Regenerating codes on graphs

Adway Patra Alexander Barg

***Abstract*—We estimate the communication complexity of node repair for regenerating codes defined on graphs. Both deterministic and random graphs are considered.**

2021 IEEE International Symposium on Information Theory (ISIT) | 978-1-5386-8209-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/ISIT45174.2021.9517983

I. Introduction

Applications of erasure-correcting codes in distributed storage are focused on recovering a single erasure under the constraint on the total amount of data "moved" from the other coordinates to correct the erased (failed) coordinate. This processing is commonly modeled by assuming that the codeword coordinates are placed on different servers (storage nodes), and aim at limiting the information communicated between them for the recovery of the failed node. In this paper we additionally assume that communication between the nodes is constrained by a (connected) graph $G(V, E)$, where V is an n-set of vertices and where the cost of sending a unit of information from v_i to v_j is determined by the graph distance $\rho(v_i, v_j)$ in G. While it is still possible to use the known methods of node repair, a natural question to study is whether there are more economical ways of accomplishing this goal given the structure of the graph G. We give an affirmative answer by showing that, if the data is encoded using a minimum storage regenerating (or MSR) code, then under some conditions it is possible to save on the communication cost of node repair compared to simple relaying of the information. We also address the same question for a random graph G from the standard Erdös-Rényi ensemble $\mathcal{G}_{n,p}$ and determine a range of parameters under which the communication cost of repair with intermediate processing is advantageous over the repair scheme based on the relaying.

For a finite field $F = \mathbb{F}_q$ we consider a code $\mathcal{C} \subset F^{nl}$ whose codewords are represented by $l \times n$ matrices over F. We assume that each coordinate (a vector in F^l) is written on a single storage node, and that a failed node amounts to having its coordinate erased. The task of node repair can be thought of as correcting a single erasure in the vector code of length n over F. Throughout this work we assume that the code is F-linear and that any k coordinates suffice to recover the codeword. Suppose moreover that for any codeword $(C_1, C_2, \dots, C_n) \in \mathcal{C}$ and any $i \in [n]$ the coordinate (node) C_i is a function of a subset of d *helper* nodes $\{C_{i_j}, j = 1, \dots, d\}$, where $d, k \le d \le n-1$ is some number, and that each of the helper nodes provides $l/(d-k+1)$ symbols for the recovery of C_i. A number of families of MSR codes with the described properties are known in the literature [4], [6]–[8], [10]–[12]. Below in our examples we consider product-matrix codes of [7] and MSR codes based on diagonal matrices [11], and we also point out that the proposed distributed repair procedure applies to any F-linear MSR code.

The authors are with Dept. of ECE and ISR, University of Maryland, College Park, MD 20742. Emails: {apatra,abarg}@umd.edu. A. Barg is also with IITP, Russian Academy of Sciences, 127051 Moscow, Russia.

This research was supported by NSF grant CCF1814487.

II. Node repair on graphs

Let $\mathcal{C}$ be an (n, k, d, l) MSR code and suppose that each coordinate of a codeword $C \in \mathcal{C}$ is written on a vertex of a graph $G(V, E), |V| = n$. Suppose further that the coordinate $C_f, f \in [n]$ is erased, or, as we will say, that the node v_f has failed. Let $D \subset V \backslash \{v_f\}, |D| = d$ be the set of helper nodes. To repair the failed node, the helper nodes provide information which is communicated to v_f over the edges in E. If one discounts the connectivity constraints, then to accomplish the repair, each of the helper nodes sends the information to the failed node over the shortest path in G, and the intermediate nodes simply relay this information further, possibly supplementing it with their own data. We call this repair strategy *accumulate and forward* (AF). A potentially more economical repair arises when the intermediate nodes are allowed to process the information.

Lower bounds on the repair bandwidth: Before proceeding, let us further specify our assumptions. In our analysis we focus only on the node repair problem and do not study the communication complexity of the "data collection" task [3]. We assume that for the failed node v_f, the helper nodes D are chosen to be the d *closest* (in terms of the graph distance) nodes from v_f. These nodes can be found by a simple breadth-first search on G starting at v_f. Denote by $G_{f,D} = (V_{f,D}, E_{f,D})$ the subgraph spanned by $\{v_f\} \cup D$. Let $t = \max_{v \in D} \rho(v, v_f)$. We will use the following notation: $\Gamma_j(v_f) = \{v \in V_{f,D} : \rho(v, v_f) = j\}, N_i(v_f) = \cup_{j=1}^{i} \Gamma_j(v_f)$. The subset $\Gamma_j(v_f)$ is formed of the helper nodes at distance j from v_f (the helper nodes in *layer* j). The case $t = 1$ corresponds to the much-studied graph-agnostic repair scenario [3], and therefore we exclude it from consideration. Observe that the graph $G_{f,D}$ is not necessarily unique; in particular, there may be multiple possible choices for the helper nodes in the t-th layer.

