Node Repair on Connected Graphs

Adway Patra, *Graduate Student Member, IEEE*, and Alexander Barg, *Fellow, IEEE*

***Abstract*— We study the problem of erasure correction (node repair) for regenerating codes defined on graphs wherein the cost of transmitting the information to the failed node depends on the graphical distance from this node to the helper vertices of the graph. The information passed to the failed node from the helpers traverses several vertices of the graph, and savings in communication complexity can be attained if the intermediate vertices process the information rather than simply relaying it toward the failed node. We derive simple information-theoretic bounds on the amount of information communicated between the nodes in the course of the repair. Next we show that Minimum Storage Regenerating (MSR) codes can be modified to perform the intermediate processing, thereby attaining the lower bound on the information exchange on the graph. We also consider node repair when the underlying graph is random, deriving conditions on the parameters that support recovery of the failed node with communication complexity smaller than required by the simple relaying.**

***Index Terms*— Distributed storage, Minimum Storage Regenerating (MSR) codes, graph-constrained storage systems, random graphs.**

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 aims at limiting the information communicated between them for the recovery of the failed node. The currently adopted framework for studying erasure correction under communication constraints was established in [6], and coding constructions that minimize the communication are known collectively as *regenerating codes*. The authors of [6] derived a lower bound on the minimum amount of information acquired from the surviving nodes for the purposes of repair.

Manuscript received August 11, 2021; revised December 30, 2021; accepted January 13, 2022. Date of publication January 25, 2022; date of current version April 21, 2022. This work was supported by NSF under Grant CCF1814487, Grant CCF2110113, and Grant CCF2104489. An earlier version of this paper was presented in part at the 2021 IEEE International Symposium on Information Theory [DOI: 10.1109/ISIT45174.2021.9517983]. *(Corresponding author: Alexander Barg.)*

Adway Patra is with the Department of Electrical and Computer Engineering and the Institute for Systems Research (ISR), University of Maryland, College Park, MD 20742 USA (e-mail: apatra@umd.edu).

Alexander Barg is with the Department of Electrical and Computer Engineering and the Institute for Systems Research (ISR), University of Maryland, College Park, MD 20742 USA, and also with the Institute for Information Transmission Problems (IITP), Russian Academy of Sciences, 127051 Moscow, Russia (e-mail: abarg@umd.edu).

Communicated by A. Thangaraj, Associate Editor for Coding Techniques.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIT.2022.3145824.

Digital Object Identifier 10.1109/TIT.2022.3145824

For a finite field $\mathbb{F} = \mathbb{F}_q$ we consider a code $\mathcal{C} \subset \mathbb{F}^{nl}$ whose codewords are represented by $l \times n$ matrices over $\mathbb{F}$. We assume that each coordinate (a vector in $\mathbb{F}^l$) 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 $\mathbb{F}$. In this paper we address 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 an n-set of vertices and 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. This group of problems is motivated by the assumption that the links between the nodes are established based on physical proximity and the associated energy constraints, limitations of the system architecture, or other features with the same effect. In the network environment such as low-power wide-area networks (LP-WAN), e.g., path loss in narrow-band lower-power IoT, the mentioned limitations arise naturally as a part of the functioning of the system. We also always assume point-to-point rather than broadcast communication. For distributed storage systems this is a natural restriction, while for IoT applications this assumption may be imposed because of energy or privacy considerations.

Under a naive approach to this problem, it is still possible to use the known methods of node repair whereby the chosen group of the helper nodes communicates some functions of their contents to the failed node. Note however that the data from the helper nodes not directly connected to the failed node will have to be relayed along some path to the failed node, increasing the bandwidth utilized for the repair. Thus, a natural question to study is whether there are more economical ways of accomplishing this goal given the structure of the graph G, under which the data from the far-off helper nodes is processed along the way and combined with the contents of the intermediate nodes, saving on the overall communication. We give an affirmative answer in Sec. III, 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. More precisely, we derive a lower bound on the repair bandwidth on a graph and show that this bound can be attained with linear MSR codes. We note that the problem of repair on graphs was considered in earlier [10] in the case when the graph is a path and in [13] for a directed ring. These papers derived lower bounds on the repair bandwidth which coincide with our bound once it is specialized to a path or a ring.

Intermediate data processing is also an essential component of a version of the node repair problem known as *cooperative repair* [21]. It is therefore of interest to examine possible

applications of cooperative MSR codes to the problem at hand, aiming again at reduced communication complexity of repair. We show one application of this idea in Sec. IV, using a family of cooperative codes to design a scheme with reduced repair bandwidth in the case of multiple failed nodes. In Sec. V-B we consider the problem of node repair in the situation when the helper nodes can exchange (and process) information before communicating with the failed nodes. We derive a framework to bound below the complexity of repair under this relaxation and use it to compute lower bounds on the repair bandwidth in several examples. One of these examples also affords a matching code construction, again inspired by cooperative codes. Since the setting of this example is rather restricted, we give the details in the appendix. Finally, in Sec. VI we address the question of node repair for random graphs from the standard Erdös-Rényi ensemble $\mathcal{G}_{n,p}$ as well as from the ensemble of random regular graphs, 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.

Throughout the paper we focus only on the node repair problem and do not study the communication complexity of the actual access to the encoded data (the "data collection" task in the terminology of [6]).

To conclude this introduction, we note that the general problem of information processing or recovery under communication constraints represented by a graph has recently been studied in a number of specific settings. Among them, locally recoverable codes on graphs [14], [15] (and the associated problems of guessing games on graphs [9] and index coding [1]), their extension to recoverable systems [7], private information retrieval on graphs [19], and others. The problem of node repair under communication constraints introduced here is another instantiation of this broadly defined theme.

II. Lower Bounds on the Repair Bandwidth

A. MSR Codes: A Reminder

Let $\mathbb{F}$ be a finite field. A *vector code* $\mathcal{C}$ of length n is an $\mathbb{F}$-linear subspace of $(\mathbb{F}^l)^n$ whose codewords can be thought of as $l \times n$ matrices. In the context of storage codes, elements of $\mathcal{C}$ are often referred to as n-words whose coordinates are l-vectors over $\mathbb{F}$. We further assume that the information contents of the codeword is kl symbols of $\mathbb{F}$, in other words, that $|\mathcal{C}| = q^{kl}$, and that any k coordinates suffice to recover the entire codeword. Thus, the code has the maximum distance separable (MDS) property, and any $n-k$ erased coordinates can be found from the remaining k ones, accounting for the optimal erasure correction capacity.

Suppose that the coordinates of the codeword are placed on n different storage nodes, and refer to the coordinates themselves as nodes. The defining property of MSR codes is related to recovering the value of an erased coordinate of the codeword, or repairing a single failed node. According to the above description, we can accomplish this by using the information from k functional nodes and downloading a total of kl symbols of the field $\mathbb{F}$. At the same time, this operation supports recovery of the entire codeword, accomplishing more than we actually need. An important finding of the work [6] was to point out that we can save on the amount of downloaded information by performing the repair based on the contents of $d > k$ helper nodes. To achieve the saving, each of the helper nodes provides a function of its contents, and [6] showed that to accomplish the repair it is necessary to download at least $\frac{dl}{d-k+1}$ field symbols. This is smaller than kl for all $d \leq n$. A code $\mathcal{C}$ with the parameters (n,k,d,l) is called MSR if it supports node recovery with *repair bandwidth* meeting the lower bound for the chosen number d of *helper nodes*. It is easy to show that for a code to have this property, each of the helper nodes necessarily provides $l/(d-k+1)$ field symbols for the recovery of the failed node (the so-called *uniform download* property).

Formally, an (n,k,d,l) linear MDS vector code $\mathcal{C}$ over $\mathbb{F}$ is called MSR if there are linear functions $h_i : \mathbb{F}^l \to \mathbb{F}^{l/(d-k+1)}, i = 1,\dots,n$ such that for any $j \in [n]$ and any subset $\{i_1,\dots,i_d\} \subset [n]\backslash\{j\}$ there exists a linear function $g_j : \mathbb{F}^{d(l/(d-k+1))} \to \mathbb{F}^l$ such that for any codeword $C = (C_1,\dots,C_n) \in \mathcal{C}$ the value C_j (the contents of the failed node) is found as

$$C_j = g_j(h_{i_1}(C_{i_1}),\dots,h_{i_d}(C_{i_d})).$$

Slightly more generally, the functions h_i could also depend on j, but this will not be important below. A number of families of MSR codes are known in the literature, among them constructions of [11], [17], [18], [22], [23], [25], see also a recent survey in [16]. In this paper we use two such families to exemplify our approach to node repair on graphs, namely product matrix codes [17] and diagonal-matrix codes [25]. It will become apparent toward the end of Sec. III that any family of $\mathbb{F}$-linear MSR codes can be incorporated in our repair scheme.

B. Problem Statement and the Lower Bounds

The problem that we consider is associating the nodes with the vertices of a graph and performing node repair by transmitting the information along the edges. 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 $G(V,E)$ with $|V| = n$, which represents a distributed storage system with a given connectivity structure. 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 a 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). To examine options for more economical repair including *intermediate processing* (IP) of the information, we begin with deriving a lower bound on the repair bandwidth.

