Compression with Unlabeled Graph Side Information

Hesam Nikpey[1], Saswati Sarkar[2], Shirin Saeedi Bidokhti[1,2]
University of Pennsylvania
[1] Department of Computer and Information Science, [2] Department of Electrical and Systems Engineering
{hesam, swati, saeedi }@seas.upenn.edu

2023 IEEE International Symposium on Information Theory (ISIT) | 978-1-6654-7554-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/ISIT54713.2023.10206622

Abstract—**With the growth of big data in the past few decades, compression has become inseparable from data generation. The data generated daily across different platforms are correlated: friend networks on Facebook and Instagram, contact networks in subsequent days, and many more. This raises the question of compressing a dataset using another correlated dataset. For instance, can we compress the Facebook graph of friends when we know Instagram's graph? This can be cast as the classical problem of source coding with side information, and the answer is known to be positive when the graphs are "labeled" and/or aligned, meaning we need to know the node corresponding to Jon Doe in both Facebook and Instagram graphs. The classical idea is to utilize joint typicality to decide whether two graphs are correlated or not. In practice, graphs are often not aligned and/or the labels are concealed to keep the identity of the users private. In these scenarios, classical ideas are no longer applicable as joint typicality highly depends on the ordering of sequences. In this work, we prove for the first time the existence of lossless graph compression schemes that utilize unlabeled side information and improve the compression rate. In order to do that, we design binning along with a novel testing criterion that relies on graph matching, the closely related quadratic assignment problem and its asymptotic properties.**

I. Introduction

Graphical representations and graph databases have emerged in all scientific disciplines, with graphs representing the interaction and relationship between objects, events, and situations. Examples include knowledge graphs in search engines and recommendation systems, protein-protein interaction networks, and genome graphs in biological systems. With the huge amount of data that is collected and generated in the past decade, their compression is of paramount importance.

Compression of graphical data goes beyond conventional image data in the distinction that may be made between the graph structure (semantics) and the graph labels. While labels are an integral part of conventional data, there are various applications where we need to work with unlabeled graphical structures. In [1], fundamental limits and efficient algorithms are devised for compressing random Erdös-Rényi (ER) graphs up to their structures.

In this work, we ask the following novel question: *In compressing graphs, can one benefit from unlabeled graph side information to reduce the rate of compression?* Consider correlated graphs G_1 and G_2; G_2 may be deemed as side information on G_1. A permuted version of G_2, say G_2^π when the permutation is unknown, is referred to as the unlabeled side information for G_1. We wish to compress G_1 and describe it to a decoder as W, the decoder has access to the unlabeled side information of G_1, G_2^π. See Fig. 1.

Let us briefly discuss some applications. Consider the knowledge graph that describes users' movie ratings. Various similar graphs with common semantics exist through Netflix, IMDB, etc and the common practice is to remove labels from the graph to ensure users' privacy [2], [3]. In order to compress the structural graph of Netflix's dataset, we ask if side information about IMDB's database could help improve rates of compression. Here, the Netflix graphical data is G_1 and the side information graph G_2^π is the IMDB graphical data. Permutation π captures the fact that the two datasets may be anonymized and not aligned. As a second example, consider the protein-protein interaction network across different species. Since there are many common protein interactions between proteins among different species, their corresponding networks are correlated but unlabeled in the sense that the underlying alignment between the two is not known [4]. We ask if such data can be jointly compressed at a rate that is less than the rate needed to compress each individually.

In this paper, we formulate the problem of graph compression with unlabeled side information and provide upper and lower bounds on the rate of compression.

A. Related Work

Graph compression. In its simplest form, graph compression can be viewed as compressing a sequence of $\binom{n}{2}$ edges that describe the graph. In particular, in the lossless settings, it can be viewed through the standard lens of compressing binary sequences and traditional results hold [5]. In recent years, three directions have been explored: (i) universality in graph compression [6]–[9] (ii) locality in graph compression [10], [11] where novel metrics of distortion are considered to ensure the local structure is preserved, and (iii) structural graph compression [1] where a graph is to be recovered up to isomorphism. Our work is closest to the last in formulation.

Side information. When side information Y about the source X is provided at the decoder, classical results such as [12] have proved that rate reduction is possible. Generalizations to distributed lossless and lossy source coding are studied in [13]–[17]. Fundamentally speaking, the rate reduction is built on the statistical knowledge of the joint distribution of (X, Y) and is captured through joint typicality of n i.i.d. realizations of X^n and Y^n. *Once one of these sequences is*

Authorized licensed use limited to: University of Pennsylvania. Downloaded on April 12,2024 at 19:39:20 UTC from IEEE Xplore. Restrictions apply.

978-1-6654-7554-9/23/$31.00 ©2023 IEEE

permuted, all existing frameworks fall short of answering the question of whether side information is still useful. To the best of our knowledge, this work is the first to tackle this question.