In the next lemma, we derive lower bounds on the amount of information contributed by a group of helper nodes for the purposes of repair. The lemma is phrased in information-theoretic terms. We assume that the information stored at the vertices is given 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 $A \subset V$ we write $W_A = \{W_i, i \in A\}$. Let S_i^f be the information provided to v_f by the ith helper node in the traditional fully connected repair setting and let $S_D^f = \{S_i^f, i \in D\}$. The RV S_i^f is a function of the contents of the node v_i, or formally, $H(S_i^f | W_i) = 0$, and the RVs $S_i^f, i \in D$ determine the contents of v_f, i.e.,

$H(W_f|S_D^f) = 0$. From the cut-set bound [3], it follows that $H(S_i^f) \geq l/(d-k+1)$ and we assume that this is achieved with equality, i.e., the codes we use have the MSR property. The next lemma is a simple consequence of the definition of MSR codes. The proofs are close to the arguments that have previously appeared in the literature, see for instance [9].

Lemma II.1. *Let $v_f, f \in [n]$ be the failed node. For a subset of the helper nodes $E \subset D$ let R_E^f be a function of S_E^f such that*

$$H(W_f|R_E^f, S_{D\setminus E}^f) = 0. \tag{1}$$

1) If $|E| \geq d-k+1$, then

$$H(R_E^f) \geq l.$$

2) If $|E| \leq d-k$, then

$$H(R_E^f) \geq \frac{|E|l}{d-k+1}.$$

Proof. 1) By the assumption (1), given the contents of all the nodes in $D\setminus E$, the information contained in R_E^f is sufficient to repair v_f, i.e.,

$$H(W_f|R_E^f, W_{D\setminus E}) = 0. \tag{2}$$

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

$$H(R_E^f, W_{D\setminus E}, W_A) = H(R_E^f, W_{D\setminus E}, W_f, W_A) \geq kl \tag{3}$$

where the equality in (3) follows from (2) and the chain rule, and the inequality follows from the MDS property of MSR codes because $|D\setminus E| + |A| + 1 = k$. Next observe that

$$\begin{aligned} H(R_E^f, W_{D\setminus E}, W_A) &\leq H(R_E^f) + H(W_{D\setminus E}, W_A) \\ &= H(R_E^f) + (k-1)l \end{aligned} \tag{4}$$

where the equality again uses the independence of any $k-1$ coordinates in an MDS code. Combining (3) and (4), we obtain the claimed inequality.

The proof of Part 2) is similar and will be omitted. □

As a consequence of this lemma, we obtain a lower bound on the amount of information transmitted between the layers in $G_{f,D}$.

Proposition II.2. *Let R_j^f be the random variable denoting the information flow from the j-th layer to the $(j-1)$-th layer. Then*

$$H(R_j^f) \geq \min\left\{l, \frac{|\cup_{i=j}^t \Gamma_i(v_f)| \cdot l}{d-k+1}\right\}$$

Proof. Follows from Lemma II.1 by taking $E = \cup_{i=j}^t \Gamma_i(v_f)$.

Note that R_j^f in the above proposition represents the joint information transmitted by all the nodes in layer j to layer $j-1$ and hence does not account for any communication occurring among the nodes in the same layer. For the special case when $G_{f,D}$ is a rooted tree, we can get a more precise statement on the total required communication for repair. Let T_f be a rooted tree with root v_f, then it defines the set of descendants of each node in T_f. Let $D(v_i)$ be the set of descendants of v_i, and let $D^*(v_i) = D(v_i) \cup \{v_i\}$.

We will be interested in the communication complexity (*repair bandwidth*) of the recovery of the erased nodes under various repair algorithms. The simplest option is to use the AF repair procedure of MSR codes, described in the beginning of this section. Its repair bandwidth can be found as

$$\beta_{\text{AF}} = \Big(t(d-|N_{t-1}(v_f)|) + \sum_{i=1}^{t-1} i|\Gamma_i(v_f)|\Big)\frac{l}{d-k+1}. \tag{5}$$

The total communication complexity of node repair using the tree T_f is bounded below in the following proposition.

