More results for regenerating codes on graphs

ADWAY PATRA and ALEXANDER BARG

***Abstract*—We study regenerating codes in heterogeneous distributed storage systems including the node repair problem in graphically constrained architectures. We show that the communication cost of repair can be decreased by downloading the amounts of data controlled by the distance of the helper to the failed node. At the same time, given the flexible choice of the repair degree, the optimal repair cost can always be attained by relying on uniform downloads. We also give a construction of codes that attain a general version of the cutset bound for heterogeneous and graphically constrained systems. The codes we construct also support data combining at intermediate nodes during repair.**

1. INTRODUCTION

Regenerating codes are designed to correct single or multiple erasures from the information obtained from the nonerased coordinates of the codeword. Using the terminology inspired by distributed storage systems, a regenerating code recovers a failed node (coordinate) by downloading data from the surviving nodes of the encoding. Suppose that the code used to protect the data has length n and that every coordinate is placed on a separate storage node. To repair a failed node, the system uses a subset of $d \leqslant n-1$ *helper nodes* that transmit information comprising some functions of their contents to be used to recover the lost data. One of the goals of the code design is to minimize the amount of data downloaded to complete the repair. Following their introduction in [4], regenerating codes have been studied in a large number of papers devoted to both constructions and impossibility bounds for the code parameters, see [11] for a recent overview of this area.

1.1. Heterogeneous and graph-constrained storage systems. Most earlier works on regenerating codes, with a few exceptions mentioned below, assume that any d of the surviving $n-1$ nodes can serve as helpers, and the choice of a particular helper set is not addressed in the code design. Existing code constructions typically also assume that downloading the same amount of data from each of the helpers minimizes the communication complexity of repair, and in many cases (e.g., for MSR codes) one can show that this is indeed true. Constructions with nonuniform download are considered when the transmission cost from different helpers to the failed node is not the same, giving rise to heterogeneous regenerating codes, studied for instance in [1], [14], [5], [2]. In [9], [10] we considered a related but different problem when the nodes of the storage system are placed on the vertices of a graph, and the repair information is downloaded along the edges of the graph. The results in [9], [10], both for MSR and non-MSR codes, rely on existing constructions of codes and are therefore based on equal contribution of every helper node (the *uniform download* assumption). While they show that it is possible to perform repair with download cost smaller than direct relaying through the intermediate nodes, sending the repair data along a path of e edges from the helper to the failed node incurs the cost proportional to e, which naturally introduces heterogeneity based on the graphical distance. It is therefore conceivable that allocating the download to the helpers depending on their distance to the failed node may further reduce the repair complexity. It is this point of view that we explore in this paper.

Another facet of repair on graphs is the possibility of replacing relaying with data processing at the intermediate nodes on the path from the helper(s) to the failed node. Exploiting this feature, [9], [10] suggested that the data may be processed at this node instead of being relayed, which in many cases results in reduced communication complexity. This general procedure, termed *intermediate processing* (IP), applies to linear regenerating codes, although using it for a specific code family requires further analysis of its structure.

1.2. Main results. We prove that optimizing the communication complexity of repair in heterogeneous systems, including graph-constrained storage architectures, involves finding the optimal number of helper nodes based on their distance to the failed node. We further show that, for a given number d of helpers, it is possible to decrease the complexity by downloading different amounts of data that depend on the distance from the helpers to the failed node. At the same time, there always exists a choice of d for which the uniform download optimizes the communication complexity of repair.

We derive an extension of the cutset bound of [4] for nonuniform downloads that generalizes results shown previously in special cases. We propose a simple stacking construction of codes that supports repair attaining this bound. While proper algebraic constructions look elusive, a stacking idea, drawing on the scheme of (multilevel) concatenated codes, fits the system architecture, allowing for a universal construction that adjusts the amount of information downloaded from the helper nodes depending on the transmission cost between the helper and the failed node.

In summary, these results contribute to a reasonably complete understanding of node repair on graphs.

1.3 Regenerating codes. An $[n, k, d, l, \beta, M]$ regenerating

The authors are with Dept. of ECE and ISR, University of Maryland, College Park, MD 20742. Emails: {apatra,abarg}@umd.edu. This research is supported in part by the US NSF through grants CCF2110113 (NSF-BSF), CCF2104489, and CCF2330909.