Graph matching. The problem of graph matching does not study compression, but it asks a question that is implicitly relevant to our problem. Given two graphs that are jointly generated according to a joint distribution but have undergone permutations of the nodes, how can we recover the mapping that aligns them? This problem is known to be related to the Quadratic Assignment Problem (QAP) [18]–[21]. In the past decade, a great body of work in information theory [22]–[27] and computer science [19], [28]–[30] have led to optimal polynomial-time algorithms for correlated random ER graphs that discover the mapping with high probability. Furthermore, [18], [19], [22] have shown that the graph matching problem has a close connection to the following hypothesis testing problem: Given two permutated graphs, determine whether they are instances of an independent probabilistic model or a joint probabilistic model.

It may appear, at first glance, that a hypothesis testing algorithm, as discussed above, along with a joint typicality test, as used in compression schemes with side information, would provide an optimal compressor for compressing graphs with unlabeled side information. However, we argue in Section III that schemes based on the above ideas would not improve the compression rate because the error in the hypothesis testing problem of [18], [19] vanishes only polynomially in n. This is why we have to resort to other methods.

II. Problem Setup

A. Notations

We show graphs with n nodes as binary vectors $G = (e_1, \dots, e_m)$ where $m = \binom{n}{2}$ and each e_k represents the edge between two nodes i, j in the graph, i.e. e_k is one if there is an edge between i and j, and the graph has no self-loop. The adjacency matrix is also denoted by a matrix $A_{n \times n} = \{a_{i,j}\}_{1 \leq i,j \leq n}$ with binary elements where $a_{i,j} = 1$ iff there is an edge between nodes i, j. We show small constants as δ and ϵ with or without subscripts. We show permutations as bijective functions $\pi : [n] \to [n]$, so there are $n!$ of them. A permutation π on a graph G is defined as $G^\pi = (e_1^\pi, \dots, e_m^\pi)$ where e_k^π is the edge after permuting nodes i, j, i.e. the edge between $\pi^{-1}(i)$ and $\pi^{-1}(j)$. We show the number of common edges of G_1 and G_2 with $|G_1 \cap G_2|$.

B. Problem formulation

Two random graphs $G_1 = (e_1, \dots, e_m)$ and $G_2 = (e'_1, \dots, e'_m)$ with n nodes are generated jointly as follows. The edges e_i, e'_i are random variables defined by $e_i = X_i Y_i$ and $e'_i = X_i Z_i$ where X_i is a Bernoulli random variable with parameter p and Y_i and Z_i are independent Bernoulli random variables with parameter γ. For two distinct i, j, all X_i, X_j, Y_i, Y_j, and Z_i, Z_j are independent of each other. We denote the resulting joint pmf on the pair of edges (e_i, e'_i) by $p^{\text{corr}}_{e,e'}$ and the marginals by p_e, p'_e which are both Bernoulli with mean $p\gamma$. This model was introduced in [31] and has since been studied vastly [18], [19], [27]. If e and e' are generated independently, we denote the corresponding pmf by $p^{\text{indep}}_{e,e'} = p_e \times p_e$.

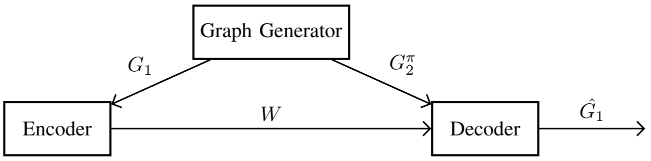

Fig. 1: Compression Scheme

The encoder (Alice) is given graph G_1 and the decoder (Bob) has access to a permutated version of G_2, i.e. G_2^π, see Fig 1. The encoder puts out message $W = f(G_1)$ of rate R where f is the encoding function and its range $\mathcal{W}$ is of cardinality $|\mathcal{W}| = 2^{\binom{n}{2}R}$. Message W is received at the decoder who also has access to G_2^π for an unknown permutation π. The decoder outputs

$$\hat{G}_1 = g(W, G_2^\pi). \tag{1}$$

The goal is to design functions f, g such that

$$\lim_{n \to \infty} \Pr(\hat{G}_1 \neq G_1) = 0. \tag{2}$$

We are interested in the minimum rate R that allows reliable recovery of G_1 per (2).

Note that π is not known by Bob. If it was, Bob would simply apply π^{-1} to his graph and reduce the problem to the classical setting of lossless source coding with side information in which case, the optimal rate is known to be

$$h(e_1 | e'_1) = p\gamma h(\gamma) + (1 - p\gamma) h(\frac{p\gamma(1-\gamma)}{1 - p\gamma}) \tag{3}$$

where h is the binary entropy function. As a matter of fact, (3) is a lower bound to our problem. As assuming π is not attainable, we believe better lower bounds can be achieved.

