $\mathcal{E}$  from Subsection 3.2. We have

$$p_\gamma = \mathbb{P}_S (\log |T_S| > m\gamma | \mathcal{E}^c) \mathbb{P}_G (\mathcal{E}^c) + \mathbb{P}_S (\log |T_S| > m\gamma | \mathcal{E}) \mathbb{P}_G (\mathcal{E}) \quad (10)$$

$$\leq \mathbb{P}_S \left( H_{\text{empirical}}(S) > H(p) + \frac{\sigma(p)}{\sqrt{m}} Q^{-1} \left( \epsilon - \frac{A}{\sqrt{m}} - e^{-m\delta_2} \right) \right) + e^{-m\delta_2} \quad (11)$$

$$= \mathbb{P} \left( S_m > Q^{-1} \left( \epsilon - \frac{A}{\sqrt{m}} - e^{-m\delta_2} \right) \right) + e^{-m\delta_2} \quad (12)$$

$$\leq Q \left( Q^{-1} \left( \epsilon - \frac{A}{\sqrt{m}} - e^{-m\delta_2} \right) \right) + \frac{A}{\sqrt{m}} + e^{-m\delta_2} \quad (13)$$

$$= \epsilon$$

where (10) follows from the law of total probability, (11) follows from upper bounding  $\mathbb{P}(\mathcal{E}^c)$  and  $\mathbb{P}(\log |T_G| > m\gamma | \mathcal{E})$  by 1 in conjunction with Lemma 4, (12) is from the definitions (9) and (8), and

finally (13) is from the Berry-Esseen theorem [15] (See Appendix C). We now bound  $M(\epsilon)$  using (4) with this choice of  $\gamma$ . Let

$$f(S) = H_{\text{empirical}}(S) - \frac{\log n!}{m} + \frac{C_U}{m}. \quad (14)$$

Similarly, with an abuse of overloaded notation <sup>2</sup>, for any Bernoulli distribution  $p \in \mathcal{B}$ , define

$$f(p) = H(p) - \frac{\log n!}{m} + \frac{C_U}{m}. \quad (15)$$

The rest of the proof is similar to [8], however, we continue the proof for completeness.

**Lemma 6.** [8] *There exists a Lipschitz constant  $K_0$  independent of  $n$ , such that for any two Bernoulli distributions  $p, \tilde{p} \in \mathcal{B}$ ,*

$$|f(p) - f(\tilde{p})| \leq K_0 \|p - \tilde{p}\|. \quad (16)$$

Fixing  $\Delta = \frac{1}{m}$ , we have

$$\begin{aligned} M(\epsilon) &\leq \sum_{\substack{S \in \mathcal{S}(n): \\ \frac{1}{m} \log |T_S| \leq \gamma}} |T_S| \\ &\leq \sum_{\substack{S \in \mathcal{S}(n): \\ f(S) - \frac{C_d}{m} \leq \gamma}} 2^{mf(S)} \\ &= \sum_{i=0}^{\infty} \sum_{\substack{S \in \mathcal{S}(n): \\ f(S) \in \mathcal{A}_i}} 2^{mf(S)} \\ &\leq \sum_{i=0}^{\infty} |\{S \in \mathcal{S}(n) : f(S) \in \mathcal{A}_i\}| \cdot 2^{m\gamma + C_d - mi\Delta} \end{aligned} \quad (17)$$

where  $C_d = C_U - C_L$  and  $\mathcal{A}_i = (\gamma + \frac{C_d}{m} - (i+1)\Delta, \gamma + \frac{C_d}{m} - i\Delta]$ . For a Bernoulli distribution  $p \in \mathcal{B}$ , define its 2-norm ball of radius  $r$  as  $B_r(p) = \{p' \in \mathcal{B} : \|p - p'\| \leq r\}$ . By extension, for a graphical structure  $S \in \mathcal{S}(n)$ , define its 2-norm ball of radius  $r$  as  $B_r(S) := B_r(q_S)$ , where  $q_S$  is the derived empirical distribution of  $S$  as defined at the beginning of this section. Note that for any two different structures  $S_1, S_2 \in \mathcal{S}(n)$ ,  $B_{\frac{1}{2m}}(S_1)$  and  $B_{\frac{1}{2m}}(S_2)$  are disjoint. Moreover, observe that

---

<sup>2</sup>The two definitions of the function  $f(\cdot)$  should be distinguished based upon their arguments.