979-8-3503-8284-6/24/$31.00 ©2024 IEEE

code over a finite field F is a subspace $\mathscr{C} \subset F^{nl}$ whose codewords are viewed as $n \times l$ matrices over F. Each column of the matrix is stored in a different storage node. It is required that the data collector have the ability to recover the original message of size M by accessing at most k nodes and downloading their stored contents. Additionally, in case any one of the nodes is erased, the l lost symbols can be recovered by contacting $d, n-1 \geqslant d \geqslant k$ surviving nodes and downloading at most β symbols from each of them (the number d of helpers is called the *repair degree*).

In this work, we consider node repair where helper nodes can contribute different amounts of data for the repair. Various special variations of this problem have been considered in the literature before, see [1], [5], [14]. However, to the best of our knowledge, a general storage vs bandwidth trade-off analysis similar to the uniform download case as well as matching code constructions have not been previously addressed.

We begin with the definition of generalized regenerating codes (GRCs).

Definition 1.1: Let $\mathcal{B} = \{\beta_i\}_{i=1}^d$ be a set of d positive integers. An $[n, k, d, l, \mathcal{B}, M]$ GRC encodes a file $\mathscr{F}$ of size M symbols over F by storing l symbols in each of the n nodes such that

1) (RECONSTRUCTION) by accessing any k out of n nodes, the original file can be recovered;
2) (REPAIR) the contents of any node $f \in [n]$ can be recovered by contacting a set $D \subseteq [n]\backslash\{f\}, |D| = d$ of nodes and downloading β_i symbols from node $\tau^{-1}(i)$ for any bijective mapping $\tau : D \to [d]$.

The mapping τ corresponds to the allocation of contributions for repair to the set of the helpers, and it highlights the fact that the assignments can be arbitrary as long they form the set $\mathcal{B}$. If all the β_i's are equal, we call such a repair scheme a *uniform download* scheme.

2. NODE REPAIR IN HETEROGENEOUS STORAGE SYSTEMS

In this section, we present general forms of the communication complexity bounds for the nonuniform download case. We begin with the cutset bound, which has previously appeared in the literature for special cases (e.g., for two different levels of download in [1]). We also present a general form of the bound for minimum communication complexity of repair for intermediate processing in graphically constrained systems.

2.1. The cutset bound. Suppose that the information stored at the nodes is described by random variables $W_i, i \in [n]$ that have some joint distribution on $(F^l)^n$ and satisfy $H(W_i) = l$ for all i, where $H(\cdot)$ is the entropy. For a subset $A \subset [n]$ we write $W_A = \{W_i, i \in A\}$. We also assume that $H(\mathscr{F}|W_K) = 0$ for any $K \subset [n], |K| = k$, which supports the data retrieval property. Denote by $D \subseteq [n]\backslash\{f\}$ the set of helper nodes. Let S_i^f be the information provided to the failed node f by the ith helper node in the traditional fully connected repair scheme, and let $S_A^f = \{S_i^f : i \in A\}$ for any $A \subseteq D$. By definition we have $H(S_i^f) = \beta_{\tau(i)}$, and

$$\begin{aligned} &H(W_K) = M, K \subset [n], |K| = k \\ &H(S_i^f|W_i) = 0, i \in D; \;\; H(W_f|S_D^f) = 0. \end{aligned} \tag{1}$$

where $\mathcal{B} = \{\beta_j\}_{j=1}^d$ and τ are introduced in Def. 1.1. Let $\Delta_r(\mathcal{B}) = \min_{R\subseteq[d],|R|=r} \sum_{i\in R} \beta_i$ denote the sum of r smallest elements from $\mathcal{B}$. The following statement gives a general bound for information transmission during repair, extending the results in [13].