Proposition II.3. *Let $J_f = \{v \in V(T_f)\setminus\{v_f\} : |D^*(v)| \geq d-k+2\}$. The total communication complexity β for the repair of node v_f on the repair tree T_f is bounded as*

$$\beta \geq \sum_{v\in J_f} l + \sum_{v\in V(T_f)\setminus(\{v_f\}\cup J_f)} \frac{|D^*(v)|l}{d-k+1}. \tag{6}$$

Proof. For every non-root node $v \notin J_f$, we have $|D^*(v)| \leq d-k$. Since T_f is a tree, any outflow of information out of the subtree spanned by $D^*(v)$ passes through the node v, so it needs to transmit at least $|D^*(v)|\cdot l/(d-k+1)$ symbols to its immediate parent in T_f by Lemma II.1. Similarly, every node $v \in J_f$ needs to transmit at least l symbols to its immediate parent by virtue of Lemma II.1. □

Note that for any node $v \notin J_f$, the AF strategy is trivially optimal. At the same time, for nodes $v \in J_f$ a better communication strategy is not a priori ruled out. This problem is addressed in the next section.

A. MSR constructions for repair of graph vertices

1) Product-matrix (PM) codes: Let us briefly recall the product-matrix construction of [7]. We begin with fixing the code length n and the dimension parameter k, and take $d = 2k-2, l = k-1$. The code $\mathcal{C} : F^{k(k-1)} \to F^{ln}$ encodes $k(k-1)$ symbols of F into a codeword of length n with each coordinate formed of $k-1$ symbols. To define this mapping, form a matrix $M = [S_1 \,|\, S_2]^T$, where S_1, S_2 are symmetric matrices of order l. The number of unique symbols in M equals $2\binom{l+1}{2} = k(k-1)$. Next let $\Psi = [\Phi, \Lambda\Phi]$ be an $n \times 2l$ matrix, where Φ is a Vandermonde matrix with rows of the form $\phi_i = (1, x_i, \ldots, x_i^{l-1})$, where $x_i, i = 1, \ldots, n$ are distinct elements of F, and $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$ with $\lambda_i = x_i^l, i = 1, \ldots, n$. A codeword of the code $\mathcal{C}$, which is an $n \times l$ matrix, is found as $C = \Psi M$, and thus the contents of the node $v_i, i = 1, \ldots, n$ is given by the product

$$C_i := [\phi_i, x_i^l\phi_i]M = \phi_i S_1 + \lambda_i \phi_i S_2. \tag{7}$$

To describe the repair procedure from [7], suppose without loss of generality that the helper nodes form the set $D = \{1, \ldots, d\}$ and that the failed node's index is $f \in [n]\setminus[D]$. The original node repair (erasure correction) procedure proposed in [7] proceeds as follows. The information downloaded by the failed node v_f is given by $(\phi_i S_1 + \lambda_i\phi_i S_2)\phi_f^T$, i.e., each helper node provides one symbol of F. Thus, the failed node downloads a d-dimensional vector $y = y_{f,D}$ given by

$$y = \Psi_D M \phi_f^T = \Psi_D \begin{bmatrix} S_1\phi_f^T \\ S_2\phi_f^T \end{bmatrix}, \tag{8}$$

where Ψ_D is the submatrix of Ψ formed of the first d rows. The matrix Ψ_D is square $d \times d$ and it is invertible by construction, so we can compute the vectors $(S_1\phi_f^T)^T = \phi_f S_1$ and $(S_2\phi_f^T)^T = \phi_f S_2$. By (7) the sum $\phi_f S_1 + \lambda_f \phi_f S_2$ equals C_f, and this completes the repair process.

Now we will modify the repair procedure in a way that supports processing the information received by the nodes in the repair tree as it is passed to the failed node v_f. Note that by (8)

$$\phi_f M^T = y^T (\Psi_D^T)^{-1}. \tag{9}$$

Using (7), (9), the contents of the node v_f can be written as

$$C_f = \phi_f M^T \begin{bmatrix} I_l \\ \lambda_f I_l \end{bmatrix} = y^T (\Psi_D^T)^{-1} \begin{bmatrix} I_l \\ \lambda_f I_l \end{bmatrix}.$$