In the classical setting, (3) is achieved by random binning and joint typicality tests within the bins [12]. This is rooted in the assumption that the binary edge sequences that represent G_1 and G_2 are aligned and hence joint typicality tests can be performed. When π is unknown and the sequences that represent G_1 and G_2 are unaligned, however, testing joint typicality seems impossible. To overcome the issue, we design an alternative testing strategy within the bins. Our approach builds on asymptotic properties of the Quadratic Assignment Problem (QAP) as briefly discussed next.

C. Preliminaries

The *quadratic assignment problem* (QAP) is a combinatorial optimization problem that appears in many applications. In this problem, for given matrices $A_{n \times n} \in \mathbb{R}^+_{n \times n}$ and $B_{n \times n} \in \mathbb{R}^+_{n \times n}$ we want to find a permutation π that maximizes

$$\sum a_{ij} b_{\pi(i),\pi(j)}. \tag{4}$$

Although the QAP problem is hard to solve [32], [33], its asymptotic behavior is well-known when the input is random

Authorized licensed use limited to: University of Pennsylvania. Downloaded on April 12,2024 at 19:39:20 UTC from IEEE Xplore. Restrictions apply.

and large enough [34]. In [34], they show the cost of the best and worst solutions get close to each other as n gets large.

More precisely, [34] considers a set of combinatorial optimization problems with parameter n defined on finite ground sets E_n. The set of feasible solutions T_n are defined as subsets of E_n, so a feasible solution $S \in T_n$ of the problem is simply a subset of E_n. Furthermore, there is a cost function $c_n : E_n \to \mathbb{R}^+$ and the cost of a solution S is $\sum_{e\in S} c_n(e)$. As an example, let's see how the QAP problem fits into this scheme. For the QAP problem, we have $E_n = \{(i,j,p,q) \mid i,j,p,q = 1,2,\dots,n\}$ and a feasible solution $S_\pi \in T_n$ of a QAP is a subset of the form $S_\pi = \{(i,j,\pi(i),\pi(j)) \mid i,j = 1,2,\dots,n\}$ for all permutations π and therefore $|S| = n^2$ and $|T_n| = n!$ [34]. If we take the matrices A and B into account, the cost function for the QAP problem is simply $c_n(i,j,\pi(i),\pi(j)) = a_{ij}b_{\pi(i)\pi(j)}$. By treating A and B as the adjacency matrices of two graphs, the QAP problem becomes equivalent to maximizing the number of common edges:

$$\sum_{i,j} c_n(i,j,\pi(i),\pi(j)) = \sum_{i,j} a_{ij}b_{\pi(i)\pi(j)}, \tag{5}$$

over all permutations π. As $\forall i : a_{ii} = 0$, we assume that the size of the solution is $m = \binom{n}{2}$ from now on. The following result is proved in [34].

Theorem 1. *[34] Let $c_n(e), e \in S, S \in T_n, n \in \mathbb{N}$ be identically distributed random variables in $[0,1]$ with expected value $E := E(c_n(e))$ and variance $\sigma^2 := \sigma^2(c_n(e)) > 0$. For given $\varepsilon > 0$, let ε_0 fulfill*

$$0 < \varepsilon_0 \le \sigma^2 \quad \textit{and} \quad 0 < \frac{E+\varepsilon_0}{E-\varepsilon_0} \le 1+\varepsilon.$$

Furthermore, let the following three conditions be satisfied:

1) *$c_n(e), e \in S$, are independently distributed for every fixed $S \in T_n, n \in \mathbb{N}$.*
2) *All $S \in T_n$ have the same cardinality for fixed n, i.e. $|S| = |\hat{S}|$ for all $S, \hat{S} \in T_n$, n fixed.*
3) *$\lambda_0|S| - \log|T_n| \to \infty$ as $n \to \infty$ where λ_0 is defined by $\lambda_0 := 2\big(\varepsilon_0\sigma/(\varepsilon_0 + 2\sigma^2)\big)^2$.*

Then

$$\mathrm{P}\left\{\frac{\max_{S\in T_n}\sum_{e\in S} c_n(e)}{\min_{S\in T_n}\sum_{e\in S} c_n(e)} < 1+\varepsilon\right\}$$
$$\ge 1 - 2|T_n|\exp(-|S|\lambda_0) \to 1 \quad \textit{as } n \to \infty. \tag{6}$$

III. Main Results

Our main result is an upper bound on the optimal rate of graph compression as defined in Section II-B as follows:

Theorem 2. *There is a compression scheme with rate $R = (1+\delta)(h(p\gamma) - \lambda_0)$, $\delta = o(1) = n^{-1/3}$, $\lambda_0 = 2\frac{s^2\sigma^2}{(s+2)^2}$, $\sigma^2 = (p\gamma)^2(1-(p\gamma)^2)$, and $s = \min(1, \frac{1-\epsilon'-p}{(1-\epsilon'+p)(1-(p\gamma)^2)})$, for a constant $1-p > \epsilon' > 0$, such that with high probability, Bob can recover G_1 losslessly.*

