A Universal Lossless Compression Method applicable to Sparse Graphs and heavy-tailed Sparse Graphs

Payam Delgosha
CS Department
University of Illinois, Urbana-Champaign
Urbana, IL 61801
Email: delgosha@illinois.edu

Venkat Anantharam
EECS Department
University of California, Berkeley
Berkeley, CA 94720
Email: ananth@eecs.berkeley.edu

2021 IEEE International Symposium on Information Theory (ISIT) | 978-1-5386-8209-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/ISIT45174.2021.9517897

***Abstract*—Graphical data arises naturally in several modern applications, including but not limited to internet graphs, social networks, genomics and proteomics. The typically large size of graphical data argues for the importance of designing universal compression methods for such data. In most applications, the graphical data is sparse, meaning that the number of edges in the graph scales more slowly than n^2, where n denotes the number of vertices. Although in some applications the number of edges scales linearly with n, in others the number of edges is much smaller than n^2 but appears to scale superlinearly with n. We call the former *sparse graphs* and the latter *heavy-tailed sparse graphs*. In this paper we introduce a universal lossless compression method which is simultaneously applicable to both classes. We do this by employing the local weak convergence framework for sparse graphs and the sparse graphon framework for heavy-tailed sparse graphs.**

I. Introduction

There has recently been an increased interest in the problem of graphical data compression [1], [2], [3], [4], [5], [6], [7], [8], [9]. For practical applications usually the real world graphical data are "sparse", where the sparsity is usually observed as a condition on the ratio of the number of the edges in the graph to the total number of potential edges. Therefore, roughly speaking, a graph with n vertices is said to be sparse in a broad sense if its number of edges is *much smaller* than n^2. One interesting sparsity regime is where the number of edges grows linearly with the number of vertices. In a series of works the authors of this paper have studied the problem of universal lossless compression [10], [11] and distributed compression [12] for sparse graphs in this sparsity regime. This was done by employing the framework of "local weak convergence", also called the "objective method", which allows one to make sense of a probabilistic framework similar to stochastic processes for sparse graphical data [13], [14], [15].

The purpose of this paper is to go beyond this sparsity regime and achieve universal compression for graphs which are still sparse, but with the number of edges growing super–linearly with the number of vertices, i.e. sparse graphs with *heavy-tailed* degree distributions. In order to address graphs with heavy-tailed degree distributions, we employ a version of the graphon theory [16], [17], [18]. This framework, which we call the *sparse graphon framework*, defines a notion of convergence for heavy-tailed sparse graphs, similar to the local weak convergence framework, but using a different metric to discuss how close a given graph seems to be to another one.

We employ the local weak convergence framework together with the sparse graphon framework to address the problem of universal compression of graphical data with possibly heavy-tailed degree distribution. More precisely, we aim to compress a graph which is either consistent with the local weak convergence framework or the sparse graphon framework. However, the universality condition requires that the encoder does not know which of the two frameworks the input graph is consistent with, neither does it know the limiting object describing the empirical statistics of the graph in each of the two frameworks. Further, we want the encoder to be information-theoretically optimal, in the sense that if we appropriately normalize the codeword length associated to the input graph it does not asymptotically exceed the entropy of the limit object, with an appropriate notion of the entropy for each of the two frameworks. In order to make sense of optimality in the local weak sense we employ the notion of BC entropy from [19]. On the other hand, in order to make sense of optimality in the sparse graphon framework we introduce a notion of entropy for this framework which can be of independent interest.

The purpose of this paper is to investigate information theoretic limits of compression in the above setting, which paves the road to seek computationally efficient compression algorithms achieving such limits (see [11] for a computationally efficient compression algorithm in the local weak convergence framework).

The structure of this paper is as follows. In Section II, we briefly review the local weak convergence framework and the BC entropy. In Section III, we briefly review the sparse graphon framework, introduce our notion of entropy for this framework, and discuss its properties. In Section IV, we discuss the properties that we expect from a universal compression scheme based on the notions of entropy for the

Notation	Meaning
$[n]$	$\{1,\dots,n\}$
$[\alpha]$	$[\lfloor\alpha\rfloor]$ for non-integer $\alpha \geq 1$
$\log(.)$	logarithm in natural basis
$\mathbb{R}_+$	nonnegative real numbers
$\{0,1\}^* - \emptyset$	the set of finite nonempty sequences of 0 and 1's
$\mathsf{nats}(x)$	length of $x \in \{0,1\}^*$ in nats: $\log 2 \times$length of x
$\mathcal{P}(\Omega)$	space of Borel probability measure on metric space Ω
$H(X)$	Shannon entropy of random variable X
$D(.\|\|.)$	relative entropy

TABLE I: List of basic notation

local weak convergence and the sparse graphon frameworks. We then state our main results on the existence of such a universal compression scheme, and a converse result. In Section V, we introduce our compression scheme and sketch our proof ideas. We finally conclude the paper in Section VI.