Introduce a $d \times l$ matrix $U := (\Psi_D^T)^{-1} \begin{bmatrix} I_l \\ \lambda_f I_l \end{bmatrix}$ and denote its rows by $U_i, i = 1, \dots, d$, then we have

$$C_f = \sum_{i=1}^{d} y_i U_i. \tag{10}$$

Note that the matrix U does not depend on the codeword, and can be precomputed. Overall this rewriting of the repair process (8) enables us to separate the contributions of the helper nodes, and offers savings in the communication cost of repair. Recalling our notation $D^*(v_i)$, suppose that, instead of transmitting the symbol y_i to its parent, the node transmits the sum $\sum_{j \in D^*(v_i)} y_j U_j$. Since we are now moving vectors rather than individual symbols along the edges of T_f, this may seem wasteful; however remember that the symbols are relayed many times, and that from some point on the repair process has to move at least l symbols along the edge by Lemma II.1. To justify the savings, suppose that $|D^*(v_i)| \geq d - k + 2 = k$, then forwarding the symbols $(y_j, j \in D^*(v_i))$ from v_i to its predecessor in T_f amounts to sending k symbols, whereas transmitting the sum $\sum_{j \in D^*(v_i)} y_j U_j$ requires $l = k - 1$ transmissions.

Therefore, the communication for repair can be summarized as follows. First, the leaf nodes in T_f send their symbols y_i one level up, then the nodes that received these symbols send them together with their symbols y_i, etc. If at any stage a node v_i has $d - k + 1$ or more descendants, then it switches to transmitting

$$\sum_{j \in D^*(v_i)} y_j U_j. \tag{11}$$

Finally if a node v_i received a vector $\sum_{j \in D(v_i)} y_j U_j$ from its immediate descendant, it adds to it the vector $y_i U_i$ and forwards it to its parent in T_f.

In summary, we have shown that, for every node $v_i \in T_f$ with $|D(v_i)| \geq d - k + 1$ descendants in T_f there exists a repair procedure under which v_i transmits exactly l symbols of F to its parent in T_f. This proves the following theorem.

Theorem II.4. *Suppose a codeword of a PM code $\mathcal{C}$ is written on the vertices of a graph G, and let T_f be the repair tree of a failed node v_f. There exists an explicit repair procedure that achieves the lower bound in (6) with equality.*

2) Examples of graphs: Let us give a few examples in which the proposed repair procedure gains in communication complexity over the AF repair. For simplicity we will assume that each helper node provides one symbol of F for the repair of v_f.

1. Suppose that the repair tree T_f is a *star* with d rays in which v_f is one of the leaves and the remaining d vertices serve as the helper nodes. Using the AF repair, each of the nonerased leaves sends its symbol to the center, which then sends d symbols to v_f, so $\beta_{\text{AF}} = 2d - 1 = 4k - 5$. At the same time, $\beta_{\text{IP}} = 3k - 4$ because the symbols of the helpers other than the center are aggregated using (11) before relaying to v_f. Another elementary example, which also shows improvement, arises when the repair tree T_f is a *path* on $d + 1$ vertices.

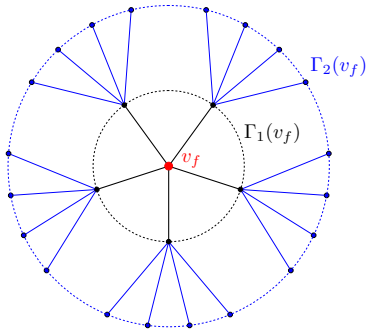
2. *Regular tree*. Suppose that G is an $(r+1)$-regular graph, and the repair tree T_f of every node is $(r+1)$-regular as shown in the figure. We need to take the depth t of the tree to satisfy $(r+1)\sum_{i=0}^{t-1} r^i \geq d$; suppose for simplicity that this holds with equality. The communication complexity of the AF repair is

$$\beta_{\text{AF}} = td - (r+1)\sum_{i=0}^{t-2}(t-i-1)r^i.$$

Suppose that $r > d - k + 1$, then from the next to last layer we can switch to uploading the linear combination of the form (11), resulting in the repair bandwidth $\beta_{\text{IP}} = d + (d-k)(r+1)\sum_{i=0}^{t-2} r^i$. The difference

$$\beta_{\text{AF}} - \beta_{\text{IP}} = (t-1)d - (r+1)\sum_{i=0}^{t-2}((d-k) + (t-i-1))r^i$$

is positive if $\frac{d-k}{d}$ is small, i.e., if $d \geq k$ is close to k.