Before proceeding, let us further specify our assumptions. We assume that for the failed node v_f, the helper nodes D are chosen to be the d *closest* nodes to v_f in terms of the graph

distance.[1] 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 for spheres and balls around v_f in $G_{f,D}$:

$$\Gamma_j(v_f) = \{v \in V_{f,D} : \rho(v, v_f) = j\},\ N_i(v_f) = \cup_{j=1}^{i} \Gamma_j(v_f),$$

and we refer to the vertices in $\Gamma_j(v_f)$ as the helper nodes in *layer* j. The case $t = 1$ corresponds to the much-studied graph-agnostic repair scenario [6], 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 $(\mathbb{F}^l)^n$ 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, and the RVs $S_i^f, i \in D$ determine the contents of v_f, or formally,

$$(i)\ H(S_i^f | W_i) = 0, \quad i \in D,$$
$$(ii)\ H(W_f | S_D^f) = 0.$$

From the cut-set bound [6] 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. In the next lemma we bound below the amount of information sent by a subset of helper nodes in an MSR code. The proof that we give is close to the arguments that have previously appeared in the literature, see for instance [20].

Lemma 1: Let $v_f, f \in [n]$ be the failed node. For a subset of the helper nodes $A \subset D$ let R_A^f be such that $H(R_A^f | W_A) = 0$ and

$$H(W_f | R_A^f, S_{D\backslash A}^f) = 0. \tag{1}$$

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

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

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

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

Proof: Part (1): By the assumption (1), given the contents of all the nodes in $D\backslash A$, the information contained in R_A^f is sufficient to repair v_f, i.e.,

$$H(W_f | R_A^f, W_{D\backslash A}) = 0. \tag{2}$$

[1]This assumption is not restrictive because, whenever the set D spans a connected subgraph, our bounds on communication complexity apply for the information processing within that subgraph.

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

$$H(R_A^f, W_{D\backslash A}, W_B) = H(R_A^f, W_{D\backslash A}, W_f, W_B) \geq kl, \tag{3}$$

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

$$\begin{aligned} H(R_A^f, W_{D\backslash A}, W_B) &\leq H(R_A^f) + H(W_{D\backslash A}, W_B) \\ &= H(R_A^f) + (k-1)l, \end{aligned} \tag{4}$$

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.

For Part (2), let $C \subseteq D\backslash A$ such that $|C| = k - 1$ and let $I = D\backslash\{A \cup C\}$. By the assumption (1), we have

$$H(W_f | R_A^f, W_C, S_I^f) = 0. \tag{5}$$

Now,

$$H(R_A^f, W_C, S_I^f) = H(R_A^f, W_f, W_C, S_I^f) \geq kl, \tag{6}$$

where the equality in (6) follows from (5) and the chain rule, and the inequality follows from the MDS property and the fact that $|C| = k - 1$. Next observe that

$$\begin{aligned} &H(R_A^f, W_C, S_I^f) \\ &\leq H(R_A^f) + H(W_C) + H(S_I^f) \\ &\leq H(R_A^f) + H(W_C) + \sum_{i \in D\backslash\{A \cup C\}} H(S_i^f) \\ &= H(R_A^f) + (k-1)l + \frac{(d - (k-1) - |A|)l}{d - k + 1} \end{aligned} \tag{7}$$

where we again use the independence of any $k-1$ coordinates in an MDS code. Combining (6) and (7), we obtain the claimed inequality. □

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

Proposition 1: 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 1 by taking $A = \cup_{i=j}^{t} \Gamma_i(v_f)$. □

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 it does not account for any other communication occurring among the helper nodes. If $G_{f,D}$ is a rooted tree, such communication does not occur, and this will be the main (but not the only) case studied below. In this case we can make our arguments more precise.

Let T_f be a rooted spanning tree of $G_{f,D}$ with root v_f (see Fig.1), 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 $D^*(v_i) = D(v_i) \cup \{v_i\}$. The total communication complexity of node repair using the tree T_f is bounded below in the following proposition.

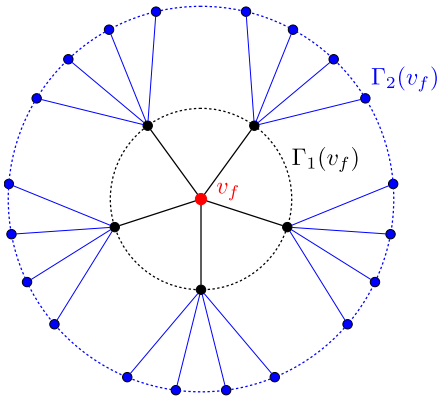

Fig. 1. Repair tree of the node v_f.

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

$$\beta \geq |J_f|l + \sum_{v\in V(T_f)\backslash(\{v_f\}\cup J_f)} \frac{|D^*(v)|l}{d-k+1}. \tag{8}$$

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 1. By the same lemma, every node $v \in J_f$ needs to transmit at least l symbols to its immediate parent. □

For comparison purposes we also write out an expression for the AF repair procedure of MSR codes, described in the beginning of this section. Its repair bandwidth can be found as

$$\beta_{\text{AF}} = \Big(t(d-|N_{t-1}(v_f)|) + \sum_{i=1}^{t-1} i|\Gamma_i(v_f)|\Big)\frac{l}{d-k+1}. \tag{9}$$

Every helper node provides $l/(d-k+1)$ symbols of information for repair, so for a node $v \notin J_f$ the AF strategy is trivially optimal by part (2) of Lemma 1. 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.

C. A Bound for Repair of Multiple Nodes

Before proceeding further, let us note a simple extension of Lemma 1 to the case of multiple failed nodes, which is often studied for regenerating codes under full node connectivity [25]. The repair of multiple nodes in a graph depends on their mutual placement and their connections to the helpers, and gives rise to several options. Denote by $F \subset V$ the set of failed nodes, and let $|F| = h \geq 1$. To keep the argument manageable, we assume that recovery of all the nodes in F relies on a *common* set D of helper nodes. With this assumption Lemma 1 affords the following extension.

Lemma 2: Let $F \subset [n], |F| = h, 1 \leq h \leq n-d$ be the set of failed nodes. For a subset of the helper nodes $A \subset D$ let R_A^F be a function of S_A^F such that

$$H(W_F|R_A^F, S_{D\backslash A}^F) = 0. \tag{10}$$

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

$$H(R_A^F) \geq hl.$$

2) If $|A| \leq d-k+h-1$, then

$$H(R_A^F) \geq \frac{h|A|l}{d-k+h}.$$

The proof follows closely the proof of Lemma 1 and will be omitted (it is included in the preprint version of this paper, arXiv:2108.00939, Appendix B). As above, in Lemma 2 we sidestepped the specific way of communicating the information from the helpers to the failed nodes, limiting ourselves to the lower bounds on the information provided by the helpers. The communication complexity of implementing the repair depends on the topology of the graph and on the relative location of the failed nodes and the helpers. In Sec. IV we present a construction that, under certain assumptions, attains the bounds of this lemma, performing the intermediate processing instead of relaying and gaining in communication complexity over the AF protocol.

III. MSR Constructions for Repair of Graph Vertices

In this section we show that linear MSR codes support a repair procedure that attains the lower bound (8) on the communication complexity. While this procedure is general, we begin with illustrating it for product-matrix codes of [17]. Then we consider several examples of graphs, estimating the savings of repair complexity compared to the AF repair. After that, we show that conceptually the same procedure applies to the diagonal-matrix codes of [25], and then briefly discuss a general version of this repair protocol as it applies to all families of $\mathbb{F}$-linear MSR codes.

A. Product-Matrix (PM) Codes

In their standard form, PM codes are described as follows. Fix the code length n and the dimension parameter k, and take $d = 2k-2, l = k-1$. The code $\mathcal{C} : \mathbb{F}^{k(k-1)} \to \mathbb{F}^{ln}$ encodes $k(k-1)$ symbols of $\mathbb{F}$ into a codeword of length n with each coordinate formed of l 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 $x_i, i = 1, \ldots, n$ be distinct elements of $\mathbb{F}$, let

$$\Phi = [\phi_1^T, \ldots, \phi_n^T]^T$$

be a Vandermonde matrix with rows of the form $\phi_i = (1, x_i, \ldots, x_i^{l-1})$ and take $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$ with $\lambda_i = x_i^l, i = 1, \ldots, n$. Now form an $n \times 2l$ matrix $\Psi = [\Phi, \Lambda\Phi]$ The encoding mapping $\mathcal{C}$ sends the matrix M to $C = \Psi M$, which is an $n \times l$ matrix, and thus the contents of the node $v_i, i = 1, \ldots, n$ is given by the product

$$C_i := [\phi_i, x_i^l\phi_i]M = \phi_i S_1 + \lambda_i \phi_i S_2. \tag{11}$$

To describe the repair procedure from [17], suppose without loss of generality that the helper nodes form the set $D = \{1, \ldots, d\}$ and that the failed node's index is $f \in [n]\backslash[D]$.

The original node repair (erasure correction) procedure proposed in [17] proceeds as follows. The information downloaded by the failed node v_f from the helper node $i \in D$ 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 $\mathbb{F}$. Thus, the failed node downloads a d-dimensional vector $y = y_{f,D}$ given by

$$y = \Psi_D M \phi_f^T = \Psi_D \begin{bmatrix} S_1 \phi_f^T \\ S_2 \phi_f^T \end{bmatrix}, \tag{12}$$

where Ψ_D is the submatrix of Ψ formed of the first $d = 2l$ 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 (11) 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 (12)

$$\phi_f M^T = y^T (\Psi_D^T)^{-1}. \tag{13}$$

Using (11),(13), 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 $U_i, i = 1, \ldots, d$, then we have

$$C_f = \sum_{i=1}^{d} y_i U_i. \tag{14}$$

Note that the matrix U does not depend on the codeword, and can be precomputed. Overall this rewriting of the repair process (12) 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 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 at least 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

$$\sum_{j \in D^*(v_i)} y_j U_j. \tag{15}$$

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 $\mathbb{F}$ to its parent in T_f. This proves the following theorem.