We close this section by introducing some notation. Table I lists our basic notation. All the graphs in this paper are simple, so we drop the prefix *simple* when referring to graphs. For a graph G and vertices v and w in G, $v \sim_G w$ denotes the existence of an edge between v and w. $\deg_G(v)$ denotes the degree of a vertex v in G. $\mathcal{G}_n$ denotes the set of graphs on the vertex set $[n]$. For a graph $G \in \mathcal{G}_n$, $A(G)$ denotes its adjacency matrix, i.e. an $n \times n$ matrix where $(A(G))_{i,j}$ is equal to one if $i \sim_G j$ and zero otherwise. For $p \geq 1$, the L^p norm of an $n \times n$ matrix A is defined as $\|A\|_p^p := \frac{1}{n^2}\sum_{1\leq i,j\leq n} |A_{i,j}|^p$.

II. The Framework of Local Weak Convergence

In this section we briefly review the framework of local weak convergence. The reader is referred to [13], [14], [15] for more details.

Given a graph G and a vertex o in G, we let $[G,o]$ denote the isomorphism class of the connected component of o in G, rooted at o. Here isomorphism is defined to preserve edges as well as the root. For $h \geq 0$, $[G,o]_h$ denotes the isomorphism class corresponding to the subgraph of G consisting of vertices with distance at most h from o, rooted at o. We can think of $[G,o]_h$ as the structure of the depth h neighborhood around o. Let $\mathcal{G}_*$ denote the space of isomorphism classes $[G,o]$ where G is connected and its vertex set is finite or countably infinite. $\mathcal{G}_*$ can be equipped with a metric, called the "local metric", where the distance between $[G,o]$ and $[G',o']$ is defined to be $1/(1+h_*)$, where h_* is the supremum over all $h \geq 0$ such that $[G,o]_h = [G',o']_h$. It can be shown that $\mathcal{G}_*$ equipped with the local metric is a Polish space [15], i.e. a complete separable metric space.

Given a finite graph G, we define $U(G) \in \mathcal{P}(\mathcal{G}_*)$ to be the law of $[G,o]$, where o is chosen uniformly at random (u.a.r) in G. We say that a sequence of finite graphs $G^{(n)}$ converges in the local weak sense to $\mu \in \mathcal{P}(\mathcal{G}_*)$ if $U(G^{(n)})$ converges weakly to μ as members of $\mathcal{P}(\mathcal{G}_*)$. Roughly speaking, this means that for every depth h, the distribution of the local neighborhood structures $[G^{(n)},o]_h$, when o is chosen u.a.r. in $G^{(n)}$, converges to the distribution of the depth h neighborhood around the root in μ as $n \to \infty$. For instance, the sequence of sparse Erdős–Rényi graphs where each edge is independently present with probability p/n converges almost surely (a.s.) in the local weak sense to a Galton-Watson limit with Poisson(p) degree distribution. Not all $\mu \in \mathcal{P}(\mathcal{G}_*)$ can show up as the local weak limit of a sequence of finite graphs, and a necessary condition called "unimodularity" must be satisfied [15], which can be thought of as a kind of stationarity condition. Let $\mathcal{T}_*$ denote the subset of the isomorphism classes of rooted trees $[T,o] \in \mathcal{G}_*$.

A. The BC Entropy

In this section we briefly review the notion of entropy introduced by Bordenave and Caputo [19] for probability distributions on $\mathcal{G}_*$, i.e. members of $\mathcal{P}(\mathcal{G}_*)$.

For integers n and m, let $\mathcal{G}_{n,m}$ be the set of graphs on the vertex set $[n]$ with m edges. Furthermore, given $\mu \in \mathcal{P}(\mathcal{G}_*)$ and $\epsilon > 0$, let $\mathcal{G}_{n,m}(\mu,\epsilon)$ denote the set of graphs $G \in \mathcal{G}_{n,m}$ such that $d_{\mathrm{LP}}(U(G),\mu) < \epsilon$, where d_{LP} denotes the Lévy–Prokhorov metric [20] on $\mathcal{P}(\mathcal{G}_*)$. Roughly speaking, it can be shown that if $m_n/n \to d/2$ where d is the expected degree at the root in μ, then we have

$$\lim_{\epsilon\downarrow 0}\lim_{n\to\infty}\frac{1}{n}\left(\log|\mathcal{G}_{n,m_n}(\mu,\epsilon)| - m_n\log n\right) = \Sigma(\mu), \qquad (1)$$

where $\Sigma(\mu)$ is a constant which depends on μ, possibly $-\infty$, and is called the BC entropy of μ. In words, $\Sigma(\mu)$ is effectively the per-vertex growth rate of the size of the typical graphs, after separating out the $m_n \log n$ leading term. It can be shown [19] that as long as $m_n/n \to d/2$ as above, $\Sigma(\mu)$ does not depend on the choice of the sequence m_n. Additionally, if μ is not unimodular, or the support of μ is not contained in $\mathcal{T}_*$, we have $\Sigma(\mu) = -\infty$. Motivated by this, from this point forward, we restrict our analysis to unimodular probability distributions on $\mathcal{T}_*$.