Theorem 2.1: For an $[n, k, d, l, \mathcal{B}, M]$ GRC,

$$M \leqslant \sum_{i=0}^{k-1} \min\{l, \Delta_{d-i}(\mathcal{B})\}. \tag{2}$$

Proof. For any $f \in [n]$, any $D \subseteq [n]\backslash\{f\}, |D| = d$ and any set $A \subset D$, we have

$$H(W_f|S_A^f, S_{D\backslash A}^f) = 0,$$

which implies

$$\begin{aligned} H(W_f|S_A^f) &= I(W_f; S_{D\backslash A}^f|S_A^f) \leqslant H(S_{D\backslash A}^f|S_A^f) \\ &\leqslant H(S_{D\backslash A}^f) \leqslant \sum_{i\in D\backslash A} H(S_i^f) \\ &= \sum_{i\in D\backslash A} \beta_{\tau(i)}. \end{aligned}$$

Since this is true for any bijective mapping τ, we conclude that

$$H(W_f|W_A) \leqslant H(W_f|S_A^f) \leqslant \min\{l, \Delta_{d-|A|}(\mathcal{B})\}. \tag{3}$$

Finally,

$$M \leqslant H(W_{[k]}) = \sum_{i=1}^{k} H(W_i|W_{[i-1]}) \leqslant \sum_{i=0}^{k-1} \min\{l, \Delta_{d-i}(\mathcal{B})\}.$$

Remark 1: The special case of Theorem 2.1 was studied in [1] for the case when $\mathcal{B}$ contains only two distinct values. It can be easily verified that in this case, (2) recovers the main result of [1], with a much shorter proof.

As in the case of homogeneous systems [4], the *Minimum Storage* (MSR) point of the bound (2), is defined by $l = \Delta_{d-k+1}(\mathcal{B})$.

2.2. The IP repair bound. We now state the generalized version of the IP bound, extending the results of [9], [10].

Lemma 2.2: Let $f \in [n]$ be the failed node. For a (nonempty) subset of helper nodes $E \subset D$, let R_E^f be a function of S_E^f such that $H(W_f|R_E^f, S_{D\backslash E}^f) = 0$. Then if $|E| \geqslant d-k+1$,

$$H(R_E^f) \geqslant M - \sum_{i=0}^{k-2} \min\{l, \Delta_{d-i}(\mathcal{B})\}.$$

Proof: Since we assumed that $H(W_f|R_E^f, S_{D\backslash E}^f) = 0$, all the more it is true that

$$H(W_f|R_E^f, W_{D\backslash E}) = 0. \tag{4}$$

We have $|D\backslash E| \leqslant k-1$. Consider a set $A \subset E$ with $|A| = k-1-|D\backslash E|$. Now,

$$H(R_E^f, W_{D\backslash E}, W_A) = H(R_E^f, W_{D\backslash E}, W_f, W_A) \geqslant M, \tag{5}$$

where the equality in (5) follows from (4) and the chain rule, and the inequality follows from the reconstruction property because $|D\backslash E| + |A| + 1 = k$. Next, observe that

$$H(R_E^f, W_{D\backslash E}, W_A) \leqslant H(R_E^f) + H(W_{D\backslash E}, W_A),$$

and so

$$H(R_E^f) \geqslant M - H(W_{D\setminus E}, W_A)$$
$$\geqslant M - \sum_{i=0}^{k-2} \min\{l, \Delta_{d-i}(\mathcal{B})\},$$

where the last inequality follows from (3). ∎

Corollary 2.3: For MSR codes, we have

$$H(R_E^f) \geqslant l = \Delta_{d-k+1}(\mathcal{B}). \quad (6)$$

Proof: At the MSR point, we have $l = \Delta_{d-k+1}(\mathcal{B})$ and $M = kl$, cf. (2). ∎

This lemma bounds below the amount of data necessarily obtained from a subset $E \subset D$ irrespective of the processing performed by the nodes in E, including the IP repair on graphs.

Note that for fixed M, the bound above does not give any improvement in terms of communication complexity over the uniform β case at the MSR point, since in both cases it is equal to the per node storage parameter l.

2.3. A stacking code construction. In this section, we present a construction of $[n, k, d, l, \mathcal{B}, M]$ GRCs. This construction generalizes the construction of [5], and is reminiscent of multilevel concatenated codes of [3]. We prove that for any set $\mathcal{B} = \{\beta_j\}_{j=1}^d$ the constructed code family is optimal in two respects, namely (i) it saturates the MSR bound (2); (ii) it is optimal for IP repair, meeting bound (6) with equality.