Remark 1. *λ_0 represents the additional compression brought about by the side information. Here $\lambda_0 \le h(p\gamma)$. Thus when $p\gamma \to 0$ or $p\gamma \to 1$, $\lambda_0 \le h(p\gamma) \approx 0$. Thus, the absolute value of the additional compression is small. This is also intuitively expected as in these cases, either the graph is very sparse or very dense, and only the edges or the pairs which do not correspond to edges can be stored in a few bits. Thus, there is not much scope of compression beyond the obvious. Thus, side information may be useful only when $p\gamma$ is away from 0 and 1.*

Our Method and Ideas: We first discuss an intuitive line of thought that did not work out but led us to the ideas behind our solution.

In the classic version of the problem where no permutation is applied to G_2, the optimal rate is achieved by random binning. In particular, the encoding codebook consists of all the typical graphs, binned into $2^{\binom{n}{2}R}$ bins. Alice finds the sequence representing G_1 in the codebook and sends the corresponding bin index. Bob would then receive the bin index and look for a unique sequence in the bin that is jointly typical with G_2. The rate, and hence the bin size, are set so that only a single sequence from the bin, namely G_1, is found with high probability and that sequence is output as the estimate.

In our problem where G_2 is permutated by an unknown permutation π, joint typicality *cannot* be applied because e_i is no longer correlated with e'_i but an unknown $e'_{\pi(i)}$. To get around joint typicality, one may argue that graph matching and graph correlation hypothesis testing ideas such as [18], [19] may help Bob to decide which sequences in the described bin are correlated with G_2^π. In particular, the graph correlation hypothesis testing problem tackles the following task. Given unaligned graphs G_1 and G_2^π, decide whether the two graphs are instances of a correlated ER model (following $p^{\text{corr}}_{e,e'}$ as introduced in Section II-B) or independent instances of the ER model (following $p^{\text{indep}}_{e,e'}$). Note that the codebook sequences are generated iid at random across different codewords and G_1, by design, is the only graph in the bin correlated with G_2^π. So one may propose the following simple idea. Use the hypothesis testing instead of typicality: for each graph in the bin, check whether it is correlated with G_2^π using the hypothesis testing algorithms of [18], [19]. It turns out that this scheme would ensure reliable recovery only if the bin sizes are of polynomial size, leading to the overall rate $h(p\gamma)$ which can also be achieved without using side information. The reason is that the failure probability of the hypothesis testing algorithms of [18], [19] goes to zero polynomially in n, whereas the bin sizes are desired to be exponential to achieve non-trivial rates.

In order to recover graph G_1 from the exponentially many elements of its bin, using the unlabeled side information graph G_2^π, we need to answer the following fundamental question: Given two independent graphs G and G_2^π, what is the probability that a permutation π^* exists such that G^{π^*} and G_2^π are jointly typical. To the best of our knowledge, little is known and there are only results on the reverse regarding the probability that two jointly typical graphs remain jointly typical under a permutation [35].

Motivated by a well-known result in graph matching that relates the alignment of graphs to maximizing the common edges between them [31, Theorem 4.1], we propose to distinguish G_1 and the rest of the graphs in the bin by finding the "best permutation" π^* that maximizes the common edges with G_2^π. Note that without permutation and in expectation, G_1 and G_2 have $\mathbb{E}[|G_1 \cap G_2|] = \binom{n}{2} p\gamma^2$ edges in common whereas independent graphs are expected to have $\binom{n}{2} p^2\gamma^2$ edges in common. The problem of finding the best permutation can be seen as solving a quadratic assignment problem as mentioned in (5), and we use its asymptotic properties [34] to analyze our scheme. The cost function in (5) captures the number of common edges between any two graphs G and G' with adjacency matrices A and B. When G and G_2 are independently generated, Theorem 1 states that with a probability that approaches 1 exponentially fast in n, all permutations return almost the same value for the cost function. So with high probability, the best permutation π^* returns almost the average, which is $\binom{n}{2} p^2\gamma^2$ common edges. On the other hand, the average for G_1 and G_2 (being correlated) is $\binom{n}{2} p\gamma^2$. As $p\gamma^2 > p^2\gamma^2$, by setting the right parameters and testing whether the number of common edges (between G_2^π and the graphs in the bin) for the best permutation is more than $(1-\epsilon')\binom{n}{2} p\gamma^2$, we rule out all the independent graphs G in the bin. On the other hand, for G_1 and G_2^π and for permutation π, by Chernoff bound we show that the number of common edges is concentrated around its mean, hence the number is more than $(1-\epsilon')\binom{n}{2} p\gamma^2$ with high probability. We give a concrete proof of the above argument in the next section.