B. A Universal Lossless Compression Scheme adapted to the Local Weak Convergence Framework

The authors of this paper have shown that in the context of the local weak convergence framework the BC entropy is the correct information-theoretical limit of compression [10]. This is done in part by introducing a universal lossless compression scheme. This is achieved in [10] for a broader class of *marked* graphs where vertices and edges can carry additional marks; however we state this result only for the setting of this paper where such marks are not present. We denote the universal compression map of [10] by $f_n^{\mathsf{lwc}} : \mathcal{G}_n \to \{0,1\}^* - \emptyset$, which assigns a prefix-free codeword to each graph on the vertex set $[n]$. Here the superscript lwc stands for "local weak convergence", and is assigned to distinguish it from the compression map we will introduce later in Section V. This compression map is lossless, i.e. there exists a decompression map g_n^{lwc} such that the composition $g_n^{\mathsf{lwc}} \circ f_n^{\mathsf{lwc}}$ is the identity map. Moreover, the compression scheme is universal in the sense that given a sequence of graphs $G^{(n)}$ converging to a

unimodular limit $\mu \in \mathcal{P}(\mathcal{G}_*)$ in the local weak sense, without prior knowledge of μ we have

$$\limsup_{n\to\infty} \frac{\mathsf{nats}(f_n^{\mathsf{lwc}}(G^{(n)})) - m^{(n)} \log n}{n} \le \Sigma(\mu), \qquad (2)$$

where $m^{(n)}$ denotes the number of the edges in $G^{(n)}$. Note that the normalization is done in a way consistent with the definition of the BC entropy. It is shown in [10] that such a universal compression scheme (comprised of a universal compression map and associated decompression map for each n) exists, satisfying the condition (2) for all unimodular $\mu \in \mathcal{P}(\mathcal{T}_*)$ with expected degree at the root in the range $(0, \infty)$, and for all sequences of finite graphs $G^{(n)}$ converging to μ in the local weak sense [10].

III. The Sparse Graphon Framework

In this section, we briefly review the graphon framework [21], with a focus on the sparse regime. Let $(\Omega, \mathcal{F}, \pi)$ be a probability space. A graphon is defined to be a nonnegative function $W : \Omega \times \Omega \to \mathbb{R}_+$ which is symmetric, i.e. $W(x,y) = W(y,x)$ for all $x, y \in \Omega$, and satisfies $\|W\|_1 := \int W(x,y) d\pi(x) d\pi(y) < \infty$. A graphon is said to be L^p for $p \ge 1$ if $\|W\|_p^p := \int (W(x,y))^p d\pi(x) d\pi(y) < \infty$.

A graph G on a finite vertex set V naturally defines a graphon W over the probability space V equipped with the uniform distribution, defined as $W(v,w) = (A(G))_{v,w}$ for $v, w \in V$.

Assume that a symmetric $n \times n$ matrix B with nonnegative entries is given together with a probability vector $p = (p_1, \ldots, p_n)$. We define the *block graphon* (p, B) to be a graphon W over the finite probability space $[n]$ equipped with the probability distribution p such that for $1 \le i, j \le n$, we have $W(i,j) = B_{i,j}$.

For two L^2 graphons W and W' defined on probability spaces $(\Omega, \mathcal{F}, \pi)$ and $(\Omega', \mathcal{F}', \pi')$ respectively, we define

$$\delta_2(W, W') := \inf_{\nu} \sqrt{\int |W(x,y) - W'(x',y')|^2 d\nu(x,x') d\nu(y,y')}, \qquad (3)$$

where the infimum is taken over all couplings ν of π and π', i.e. ν is a probability measure over $\Omega \times \Omega'$ with marginals π and π', respectively. In fact, δ_2 yields a metric on the space of equivalence classes of L^2 graphons, with reference to the notion of equivalence defined in [18, Definition 2.5] (see [18, Theorem 2.11 and Appendix A] and [22]). Moreover, for two graphons W and W' on two probability spaces $(\Omega, \mathcal{F}, \pi)$ and $(\Omega', \mathcal{F}', \pi')$, respectively we define the cut norm as

$$\delta_\square(W, W') := \inf_{\nu} \sup_{S,T \subseteq \Omega \times \Omega'} \left| \int_{S \times T} \left(W(x,y) - W'(x',y')\right) d\nu(x,x') d\nu(y,y') \right|, \qquad (4)$$

where the integration is over $(x,x') \in S$ and $(y,y') \in T$, and the supremum is over measurable subsets S and T of $\Omega \times \Omega'$. Moreover, the infimum is taken over all couplings ν of π and π'. Note that every graphon is by definition an L^1 function, hence the cut norm is well defined. In fact, $\delta_\square$ yields a metric on the space of equivalence classes of graphons, with reference to the notion of equivalence defined in [18, Definition 2.5] (see [18, Theorem 2.11 and Appendix A] and [22]).