Theorem 1: 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 (8) with equality.

To match the above procedure to the bound (8), recall that each helper node in the PM code construction provides one symbol of $\mathbb{F}$ for repair.

B. 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 $\mathbb{F}$ 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, the repair bandwidth with intermediate processing equals $\beta_{\text{IP}} = 3k-4$ because the symbols of the helpers other than the center are aggregated using (15) before relaying to v_f. Another elementary example, which results in a similar 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 Fig.1. 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 (15), 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. Note that the regime of small $d-k$ arises also as a sufficient condition of repair bandwidth savings for random graphs in Sec. VI.

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 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

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

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

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

and so

$$\mathbb{E}[\beta_{\rm AF} - \beta_{\rm IP}] = sd - \sum_{i=1}^{s}(2-p)^i(d-k+1+s-i)$$

which is positive for small values of $d-k$ and large d.

C. Diagonal-Matrix MSR Codes

While the product-matrix codes are limited by the code rate $k/n < 1/2$, the construction of [25] 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 [25] 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 $\mathbb{F}$ of size at least rn. Let $\{\lambda_{i,j}\}_{i\in[n],j=0,1,\ldots,r-1}$ be rn distinct elements of $\mathbb{F}$. 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 the matrix $A_i = \mathrm{diag}(\lambda_{i,a_i}, a = 0,\ldots,l-1)$. The code $\mathcal{C}$ is formed of the codewords $C = (C_1,\ldots,C_n) \in (\mathbb{F}^l)^n$ that satisfy the following set of r parity-check equations:

$$\sum_{i=1}^{n} A_i^{t-1} C_i = 0, \quad t = 1,\ldots,r. \tag{16}$$

Let $C_i = (c_{i,a}, a = 0,\ldots,l-1)^T$. Since the matrices A_i are diagonal, the parity check equations (16) take the form

$$\sum_{i=1}^{n} \lambda_{i,a_i}^{t-1} c_{i,a} = 0, \quad t = 1,\ldots,r, \; a = 0,1,\ldots,l-1. \tag{17}$$

The node repair with no communication constraints proceeds as follows. Assume that the node $i \in [n]$ has failed. We partition the set of coordinates $(c_{i,a})$ into groups of size r whose indices differ only in the ith entry. Namely, start with some $a \in \{0,\ldots,l-1\}$ and consider the set of indices $a(i,u) = (a_n,\ldots,a_{i+1},u,a_{i-1},\ldots,a_1)$, $u = 0,1,\ldots,r-1$. The information downloaded 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 (17) 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

$$\begin{bmatrix} 1 & 1 & \ldots & 1 \\ \lambda_{i,0} & \lambda_{i,1} & \ldots & \lambda_{i,r-1} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{i,0}^{r-1} & \lambda_{i,1}^{r-1} & \ldots & \lambda_{i,r-1}^{r-1} \end{bmatrix} \begin{bmatrix} c_{i,a(i,0)} \\ c_{i,a(i,1)} \\ \vdots \\ c_{i,a(i,r-1)} \end{bmatrix} = - \begin{bmatrix} \sum_{j\ne i} \mu_{j,i}^{(a)} \\ \sum_{j\ne i} \lambda_{j,a_j} \mu_{j,i}^{(a)} \\ \vdots \\ \sum_{j\ne i} \lambda_{j,a_j}^{r-1} \mu_{j,i}^{(a)} \end{bmatrix}. \tag{18}$$

This equation permits recovery of the symbols $c_{i,a(i,u)}, 0 \le u \le r-1$ of the failed coordinate, and varying a, we recover the other groups of coordinates in the same manner.

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 (18) 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 (18) is also Vandermonde, denote it by V_2. With these notations, (18) can be rewritten as

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

or

$$\begin{aligned} &[c_{n,a(n,0)}, c_{n,a(n,1)}, \ldots, c_{n,a(n,r-1)}] \\ &\qquad = [\mu_{1,n}^{(a)}, \mu_{2,n}^{(a)}, \ldots, \mu_{n-1,n}^{(a)}] U \\ &\qquad = \sum_{j=1}^{n-1} \mu_{j,n}^{(a)} U_j \end{aligned} \tag{19}$$

where we denoted $U := -V_1^T (V_2^T)^{-1}$ and U_j is the jth row of U. This representation is essentially the same as (14), and hence the generic distributed repair scheme described in Sec. III applies to the codes considered in this section. Specifically, the matrix U is independent of the codeword, and can be computed in advance, and once a node v in the repair tree has $d-k+1$ or more descendants, it switches to transmitting $\sum_{j\in D^*(v)} \mu_{j,n}^{(a)} U_j$. This procedure supports repair bandwidth gains over the AF strategy for each of the groups of the node components mentioned above.

D. Node Repair for General Linear Array Codes

From the examples in the previous sections it is clear that the graph-based repair procedure defined in (15) applies to any $\mathbb{F}$-linear MSR code for which the information downloaded from the helper nodes is an $\mathbb{F}$-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 (15). 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. The savings in repair bandwidth will be the same as for the two constructions considered above in this section.

Remark (MBR Codes): For the other extremal point of the storage-bandwidth trade-off [6], i.e., the Minimum Bandwidth Regenerating 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.

IV. Node Repair for Multiple Failures

In this section we present a code construction for the repair of multiple nodes that attains the lower bound of Lemma 2. We begin with specifying our assumptions. Suppose that the data is stored on a connected graph $G(V,E)$, and $F \subset V$ is a set of failed vertices of size h. Further, let $D, |D| = d$ be the subset of helper nodes. The data is encoded using an (n,k,d,l) MSR code, where $n = |V|$ is the number of vertices. The encoding scheme that we present below further assumes that the communication from D to F passes through some fixed node $w \in D$ as shown in Fig. 2 for $h = 2$ and $F = \{v_1, v_2\}$. This assumption, taken to fit the structure behind Lemma 2, suggests that we perform simple relaying along the path(s) from w to the failed vertices. The repair process becomes more complicated if the failed vertices have different access points to D, and we do not consider it here. We further assume that the set D spans a connected subgraph $G_D \subset G$ and denote by T_w a (rooted) spanning tree of G_D with root w. Finally, denote by $D_w(v)$ the set of descendants of $v \in V(T_w)$ in the tree T_w and let $D^*_w(v) = D_w(v) \cup \{v\}$.

Under these assumptions it is possible to write out a bound on the communication complexity of repair within the set of the helper nodes until the data reaches the node w (after that the data is no longer processed until it reaches the nodes in F). The following proposition is an obvious extension of the bound (8).

Proposition 3: Let $J_{w,h} = \{v \in V(T_w) : D_w(v) \geq d-k+h\}$. The total communication along the edges of T_w for repair of the nodes in F is bounded below as

$$\beta(D) \geq |J_{w,h}|l + \sum_{v \in V(T_w)\setminus J_{w,h}} \frac{|D^*_w(v)|l}{d-k+h}.$$

Below we present a construction of codes and a repair scheme that meets this bound with equality, attaining the minimum possible communication complexity of repair of the nodes in F under the assumptions discussed above (it is possible that removing these assumptions enables one to further lower the communication cost). The scheme relies on the idea of *cooperative repair* [21]. In this setting, under the full connectivity assumption, two or more failed nodes connect directly to the same set of helpers, evaluate partial information about their contents, and then exchange the results to complete the repair.

We use this idea for repair on graphs wherein the information from helpers is transmitted along some path to the failed node, relying on a family of cooperative codes constructed recently in [24]. The savings come from the fact that in the course of this transmission we can perform intermediate processing rather than simple relaying. In our presentation for simplicity we assume that $d = k+1, h = 2$ as in Fig. 2. At the same time it will be obvious that the technique applies to all other feasible parameter regimes.

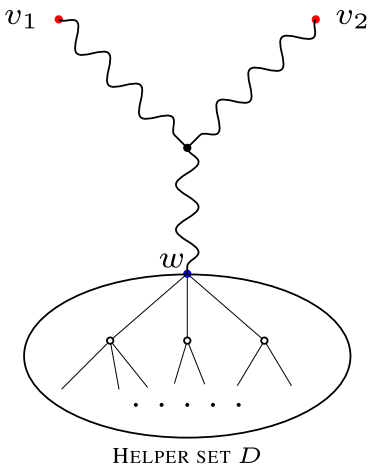

Fig. 2. Graph topology for repair of multiple nodes.