3. *Galton-Watson tree.* Having in mind a scenario in which the helper nodes are chosen randomly and independently by the nodes already included in the repair tree T_f, suppose that it is constructed following a branching process with the root v_f, resulting in a Galton-Watson ensemble of random trees $\mathcal{T}_f$. In this example we choose a simple "offspring pmf" under which a node in layer i has 1 or 2 descendants with probability p and $1 - p$, respectively. Let $Z_i = |\Gamma_i(v_f)|$ be the total number of vertices in layer i of $\mathcal{T}_f$. Thus, $\Pr(Z_1 = 1) = p = 1 - \Pr(Z_1 = 2)$ where $p \in (0, 1)$ is chosen to satisfy $m := \mathbb{E}(Z_1) = 2 - p > 1$ so that we are operating in the *supercritical regime.* Assuming that a tree of

depth t suffices for repair, we have

$$\beta_{AF} = td - \sum_{i=1}^{t-1}(t-i)Z_i; \;\; \mathbb{E}[\beta_{\text{AF}}] = td - \sum_{i=1}^{t-1}(t-i)m^i.$$

If we assume that the intermediate processing technique can be applied to layers $i : 1 \le i \le s$, then an easy calculation yields

$$\mathbb{E}[\beta_{\text{IP}}] = (t-s)d + (d-k+1-t+s)\sum_{i=1}^{s} m^i - \sum_{i=s+1}^{t-1}(t-i)m^i$$

and so $\mathbb{E}[\beta_{\text{AF}} - \beta_{\text{IP}}] = sd - \sum_{i=1}^{s}(2-p)^i(d-k+1+s-i)$ which is positive for small values of $d-k$ and large d.

3) Diagonal-matrix MSR codes: While the product-matrix codes are limited by the code rate $k/n < 1/2$, the construction of [11] removes this limitation, providing explicit families of exact-repair MSR codes for all possible values of $n-1 \ge d \ge k$.

The codes in [11] are defined in terms of the parity-check matrix which has a block diagonal structure. Below we assume that the parameters of the (n,k,l) array code $\mathcal{C}$ are fixed, and that $d = n-1, l = r^n$, where $r := n-k$. The code is defined over a finite field F of size at least rn. Let $\{\lambda_{i,j}\}_{i\in[n],j=0,1,\ldots,r-1}$ be rn distinct elements of F. For an integer $a \in \{0,1,\ldots,l-1\}$ let a_i be the i-th digit of its r-ary expansion. For $i = 1,2,\ldots,n$ define a matrix $A_i = \text{diag}(\lambda_{i,a_i}, a = 0,\ldots,l-1)$. The code $\mathcal{C}$ is formed of the codewords $C = (C_1,\ldots,C_n) \in (F^l)^n$ that satisfy the following set of r parity-check equations:

$$\sum_{i=1}^{n} A_i^{t-1} C_i = 0, \quad t = 1,\ldots,r. \tag{12}$$

Let $C_i = (c_{i,a}, a = 0,\ldots,l-1)^T$. Since the matrices A_i are diagonal, the parity check equations (12) take the form

$$\sum_{i=1}^{n} \lambda_{i,a_i}^{t-1} c_{i,a} = 0, \quad t = 1,\ldots,r, \; a = 0,1,\ldots,l-1. \tag{13}$$

The node repair with no communication constraints proceeds as follows. Assume that the node $i \in [n]$ has failed. Consider the set of indices $a(i,u) = (a_n,\ldots,a_{i+1},u,a_{i-1},\ldots,a_1)$, where $u = 0,1,\ldots,r-1$. The information uploaded from the helper node $j \in [n]\backslash\{i\}$ is given by $\mu_{j,i}^{(a)} = \sum_{u=0}^{r-1} c_{j,a(i,u)}$. Writing (13) for each of the indices $a(i,u)$, we obtain

$$\lambda_{i,u}^t c_{i,a(i,u)} + \sum_{j\ne i} \lambda_{j,a_j}^t c_{j,a(i,u)} = 0, \quad t = 0,1,\ldots,r-1.$$

Summing these equations on u and writing the result in matrix form, we obtain the relation

$$\begin{bmatrix} 1 & 1 & \ldots & 1 \\ \lambda_{i,0} & \lambda_{i,1} & \ldots & \lambda_{i,r-1} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{i,0}^{r-1} & \lambda_{i,1}^{r-1} & \ldots & \lambda_{i,r-1}^{r-1} \end{bmatrix} \begin{bmatrix} c_{i,a(i,0)} \\ c_{i,a(i,1)} \\ \vdots \\ c_{i,a(i,r-1)} \end{bmatrix} = - \begin{bmatrix} \sum_{j\ne i} \mu_{j,i}^{(a)} \\ \sum_{j\ne i} \lambda_{j,a_j} \mu_{j,i}^{(a)} \\ \vdots \\ \sum_{j\ne i} \lambda_{j,a_j}^{r-1} \mu_{j,i}^{(a)} \end{bmatrix}. \tag{14}$$

This equation permits recovery of the symbols $c_{i,a(i,u)}, 0 \le u \le r-1$ of the failed coordinate, and the other symbols are recovered similarly [11].