A graphon W is said to be normalized if $\|W\|_1 = 1$. Given a normalized graphon W on a probability space $(\Omega, \mathcal{F}, \pi)$ and a sequence of *target densities* ρ_n, the sequence of W-random graphs with target density ρ_n is the sequence of random graphs $G^{(n)}$ on the vertex set $[n]$ defined as follows. We first generate random variables $(X_i : i \ge 1)$ i.i.d. from distribution π. Then, for each n and each pair of vertices $1 \le v, w \le n$, we independently place an edge between v and w in $G^{(n)}$ with probability $\min\{1, \rho_n W(X_v, X_w)\}$. We denote the law of $G^{(n)}$ generated according to this procedure by $\mathcal{G}(n; \rho_n W)$.

Given a graph G on n vertices carrying m edges, we define the *density* of G to be $2m/n^2$ and denote it by $\rho(G)$. It can be seen that if $G^{(n)} \sim \mathcal{G}(n; \rho_n W)$, under some conditions formalized in the following theorem, we have $\rho(G^{(n)})/\rho_n \to 1$ a.s. as $n \to \infty$. This justifies the terminology *target density*.

Theorem 1 (Theorem 2.14 in [18]). *Let $G^{(n)} \sim \mathcal{G}(n; \rho_n W)$ be a sequence of W-random graphs with target density ρ_n, where W is a normalized graphon over an arbitrary probability space, and ρ_n is such that $n\rho_n \to \infty$ and $\rho_n \to 0$. Then, as $n \to \infty$, we have $\rho(G^{(n)})/\rho_n \to 1$ a.s. and*

$$\delta_\square \left(\frac{1}{\rho(G^{(n)})} G^{(n)}, W \right) \to 0 \qquad \text{a.s.}.$$

Note that, as we discussed above, $G^{(n)}$ naturally defines a graphon, and $G^{(n)}/\rho(G^{(n)})$ refers to the scaled graphon corresponding to $G^{(n)}$. In fact the theorem above implies that if $\rho_n \to 0$ and $n\rho_n \to \infty$, with $m^{(n)}$ denoting the number of edges in $G^{(n)} \sim \mathcal{G}(n; \rho_n W)$, we have $m^{(n)}/n^2 \to 0$ a.s., i.e. $G^{(n)}$ is sparse, but $2m^{(n)}/n \to \infty$ a.s., i.e. the average degree of $G^{(n)}$ is not bounded. Therefore, this sparse graphon framework allows us to study *heavy-tailed* sparse graphs, as opposed to the local weak convergence framework, where there is a well-defined limit degree distribution at the root.

A. Sparse Graphon Estimation

Borgs et al. have introduced several methods for estimating the graphon W upon observing a sequence of W-random graphons [18]. Here we introduce one of their methods, called the *least squares algorithm*, which will be useful in our discussion. Given integers n and k, a function $\pi : [n] \to [k]$, and a $k \times k$ matrix B, we define B^π as the $n \times n$ matrix such that $(B^\pi)_{i,j} = B_{\pi(i),\pi(j)}$ for $1 \le i, j \le n$.

Least Squares Algorithm: Given a graph G on n vertices, and a parameter β such that $1 \le \beta \le n$, let

$$(\hat{\pi}, \hat{B}) \in \operatorname*{arg\,min}_{\pi:[n]\to[\beta], B \in \mathbb{R}_+^{[\beta]\times[\beta]}} \|A(G) - B^\pi\|_2, \qquad (5)$$

where the minimization is taken over $[\beta] \times [\beta]$ matrices B and $\pi : [n] \to [\beta]$ such that for $1 \le i \le \lfloor \beta \rfloor$, either $\pi^{-1}(\{i\}) = \emptyset$ or $|\pi^{-1}(\{i\})| \ge \lceil n/\beta \rceil$. Recall that $[\beta]$ is a shorthand for $[\lfloor \beta \rfloor]$. Assume that we have solved the optimization problem in (5), and $\hat{\pi}$ and $\hat{B}$ are its optimizers.

Then, we define the output of the least squares estimation algorithm to be the block graphon $(\hat{p}, \hat{B})$ where the probability vector $\hat{p} = (\hat{p}_1, \ldots, \hat{p}_{\lfloor\beta\rfloor})$ is such that $\hat{p}_i = |\hat{\pi}^{-1}(\{i\})|/n$ for $1 \le i \le \lfloor\beta\rfloor$.