Let $\mathcal{C}$ be the $[n, k, d = k+1, l = 3 \times 2^n]$ *cooperative repair* MSR code from the family constructed in [24].[2] Every coordinate C_i of a codeword $C = (C_1, \ldots, C_n) \in (\mathbb{F}^l)^n$ is a vector $\{c_{i,b,a} : b \in \{1,2,3\}, a \in \{0,1,\ldots 2^n-1\}\}$. For $2n$ distinct field elements $\{\lambda_{i,j} : i \in [n], j \in \{0,1\}\}$, the parity check equations that define the code are

$$\sum_{i=1}^{n} \lambda^t_{i,a_i} c_{i,b,a} = 0 \quad \forall \quad t \in \{0,1,\ldots,n-k-1\}, \\ a \in \{0,1,\ldots,2^n-1\}, \quad b \in \{1,2,3\}, \qquad (20)$$

where a_i is the i-th coordinate in the binary representation of a. Below we use the notation $a(i, a_i \oplus 1)$ to denote the number obtained from a by flipping the ith bit in its binary expansion. Assume that the failed nodes correspond to coordinates 1 and 2 and fix a value of $a \in \{0,1,\ldots,2^n-1\}$. The standard cooperative repair under direct connectivity (on a complete graph) proceeds in two steps. In step 1, helper node i sends $\{c_{i,1,a} + c_{i,2,a(1,a_1\oplus 1)} : a \in \{0,1,\ldots,2^n-1\}\}$ to node 1 and $\{c_{i,1,a} + c_{i,3,a(2,a_2\oplus 1)} : a \in \{0,1,\ldots,2^n-1\}\}$ to node 2. Using a with $b = 1$ and $a(1, a_1 \oplus 1)$ with $b = 2$ in (20) and summing the corresponding equations, we obtain

$$\lambda^t_{1,a_1} c_{1,1,a} + \lambda^t_{1,a_1\oplus 1} c_{1,2,a(1,a_1\oplus 1)} \\ + \lambda^t_{2,a_2}(c_{2,1,a} + c_{2,2,a(1,a_1\oplus 1)}) \\ + \sum_{i=3}^{n} \lambda^t_{i,a_i}(c_{i,1,a} + c_{i,2,a(1,a_1\oplus 1)}) = 0 \qquad (21)$$

for all $t \in \{0,1,\ldots,n-k-1\}$. Equations (21) form a set of parity checks of an $(n+1, k+1)$ Reed-Solomon code, and hence knowing $c_{i,1,a} + c_{i,2,a(1,a_1\oplus 1)}$ at $k+1$ positions allows node 1 to recover $c_{1,1,a}, c_{1,2,a(1,a_1\oplus 1)}$ and $(c_{2,1,a} + c_{2,2,a(1,a_1\oplus 1)})$. A similar argument shows that node 2 can recover $c_{2,1,a}, c_{2,3,a(2,a_2\oplus 1)}$ and $(c_{1,1,a} + c_{1,3,a(2,a_2\oplus 1)})$. In step 2 of the repair, node 1 sends $(c_{2,1,a} + c_{2,2,a(1,a_1\oplus 1)})$ to node 2 and node 2 sends $(c_{1,1,a} + c_{1,3,a(2,a_2\oplus 1)})$ to node 1, which completes the repair of both node 1 and 2; for details see [24].

[2] We could use other code families, for instance, the codes from [26].

To see how intermediate processing at the nodes of the tree T_w can simplify repair on a graph of the type shown in Fig. 2, observe that the first step above, node 1 seeks to learn three code symbols (namely $c_{1,1,a}, c_{1,2,a(1,a_1\oplus 1)}$ and $(c_{2,1,a}+c_{2,2,a(1,a_1\oplus 1)})$) of the $(n+1,k+1)$ RS codeword, and it does so by collecting $k+1$ symbols from $k+1$ helper nodes. In an RS code, once we know any $k+1$ coordinates, all the other coordinates of the codeword can be computed via Lagrange interpolation and subsequent evaluation. This can be expressed in matrix form as follows:

$$\begin{bmatrix} c_{1,1,a} \\ c_{1,2,a(1,a_1\oplus 1)} \\ (c_{2,1,a}+c_{2,2,a(1,a_1\oplus 1)}) \end{bmatrix} = [U_1\ U_2\ \ldots\ U_{k+1}] \begin{bmatrix} (c_{i_1,1,a}+c_{i_1,2,a(1,a_1\oplus 1)}) \\ (c_{i_2,1,a}+c_{i_2,2,a(1,a_1\oplus 1)}) \\ \vdots \\ (c_{i_{k+1},1,a}+c_{ik+1,2,a(1,a_1\oplus 1)}) \end{bmatrix},$$

where $i_1, i_2, \ldots, i_{k+1}$ are the helper nodes and the matrix $U = [U_1\ U_2\ \ldots\ U_{k+1}]$ is obtained by multiplying an inverse Vandermonde matrix (Lagrange interpolation) and a matrix corresponding to evaluating the obtained polynomial at the three coordinates being sought. Since the matrix U can again be pre-computed, a node that has collected the values $(c_{i,1,a}+c_{i,2,a(1,a_1\oplus 1)})$ from three or more helper nodes, can start transmitting the corresponding linear combinations, much in the same way as was done in Section III. The above procedure is repeated for node 2 with appropriate adjustments to the subscripts in the last displayed equation. Step 2 of the repair process is unchanged from that of the standard cooperative repair, and it yields no communication savings. Exactly as in the case of a single failed node, viz., Theorem 1, we can show that this procedure meets the bound of Lemma 2.

V. Repair With Information Exchange Among the Helpers

The bounds and constructions presented earlier in this paper are focused on communication from the helper nodes to the failed node. In this section we consider a more general problem (and potential savings in the repair cost) when the helper nodes may communicate with each other before transmitting the information to the failed node. Recall that a variant of this problem was considered earlier in the literature under very specific assumptions: The nodes in the storage cluster are organized in subsets, called *racks*, and communication between the nodes in the rack does not count toward the repair bandwidth. This model enables one to derive tighter bounds on the cost of node repair [12], and there are families of codes that attain these bounds [4].

Another version of information exchange in the context of erasure recovery appeared earlier in the problem of *cooperative repair*, already mentioned in the previous section. In this setting (assuming full connectivity) several failed nodes contact the same set of helpers and process the received information, gaining some knowledge about their contents and about the contents of the other failed nodes. They then exchange information to complete the repair. This problem, introduced in [21], is vaguely reminiscent of repair on graphs because different nodes of the encoding acquire partial information about the contents of other nodes. Below we make this link more precise by presenting an example of node repair on graphs motivated by cooperative repair (albeit in a rather restricted setting).

We begin with establishing a framework for finding a lower bound on the total communication complexity of repair for general graphs. Let us define some additional notation. Let an (n,k,d,l) MSR code be defined on a connected graph $G=(V,E)$. Assume that the subgraph $G_{f,D}=(V_{f,D},E_{f,D})$ spanned by the failed node v_f and the set of helper nodes D is connected. Construct a directed graph $\bar{G}_{f,D}=(V_{f,D},\bar{E}_{f,D})$ as follows:

- For every edge $(u,v)\in E_{f,D}$ with $u,v\in D$, add the two directed edges (u,v) and (v,u) to $\bar{E}_{f,D}$.
- For every edge $(u,v_f)\in E_{f,D}$, add the directed edge (u,v_f) to $\bar{E}_{f,D}$.

For an arbitrary communication protocol that repairs the failed node v_f, let $X_{u,v}$, for $(u,v)\in\bar{E}_{f,D}$, be the total number of symbols sent along the edge (u,v) during the complete protocol. Fix an order of the edges in $\bar{E}=\bar{E}_{f,D}$ and let $\bar{X}$ be the vector of $X_{u,v}$'s. Let $\mathcal{P}^*(D)$ be the set of all non-empty subsets of D. Define a binary matrix M of size $(2^d-1)\times|\bar{E}|$ by setting $M_{S,(u,v)}=\mathbb{1}(u\in S\wedge v\in S^c)$, where $S^c=V_{f,D}\backslash S$. The rows of M are characteristic vectors of the cuts (S,S^c). Let $\bar{b}\in\mathbb{R}^{|2^d-1|}$ with $\bar{b}_S=\beta\cdot\min\{d-k+1,|S|\}$ for all $S\in\mathcal{P}^*(D)$.

Proposition 4: For the failed node v_f and helper nodes D, the total communication complexity of repair is bounded below by the solution to the following linear program with $|\bar{E}|$ variables and 2^d-1 constraints:

$$\begin{aligned} \text{minimize}\quad & \mathbf{1}^T\bar{X} \\ \text{subject to}\quad & M\bar{X}\geq\bar{b}, \\ & \bar{X}\geq 0. \end{aligned}$$

Proof: We only need to justify the inequality $M\bar{X}\geq\bar{b}$. For any set $S\in\mathcal{P}^*(D)$, Lemma 1 implies

$$\sum_{\substack{(u,v)\in\bar{E}_{f,D}\\ u\in S, v\in S^c}} X_{u,v}\geq R_S^f\geq\min\{d-k+1,|S|\}\beta.$$

Collecting these inequalities for all $S\in\mathcal{P}^*(D)$, we obtain the claimed set of constraints. □

The key observation here is that the quantity R_A^f in Lemma 1 represents the total outflow of information transmitted from the set of nodes A for the repair, and hence the bounds still hold irrespective of the communication among the nodes in set A.

In the remainder of this section we consider two settings in which the bound of this proposition enables one to prove optimality of communication for recovery while allowing communication between the helper nodes, namely when the failed node has the largest and the smallest possible number of helpers, respectively, as immediate neighbors. In both cases we allow arbitrary communication among the helper set.