To adapt this procedure to repair on graphs, assume that the failed node is $i = n$ and write the vector on the right-hand side of (14) as $[\mu_{1,n}^{(a)}, \mu_{2,n}^{(a)}, \ldots, \mu_{n-1,n}^{(a)}]V_1^T$, where

$$V_1 := \text{Vandermonde}(\lambda_{1,a_1}, \lambda_{2,a_2}, \ldots, \lambda_{n-1,a_{n-1}})$$

is an $r \times (n-1)$ Vandermonde matrix with columns defined by the arguments. The matrix on the left in (14) is also Vandermonde, denote it by V_2. With these notations, (14) can be rewritten as

$$[c_{n,a(n,0)}, c_{n,a(n,1)}, \ldots, c_{n,a(n,r-1)}]V_2^T = -[\mu_{1,n}^{(a)}, \mu_{2,n}^{(a)}, \ldots, \mu_{n-1,n}^{(a)}]V_1^T$$

or

$$[c_{n,a(n,0)}, c_{n,a(n,1)}, \ldots, c_{n,a(n,r-1)}] = -[\mu_{1,n}^{(a)}, \mu_{2,n}^{(a)}, \ldots, \mu_{n-1,n}^{(a)}]U,$$

with $U := V_1^T(V_2^T)^{-1}$. This representation is essentially the same as (10), and hence the generic distributed repair scheme described in Sec. II-A applies to the codes considered in this section. This scheme will result in repair bandwidth gains for each of the groups of the node components mentioned above.

4) Node repair for general linear array codes: From the examples in the previous section it is clear that the graph-based repair procedure defined in (11) applies to any F-linear MSR code for which the information downloaded from the helper nodes is an F-linear function of their contents (all the known MSR codes are such). Indeed, the download operation can be written as $C(D)U$, where $C(D)$ is the contents of the helper nodes and U represents the linear transformation of the form (11). Once we reach the helper nodes in T_f with at least $d-k+1$ descendants, then we can switch to relaying linear combinations rather than the contents of the helper nodes.

Remark (MBR codes): For the other extremal point of the storage-bandwidth trade-off [3], i.e., the Minimum Bandwidth Regenerating (MBR) codes, the AF repair strategy is optimal in terms of the repair bandwidth because the amount of downloaded information is minimized by the code design.

III. Node repair on random graphs

In this section we analyze the distributed repair procedure in the case when the underlying graph $G(V,E)$ is a random graph from the $\mathcal{G}_{n,p}$ ensemble, where $0 < p < 1$. As before, we assume that the coordinates $C_1,\ldots,C_n$ of a codeword of an (n,k,d) MSR code are placed on the vertices $v_1,\ldots,v_n$. The main question that we address is to find the relation between the parameters of the problem p,n,k,d such that graph-based repair of the failed node with high probability results in lower repair bandwidth than the AF strategy.

We will assume that $p \gg \frac{\log n}{n}$ because if G is not connected then with positive probability the node v_f is isolated, and repair is not possible (the notation $f(n) \gg g(n)$ means that $g(n) = o(f(n))$). Furthermore, $\mathbb{P}_{\mathcal{G}_{n,p}}(\deg(v_f) \ge d) = \sum_{i=d}^{n}\binom{n}{i}p^i(1-p)^{n-i}$, which goes to zero for $n \to \infty$ if $d \gg np$. Thus, overall this is the parameter regime that may make the graph-based repair (possible and) advantageous over the agnostic AF repair procedure.

We will use the following two results regarding the random Erdös-Rényi graphs (below $\mathbb{P} = \mathbb{P}_{\mathcal{G}_{n,p}}$).

Lemma III.1 ([1], p.50; [5], Sec.7.1)**.** *(i) If* $p^2n - 2\log n \to \infty$, *and* $n^2(1-p) \to \infty$, *then* $\mathbb{P}(\mathsf{diam}(G) = 2) \to 1$.