IV. Proof of Theorem 2

We first use Theorem 1 to get a concentration result for the number of common edges when the graphs are generated independently (Lemma 1). Then we use it to provide a compression scheme and finally optimize the set of parameters.

Lemma 1. *For two randomly generated ER graphs G_1 and G_2 with parameter $p\gamma$, the probability that there exists a permutation π so that G_1^π and G_2 share more than $mp\gamma^2(1-\epsilon')$, where $1-p > \epsilon' > 0$ is a constant , is at most*

$$q = 2^{-(\lambda_0 m - 2n\log n)} \tag{7}$$

where $\lambda_0 = 2\frac{s^2\sigma^2}{(s+2)^2}$ *and* $s = \min(1, \frac{1-\epsilon'-p}{(1-\epsilon'+p)(1-(p\gamma)^2)})$.

Proof. We prove the lemma using Theorem 1. Let e_i's be the edges of G_1 and e_i''s be the edges of G_2. First, we simplify the parameters used in the theorem. For any π, we have $E = \mathbb{E}[e_i e'_{\pi(i)}] = (p\gamma)^2$, and $\sigma^2(e_i e'_{\pi(i)}) = \sigma^2 = (p\gamma)^2(1-(p\gamma)^2)$. Set $\epsilon_0 = s\sigma^2, 0 < s \leq 1$, then

$$\frac{E+\epsilon_0}{E-\epsilon_0} = \frac{1+s(1-(p\gamma)^2)}{1-s(1-(p\gamma)^2)} \leq 1+\epsilon.$$

The first condition of Theorem 1 holds as $c_n(e) = e_i e_{\pi(j)}$ and they are independent of each other by ER property of the graphs. The second condition also holds as $|S| = \binom{n}{2} = m$ in our setting. In the third condition, we have

$$\lambda_0 = 2(\frac{s\sigma^2 \cdot \sigma}{s\sigma^2 + 2\sigma^2})^2 = 2\frac{s^2\sigma^2}{(s+2)^2}.$$

Set $1+\epsilon = (1-\epsilon')/p$, this choice becomes clear when we use it in Theorem 2. Note that as $\epsilon' < 1-p$, we have $\epsilon > 0$. Then we have

$$\frac{1+s(1-(p\gamma)^2)}{1-s(1-(p\gamma)^2)} \leq 1+\epsilon = (1-\epsilon')/p$$

From the above, we get $s \leq \frac{1-\epsilon'-p}{(1-\epsilon'+p)(1-(p\gamma)^2)}$. Note that $s \leq 1$, and $(\frac{s}{s+2})^2$ is increasing in $[0,1]$. The largest λ_0 is achieved by setting $s = \min(1, \frac{1-\epsilon'-p}{(1-\epsilon'+p)(1-(p\gamma)^2)})$ and, and by the fact that $\log n! < 2n\log n$ and replacing it in the formula from Theorem 1 we get the above lemma. □

We are now ready to prove Theorem 2.

Proof. The proof is by random coding. Below, we first describe the codebook generation, and encoding/decoding scheme, and then provide the corresponding error analysis.

Codebook construction: generate $2^{m(1+\delta)h(p\gamma)}$ codewords $G(w,v), w = 1,\ldots 2^{mR}, v = 1,\ldots, 2^{m(h(p\gamma)(1+\delta)-R)}$ independently each consisting m iid binary values following Bernoulli($p\gamma$). Note that codewords are graphs.

Encoder: Alice finds a pair (w,v) so that $G_1 = G(w,v)$. If there is more than one graph, she chooses from a predefined function $(w,v) = f_1(G_1, G(.))$. If there is no graph, she chooses $(w,v) = (1,1)$. She then sends w using R bits.

Decoder: For a received message w, Bob finds all graphs $G = G(w,v), v = 1,\ldots 2^{m((1+\delta)h(p\gamma)-R)}$ and for each graph $G(w,v)$, finds π^* that maximizes $t(v) = |G^{\pi^*}(w,v) \cap G^\pi|$. If there is a unique v satisfying $t(v) \geq mp\gamma^2(1-\epsilon')$ return $G(w,v)$. Otherwise return $G(w,1)$.

Analysis: We will bound $Pr(G_1 \neq G(w,v))$. We will bound the following errors:

1) Alice cannot find an index (w,v) such that $G_1 = G(w,v)$.
2) There is an index (w,v) such that $G(w,v) \neq G_1$ and $t(v) > mp\gamma^2(1-\epsilon')$.
3) For index (w,v) where $G(w,v) = G_1$, Bob gets $t(v) < mp\gamma^2(1-\epsilon')$.

The first error is $o(1)$ by the definition of typical sets and their properties.