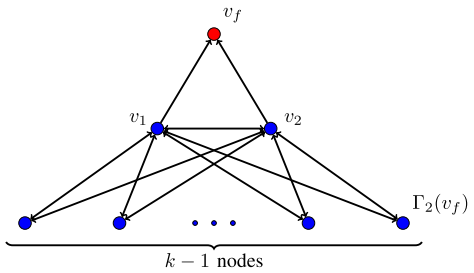

Fig. 3. Repair graph of the node v_f that attains the LP lower bound.

A. The Case of the Complete Graph

This case corresponds to the original repair problem of [6], and the cut-set bound provides the minimum required download per helper node for the repair of a failed node. In this model, the transmitted data of each helper node is a function of its own stored content only. Can communication complexity be reduced if the helper nodes are allowed to exchange information before communicating with the failed node? An easy corollary of Proposition 4 and Lemma 1 implies that in case of MSR codes the answer is negative.

Proposition 5: For the complete graph K_n, the communication complexity is bounded below by $d\beta$ and is achieved by having all the helper nodes directly transmit β symbols to the failed node.

Proof: Consider the assignment of variables of the LP problem X^* with $X^*_{u,v} = \beta\mathbb{1}(v = v_f)$. It is clearly feasible because it corresponds to all the helper nodes transmitting β symbols to the failed node. Indeed, this assignment satisfies the bounds of Lemma 1 and thus also the inequality constraints of Proposition 4. Next we show that X^* is optimal by considering the dual LP problem, which has the form

$$\begin{aligned} \text{maximize} \quad & \bar{b}^T\bar{Y} \\ \text{subject to} \quad & M^T\bar{Y} \le \mathbf{1}, \\ & \bar{Y} \ge 0. \end{aligned}$$

Take the assignment of variables Y^* with $Y^*_S = \mathbb{1}(|S| = 1)$ for all $S \subset D$. Since for two different $S_1 = \{v_1\}$ and $S_2 = \{v_2\}$ any edge $(u, v) \in \bar{E}$ can belong to at most one of the cuts (S_1, S_1^c) or (S_2, S_2^c), we have that $M^TY^* \le \mathbf{1}$. Since $\mathbf{1}^TX^* = \bar{b}^TY^* = d\beta$, we conclude that X^* is indeed optimal. □

B. The Case of Two Neighbors

Assume that the information is encoded with an $[n, k, d = k+1, l]$ MSR code and stored on a graph $G(V, E)$ with $|V| = n$. Consider the repair graph (no longer a tree) shown in Fig.3 with the failed node v_f connected to two helper nodes which connect to the remaining subset of the helper set.

We will prove that for this graph the minimum required communication for repair equals $(d+1)\beta = (k+2)\beta$. To show this, assume that the failed node v_f relies on a set D of $k+1$ helpers for repair, and that it is connected to two of them, denoted v_1 and v_2. Assume further that the $k+1$ helpers span a complete graph K_{k+1}, where k is the dimension of the MSR code used for the encoding of the data. To link this graph to the LP problem of Prop. 4, construct a directed graph by replacing every edge between a pair of helpers with a pair of opposing directed edges, and make a directed edge from each of v_1, v_2 to v_f. Thus, the new set of *directed* edges is

$$\bar{E} = \{((v_i, v_j), v_i, v_j \in D), \text{ and } (v_1, v_f), (v_2, v_f)\}.$$

To construct a primal LP program, assign

$$X^*_{(u,v)} = \begin{cases} \beta & \text{if } u \in \Gamma_2(v_f) \cup \{v_2\}, v = v_1 \\ 2\beta & \text{if } u = v_1, v = v_f \\ 0 & \text{otherwise.} \end{cases}$$

This assignment defines a valid repair protocol, so it's a feasible solution of the LP problem which gives the value of the objective function to be

$$\mathbf{1}^TX^* = (d-1)\beta + 2\beta = (d+1)\beta \tag{22}$$

Construct a dual program $Y^* = (Y^*_S)_S$ by setting

$$Y^*_S = \begin{cases} \frac{1}{d-2} & \text{if } |S| = 2, S \ne \{v_1, v_2\} \\ 0 & \text{otherwise.} \end{cases}$$

The vector Y^* is a feasible assignment of the dual program. To show this, consider an edge (u, v) with $u \in D$. Our argument depends on whether $v \in D$ or $v = v_f$. In the first case, the row of M^T contains $d-2$ ones which correspond to the cuts in $G_{f,D}$ that contain the edge (u, v) (there are exactly $d-2$ such cuts), so this row times Y^* equals one. If $v = v_f$, then u is either v_1 or v_2. Say it is v_1, then the row (v_1, v_f) contains $d-1$ ones which correspond to the cuts that contain the edge (v_1, v_f). Further, $Y^* = 0$ in the coordinate $S = \{v_1, v_2\}$, so the nonzeros in the vector Y^* and the $(M^T)_{v_1,v_f}$ overlap in $d-2$ places, again satisfying the constraints of the dual program.

To compute the value of the dual problem, note that $Y^* \ne 0$ in $\binom{d}{2} - 1 = \frac{(d-2)(d+1)}{2}$ coordinates, and the corresponding entries in $\bar{b}$ are set to 2β. Thus, $\bar{b}^TY^* = \frac{(d-2)(d+1)}{2} \cdot \frac{1}{d-2} \cdot 2\beta = (d+1)\beta$, which equals the value of the primal problem, proving that the repair protocol defined by it yields the minimum possible communication complexity. Finally, we argue that if there does not exist a repair protocol that performs better in terms of complexity when the helper nodes form the complete graph, then there cannot exist a repair protocol that performs better for any sub-graph of the complete graph.

The repair bandwidth $(d+1)\beta$ can be attained by sending the data from all the helper nodes but v_1 to the node v_1, combining them and passing the result to v_f (which is the IP repair discussed earlier). This repair protocol does not involve two-way communication between the neighboring helper nodes. In the appendix we construct another protocol that does involve it, while still having the same communication complexity of repair.

C. Can the Repair Bandwidth Be Lower Than the IP Protocol?

So far we have not identified cases in which communication among the helper nodes reduces the complexity of repair compared to the IP protocol. That this may be possible is

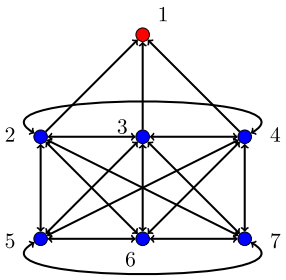

Fig. 4. The repair graph.

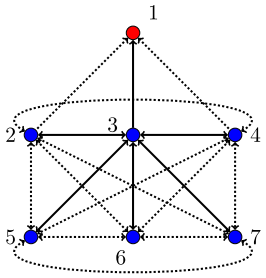

Fig. 5. Repair using IP.

demonstrated in the next numerical example in which the value of the linear program is below the repair bandwidth of the IP scheme. Note that we still stop short of constructing an actual node repair scheme that would have this value of the communication complexity.

Consider the graph $G_{f,D}$ in Fig. 4 with three direct neighbors of the failed node, and let $d = 6, k = 5$. We assume that $v_f = 1$, and it is directly connected to helper nodes 2, 3 and 4. The six helper nodes form a complete graph K_6. The IP technique can achieve the complexity of 7 units by transmitting along the spanning tree shown in Fig. 5 and performing IP (combining the data) at node 3.

To define the LP problem we construct a directed graph $\bar{G}_{f,D}$ as explained in the beginning of this section. The linear program of Prop. 4 in this case has value 6.75 and the assignments of variables are: the primal program

$$X^*_{(u,v)} = \begin{cases} 1 & \text{if } u \in \{2,3,4\}, v = 1, \\ 0.5 & \text{if } u, v \in \{5,6,7\}, u \neq v, \\ 0.25 & \text{if } u \in \{5,6,7\}, v \in \{2,3,4\}, \\ 0 & \text{otherwise;} \end{cases}$$

the dual program:

$$Y^*_S = \begin{cases} 0.125 & \text{if } |S| = 2, S \subset \{2,3,4\}, \\ 0.25 & \text{if } |S| = 2, S \not\subset \{2,3,4\}, \\ 0 & \text{otherwise} \end{cases}$$

Many more similar examples can be constructed for small-size graphs.

VI. 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 sampled from the $\mathcal{G}_{n,p}$ ensemble, where $0 < p < 1$. We denote such a random element from the ensemble as $\mathbb{G}_{n,p}$. 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 finding relations between the parameters 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. Throughout this section we assume that each helper node provides one field symbol for the repair of the (single) failed node.

We will assume that $p \gg \frac{\log n}{n}$ because if $\mathbb{G}_{n,p}$ 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, $\mathbb{P}_{\mathcal{G}_{n,p}}(\deg(v_f) \geq d) = \sum_{i=d}^{n} \binom{n}{i} p^i (1-p)^{n-i}$, 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.

Throughout we will assume that k and $d \in \Theta(n)$, and that $\chi(n) := d - k$ is $o(n)$, i.e., d is close to k. For simplicity (without loss of generality) we also assume that each helper node provides only one symbol of $\mathbb{F}$ for the repair of the failed node.

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

Lemma 3 ([3], p. 50; [8], Sec.7.1): (i) If $p^2 n - 2\log n \to \infty$, and $n^2(1-p) \to \infty$, then

$$\mathbb{P}(\mathsf{diam}(\mathbb{G}_{n,p}) = 2) \to 1.$$

(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 $\mathbb{P}(\mathsf{diam}(\mathbb{G}_{n,p}) = t) \to 1$.