(ii) Suppose that the functions $t = t(n) \geq 3$ *and* $0 < p = p(n) < 1$ *satisfy*

$$(\log n)/t - 3\log\log n \to \infty, \quad p^t n^{t-1} - 2\log n \to \infty,$$
$$p^{t-1}n^{t-2} - 2\log n \to -\infty,$$

then $\mathbb{P}(\mathsf{diam}(G) = t) \to 1$.

Lemma III.2 ([2], Lemma 3)**.** *Suppose that* $p \geq \frac{\log n}{n}$. *For any* $\epsilon > 0$ *and all* $i = 1, \ldots, \lfloor \log n \rfloor$

$$\mathbb{P}(|\Gamma_i(x)| \leq (1+\epsilon)(np)^i) \geq 1 - 1/\log^2 n \tag{15}$$
$$\mathbb{P}(|N_i(x)| \leq (1+2\epsilon)(np)^i) \geq 1 - 1/\log^2 n. \tag{16}$$

1) Repair threshold: We say that t*-layer repair* of the failed node v *is possible* if

$$\mathbb{P}(|N_t(v)| \geq d) \to 1 \text{ as } n \to \infty.$$

The next proposition establishes a threshold for t-layer repair in terms of p for *linear* number of helpers.

Proposition III.3. *Let* $d = \delta n, 0 < \delta < 1$ *and let* t *be a fixed integer. Then* t *is the threshold depth for repair if*

$$(np)^{t-1} = o(n), \quad p^t n^{t-1} - 2\log n \to \infty. \tag{17}$$

Proof. To show that t-layer repair is possible, we observe that from Lemma III.1, $\mathbb{P}(\mathsf{diam}(G) = t) \to 1$. This implies that for any failed node v, all the other nodes in the graph are reachable in at most t steps, and in particular, $|N_t(v)| = n > d$. To show that t is the smallest radius that supports repair, observe that by (16) for any $\epsilon > 0$

$$\mathbb{P}(|N_{t-1}(v)| \leq (1+2\epsilon)(np)^{t-1}) \to 1. \tag{18}$$

Since d is a linear function of n, the event

$$\{|N_{t-1}(v)| \geq d\} \subset \{|N_{t-1}(v)|/n > 0\} \quad (n \to \infty).$$

Together with (18) this implies that $\mathbb{P}(|N_{t-1}(v)|/n \geq \gamma) \to 0$ for any $\gamma > 0$. □

Remark: Given t, the conditions (17) are satisfied if

$$n^{-(t-1)/t} g(n) \ll p(n) \ll n^{-(t-2)/(t-1)},$$

where $g(n) \gg (2\log n)^{1/t}$.

2) Repair bandwidth: In this section we estimate the communication complexity of node recovery on a random graph. Throughout this section we assume that t is the threshold for repair, i.e., conditions (17) hold for $t, n,$ and p.

Proposition III.4. *The repair bandwidth* β_{AF} *satisfies* $\mathbb{P}(\beta_{\text{AF}} \geq td - o(n)) \to 1$

Remark: This proposition implies that for large n, "most" of the helper nodes are at distance t from the failed node. Note that, assuming (17), Lemma III.2 along with Lemma 8 in [2] imply that the neighborhood $\Gamma_t(v)$ with high probability grows as $c(np)^t$. for some constant $c < 1$. This provides an intuitive explanation of the claim of Prop. III.4 for $d = \Theta(n)$ and $(np)^{t-1} = o(n)$, implying that the AF repair strategy results in a t-fold increase of repair bandwidth compared to full connectivity.

Proof. Rewriting the expression for β_{AF} in (5), we obtain

$$\beta_{\text{AF}} = td - \sum_{i=1}^{t-1} (t-i)|\Gamma_i(f)|.$$

Let $E_i = \{|\Gamma_i(f)| \leq (1+\epsilon)(np)^i\}$ and notice that $E := \cap_{i=1}^{t-1} E_i \subseteq \{\beta_{\text{AF}} \geq (td - o(n))\}$. From Lemma III.2 we know that $\mathbb{P}(E_i^c) \leq 1/\log^2 n$ for all i, and thus

$$Pr(\cup_{i=1}^{t-1} E_i^c) \leq \sum_{i=1}^{t-1} Pr(E_i^c) \leq t/\log^2 n.$$

Finally, $\mathbb{P}(\beta_{\text{AF}} \geq td - o(n)) \geq Pr(E) \geq 1 - \frac{t}{\log^2 n} \to 1$. □

The next proposition gives further insights into the relationship between β_{AF} and t. Its proof is similar to the proof of Prop. III.4 and will be omitted.