For the second error, by Lemma 1 and the fact that G_2 and $G \neq G_1$ are generated independently at random, we get for a single index (w,v) where $G(w,v) \neq G_1$ the probability that $t(v) > mp\gamma^2(1-\epsilon')$ is at most $2^{-(\lambda_0 m - 2n\log n)}$ where $\lambda_0 = 2\frac{s^2\sigma^2}{(s+2)^2}$ and $s = \min(1, \frac{1-\epsilon'-p}{(1-\epsilon'+p)(1-(p\gamma)^2)})$. Then the probability that no such graphs $G(u,w)$ get $t(v) > mp\gamma^2(1-\epsilon')$ is

$$(1-2^{-(\lambda_0 m - 2n\log n)})^{2^{m(h(p\gamma)(1+\delta)-R))}}$$

$$\simeq \exp(-2^{-n^{5/3}+2n\log n}) \to 1 \quad \text{as} \quad n \to \infty.$$

So the probability that there is an index (w,v) with $t(v) > mp\gamma^2(1-\epsilon)$ is $o(1)$.

For the third and last error, note that by definition of π^*, $|G_1^{\pi^*} \cap G_2^\pi| \geq |G_1^\pi \cap G_2^\pi|$. We will show that $|G_1^\pi \cap G_2^\pi| \geq mp\gamma^2(1-\epsilon')$ with high probability, hence $|G_1^{\pi^*} \cap G_2^\pi| \geq mp\gamma^2(1-\epsilon')$ with high probability. Note that $|G_1^\pi \cap G_2^\pi| =$

Authorized licensed use limited to: University of Pennsylvania. Downloaded on April 12,2024 at 19:39:20 UTC from IEEE Xplore. Restrictions apply.

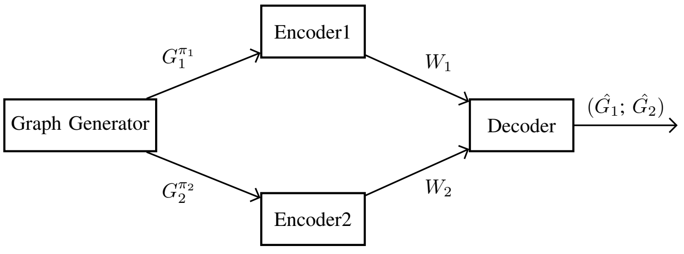

Fig. 2: Generalized Compression Scheme

$|G_1 \cap G_2|$, and by a direct application of Chernoff bound for $S_i = e_i e_i'$, where S_i's are independent by ER, and $\mathbb{E}[S_i] = p\gamma^2$, we get

$$\Pr(\sum S_i < (1-\epsilon')mp\gamma^2) < e^{\Theta(-mp\gamma^2)} = o(1).$$

Observe that $\sum S_i = |G_1^\pi \cap G_2^\pi|$, hence the probability that $t(v) < mp\gamma^2(1-\epsilon')$ is $o(1)$. Putting them all together would result in the theorem. □

V. Extensions

A. Distributed Compression of Unlabeled Graphs

In Section II-B, we presented a compression scheme, utilizing the unlabeled side information graph G_2^π at the decoder. A more general scenario can be considered in which correlated graphs G_1 and G_2 are to be compressed in a distributed manner. More precisely, encoder 1 (Alice) has $G_1^{\pi_1}$ and encoder 2 (Carol) has $G_2^{\pi_2}$ where G_1 and G_2 are generated jointly as before and the permutations π_1 and π_2 are unknown, see Fig. 2. Alice and Carol send messages W_1 and W_2 of rates R_1, and R_2, respectively, to the decoder (Bob). Bob aims to recover (G_1, G_2) losslessly, i.e.

$$\lim_{n\to\infty} \Pr((\hat{G}_1, \hat{G}_2) \neq (G_1^{\pi_1}, G_2^{\pi_2})) = 0.$$

We are interested in the minimum rates R_1, R_2 that allow reliable recovery of (G_1, G_2) up to the permutations of nodes. We will draw a proof sketch of the following theorem using the techniques we have developed so far.

Theorem 3. *There is a compression scheme with rate* $R_1 = R_2 = R = (1+\delta)(h(p\gamma) - \lambda_0/2)$, $\delta = o(1) = n^{-1/3}$, $\lambda_0 = 2\frac{s^2\sigma^2}{(s+2)^2}$, $\sigma^2 = (p\gamma)^2(1-(p\gamma)^2)$, *and* $s = \min(1, \frac{1-\epsilon'-p}{(1-\epsilon'+p)(1-(p\gamma)^2)})$, *for any constant* $1-p > \epsilon' > 0$, *such that with high probability, Bob can recover* $(G_1^{\pi_1}, G_2^{\pi_2})$.