Lemma 4 ([5], Lemma 3): Suppose that $p \geq \frac{\log n}{n}$. For any $\epsilon > 0$ and all $i = 1, \ldots, \lfloor \log n \rfloor$

$$\mathbb{P}(|\Gamma_i(x)| \leq (1+\epsilon)(np)^i) \geq 1 - 1/\log^2 n \tag{23}$$
$$\mathbb{P}(|N_i(x)| \leq (1+2\epsilon)(np)^i) \geq 1 - 1/\log^2 n. \tag{24}$$

A. Repair Threshold

Let t be a fixed integer. We say that *t-layer repair* of the failed node v *is possible* if

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

and call the minimum t for which this holds the *threshold depth* for repair. Note that such a t is a function of n and p. The next proposition establishes a threshold for t-layer repair in terms of p.

Proposition 6: If

$$(np)^{t-1} = o(n), \quad p^t n^{t-1} - 2\log n \to \infty, \tag{25}$$

then t is the threshold depth for repair.

Proof: To show that t-layer repair is possible, we observe that from Lemma 3, $\mathbb{P}(\mathsf{diam}(\mathbb{G}_{n,p}) = t) \to 1$ for all $t \geq 2$. 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 (24) for any $\epsilon > 0$

$$\mathbb{P}(|N_{t-1}(v)| \le (1+2\epsilon)(np)^{t-1}) \to 1. \tag{26}$$

Since d is a linear function of n, we have the inclusion

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

Together with (26) this implies that $\mathbb{P}(|N_{t-1}(v)|/n \ge \gamma) \to 0$ for any $\gamma > 0$. □

Remark: Given t, the conditions (25) are satisfied if

$$n^{-(t-1)/t} g(n) \ll p(n) \ll n^{-(t-2)/(t-1)}, \tag{27}$$

where $g(n) \gg (2\log n)^{1/t}$. Rephrasing Prop. 6, we could say that for a given repair depth t the probability $p(n)$ that satisfies conditions (27) is a threshold for repair of depth t in the ensemble $\mathcal{G}_{n,p}$.

B. 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 (25) hold for t, n, and p. Recall that by our assumption, $l = d - k + 1$.

Proposition 7: The repair bandwidth β_{AF} satisfies

$$\mathbb{P}(\beta_{\text{AF}} \ge td - o(n)) \to 1$$

where t is the threshold for repair as given by (25).

Proof: Rewriting the expression for β_{AF} in (9), we obtain

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

Define the events $E_i = \{|\Gamma_i(v_f)| \le (1+\epsilon)(np)^i\}$ and notice that $E := \cap_{i=1}^{t-1} E_i \subseteq \{\beta_{\text{AF}} \ge (td - o(n))\}$. From Lemma 4 we know that $\mathbb{P}(E_i^c) \le 1/\log^2 n$ for all i, and thus

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

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

Remark: This proposition implies that for large n, most of the helper nodes are at distance t from the failed node. Note that, assuming (25), Lemma 4 along with Lemma 8 in [5] imply that the size of 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. 7 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.

The next proposition gives further insights into the relationship between β_{AF} and t.

Proposition 8: Let $d = \delta n, 0 < \delta < 1$, let $\kappa(n)$ be a function of n such that $\underline{c} \le \kappa(n)/n \le \bar{c}$ for some constants $\underline{c}, \bar{c}$ and all sufficiently large n, and let t be the threshold for repair as given by (25). We have

$$\mathbb{P}(\beta_{\text{AF}} \le \kappa(n)) \to \begin{cases} 0 & \text{if } t > \bar{c}/\delta \\ 1 & \text{if } t \le \underline{c}/\delta. \end{cases}$$

Proof: To prove the first claim in the proposition, compute

$$\begin{aligned}
&\mathbb{P}(\beta_{\text{AF}} \le \kappa(n)) \le \mathbb{P}(\beta_{\text{AF}} \le \overline{c}n) \\
&= \mathbb{P}\Big(td - \sum_{i=1}^{t-1}(t-i)|\Gamma_i(v_f)| \le \overline{c}n\Big) \\
&= \mathbb{P}\Big(\sum_{i=1}^{t-1}(t-i)|\Gamma_i(v_f)| \ge (t\delta - \overline{c})n\Big) \\
&\le \mathbb{P}\Big(\sum_{i=1}^{t-1}(t-i)|\Gamma_i(v_f)| \ge (t\delta - \overline{c})n \;\Big|\; E\Big)\mathbb{P}(E) + \mathbb{P}(E^c),
\end{aligned}$$

where the event E is defined above in Prop. 7. Conditional on E we have $\sum_{i=1}^{t-1}(t-i)(1+\epsilon)(np)^i = \Theta((np)^{t-1})$, and (25) implies that the first term $\to 0$ w.h.p. To complete the proof notice that $\mathbb{P}(E^c) \le t/\log^2 n \to 0$.

For the second claim, observe that

$$\mathbb{P}(\beta_{\text{AF}} \le \kappa(n)) \ge \mathbb{P}(\beta_{\text{AF}} \le \underline{c}n) \ge \mathbb{P}(\beta_{\text{AF}} \le td) = 1$$

concluding the proof. □

Now let us show that the graph-based repair as defined in (15) or (19) with high probability has smaller repair bandwidth.

Theorem 2: Let t be the threshold given in Prop. 6. Let $\chi(n)$ be such that $\chi(n)n^{s-1}p^s \to 0$ where $s \le t-1$ is the largest integer for which this condition holds. Then $\mathbb{P}(\beta_{\text{IP}} \le (t-s)d + o(n)) \to 1$.

Remark: Since $pn \to \infty$, it is easy to check that the assumptions of the theorem are non-vacuous, i.e., the largest s satisfying the condition exists and is well defined.

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 (15).

Then, for the failed node v_f, we have

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

Proceeding similarly to the proof of Prop. 7, we obtain

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

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

To conclude, we have shown a strict separation between the typical communication cost of node recovery using the IP repair procedure and the graph-agnostic AF protocol when the number of helpers d is only slightly more than k. Note the following simple corollary:

Corollary 1: Let t be the threshold given in Prop. 6. For $\chi(n) = O(1)$,

$$\mathbb{P}(\beta_{\text{IP}} \leq d + o(n)) \to 1.$$

Proof: By (25), for $\chi(n) = O(1)$, the condition $\chi(n) n^{s-1} p^s \to 0$ is satisfied for $s = t - 1$. □

Corollary 1 supports the following intuition. Since $\chi(n)$ is a constant, nodes in all layers but the last can do a very high amount of compression and hence the contribution of those layers to the total communication complexity becomes insignificant; the complexity is primarily determined by the number of helper nodes in the last layer which is approximately d. This implies that even in random graphs where we do not have direct connectivity among all the helper nodes and the failed node, it is possible to bring down the communication complexity to the same order as that of the case of direct connectivity using IP.

The above theorem and corollary suggest that the communication complexity is primarily controlled by the two parameters $p(n)$ and $\chi(n)$. One can ask the question, for what values of these parameters, does the complexity become significantly higher than that of the complexity of repair under full connectivity, i.e., d. In other words, we wish to study the behavior of $\beta^* - d$ where β^* is the minimum complexity over all possible repair schemes.[3] In this regard, Corollary 1 says that for $\chi(n) = O(1)$, we have $\beta^* - d = o(n)$ with high probability. We will now show that for sparse graphs with high probability the repair becomes significantly more complex than sending d symbols. The following theorem quantifies this claim. Its proof relies on Lemma 1 together with another lemma from [5].

Lemma 5 ([5], Lemma 2): Suppose that $p > \frac{c \log n}{n}$ for a constant $c \leq 2$. Then with probability at least $1 - o(\frac{1}{n})$, we have for all $1 \leq i \leq n$

$$|\Gamma_i(x)| \leq \frac{9}{c}(np)^i \tag{28}$$

$$|N_i(x)| \leq \frac{10}{c}(np)^i. \tag{29}$$

Theorem 3: For $p(n) = o(\frac{\chi(n)}{n})$, we have $\mathbb{P}(\beta^* - d = \Theta(n)) \to 1$.