Given a repair degree d and a set of $\mathcal{B} = \{\beta_j\}_{j=1}^d$ integers, we aim to construct a regenerating code that repairs any failed node f by downloading at most β_j symbols from node $\tau^{-1}(j)$ for any subset of helper nodes $D \subseteq [n]\setminus\{f\}$ and any permutation $\tau : D \to [d]$. Without loss of generality, we assume that the set $\{\beta_j\}$ is sorted in nondecreasing order. Let $\mu = (\mu_1, \ldots, \mu_{d-k+1})$ be the binary vector with $\mu_j = \mathbb{1}_{(\beta_j > \beta_{j-1})}$, where $\beta_0 := 0$. Let $\mathrm{supp}(\mu)$ be the set of indices j with $\mu_j = 1$.

Construction 2.1: Suppose that $\mathcal{B} = \{\beta_j\}_{j=1}^d$ and μ are as described above, and let $S := \{j : \mu_j = 1\}$. For each $j \in S$ take an MSR code $\mathscr{C}_j$ with parameters
$[n, k, d-j+1, l_j = (d-j-k+2)(\beta_j - \beta_{j-1}), (\beta_j - \beta_{j-1}), M_j = kl_j]$.
The $[n, k, d, l, \mathcal{B}, M]$ GRC code is formed by stacking the codes $\{\mathscr{C}_j\}_{j\in S}$, where $l = \sum_{j\in S} l_j$ and $M = \sum_{j\in S} M_j$.

The intuition behind the construction is as follows. Upon arranging the β_js in nondecreasing order, for every j such that $\beta_j > \beta_{j-1}$, we add to the stack an MSR code with per node download equal to the gap $\beta_j - \beta_{j-1}$ and repair degree $d - j + 1$.

Theorem 2.4: The code in Construction 2.1 is an $[n, k, d]$ regenerating code that supports the repair of any node f from a helper set $D \subseteq [n]\setminus\{f\}, |D| = d$ by downloading at most β_j symbols of F from node $\tau^{-1}(j)$ for any bijection $\tau : D \to [d]$.

Proof: Observe that the length n and the data reconstruction parameter k are the same for all the component codes and hence inherited directly. To prove the repair property, fix a helper set D and a permutation τ such that $\{\beta_j\}$ are in nondecreasing order. Node $\tau^{-1}(j)$ participates in the recovery of node f only in the component codes $\{\mathscr{C}_p : p \leqslant j\}$ and hence it sends a total of $\sum_{p=1}^j (\beta_p - \beta_{p-1}) = \beta_j$ symbols. ∎

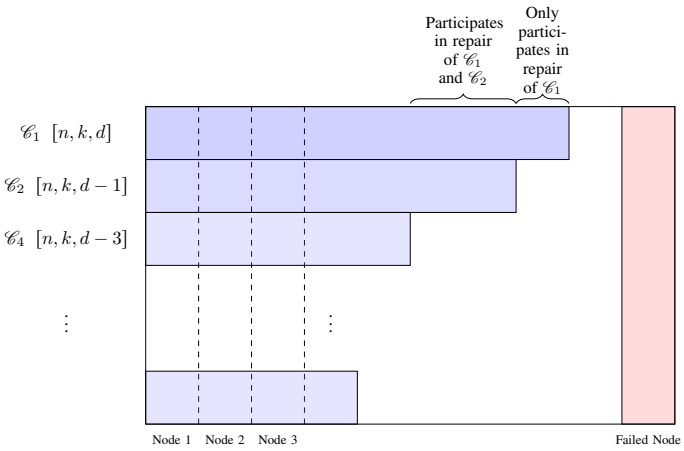

Fig. 1. Stacked MSR construction for $S = \{1, 2, 4, \cdots\}$ (an example)

The component codes can be chosen from a variety of known MSR constructions. Since the parameters of these outer codes depend on the given set $\{\beta_j\}$, a convenient choice is the Product-Matrix codes [12], which in their basic version work by downloading a single symbol from every helper. By stacking several codewords of these codes we can obtain any of the component codes $\mathscr{C}_j$ as in Fig. 1, thereby matching any set of per-node download values as required by the construction.

Next, we show that this construction attains the minimum possible node size for the given parameters.

Proposition 2.5: Codes of Construction 2.1 meet the bound (2) with equality at the MSR point.