With an appropriate choice of the parameter β_n for each n the above algorithm yields a consistent graphon estimation scheme, in the following sense.

Theorem 2 (Theorem 3.1 in [18]). *Let W be an L^2 graphon, normalized so that $\|W\|_1 = 1$, and let $G^{(n)}$ be a sequence of W-random graphs with target densities $(\rho_n : n \ge 1)$. Furthermore, let $\widehat{W}^{(n)} := (\hat{p}^{(n)}, \hat{B}^{(n)})$ be the output of the above least squares algorithm for $G^{(n)}$ with parameter β_n. If ρ_n and β_n are such that as $n \to \infty$ we have $\rho_n \to 0$, $n\rho_n \to \infty$, $\beta_n \to \infty$, and $\beta_n^2 \log \beta_n = o(n\rho_n)$, then we have with probability 1 that*

$$\lim_{n\to\infty} \delta_2\left(\frac{1}{\rho_n}\widehat{W}^{(n)}, W\right) = 0.$$

B. Sparse Graphon Entropy

For an L^2 graphon W over a probability space $(\Omega, \mathcal{F}, \pi)$, we define

$$\mathrm{Ent}(W) := \mathbb{E}[W \log W] - \mathbb{E}[W] \log \mathbb{E}[W], \tag{6}$$

where the expectations are taken with respect to the product measure $\pi \times \pi$. Note that when W is a normalized graphon we have $\mathbb{E}[W] = 1$ and $\mathrm{Ent}(W) = \mathbb{E}[W \log W]$. In fact, every normalized graphon W corresponds to a probability measure ν on $\Omega \times \Omega$ which is defined through the relation $\frac{d\nu}{d(\pi\times\pi)} = W$. With this, for such a normalized graphon, we may write

$$\mathrm{Ent}(W) = D(\nu \| \pi \times \pi). \tag{7}$$

Thus $\mathrm{Ent}(W)$ is a conic version of relative entropy, just as it is for nonnegative random variables [23, pg. 94].

We can prove the following theorem. The last part of this theorem gives an operational meaning for $\mathrm{Ent}(W)$ in terms of the asymptotic behavior of the entropy of W-random graphs.

Theorem 3. *Assume that W is an L^2 graphon on a probability space $(\Omega, \mathcal{F}, \pi)$. Then, the following hold:*

1) *$\mathrm{Ent}(W)$ is well defined and $\mathrm{Ent}(W) < \infty$.*
2) *$\mathrm{Ent}(W) \ge 0$.*
3) *For $\alpha > 0$, we have $\mathrm{Ent}(\alpha W) = \alpha\,\mathrm{Ent}(W)$.*
4) *Assume that a sequence W_n of L^2 graphons over $(\Omega_n, \mathcal{F}_n, \pi_n)$ is given such that $\delta_2(W_n, W) \to 0$ as $n \to \infty$. Then we have $\mathrm{Ent}(W_n) \to \mathrm{Ent}(W)$ as $n \to \infty$.*
5) *Assume that $G^{(n)} \sim \mathcal{G}(n; \rho_n W)$ is a sequence of W-random graphs with target density ρ_n such that $n\rho_n \to \infty$ and $\rho_n \to 0$. Then, with $\overline{m}_n := \binom{n}{2}\rho_n$, we have*

$$\lim_{n\to\infty} \frac{H(G^{(n)}) - \overline{m}_n \log \frac{1}{\rho_n}}{\overline{m}_n} = 1 - \mathrm{Ent}(W).$$

IV. Problem Statement and Main Results

In this section we formalize the problem of finding a universal compression scheme which is capable of compressing a sequence comprised of either sparse graphs which are convergent in the local weak sense as discussed in Section II, or heavy-tailed sparse graphs generated as a sequence of W-random graphs as discussed in Section III.

More precisely, for each n, we want to design a compression map $f_n : \mathcal{G}_n \to \{0,1\}^* - \emptyset$ which assigns a prefix–free codeword to every graph on the vertex set $[n]$, as well as a decompression map g_n, such that $g_n \circ f_n$ is identity, i.e. lossless compression. Additionally, we want this compression scheme to be *universally optimal* in the following sense:

1) If we have a sequence of graphs $G^{(n)}$ converging in the local weak sense to some unimodular $\mu \in \mathcal{P}(\mathcal{T}_*)$, then

$$\limsup_{n\to\infty} \frac{\mathsf{nats}(f_n(G^{(n)})) - m^{(n)} \log n}{n} \le \Sigma(\mu).$$

Here, $m^{(n)}$ is the number of edges in $G^{(n)}$, and the normalization of the codeword length is done in a way consistent with the definition of the BC entropy.