Proof: Given $p(n)$, let t be the threshold for repair as given in Prop. 6. Clearly, any helper node $v \in N_t(v_f)$ needs to transmit at least one unit of information, so $\beta^* \geq d - |N_{t-1}(v_f)|$. Let $F_n := \{\Gamma_{t-1}(v_f) \leq \frac{9}{c}(np)^{t-1}\}$. Now consider a node $v \in N_{t-1}(v_f)$. From Lemma 5, the immediate neighborhood of this node satisfies

$$\mathbb{P}\Big(|\Gamma_1(v)| \leq \frac{9}{c} np\Big) \geq 1 - o\Big(\frac{1}{n}\Big)$$

for some constant $c \leq 2$. Let $D(v)$ be the *immediate* descendants of node v in the repair tree. For every $\delta > 0$, there exists an n_1 such that $|D(v)| \leq |\Gamma_1(v)| \leq \frac{9}{c} np$ with probability at least $1 - \frac{\delta}{n}$ for every $n \geq n_1$. Further, since $np = o(\chi(n))$, for every $\epsilon > 0$, there exists an n_2 such that $np \leq \epsilon \chi(n)$ for every $n \geq n_2$. Combining these two statements, we claim that the event $E_{v,\epsilon,n} := \{|D(v)| \leq \epsilon \chi(n)\}$ satisfies

$$\mathbb{P}(E^c_{v,\epsilon,n}) \leq \delta/n \text{ for all } \epsilon, \delta > 0, n \geq \max(n_1, n_2).$$

Since this is true for all $v \in \Gamma_{t-1}(v_f)$, we have

$$\begin{aligned}\mathbb{P}(\cup_{v \in \Gamma_{t-1}(v_f)} E^c_{v,\epsilon,n}) &= \mathbb{P}(\cup_{v \in \Gamma_{t-1}(v_f)} E^c_{v,\epsilon,n} | F_n)\mathbb{P}(F_n) \\ &\quad + \mathbb{P}(\cup_{v \in \Gamma_{t-1}(v_f)} E^c_{v,\epsilon,n} | F^c_n)\mathbb{P}(F^c_n) \\ &\leq \frac{9\delta}{c} \frac{(np)^{t-1}}{n} + o\Big(\frac{1}{n}\Big) \to 0,\end{aligned}$$

where the last step follows because by the definition of the threshold t, $(np)^{t-1} = o(n)$. This implies that $\mathbb{P}(\cap_{v \in \Gamma_{t-1}(v_f)} E_{v,\epsilon,n}) \to 1$ for all $\epsilon > 0$. Now by Lemma 1, for $|D(v)| \leq \chi(n)$ the outflow of communication from the set $D(v) \cup \{v\}$ has to be at least $|D(v)| + 1$ and by the above analysis this is true for all $v \in \Gamma_{t-1}(v_f)$ with high probability. This implies that with high probability

$$\begin{aligned}\beta^* &\geq d - |N_{t-1}(v_f)| + \sum_{v \in \Gamma_{t-1}(v_f)} (|D(v)| + 1) \\ &\geq d - |N_{t-1}(v_f)| + \Gamma_t(v_f) = 2(d - |N_{t-1}(v_f)|).\end{aligned}$$

Finally, noting that $\mathbb{P}(|N_{t-1}(v_f)| = o(n)) \to 1$ gives the desired claim. □

This theorem gives a sufficient condition for the separation of complexity of repair on a complete graph and a sparse random graph.

C. Random Regular Graphs

In this section we briefly address node repair on random regular graphs. We single out this ensemble from a multitude of other options because it is conceivable that the architecture of the storage system places the same number of servers in close proximity to any single server, and this is modeled by a regular graph. Let $\mathcal{G}_{n,r}$ be the set of all r-regular n-vertex graphs with a uniform distribution on it. We denote a random element from this ensemble by $\mathbb{G}_{n,r}$. Assume again that the data is encoded with an (n, k, d, l) MSR code, and the coordinates $C_1, \ldots, C_n$ of the codeword are placed on the vertices $v_1, \ldots, v_n$.

We will derive conditions on the parameters $k = k(n), d = d(n)$ and $r = r(n)$ such that as $n \to \infty$, with high probability the graph-based repair process is advantageous over the AF strategy. We again assume that $d = \Theta(n)$. Denote by $\mathcal{G}_{n,m}$ the ensemble of graphs with n vertices and m edges and let $\mathbb{G}_{n,m}$ be a random graph sampled from it.

For the purposes of node repair we need the graph to be connected. In [2], Bollobás showed that $\mathbb{G}_{n,r}$ is r-connected with high probability.

Recall that a property of graphs is called *increasing* if it is inherited from a subgraph to any graph that contains it. The following equivalence between properties of $\mathbb{G}_{n,m}$ and $\mathbb{G}_{n,r}$ will be used below.

[3]Our arguments rely on the information-theoretic lower bounds, so they indeed apply to all possible repair schemes.

Lemma 6 ([8], Corollary 10.11): Let $\mathcal{L}$ be an increasing property of graphs such that $\mathbb{G}_{n,m}$ satisfies $\mathcal{L}$ with high probability for some $m = m(n)$, where $n \log n \ll m \ll n^2$. Then $\mathbb{G}_{n,r}$ satisfies $\mathcal{L}$ with high probability for $r = r(n) \sim \frac{2m}{n}$.

The following proposition is a counterpart to Prop. 6 for random regular graphs. Here the definition of the threshold depth of repair is the same as in Sec. VI-A.

Proposition 9: Let $d = \delta n, 0 < \delta < 1$ and let t be a fixed integer. Then t is the threshold depth for repair if

$$r^{t-1} = o(n), \quad \frac{r^t}{n} - 2\log n \to \infty. \tag{30}$$

Proof: For finite t, $|N_{t-1}(v)| \le \sum_{i=1}^{t-1} r^i = \Theta(r^{t-1}) = o(n) \ll d$ and so $(t-1)$-layer repair is not possible.

For the other direction, let $r(n)$ satisfy relations (30). Take $p(n) = \frac{r(n)}{n}$, then by Prop. 6 $\mathbb{G}_{n,p}$ satisfies t-layer repair with high probability. Recall a basic fact that the ensembles $\mathcal{G}_{n,p}$ and $\mathcal{G}_{n,m}$ are equivalent for all monotone properties (i.e., graphs sampled from them either both have the property w.h.p. or they both do not). Therefore, the graph $\mathbb{G}_{n,m}$ with $m(n) \sim \frac{n^2}{2} p(n) = \frac{nr(n)}{2}$ affords t-layer repair with high probability. Now, satisfying t-layer repair is a monotone increasing graph property and so by Lemma 6 we have that $\mathbb{G}_{n,r}$ affords t-layer repair with high probability. □

With the threshold conditions established, counterparts of Prop. 7, 8 and Theorem 2 can be proved for $\mathbb{G}_{n,r}$ by simply replacing np with r and proceeding along similar arguments (which are in fact simpler because the neighborhood sizes of a vertex afford uniform bounds). In the following statements, given without proofs, t is the repair threshold as defined in (30).

Proposition 10: The repair bandwidth β_{AF} satisfies

$$\mathbb{P}(\beta_{\text{AF}} \ge td - o(n)) \to 1.$$

Proposition 11: Let $d = \delta n, 0 < \delta < 1$ and let $\kappa(n)$ be a function of n such that $\underline{c} \le \kappa(n)/n \le \bar{c}$ starting with some n. Then

$$\mathbb{P}(\beta_{\text{AF}} \le \kappa(n)) \to \begin{cases} 0 & \text{if } t > \bar{c}/\delta \\ 1 & \text{if } t \le \underline{c}/\delta. \end{cases}$$

for $t > \bar{c}/\delta$ we have $\mathbb{P}(\beta_{\text{AF}} \le \kappa(n)) \to 0$.

Theorem 4: Let $d - k = \chi(n)$ be a function of n such that $\frac{\chi(n) r(n)^s}{n} \to 0$ where $s \le t-1$ is the largest integer for which this condition holds. Then $\mathbb{P}(\beta_{\text{IP}} \le (t-s)d + o(n)) \to 1$.

Similarly to the case of $\mathcal{G}_{n,p}$, this shows a strict separation between the typical communication cost of node recovery using the IP repair procedure and the graph-agnostic AF protocol under a certain assumption on $\chi(n)$. A theorem that parallels Theorem 3 can be also easily established (note that the bounds of the form given in Lemma 5 for regular graphs come for free).

VII. Conclusion

In this paper we posed and advanced the problem of erasure correction (node repair) when the elements of encoded information are placed on the nodes of a graph, adopting the total amount of communication for repair as the figure of merit in the analysis. The main difference of this problem from the standard setting of regenerating codes stems from the fact that most helpers are not directly connected to the failed node, and the information transmitted by them can be processed by the intermediate nodes or combined with the contents of these nodes. We showed that the intermediate processing scheme can be implemented by linear MSR codes, attaining the general lower bounds on complexity derived in the paper. These results were also extended to the case of multiple failed nodes. We also established a framework for the analysis of repair schemes when the helpers communicate among themselves before contributing data toward the repair task, and gave simple examples when the arising complexity bounds are attained with equality. Finally, we studied the repair problem when the underlying graph is random, establishing bounds on the edge probability under which the intermediate processing scheme provides complexity savings compared to simple relaying.

Among the problems that so far have resisted analysis is the case when some of the helper nodes provide incorrect information. Since the erroneous information is propagated along the edges and potentially combined with the contents of the intermediate nodes, the repair procedures proposed in this paper do not support node recovery. This is unlike the case of complete connectivity, where optimal repair is possible even in the presence of errors [25]. Another version of repair with errors assumes that the edges form noisy channels, so the information propagated along them is received as realizations of random variables. A simple way to address this problem suggests to add redundancy to the data transmitted over the edges, and it combines channel coding and the repair task. Optimizing the tradeoffs that arise as a result presents an open problem. Yet another challenge is to construct codes and nontrivial repair protocols when the helper nodes communicate among themselves in the process of repair, extending the approach of Sec. V.

Appendix

We present an encoding/repair scheme for the example in Sec. V-B, Fig.3 that enables recovery of the contents of the node v_f which assumes that the nodes in $\Gamma_1(v_f)$ exchange information before passing the repair data to the node v_f. The construction shares some features of cooperative repair, and it relies on a code family constructed earlier for the case of the complete graph K_n [25, Sec.IV]. Namely, suppose that the information is encoded with an $[n, k, d = k+1, l = 2^n]$ MSR array code $\mathcal{C}$ and the codeword coordinates are placed on the vertices of a graph $G(V, E)$ with $|V| = n$. Suppose further that the repair graph of the failed node v_f is as shown in the figure. In accordance with [25], Construction 2, we will assume that each helper node provides $\beta = 2^{n-1}$ symbols for the repair of v_f.