Proof: The node size for the code of this construction equals

$$l = \sum_{j=1}^{d-k+1} (d-j-k+2)(\beta_j - \beta_{j-1})$$
$$= \sum_{i=1}^{d-k+1} i(\beta_{d-i-k+2} - \beta_{d-i-k+1}) = \sum_{i=1}^{d-k+1} \beta_i.$$

Since $\beta_j \geqslant \beta_{j-1}$ for all j, summing the first d of the β_js gives the minimum value over all mappings τ. Thus, the sum on the last line equals $\Delta_{d-k+1}(\mathcal{B})$, matching (2). ∎

Moreover, since the codes in Construction 2.1 are formed of F-linear MSR codes, they are themselves F-linear and therefore support IP repair. We will show that they minimize the amount of data sent by any subset of $d - k + 1$ nodes.

Proposition 2.6: Codes of Construction 2.1 meet the lower bound (6) with equality.

Proof: For a given j and a code $\mathscr{C}_j$ it is possible to perform IP repair. Specifically, any subset of at least $d-j-k+2$ nodes can perform intermediate processing for $\mathscr{C}_j$ to compress their repair data to l_j symbols of F. Therefore overall, the

subset nodes of size $d-k+1$ or more can perform IP repair, compressing their data to $\sum_{j\in S} l_j = l$ symbols. ∎

3. Repair on Graphs with Nonuniform Contributions

In this section, we highlight the advantages of the above code construction by specializing it to repair on graphs. The basic problem addresses the communication complexity of node repair under the assumption that communication between the nodes is constrained by a (connected) graph $G(V, E)$ where V is a set of n distinct vertices and the cost of sending a unit of information from any node i to any node j is determined by the graph distance $\rho(i, j)$ in G. The nontrivial situation arises when the helper nodes contacted during the repair process of a failed node are not in the immediate neighborhood of that node. In [9], it was shown that it is possible to do better than simple relaying of helper data in case of such multi-layered repair under certain conditions. Here we address the general situation of possibly unequal contributions of the nodes, extending the earlier uniform download analysis.

For a failed node $f \in [n]$, let D be a set of d helper nodes closest to it in terms of the graph distance. Let $G_{f,D}$ be the subgraph spanned by $\{f\} \cup D$ in G and let $T_{f,D}$ be a spanning tree of this subgraph with f as the root. Let $t = \max_{h\in D} \rho(f, h)$ be the height of the tree. Since there can be multiple possible choices of $G_{f,D}$ and $T_{f,D}$, to make the analysis general, we assume certain regularity in the underlying graph G. More precisely, we assume that for every node $f \in [n]$, there exists a spanning tree $\mathcal{T}$ of $G_{f,D}$ for some choice of D.

Example 1: As an example, suppose $G(V, E)$ is a connected t-regular graph. One way to guarantee the existence of the tree $\mathcal{T}$ is to consider graphs with girth g, in which case a ball of radius $\lfloor g/2 \rfloor - 1$ around any vertex is a tree with t immediate neighbors of the center and $t(t-1)^{i-1}$ vertices in layer i. A line of work starting with Margulis's paper [7] yielded constructions of such graph families with $g \geqslant C(n, t) \log_{t-1} n$, where n is the number of vertices and $C(n, t)$ is a constant.

Suppose that $\mathcal{T}$ contains d_i helper nodes at depth i from the root with $\sum_i d_i = d$. With this assumption, the repair procedure for MSR codes that optimizes the overall communication complexity of repair with uniform download was found in [9]. It involves transporting the helper data towards the failed node, i.e., the root of the tree, along the edges of the tree, whereby nodes having more than $d-k+1$ children process the information and send l symbols relying on the IP technique. Let J be the set of nodes in $\mathcal{T}$ with at least $d-k+1$ children and let J_i be the set of nodes in J at distance i from the root, so that $J = \cup_i J_i$. For an $i \notin J$, let $\mathcal{P}(i)$ denote the nearest parent of node i in J, and if no such parent exists, then let $\mathcal{P}(i) = f$. Define an $[n, k, l, d, \beta, M]$ MSR code on the graph, then the total communication complexity of the repair process is:

$$\Lambda_{\mathrm{U}}^{\mathcal{T}} = \sum_{i\in J\setminus\{f\}} l + \sum_{i\in D\setminus J} \rho(i, \mathcal{P}(i))\beta. \tag{7}$$

We begin with the following observation.

Claim: The set J does not change when we switch from the uniform download model to the nonuniform one and vice versa. Furthermore, every node in J keeps transmitting l symbols by relying on the IP procedure.