Proof sketch. Generate a random codebook per encoder and randomly bin them into 2^{mR_1} and 2^{mR_2} bins where $m = \binom{n}{2}$. Similar to the side information problem, Alice and Carol find their corresponding graphs $G_1^{\pi_1}$ and $G_2^{\pi_2}$ in the bins and send the bin's indices W_1 and W_2 to Bob. Now Bob has to find the pair $(G_1^{\pi_1}, G_2^{\pi_2})$ where $G_1^{\pi_1}$ is in the first bin (correspondent to W_1) and $G_2^{\pi_2}$ is in the second bin (correspondent to W_2). We propose a scheme where Bob iterates over all pairs (G, H) where G is in the first bin and H is in the second bin and finds π^* that maximizes the common edges $|G \cap H^{\pi^*}|$ for each pair. Setting the threshold $(1-\epsilon')mp\gamma^2$ on the number of common edges as in Theorem 2, we look for the unique pair of graphs whose number of common edges passes the threshold $|G \cap H^{\pi^*}| \geq (1-\epsilon')mp\gamma^2$. If such a unique pair is found, Bob outputs the pair. Otherwise, error is declared.

Since the probability of two independent random graphs passing the threshold $(1-\epsilon')mp\gamma^2$ is q (see (7)), we set the bin sizes equal to s so that $(1-q)^{s^2} \to 1$. Here, we design schemes where $R_1 = R_2$, but one can parameterize based on R_1 and R_2. As $R_1 = R_2$, the bins that Carol and Alice send have the same size s. It is not hard to see that by considering the square root of the bin sizes set in Theorem 2, we again get a proper lossless compression with high probability. By reducing the bin size, the improvement in rate would be halved, meaning the compression scheme still improves the rate $R_1 = R_2$ by a constant compared with no side information.

B. Other Graph Generation Models

In Section II-B, we assumed that the input graphs (G_1, G_2) are generated by a correlated ER model. However, one can consider other models for generating (G_1, G_2) where the edges are jointly generated according to a joint pmf, an example can be Bernoulli random variables with correlation factor ρ as the edges of the graphs. As long as the generating model satisfies the conditions of Theorem 1, particularly the first condition, and there is a proper gap between the average of the cost function (5) between the correlated and independent pairs of graphs, one can potentially use the same ideas and scheme to improve the rate. The threshold can be set in the gap between the two averages and using concentration tools and Theorem 1 Bob can distinguish G_1 from the other graphs in the bin.

VI. Conclusion and Future Work

Inspired by the real-world application of compressing unlabeled graphs in which users' identities are concealed, we introduced the novel problem of graph compression with unlabeled side information at the decoder. Previous compression techniques that utilize the statistical knowledge of side information fall short when graphs are unlabeled because joint typicality is rendered inapplicable as soon as the labels are removed. For the case where the original graph and the side information graph are generated from a correlated Erdös Rènyi model, we provided novel results. In particular, by incorporating ideas from source coding with side information, graph matching, and results from classic combinatorial optimization problems, we proposed novel compression schemes that benefit from the existence of unlabeled graph side information at the decoder and we further derived upper bounds on the optimal rate of compression. We discussed various extensions including the problem of distributed source coding of unlabeled graphs.

Several questions remain open for further investigation. First and foremost, it is unclear whether the rates that are achieved in this work are near-optimal. Another interesting line of work is to seek compression algorithms of moderate complexity that can benefit from unlabeled side information.

Authorized licensed use limited to: University of Pennsylvania. Downloaded on April 12,2024 at 19:39:20 UTC from IEEE Xplore. Restrictions apply.

References

[1] 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.

[2] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *2008 IEEE Symposium on Security and Privacy (sp 2008)*. IEEE, 2008, pp. 111–125.

[3] ——, "De-anonymizing social networks," in *2009 30th IEEE symposium on security and privacy*. IEEE, 2009, pp. 173–187.

[4] E. Kazemi, H. Hassani, M. Grossglauser, and H. Pezeshgi Modarres, "Proper: global protein interaction network alignment through percolation matching," *BMC bioinformatics*, vol. 17, no. 1, pp. 1–16, 2016.

[5] C. E. Shannon, "A mathematical theory of communication," *The Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[6] P. Delgosha and V. Anantharam, "A universal low complexity compression algorithm for sparse marked graphs," in *2020 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2020, pp. 2349–2354.

[7] A. Bhatt, Z. Wang, C. Wang, and L. Wang, "Universal graph compression: Stochastic block models," in *2021 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2021, pp. 3038–3043.

[8] P. Delgosha and V. Anantharam, "Distributed compression of graphical data," in *2018 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2018, pp. 2216–2220.

[9] L. Wang and O. Shayevitz, "Graph information ratio," *SIAM Journal on Discrete Mathematics*, vol. 31, no. 4, pp. 2703–2734, 2017.