$\text{Vol}\left(B_{\frac{1}{2m}}(S)\right) = \frac{1}{m}$ . We have

$$\begin{aligned} \left| \{S \in \mathcal{S}(n) : f(S) \in \mathcal{A}_i\} \right| &= \sum_{\substack{S \in \mathcal{S}(n): \\ f(S) \in \mathcal{A}_i}} \frac{\text{Vol}\left(B_{\frac{1}{2m}}(S)\right)}{\frac{1}{m}} \\ &= m \sum_{\substack{S \in \mathcal{S}(n): \\ f(S) \in \mathcal{A}_i}} \text{Vol}\left(B_{\frac{1}{2m}}(S)\right) \\ &= m \text{Vol}\left( \bigcup_{\substack{S \in \mathcal{S}(n): \\ f(S) \in \mathcal{A}_i}} B_{\frac{1}{2m}}(S) \right) \end{aligned} \quad (18)$$

$$\leq m \text{Vol}\left( \bigcup_{\substack{p \in \mathcal{B}: \\ f(p) \in \mathcal{A}_i}} B_{\frac{1}{2m}}(p) \right) \quad (19)$$

where (18) is from disjointness of the balls. Let  $\rho(\lambda) = \text{Vol}\{p \in \mathcal{B} : f(p) \leq \lambda\}$ . The following lemma from [8], shows the Lipschitzness of  $\rho(\cdot)$ .

**Lemma 7.** [8] *There exists a Lipschitz constant  $K_1$  such that for all  $a, b$ ,*

$$|\rho(a) - \rho(b)| \leq K_1 |a - b|.$$

We continue from (19). Let  $a = \gamma + \frac{C_d}{m} - (i+1)\Delta$ . We have

$$\begin{aligned} \left| \{S \in \mathcal{S}(n) : f(S) \in \mathcal{A}_i\} \right| &\leq m \cdot \text{Vol}\left( \bigcup_{a < f(p) \leq a + \Delta} B_{\frac{1}{2m}}(p) \right) \\ &\leq m \text{Vol}\left( \left\{ p : f(p) \in \left( a - \frac{K_0}{2m}, a + \Delta + \frac{K_0}{2m} \right) \right\} \right) \end{aligned} \quad (20)$$

$$\begin{aligned} &= m \left( \rho\left(a + \Delta + \frac{K_0}{2m}\right) - \rho\left(a - \frac{K_0}{2m}\right) \right) \\ &\leq m K_1 K_0 \cdot \left( \Delta + \frac{K_0}{m} \right) \end{aligned} \quad (21)$$

where (20) is from the observation that for any  $\tilde{p} \in B_{\frac{1}{2m}}(p)$ ,  $|f(\tilde{p}) - f(p)| \leq \frac{K_0}{2m}$  and (21) is from Lemma 7. Applying (21) to (17), we obtain

$$\begin{aligned} M(\epsilon) &\leq \sum_{i=0}^{\infty} m K_1 K_0 \cdot \left( \Delta + \frac{K_0}{m} \right) \cdot 2^{m\gamma + C_d - mi\Delta} \\ &= m K_1 K_0 \cdot \left( \Delta + \frac{K_0}{m} \right) \cdot 2^{m\gamma + C_d} \cdot \frac{1}{1 - 2^{-m\Delta}}. \end{aligned}$$

Since  $\Delta = \frac{1}{m}$ , we have

$$\begin{aligned} \log M(\epsilon) &\leq \log(K_1 K_0 (K_0 + 1)) + m\gamma + C_d + 1 \\ &= mH(p) + \sigma\sqrt{m}Q^{-1}\left(\epsilon - \frac{A}{\sqrt{m}} - e^{-m\delta_2}\right) - \log n! + C_U + C_1 \\ &\leq mH(p) + \sigma\sqrt{m}Q^{-1}(\epsilon) - \log n! + C \end{aligned}$$

for constants  $C_1, C$  independent of  $n$ .