2) If $G^{(n)} \sim \mathcal{G}(n; \rho_n W)$ for a normalized L^2 graphon W and a sequence of target densities ρ_n with $\rho_n \to 0$ and $n\rho_n \to \infty$, then with probability 1 we have

$$\limsup_{n\to\infty} \frac{\mathsf{nats}(f_n(G^{(n)})) - \overline{m}_n \log \frac{1}{\rho_n}}{\overline{m}_n} \le 1 - \mathrm{Ent}(W),$$

where $\overline{m}_n := \binom{n}{2}\rho_n$. Note that here the normalization is consistent with the asymptotics of the sparse graphon entropy as discussed in the last part of Theorem 3 in Section III-B.

In this setup the encoder only observes the graph realization $G^{(n)}$, and not the whole sequence $(G^{(n)} : n \ge 1)$. Moreover, the encoder does not a priori know from which of the two ensemble types the realization $G^{(n)}$ comes, nor does it know the limit objects for each of the two sequence of ensembles.

We address this problem by introducing such a universal compression scheme, and will further discuss a converse result. Our compression scheme employs a *splitting* method. More precisely, given a graph $G^{(n)}$, we choose a splitting parameter Δ_n and split $G^{(n)}$ into two graphs, denoted by $G^{(n)}_{\Delta_n}$ and $G^{(n)}_*$. These two graphs are both on the vertex set $[n]$, and each edge in $G^{(n)}$ appears in precisely one of them. More precisely, $G^{(n)}_{\Delta_n}$ consists of the edges (v, w) in $G^{(n)}$ where the degrees of both of their endpoints are at most Δ_n. We then define $G^{(n)}_*$ to include the remaining of edges in $G^{(n)}$. We encode each of these two graphs separately, as discussed in Section V. Roughly speaking, the splitting parameter is chosen so that when $G^{(n)}$ is coming from a sequence in the local weak convergence regime, $G^{(n)}_{\Delta_n}$ contains most of the edges in $G^{(n)}$, while when $G^{(n)}$ is coming from a sparse graphon ensemble $G^{(n)}_*$ contains most of the edges in $G^{(n)}$. To emphasize the dependence of the compression and the decompression maps on the parameter Δ_n, we denote these mappings by $f_n^{\Delta_n}$ and $g_n^{\Delta_n}$, respectively. We can prove that, with an appropriate choice of Δ_n, a universal compression scheme exists in the sense of the following theorem. The sequence a_n here ensures that $n\rho_n$ does not converge to infinity arbitrarily slowly.

Theorem 4. *Assume that $n\rho_n \geq a_n$ where $(a_n : n \geq 1)$ is known to both the encoder and the decoder, with $a_n \to \infty$ as $n \to \infty$. Then, we can choose the sequence of splitting parameters $(\Delta_n : n \geq 1)$ such that our compression scheme $((f_n^{\Delta_n}, g_n^{\Delta_n}) : n \geq 1)$ achieves optimal universal compression, in the sense discussed above. Namely, we have*

1) *If $G^{(n)}$ is a sequence of random graphs converging a.s. in the local weak sense to some unimodular $\mu \in \mathcal{P}(\mathcal{T}_*)$ with $\deg(\mu) \in (0, \infty)$, then with probability 1*

$$\limsup_{n\to\infty} \frac{\mathsf{nats}(f_n^{\Delta_n}(G^{(n)})) - m^{(n)} \log n}{n} \leq \Sigma(\mu),$$

where $m^{(n)}$ denotes the number of edges in $G^{(n)}$.

2) *On the other hand, if $G^{(n)} \sim \mathcal{G}(n; \rho_n W)$ is a sequence of W-random graphs with target densities ρ_n, where W is a normalized L^2 graphon, assuming that $\rho_n \to 0$ as $n \to \infty$ and $n\rho_n \geq a_n$, with probability 1 we have*

$$\limsup_{n\to\infty} \frac{\mathsf{nats}(f_n^{\Delta_n}(G^{(n)})) - \overline{m}_n \log \frac{1}{\rho_n}}{\overline{m}_n} \leq 1 - \mathrm{Ent}(W),$$

where $\overline{m}_n := \binom{n}{2}\rho_n$.

In the above setting, the encoder and the decoder only know the sequence a_n, and do not know from which of the two settings the input graph $G^{(n)}$ is generated, neither do they know the limit objects μ nor W in each setting, respectively.

Note that the first part of this theorem also allows for a fixed sequence $G^{(n)}$. We also have the following converse result:

Theorem 5. *Assume that $((f_n, g_n) : n \geq 1)$ is a sequence of lossless compression/decompression maps (i.e. $g_n \circ f_n$ is the identity map). Then we have the following.*

1) *For any unimodular $\mu \in \mathcal{P}(\mathcal{T}_*)$ with $\deg(\mu) \in (0, \infty)$, there exists a sequence of random graphs $G^{(n)}$ defined on a joint probably space such that $U(G^{(n)})$ converges a.s. to μ in the local weak sense, and for all $t < \Sigma(\mu)$*

$$\mathbb{P}\left(\limsup_{n\to\infty} \frac{\mathsf{nats}(f_n(G^{(n)})) - m^{(n)} \log n}{n} \leq t\right) < 1,$$

where $m^{(n)}$ denotes the number of edges in $G^{(n)}$.