Indeed, if a node has $d-k+1$ or more children in the tree, they jointly must transmit at least l symbols for repair because of the bound (6), irrespective of whether the β_i's are equal or different. Lemma 2.6 further implies that this bound is achievable by the stacking construction.

Assume now that nodes in layer i each contribute β_i symbols for repair with $\beta_1 \geqslant \beta_2 \geqslant \cdots \geqslant \beta_t$. This can be accomplished by using an $[n, k, d, l, \mathcal{B}, M]$ code from Construction 2.1 with the set $\mathcal{B}$ formed of β_i's, each appearing d_i times, for all $i \in [t]$. Let $\delta_i = \beta - \beta_i$. Furthermore, let t' be the largest number such that $\sum_{i=t}^{t'} d_i \geqslant d-k+1$. It can be checked that the vector μ of Construction 2.1 in this case is given by

$$\mu = e_1 + \sum_{i=t'}^{t} e_{x_i},$$

where $x_i = \sum_{j=i}^{t} d_j + 1$ and $e_x \in \{0, 1\}^{d-k+1}$ is a vector with a single 1 in position x (this is the indicator of the set of growth points in the sequence of the smallest $d-k+1$ entries in the set $\mathcal{B}$.).

The total communication complexity in the nonuniform contribution model is given by the following expression:

$$\Lambda_{\mathrm{NU}}^{\mathcal{T}} = \sum_{i\in J\setminus\{f\}} l + \sum_{i=1}^{t} \sum_{j\in D_i\setminus J_i} \rho(j, \mathcal{P}(j))\beta_i. \tag{8}$$

Theorem 3.1: Using Construction 2.1 for the repair tree $\mathcal{T}$, the nonuniform contribution scheme achieves the overall repair bandwidth given in Eq. (8). It attains savings over the uniform contribution model whenever

$$\sum_{i=1}^{t} \sum_{j\in D_i\setminus J_i} \rho(j, \mathcal{P}(j))\delta_i > 0 \tag{9}$$

subject to

$$\sum_{i=t}^{t'+1} d_i\delta_i + \big(d-k+1-\sum_{i=t}^{t'+1} d_i\big)\delta_{t'} = 0. \tag{10}$$

Proof: Condition (9) follows by comparing expressions (8) and (7). To obtain (10), recall our notation Δ_{d-k+1} defined before Theorem 2.1. For the graph case considered, it has the following form:

$$\Delta_{d-k+1} = \sum_{i=t'+1}^{t} d_i\beta_i + \big(d-k+1-\sum_{i=t'+1}^{t} d_i\big)\beta_{t'} = l,$$

where the last equality follows from Theorem 2.1. Rewriting this using the δ_i's, we obtain Eq. (10). ∎

We illustrate this theorem in Example 2 below. Note that the set J may not include all the nodes capable of performing IP. Indeed, for a choice of $\mathcal{B} = \{\beta_i\}$, any node in the repair tree that accumulates the repair data of a set A such that $\sum_{i\in A} \beta_i \geqslant l$ can gainfully perform IP. Hence, the minimum communication complexity of repair can potentially be even lower than Eq. (8).

4. Optimizing the helper data and the repair degree

The above analysis suggests that for the case of repair on graphs, lowering the contribution of the farthest away nodes at the expense of increasing the contributions from the nearer nodes may reduce the amount of communication. This gives rise to the question of the limits of this exchange. In the limiting case, one might stop accessing data from the farthest nodes altogether, effectively decreasing the repair degree d of the repair process. The universal constructions of regenerating codes proposed in [15], [6] support the option of dynamically adjusting the repair degree d. In this section, we show that this added flexibility indeed minimizes the overall communication bandwidth.

The optimal choice of β_i's can be found by formulating the graph repair problem as an optimization problem, as suggested in [5] for heterogeneous systems. For this, without loss of generality, we take n to be the failed node and set $D = [n-1]$. Assume that node $i \in D$ contributes β_i symbols for the repair of f, adding that now some of the β_i's can be 0. By Theorem 2.1, we have that $l = \Delta_{n-k}(\mathcal{B})$, which imposes constraints on our choice of β_i's. The objective function of our minimization problem is given by the expression in Eq. (8). Note that letting some β_i's to be 0 does not change the set J, since due to the constraint $l = \Delta_{n-k}(\mathcal{B})$, each node in J can still perform IP. Since the first term in Eq. (8) is independent of the choice of β_i's, the optimization problem can be stated in the following simple form:

$$\begin{aligned} &\min \sum_{i\in D\setminus J} b_i\beta_i \\ &\text{subject to} \sum_{i\in A}\beta_i \geqslant l, \quad \forall A \subseteq [n-1], |A| = n-k \\ &\qquad 0 \leqslant \beta_i \leqslant l, \quad i \in [n-1], \end{aligned} \tag{11}$$

where the costs b_i can be calculated from the structure of $\mathcal{T}$. Without loss of generality, we shall assume that the costs b_i's are arranged in non-increasing order, i.e., $b_1 \geqslant b_2 \geqslant \cdots \geqslant b_{n-1}$. With these assumptions, we restate Lemma 1 from [5] for our purposes.

Lemma 4.1: If $\{\beta_i^*\}$ is an optimal solution for the above optimization problem, then

1) $\beta_1^* \leqslant \beta_2^* \leqslant \cdots \leqslant \beta_{n-1}^*$, and
2) $\beta_j^* = \beta_{n-k}^*$ for all $i > n-k$.

The authors in [5] further claimed that the optimal solution of the above optimization problem takes the form given in the next theorem. The proof does not seem to appear in the published literature, so we have included it in the preprint version [8].

Theorem 4.2: ([5], Theorem 1]) There exists an optimal solution of the above LP such that

$$\beta_i^* = \begin{cases} 0 & 1 \leqslant i \leqslant n-d-1 \\ \frac{l}{d-k+1} & n-d \leqslant i \leqslant n-1 \end{cases} \tag{12}$$

for some d in the range $k \leqslant d \leqslant n-1$.

Example 2: We give an example to show that for a given repair degree d, the nonuniform assignment yields communication savings guaranteed by Theorem 3.1, and that at the same time, the maximum savings can be attained by adjusting the repair degree and switching to the uniform assignment.

Consider the t-regular Cayley graphs mentioned in Example 1. Suppose that the repair tree $\mathcal{T}$ is formed of a layers, where $a < \lfloor g/2 \rfloor - 1$, then $d_i = t(t-1)^{i-1}, i \leqslant a-1$ and $d_a = d - \sum_{i=1}^{a-1} t(t-1)^{i-1}$. Suppose further that $d_a + d_{a-1} \geqslant d-k+1$. To simplify the analysis, we are not including IP since it is somewhat independent of the current discussion and can be easily incorporated into it. The overall repair bandwidth for a uniform contribution repair scheme for an $[n,k,d,l,\beta = \frac{l}{d-k+1}, M]$ MSR code is $\Lambda_{\mathrm{U}}^{\mathcal{T}} = \beta \sum_{i=1}^{a} i d_i$. Now let us switch to the nonuniform contribution repair scheme with helper nodes at layer i contributing β_i symbols each, with β_i's nonincreasing. From Theorem 3.1, we have that $d_a\delta_a + (d-k+1-d_a)\delta_{a-1} = 0$, with $\delta_i = \beta - \beta_i$, and repair bandwidth under this scheme is $\Lambda_{\mathrm{NU}}^{\mathcal{T}} = \sum_{i=1}^{a} i d_i \beta_i$. Note that if $\delta_a > 0$ then $\delta_{a-1} < 0$ and $\delta_i \leqslant \delta_{a-1}$ for all $i \leqslant a-2$, so we let $\delta_i = -\frac{d_a}{d-k+1-d_a}\delta_a$ for all $i \leqslant a-1$ and observe that the savings in the nonuniform setting are

$$\begin{aligned} \Lambda_{\mathrm{U}}^{\mathcal{T}} - \Lambda_{\mathrm{NU}}^{\mathcal{T}} &= \sum_{i=1}^{a} i d_i \delta_i \\ &= \frac{d_a\delta_a}{d-k+1-d_a}\Big(\sum_{i=1}^{a-1}(a-i)d_i - a(k-1)\Big). \end{aligned}$$

In summary, using the nonuniform scheme results in savings whenever the expression in the parentheses is positive, which is possible for small k.