## 6 Conclusion and Future Work

We proposed a variant of the Type Size code for compressing graphical structures. Erdős-Rényi model is adopted as the underlying mechanism for generating the random graphs. We provided an analysis to derive the fine asymptotics of the overflow rate of the proposed Type Size code for compressing such structures. However, Erdős-Rényi model fails to fully represent the real-world networks such as the world wide web. The alternative models include the power law and the preferential attachment models. We study the finite blocklength compression of graphical structures not generated by the Erdős-Rényi model as a future work. The multigraph version of the problem, which permits multiple edges between pairs of nodes and the lossy version of the problem are also interesting future directions of this research.

## References

- [1] Y. Liu, T. Safavi, A. Dighe, and D. Koutra, “Graph summarization methods and applications: A survey,” *ACM Computing Surveys (CSUR)*, vol. 51, no. 62, 2018.
- [2] F. Zhou, “Graph compression.” *Department of Computer Science and Helsinki Institute for Information Technology HIIT*, pp. 1–12, 2015.
- [3] Y. Choi and W. Szpankowski, “Compression of graphical structures: Fundamental limits, algorithms, and experiments,” *IEEE Transactions on Information Theory*, vol. 58, pp. 620–638, Feb 2012.
- [4] M. Mohri, M. Riley, and A. Suresh, “Automata and graph compression,” in *Information Theory (ISIT), 2015 IEEE International Symposium on*, June 2015, pp. 2989–2993.
- [5] A. M. Luczak, Tomasz and W. Szpankowski, “Structural information and compression of scale-free graphs.” Urbana, 2017.
- [6] M. Besta and H. Torsten, “Survey and taxonomy of lossless graph compression and space-efficient graph representations,” *arxiv.org:1806.01799*.
- [7] W. Szpankowski and S. Verdu, “Minimum expected length of fixed-to-variable lossless compression without prefix constraints,” *Information Theory, IEEE Transactions on*, vol. 57, no. 7, pp. 4017–4025, July 2011.
- [8] O. Kosut and L. Sankar, “Asymptotics and non-asymptotics for universal fixed-to-variable source coding,” *Information Theory, IEEE Transactions on*, vol. 63, pp. 3757–3772, 2017.
- [9] N. Iri and O. Kosut, “Universal coding with point type classes,” in *2017 51st Annual Conference on Information Sciences and Systems (CISS)*, 2017.
- [10] —, “Third-order coding rate for universal compression of markov sources,” in *Information Theory (ISIT), 2015 IEEE International Symposium on*, June 2015, pp. 1996–2000.



- [11] —, “Fine asymptotics for universal one-to-one compression of parametric sources,” *IEEE Transactions on Information Theory*, vol. 65, no. 4, pp. 2442–2458, Apr 2019.
- [12] B. Bollobas, *Random graphs*. Cambridge university press, 2001.
- [13] E. M. Wright, “The number of unlabelled graphs with many nodes and edges,” *Bulletin of the American Mathematical Society*, vol. 78, no. 6, pp. 1032–1034, 1972.
- [14] W. Hoeffding, *Probability inequalities for sums of bounded random variables*. New York: Springer, 1994.
- [15] I. Kontoyiannis and S. Verdú, “Optimal lossless data compression: Non-asymptotics and asymptotics,” *Information Theory, IEEE Transactions on*, vol. 60, no. 2, pp. 777–795, Feb 2014.
- [16] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Orlando, FL: Academic Press, 1982.