2) *For any normalized L^2 graphon W and any sequence of target densities ρ_n such that $n\rho_n \to \infty$ and $\rho_n \to 0$, if $G^{(n)} \sim \mathcal{G}(n; \rho_n W)$ is the sequence of W-random graphs with target densities ρ_n, then for all $t < 1 - \mathrm{Ent}(W)$,*

$$\mathbb{P}\left(\limsup_{n\to\infty} \frac{\mathsf{nats}(f_n(G^{(n)})) - \overline{m}_n \log \frac{1}{\rho_n}}{\overline{m}_n} \leq t\right) < 1.$$

V. Compression Scheme and Proof Sketch

In this section, we introduce the compression and decompression maps $f_n^{\Delta_n}$ and $g_n^{\Delta_n}$ and discuss the proof ideas of our main results. Given the sequence a_n satisfying $n\rho_n \geq a_n$ and $a_n \to \infty$ as $n \to \infty$ as in Theorem 4, we choose the splitting parameter $\Delta_n := \min\{\log a_n, \log\log n\}$ and find $G_{\Delta_n}^{(n)}$ and $G_*^{(n)}$ as defined in Section IV. We first compress $G_{\Delta_n}^{(n)}$ using the compression method f_n^{lwc} of Section II-B. Let R_n denote the set of vertices $v \in [n]$ such that either $\deg_{G^{(n)}}(v) > \Delta_n$ or $\deg_{G^{(n)}}(w) > \Delta_n$ for some $w \sim_{G^{(n)}} v$. Clearly, both endpoints of every edge in $G_*^{(n)}$ are in R_n. We may encode the set R_n using $\log n + \log\binom{n}{|R_n|}$ nats. We then apply the least squares algorithm of Section III-A to $G^{(n)}$ with parameter β_n, defined as follows, to obtain $\hat{\pi}_n$ and $\hat{B}_n$. To define β_n, let $\alpha_n := \exp(\lfloor \log m_*^{(n)}/n \rfloor)$ where $m_*^{(n)}$ denotes the number of edges in $G_*^{(n)}$. Moreover, let $\beta_n := \sqrt{\alpha_n}/\log\alpha_n$ if $\alpha_n > e^2$ and 1 otherwise. By rearranging the rows and columns in the adjacency matrix of $G^{(n)}$, we can think of $\hat{\pi}_n$ as partitioning this adjacency matrix into at most β_n blocks. Since $G_*^{(n)}$ is a subgraph of $G^{(n)}$, we encode $G_*^{(n)}$ by going over each block, and for the block $(i, j), 1 \leq i \leq j \leq \beta_n$, we encode the edges of $G_*^{(n)}$ that fall in block (i, j). We do this by first encoding the number of edges of $G_*^{(n)}$ in block (i, j) followed by encoding the positions of ones in that block.

For the proof of Theorem 4, if $G^{(n)}$ is convergent in the local weak sense to some μ, it can be shown that, since $\Delta_n \to \infty$, $G_{\Delta_n}^{(n)}$ converges in the local weak sense to the same limit. Therefore, using Eq. (2), the asymptotic normalized codeword length associated to compressing $G_{\Delta_n}^{(n)}$ does not exceed $\Sigma(\mu)$. Also, it can be shown that in this case the number of nats used to encode $G_*^{(n)}$ is $o(n)$. This shows the first part of Theorem 4. To prove the second part of the theorem we first show that the above choice of β_n satisfies the conditions of Theorem 2. This means $\delta_2(\widehat{W}^{(n)}/\rho_n, W) \to 0$ a.s., where $\widehat{W}^{(n)}$ is the block graphon obtained by applying the least squares algorithm on $G^{(n)}$ with parameter β_n. Moreover, it can be shown that the number of nats used to encode $G_{\Delta_n}^{(n)}$ is $o(\overline{m}_n)$, and the number of nats used to encode $G_*^{(n)}$ is $\overline{m}_n \log\frac{1}{\rho_n} + \overline{m}_n(1 - \mathrm{Ent}(\widehat{W}^{(n)}/\rho_n)) + o(\overline{m}_n)$. But, using part 4 of Theorem 3, $\mathrm{Ent}(\widehat{W}^{(n)}/\rho_n) \to \mathrm{Ent}(W)$ a.s. since, with probability 1, we have $\delta_2(\widehat{W}^{(n)}/\rho_n, W) \to 0$.