Proposition III.5. *Let* $d = \delta n, 0 < \delta < 1$ *and let* $\kappa(n)$ *be a function of* n *such that* $\underline{c} \leq \kappa(n)/n \leq \bar{c}$ *starting with some* n. *Then for* $t > \bar{c}/\delta$ *we have* $\mathbb{P}(\beta_{\text{AF}} \leq \kappa(n)) \to 0$. *Further, if* $t \leq \underline{c}/\delta$, *then* $\mathbb{P}(\beta_{\text{AF}} \leq \kappa(n)) \to 1$.

Now let us show that the graph-based repair as given in (11) with high probability has smaller repair bandwidth.

Theorem III.6. *Let* t *be the threshold given in Prop III.3. For* $d = \Theta(n)$ *let* $d - k = \chi(n)$ *be a function of* n *such that* $\chi(n)n^{s-1}p^s \to 0$ *where* $s \leq t-1$ *is the largest integer for which this condition holds. Then* $\mathbb{P}(\beta_{\text{IP}} \leq (t-s)d + o(n)) \to 1$.

Proof. Let T_f be the repair tree with the root v_f. By assumption, the distance from the root to the leaves is t, and we will assume that the helper nodes in $\Gamma_i(v_f), i = t, t-1, \ldots, s+1$ simply relay their information along the edges, while the nodes in $N_s(v_f)$ transmit $l = d - k + 1$ symbols given by a linear combination of the form given in (11).

Then, for the failed node v_f, we have

$$\begin{aligned}
\beta_{\text{IP}} &= (t-s)(d - |N_{t-1}(f)|) + \sum_{i=1}^{t-s-1} (t-s-i)|\Gamma_{t-i}(f)| \\
&\quad + (d-k+1)\sum_{i=1}^{s} |\Gamma_i(f)| \\
&= (t-s)d + \sum_{i=1}^{s} |\Gamma_i(f)|(d-k+1-(t-s)) \\
&\quad - \sum_{i=s+1}^{t-1} |\Gamma_i(f)|(t-i) \\
&\leq (t-s)d + \sum_{i=1}^{s} |\Gamma_i(f)|(d-k+1-t+s)).
\end{aligned}$$

Proceeding similarly to the proof of Proposition III.4, we obtain

$$\mathbb{P}\Big(\beta_{\text{IP}} \leq (t-s)d + \sum_{i=1}^{s} (np)^i(\chi(n)+1-t+s)\Big) \geq 1 - s/\log^2 n \to 1.$$

Now using the assumption $\chi(n)(np)^s = o(n)$ finishes the proof. □

REFERENCES

[1] B. Bollobás, "The diameter of random graphs," *Trans. AMS*, vol. 267, no. 1, pp. 41–52, 1981.

[2] F. Chung and L. Lu, "The diameter of sparse random graphs," *Advances in Applied Mathematics*, vol. 26, no. 4, pp. 257 – 279, 2001.

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

[4] M. Elyasi and S. Mohajer, "Cascade codes for distributed storage systems," *IEEE Trans. Inf. Theory*, vol. 66, no. 12, pp. 7490–7527, 2020.

[5] A. Frieze and M. Karońsky, *Introduction to Random Graphs*. Cambridge University Press, 2016.

[6] S. Goparaju, A. Fazeli, and A. Vardy, "Minimum storage regenerating codes for all parameters," *IEEE Trans. Inf. Theory*, vol. 63, no. 10, pp. 6318–6328, 2017.

[7] 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.

[8] N. Raviv, N. Silberstein, and T. Etzion, "Constructions of high-rate minimum storage regenerating codes over small fields," *IEEE Trans. Inf. Theory*, vol. 63, no. 4, pp. 2015–2038, 2017.

[9] 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.

[10] I. Tamo, Z. Wang, and J. Bruck, "Access versus bandwidth in codes for storage," *IEEE Trans. Inf. Theory*, vol. 60, no. 4, pp. 2028–2037, 2014.

[11] M. Ye and A. Barg, "Explicit constructions of high-rate MDS array codes with optimal repair bandwidth," *IEEE Trans. Inf. Theory*, vol. 63, no. 4, pp. 2001–2014, 2017.

[12] 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.