# Appendices

## A Proof of Lemma 3

Observe that the number of edges in the Erdős-Rényi graph can be derived as the sum of  $m$  i.i.d. random variables  $X_i$ , defined in Subsection 3.1, i.e.  $J(S) = \sum_{i=1}^m X_i$ . For positive  $\alpha, t > 0$ , from Chernoff bound [14], we have

$$\mathbb{P}_S(J(S) \leq \alpha) \leq \min_{t>0} e^{t\alpha} \prod_i \mathbb{E}(e^{-tX_i}). \quad (22)$$

On the other hand

$$\begin{aligned} \mathbb{E}(e^{-tX_i}) &= pe^{-t} + (1-p) \\ &= 1 + p(e^{-t} - 1) \\ &\leq e^{p(e^{-t}-1)} \end{aligned} \quad (23)$$

where (23) follows from  $\ln x \leq x - 1$ , for any positive  $x > 0$ . Hence

$$\mathbb{P}_S(J(S) \leq \alpha) \leq \min_{t>0} e^{t\alpha} \cdot e^{mp(e^{-t}-1)}. \quad (24)$$

Since (24) holds for any  $\alpha, t > 0$ , therefore, we may take  $\alpha = (1 - \delta_1)mp$  and  $t = -\ln(1 - \delta_1) > 0$ , for an arbitrary  $0 < \delta_1 < 1$ . Subsequently, we obtain

$$\begin{aligned} \mathbb{P}_S(J(S) \leq (1 - \delta_1)mp) &\leq \left(\frac{1}{1 - \delta_1}\right)^{(1-\delta_1)mp} e^{-mp\delta_1} \\ &= e^{-m\delta_2} \end{aligned} \quad (25)$$

where  $\delta_2 = -\delta_1 - (1 - \delta_1)\ln(1 - \delta_1) > 0$  is a positive constant.

## B Proof of Lemma 4

Utilizing Theorem 2, we proceed by providing tight upper and lower bounds for  $\Lambda(n, j)$  in (5).

**Upper Bound:**

Recall  $j^c = \binom{n}{2} - j$ . Using the Stirling's formula [16], we have

$$\begin{aligned}
\log \binom{\binom{n}{2}}{j} &= \log \frac{\left(\binom{n}{2}\right)!}{j! \left(\binom{n}{2} - j\right)!} \\
&\leq \log \frac{\sqrt{2\pi} \left(\binom{n}{2}\right)^{\binom{n}{2} + \frac{1}{2}} 2^{-\left(\binom{n}{2}\right) + \frac{1}{12\left(\binom{n}{2}\right)}}}{\left(\sqrt{2\pi} j^{\binom{n}{2} + \frac{1}{2}} 2^{-j + \frac{1}{12(j+1)}}\right) \left(\sqrt{2\pi} j^{j^c + \frac{1}{2}} e^{-j^c + \frac{1}{12(j^c+1)}}\right)} \\
&= \left(\left(\binom{n}{2} + \frac{1}{2}\right) \log \binom{n}{2} - \binom{n}{2} + \frac{1}{12\binom{n}{2}} - \left(j + \frac{1}{2}\right) \log j + j - \frac{1}{12(j+1)} - \left(j^c + \frac{1}{2}\right) \log j^c + j^c - \frac{1}{12(j^c+1)} - \log \sqrt{2\pi}\right) \\
&\leq \left(\left(\binom{n}{2} + \frac{1}{2}\right) \log \binom{n}{2} - (j \log j + j^c \log j^c) - \frac{1}{2} (\log j + \log j^c) + \frac{1}{12} - \log \sqrt{2\pi}\right). \tag{26}
\end{aligned}$$

Recall  $q_S \in \mathcal{B}$ , a derived Bernoulli distribution with  $q_S(X=0) = \frac{j^c}{\binom{n}{2}}$  and  $q_S(X=1) = \frac{j}{\binom{n}{2}}$ , with an entropy  $H(q_S) = H_{\text{empirical}}(S)$ . Note that

$$\begin{aligned}
-j \log j - j^c \log j^c &= -\binom{n}{2} \frac{j}{\binom{n}{2}} \log \frac{j}{\binom{n}{2}} - \binom{n}{2} \frac{j^c}{\binom{n}{2}} \log \frac{j^c}{\binom{n}{2}} - j \log \binom{n}{2} - j^c \log \binom{n}{2} \\
&= \binom{n}{2} H_{\text{empirical}}(S) - \binom{n}{2} \log \binom{n}{2}. \tag{27}
\end{aligned}$$

Without loss of generality, we assume  $0 < j, j^c < \binom{n}{2}$ . This can be accommodated by adding two extra header bits to the compressed data stream, to indicate if the graph is empty, complete, or else. Hence, we have

$$\log j + \log j^c \geq \log \binom{n}{2} - 1. \tag{28}$$

(26) in conjunction with (27, 28), gives the upper bound.