Now suppose that we have the freedom of choosing the repair degree d. Observe that as we increase $\delta_a = \beta - \beta_a$ above, the savings in overall bandwidth increase with the maximum attained when $\delta_a = \beta$ or $\beta_a = 0$. At this point,

$$\begin{aligned} \beta_i = \beta - \delta_{a-1} &= \beta + \frac{d_a\delta_a}{d-k+1-d_a} \\ &= \frac{l}{d-d_a-k+1} = \frac{l}{d'-k+1}, 1 \leqslant i \leqslant a-1, \end{aligned}$$

where $d' = d - d_a$ is the new repair degree, since the nodes in the ath layer are effectively excluded, and every helper node in layers $a-1$ and below contributes equally.

5. Conclusion

In this paper, we have analyzed the repair problem with the assumption that different helper nodes may contribute differently towards the repair process. We established performance bounds for such a nonuniform contribution model and constructed a matching family of regenerating codes. In some cases, node repair with nonuniform contributions results in smaller communication complexity than the previously analyzed uniform model [9]. At the same time, if the repair degree can be dynamically adjusted (or if the graph is sufficiently regular), then the added complexity of adopting a nonuniform contribution model can be avoided. An interesting open question is to design a code construction for the nonuniform contribution model without using the stacking method which suffers from a large per-node storage.

REFERENCES

[1] S. Akhlaghi, A. Kiani, and M. R. Ghanavati, "Cost-bandwidth tradeoff in distributed storage systems," *Computer Communications*, vol. 33, no. 17, pp. 2105–2115, 2010.

[2] K. Benerjee and M. Gupta, "Tradeoff for heterogeneous distributed storage systems between storage and repair cost," *Problems of Information Transmission*, vol. 57, 03 2015.

[3] È. L. Blokh and V. V. Zyablov, "Coding of generalized concatenated codes," *Problemy Peredachi Informatsii*, vol. 10, no. 3, pp. 45–50, 1974.

[4] A. G. Dimakis, P. B. Godfrey, Y. Wu, M. J. Wainwright, and K. Ramchandran, "Network coding for distributed storage systems," *IEEE Trans. Inf. Theory*, vol. 56, no. 9, pp. 4539–4551, 2010.

[5] Z. Li, W. H. Mow, L. Deng, and T.-Y. Wu, "Optimal-repair-cost MDS array codes for a class of heterogeneous distributed storage systems," in *2022 IEEE International Symposium on Information Theory (ISIT)*, 2022, pp. 2379–2384.

[6] Y. Liu, J. Li, and X. H. Tang, "Optimal repair/access MDS array codes with multiple repair degrees," 2022, eprint arXiv:2205.13446.

[7] G. A. Margulis, "Explicit constructions of graphs without short cycles and low density codes," *Combinatorica*, vol. 2, no. 1, pp. 71–78, 1982.

[8] A. Patra and A. Barg, "Generalized regenerating codes and node repair on graphs," eprint, arXiv, May 2024.

[9] ——, "Node repair on connected graphs," *IEEE Transactions on Information Theory*, vol. 68, no. 5, pp. 3081–3095, 2022.

[10] ——, "Node repair on connected graphs, Part II," 2022, eprint arXiv:2211.00797.

[11] V. Ramkumar, S. B. Balaji, B. Sasidharan, M. Vajha, M. N. Krishnan, and P. V. Kumar, "Codes for distributed storage," *Foundations and Trends in Communications and Information Theory*, vol. 19, no. 4, pp. 547–813, 2022.

[12] K. V. Rashmi, N. B. Shah, and P. V. Kumar, "Optimal exact-regenerating codes for distributed storage at the MSR and MBR points via a product-matrix construction," *IEEE Trans. Inf. Theory*, vol. 57, no. 8, pp. 5227–5239, 2011.

[13] N. B. Shah, K. V. Rashmi, P. V. Kumar, and K. Ramchandran, "Distributed storage codes with repair-by-transfer and nonachievability of interior points on the storage-bandwidth tradeoff," *IEEE Trans. Inf. Theory*, vol. 58, no. 3, pp. 1837–1852, 2012.

[14] J. Wang, Y. Luo, and K. W. Shum, "Storage and repair bandwidth tradeoff for heterogeneous cluster distributed storage systems," *Science China Information Sciences*, vol. 63, pp. 1–15, 2020.

[15] M. Ye and A. Barg, "Explicit constructions of optimal-access MDS codes with nearly optimal sub-packetization," *IEEE Trans. Inf. Theory*, vol. 63, no. 10, pp. 6307–6317, 2017.