[10] R. Bustin and O. Shayevitz, "On lossy compression of directed graphs," *IEEE Transactions on Information Theory*, vol. 68, no. 4, pp. 2101–2122, 2021.

[11] ——, "On lossy compression of binary matrices," in *2017 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2017, pp. 1573–1577.

[12] D. Slepian and J. Wolf, "Noiseless coding of correlated information sources," *IEEE Transactions on information Theory*, vol. 19, no. 4, pp. 471–480, 1973.

[13] A. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Transactions on information Theory*, vol. 22, no. 1, pp. 1–10, 1976.

[14] Y. Oohama, "Gaussian multiterminal source coding," *IEEE Transactions on Information Theory*, vol. 43, no. 6, pp. 1912–1923, 1997.

[15] S.-Y. Tung, *MULTITERMINAL SOURCE CODING*. Cornell University, 1978.

[16] Y. Oohama, "Rate-distortion theory for gaussian multiterminal source coding systems with several side informations at the decoder," *IEEE Transactions on Information Theory*, vol. 51, no. 7, pp. 2577–2593, 2005.

[17] A. B. Wagner, S. Tavildar, and P. Viswanath, "Rate region of the quadratic gaussian two-encoder source-coding problem," *IEEE Transactions on Information Theory*, vol. 54, no. 5, pp. 1938–1961, 2008.

[18] C. Mao, Y. Wu, J. Xu, and S. H. Yu, "Testing network correlation efficiently via counting trees," *arXiv preprint arXiv:2110.11816*, 2021.

[19] B. Barak, C.-N. Chou, Z. Lei, T. Schramm, and Y. Sheng, "(nearly) efficient algorithms for the graph matching problem on correlated random graphs," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[20] J. T. Vogelstein, J. M. Conroy, L. J. Podrazik, S. G. Kratzer, E. T. Harley, D. E. Fishkind, R. J. Vogelstein, and C. E. Priebe, "Large (brain) graph matching via fast approximate quadratic programming," *arXiv preprint arXiv:1112.5507*, 2011.

[21] E. L. Lawler, "The quadratic assignment problem," *Management science*, vol. 9, no. 4, pp. 586–599, 1963.

[22] C. Mao, Y. Wu, J. Xu, and S. H. Yu, "Random graph matching at otter's threshold via counting chandeliers," *arXiv preprint arXiv:2209.12313*, 2022.

[23] D. E. Fishkind, S. Adali, H. G. Patsolic, L. Meng, D. Singh, V. Lyzinski, and C. E. Priebe, "Seeded graph matching," *Pattern recognition*, vol. 87, pp. 203–215, 2019.

[24] F. Shirani, S. Garg, and E. Erkip, "Typicality matching for pairs of correlated graphs," in *2018 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2018, pp. 221–225.

[25] ——, "Seeded graph matching: Efficient algorithms and theoretical guarantees," in *2017 51st Asilomar Conference on Signals, Systems, and Computers*. IEEE, 2017, pp. 253–257.

[26] E. Kazemi, "Network alignment: Theory, algorithms, and applications," EPFL, Tech. Rep., 2016.

[27] D. Cullina and N. Kiyavash, "Improved achievability and converse bounds for erdos-renyi graph matching," *ACM SIGMETRICS performance evaluation review*, vol. 44, no. 1, pp. 63–72, 2016.

[28] L. Babai, "Graph isomorphism in quasipolynomial time," in *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, 2016, pp. 684–697.

[29] B. Luo and E. R. Hancock, "Structural graph matching using the em algorithm and singular value decomposition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 10, pp. 1120–1136, 2001.

[30] E. Kazemi, S. H. Hassani, and M. Grossglauser, "Growing a graph matching from a handful of seeds," *Proceedings of the VLDB Endowment*, vol. 8, no. 10, pp. 1010–1021, 2015.

[31] P. Pedarsani and M. Grossglauser, "On the privacy of anonymized networks," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2011, pp. 1235–1243.

[32] J. Hartmanis, "Computers and intractability: a guide to the theory of np-completeness (michael r. garey and david s. johnson)," *Siam Review*, vol. 24, no. 1, p. 90, 1982.

[33] S. Sahni and T. Gonzalez, "P-complete approximation problems," *Journal of the ACM (JACM)*, vol. 23, no. 3, pp. 555–565, 1976.

[34] R. E. Burkard and U. Fincke, "Probabilistic asymptotic properties of some combinatorial optimization problems," *Discrete Applied Mathematics*, vol. 12, no. 1, pp. 21–29, 1985.

[35] F. Shirani, S. Garg, and E. Erkip, "On the joint typicality of permutations of sequences of random variables," *arXiv preprint arXiv:2001.06962*, 2020.

Authorized licensed use limited to: University of Pennsylvania. Downloaded on April 12,2024 at 19:39:20 UTC from IEEE Xplore. Restrictions apply.