**Lower Bound:**

Using the Stirling's formula [16], we have

$$\begin{aligned}
\log \binom{\binom{n}{2}}{j} &= \log \frac{\left(\binom{n}{2}\right)!}{j! \left(\binom{n}{2} - j\right)!} \\
&\geq \log \frac{\sqrt{2\pi} \left(\binom{n}{2}\right)^{\binom{n}{2} + \frac{1}{2}} 2^{-\left(\binom{n}{2}\right) + \frac{1}{12\left(\binom{n}{2}\right) + 1}}}{\left(\sqrt{2\pi} j^{\binom{n}{2} + \frac{1}{2}} 2^{-j + \frac{1}{12j}}\right) \left(\sqrt{2\pi} j^{j^c + \frac{1}{2}} 2^{-j^c + \frac{1}{12j^c}}\right)} \\
&= \left(\left(\binom{n}{2} + \frac{1}{2}\right) \log \binom{n}{2} - \binom{n}{2} + \frac{1}{12\left(\binom{n}{2} + 1\right)} - \left(j + \frac{1}{2}\right) \log j + j - \frac{1}{12j} - \left(j^c + \frac{1}{2}\right) \log j^c + j^c - \frac{1}{12j^c} - \log \sqrt{2\pi}\right) \\
&= \left(\left(\binom{n}{2} + \frac{1}{2}\right) \log \binom{n}{2} - (j \log j + j^c \log j^c) - \frac{1}{2} (\log j + \log j^c) - \frac{1}{6} - \log \sqrt{2\pi}\right).
\end{aligned}$$

Moreover, note that

$$\log j + \log j^c \leq \frac{1}{2} \log \binom{n}{2}. \tag{29}$$

The rest of the proof is similar to the proof of the upper bound.

## C Berry-Esséen Bound

**Lemma 8.** [15] *Let  $\{Z_i\}$  be independent and identically distributed random variables with zero mean and unit variance, and let  $\tilde{Z}$  be a standard normal. Then, for a constant  $A > 0$  independent of  $n$ , all  $n \geq 1$  and all  $z$  we have*

$$\left| \mathbb{P} \left( \frac{1}{\sqrt{n}} \sum_{i=1}^n Z_i \leq z \right) - \mathbb{P}(\tilde{Z} \leq z) \right| \leq \frac{A}{\sqrt{n}}$$

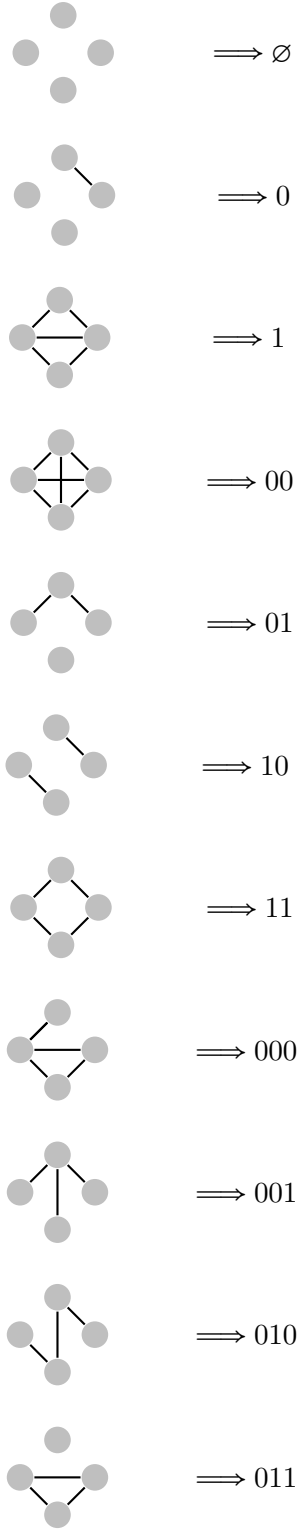


Figure 1: TS code for graphical structures of size  $n = 4$ .