The proof of Theorem 5 follows from the fact that in the context of local weak convergence no compression scheme can achieve an asymptotic rate below the BC entropy [10, Thoerem 4]. The second part follows from the asymptotic behavior of the entropy of W-random graphs, i.e. part 5 of Theorem 3.

VI. Conclusion

We introduced a universal lossless compression method simultaneously applicable to both sparse graphs and heavy-tailed sparse graphs. We employed the framework of local weak convergence for sparse graphs, and the sparse graphon framework for heavy-tailed sparse graphs.

Acknowledgements

Thanks for support from the NSF grants CNS–1527846, CCF–1618145, CCF-1901004, CIF-2007965, the NSF Science & Technology Center grant CCF–0939370 (Science of Information), and the William and Flora Hewlett Foundation supported Center for Long Term Cybersecurity at Berkeley.

REFERENCES

[1] P. Boldi and S. Vigna, "The webgraph framework i: compression techniques," in *Proceedings of the 13th international conference on World Wide Web*. ACM, 2004, pp. 595–602.

[2] P. Boldi, M. Rosa, M. Santini, and S. Vigna, "Layered label propagation: A multiresolution coordinate-free ordering for compressing social networks," in *Proceedings of the 20th international conference on World wide web*. ACM, 2011, pp. 587–596.

[3] 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.

[4] D. J. Aldous and N. Ross, "Entropy of some models of sparse random graphs with vertex-names," *Probability in the Engineering and Informational Sciences*, vol. 28, no. 02, pp. 145–168, 2014.

[5] E. Abbe, "Graph compression: The effect of clusters," in *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2016, pp. 1–8.

[6] A. Lempel and J. Ziv, "Compression of two-dimensional data," *IEEE transactions on information theory*, vol. 32, no. 1, pp. 2–8, 1986.

[7] A. R. Asadi, E. Abbe, and S. Verdú, "Compressing data on graphs with clusters," in *2017 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2017, pp. 1583–1587.

[8] T. Łuczak, A. Magner, and W. Szpankowski, "Compression of preferential attachment graphs," in *2019 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2019, pp. 1697–1701.

[9] A. Bhatt, Z. Wang, C. Wang, and L. Wang, "Universal graph compression: Stochastic block models," *arXiv preprint arXiv:2006.02643*, 2020.

[10] P. Delgosha and V. Anantharam, "Universal lossless compression of graphical data," *IEEE Transactions on Information Theory*, vol. 66, no. 11, pp. 6962–6976, 2020.

[11] ——, "A universal low complexity compression algorithm for sparse marked graphs," in *2020 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2020, pp. 2349–2354.

[12] ——, "Distributed compression of graphical data," *arXiv preprint arXiv:1802.07446*, 2018.

[13] I. Benjamini and O. Schramm, "Recurrence of distributional limits of finite planar graphs," *Electron. J. Probab.*, vol. 6, pp. no. 23, 13 pp. (electronic), 2001. [Online]. Available: http://dx.doi.org/10.1214/EJP.v6-96

[14] D. Aldous and J. M. Steele, "The objective method: probabilistic combinatorial optimization and local weak convergence," in *Probability on discrete structures*. Springer, 2004, pp. 1–72.

[15] D. Aldous and R. Lyons, "Processes on unimodular random networks," *Electron. J. Probab*, vol. 12, no. 54, pp. 1454–1508, 2007.

[16] C. Borgs, J. Chayes, H. Cohn, and Y. Zhao, "An L^p theory of sparse graph convergence I: Limits, sparse random graph models, and power law distributions," *Transactions of the American Mathematical Society*, vol. 372, no. 5, pp. 3019–3062, 2019.

[17] C. Borgs, J. T. Chayes, H. Cohn, Y. Zhao *et al.*, "An L^p theory of sparse graph convergence II: LD convergence, quotients and right convergence," *The Annals of Probability*, vol. 46, no. 1, pp. 337–396, 2018.

[18] C. Borgs, J. T. Chayes, H. Cohn, and S. Ganguly, "Consistent nonparametric estimation for heavy-tailed sparse graphs," *arXiv preprint arXiv:1508.06675*, 2015.

[19] C. Bordenave and P. Caputo, "Large deviations of empirical neighborhood distribution in sparse random graphs," *Probability Theory and Related Fields*, vol. 163, no. 1-2, pp. 149–222, 2015.

[20] P. Billingsley, *Convergence of probability measures*. John Wiley & Sons, 2013.

[21] L. Lovász, *Large networks and graph limits*. American Mathematical Soc., 2012, vol. 60.

[22] S. Janson, "Graphons, cut norm and distance, couplings and rearrangements," *arXiv preprint arXiv:1009.2376*, 2010.

[23] S. Boucheron, G. Lugosi, and P. Massart, *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.