Let $C = (c_1, c_2, \ldots, c_n) \in \mathcal{C}$ be a codeword with $c_i = (c_{i,0}, c_{i,1}, \ldots, c_{i,l-1}) \in \mathbb{F}^l$. The code is defined by the following parity-check equations:

$$\sum_{i=1}^{n} \lambda_{i,a_i}^t c_{i,a} = 0 \quad \text{for all } a \in \{0, 1, \ldots, l-1\}, \\ t \in \{0, 1, \ldots, n-k-1\},$$

where $(a_1, a_2, \ldots, a_n)$ is the binary representation of a. Below we assume that $v_f = 1$, that $\Gamma_1(1) = \{2, 3\}$. and that $\Gamma_2(1) = \{4, 5, \ldots, k+2\}$. For a string $s = (s_1, s_2, \ldots, s_{n-k-2}) \in \{0,1\}^{n-k-2}$, consider the set of 2^{k+2} values of a for which $(a_{k+3}, a_{k+4}, \ldots, a_n) = s$. Isolate a subset of this set by fixing a string $\hat{s} = (\hat{s}_1, \hat{s}_2, \ldots, \hat{s}_{k-1}) \in \{0,1\}^{k-1}$ and collecting only those values of a for which $(a_4, a_5, \ldots, a_{k+2}) = \hat{s}$. Having fixed s and $\hat{s}$, we are left with 8 parity check equations which can be labeled by a binary vector $\tilde{s} \in \{0,1\}^3$:

$$\begin{aligned}\lambda^t_{1,\tilde{s}_1} c_{1,(\tilde{s},\hat{s},s)} + \lambda^t_{2,\tilde{s}_2} c_{2,(\tilde{s},\hat{s},s)} + \lambda^t_{3,\tilde{s}_3} c_{3,(\tilde{s},\hat{s},s)} \\ + \sum_{i=4}^{k+2} \lambda^t_{i,\hat{s}_{i-3}} c_{i,(\tilde{s},\hat{s},s)} + \sum_{i=k+3}^{n} \lambda^t_{i,s_{i-k-2}} c_{i,(\tilde{s},\hat{s},s)} = 0 \\ \forall\ \tilde{s} \in \{0,1\}^3,\ \ t \in \{0,1,\ldots,n-k-1\} \qquad (31)\end{aligned}$$

For fixed $\hat{s}$ and s, the λ's in the last two sums the same in all the equations. Define

$$\begin{aligned}\mu^{(\hat{s},s)}_{2,1,i} &= c_{i,(000,\hat{s},s)} + c_{i,(010,\hat{s},s)} + c_{i,(100,\hat{s},s)} \\ \mu^{(\hat{s},s)}_{3,1,i} &= c_{i,(000,\hat{s},s)} + c_{i,(100,\hat{s},s)} + c_{i,(101,\hat{s},s)} \\ \mu^{(\hat{s},s)}_{2,2,i} &= c_{i,(001,\hat{s},s)} + c_{i,(011,\hat{s},s)} + c_{i,(111,\hat{s},s)} \\ \mu^{(\hat{s},s)}_{3,2,i} &= c_{i,(011,\hat{s},s)} + c_{i,(110,\hat{s},s)} + c_{i,(111,\hat{s},s)}\end{aligned}$$

For $i \in \{4, 5, \ldots, k+2\}$ the helper node i sends $\mu^{(\hat{s},s)}_{2,1,i}, \mu^{(\hat{s},s)}_{2,2,i}$ to node 2 and $\mu^{(\hat{s},s)}_{3,1,i}, \mu^{(\hat{s},s)}_{3,2,i}$ to node 3. Additionally node 3 sends $\mu^{(\hat{s},s)}_{2,1,3}, \mu^{(\hat{s},s)}_{2,2,3}$ to node 2 and node 2 sends $\mu^{(\hat{s},s)}_{3,1,2}, \mu^{(\hat{s},s)}_{3,2,2}$ to node 3.

Node 2, having $\mu^{(\hat{s},s)}_{2,1,i}$ for all $i \in \{3, \ldots, k+2\}$, can recover $c_{1,(000,\hat{s},s)} + c_{1,(010,\hat{s},s)}, c_{1,(100,\hat{s},s)}$. To see this, sum Eqns. (31) for $\tilde{s} \in \{000, 010, 100\}$ to obtain

$$\begin{aligned}\lambda^t_{1,0}(c_{1,(000,\hat{s},s)} + c_{1,(010,\hat{s},s)}) + \lambda^t_{1,1} c_{1,(100,\hat{s},s)} \\ + \lambda^t_{2,0}(c_{2,(000,\hat{s},s)} + c_{2,(100,\hat{s},s)}) + \lambda^t_{2,1} c_{2,(010,\hat{s},s)} \\ + \lambda^t_{3,0} \mu^{(\hat{s},s)}_{2,1,3} + \sum_{i=4}^{k+2} \lambda^t_{i,\hat{s}_{i-3}} \mu^{(\hat{s},s)}_{2,1,i} + \sum_{i=k+3}^{n} \lambda^t_{i,s_{i-k-2}} \mu^{(\hat{s},s)}_{2,1,i} = 0.\end{aligned}$$

The multiplies of the λ's in this equation form a codeword of an $(n+2, k+2 = d+1)$ Reed-Solomon code. Node 2 collects $\mu^{(\hat{s},s)}_{2,1,i}$ for all $i \in \{3, \ldots, k+2\}$ and it already knows $(c_{2,(000,\hat{s},s)} + c_{2,(100,\hat{s},s)}), c_{2,(010,\hat{s},s)}$, and so it can recover $(c_{1,(000,\hat{s},s)} + c_{1,(010,\hat{s},s)})$ and $c_{1,(100,\hat{s},s)}$. Similarly, it can be shown that with $\mu^{(\hat{s},s)}_{2,2,i}$ for all $i \in \{3, \ldots, k+2\}$, node 2 can recover $(c_{1,(001,\hat{s},s)} + c_{1,(011,\hat{s},s)})$ and $c_{1,(111,\hat{s},s)}$.

Node 3, using $\mu^{(\hat{s},s)}_{3,1,i}$ and $\mu^{(\hat{s},s)}_{3,2,i}$ for all $i \in \{2, 4, 5 \ldots, k+2\}$, can recover $(c_{1,(100,\hat{s},s)} + c_{1,(101,\hat{s},s)})$ and $c_{1,(000,\hat{s},s)}$ and $(c_{1,(110,\hat{s},s)} + c_{1,(111,\hat{s},s)})$ and $c_{1,(011,\hat{s},s)}$. Nodes 2 and 3 send these recovered linear combinations to the failed node, which can recover all $c_{1,(\tilde{s},\hat{s},s)}$ for $\tilde{s} \in \{0,1\}^3$. Finally this is done for all $\hat{s}$ and s, and this recovers the entire of node 1.

Communication Complexity: Each helper node $i \in \{4, 5, \ldots, k+2\}$ sends two symbols to node 2 and 2 symbols to node 3 for each fixed $\hat{s}, s$. Hence they transmit 4 symbols each resulting in a total transmission of $4(k-1)$ for each fixed $\hat{s}, s$. Similarly node 3 transmits 2 symbols to node 2 and node 2 transmits 2 symbols to node 3 for each fixed $\hat{s}, s$. Finally node 2 and node 3 total transmit 8 symbols to node 1 for each fixed $\hat{s}, s$. This is repeated for every possible $\hat{s}$ and s. Hence the total communication complexity is

$$\begin{aligned}B &= (4(k-1) + 4 + 8) \cdot 2^{k-1} \cdot 2^{n-k-2} \\ &= 2^{n-1}k + 2^n = \beta(d-1) + l = (d+1)\beta\end{aligned}$$

(since $l = 2\beta$). This matches the communication complexity of repair attainable with an IP protocol of Sec. V-B. □

REFERENCES

[1] A. Blasiak, R. Kleinberg, and E. Lubetzky, "Broadcasting with side information: Bounding and approximating the broadcast rate," *IEEE Trans. Inf. Theory*, vol. 59, no. 9, pp. 5811–5823, Sep. 2013.

[2] B. Bollobás, "A probabilistic proof of an asymptotic formula for the number of labelled regular graphs," *Eur. J. Combin.*, vol. 1, no. 4, pp. 311–316, Dec. 1980.

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

[4] Z. Chen and A. Barg, "Explicit constructions of MSR codes for clustered distributed storage: The rack-aware storage model," *IEEE Trans. Inf. Theory*, vol. 66, no. 2, pp. 866–879, Sep. 2020.

[5] F. Chung and L. Lu, "The diameter of sparse random graphs," *Adv. Appl. Math.*, vol. 26, no. 4, pp. 257–279, May 2001.

[6] 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, Sep. 2010.

[7] O. Elishco and A. Barg, "Recoverable systems," 2020, *arXiv:2010.00589*.

[8] A. Frieze and M. Karońsky, *Introduction to Random Graphs*. Cambridge, U.K.: Cambridge Univ. Press, 2016.

[9] M. Gadouleau, "Finite dynamical systems, hat games, and coding theory," *SIAM J. Discrete Math.*, vol. 32, no. 3, pp. 1922–1945, Jan. 2018.

[10] M. Gerami and M. Xiao, "Exact optimized-cost repair in multi-hop distributed storage networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2014, pp. 4120–4124.

[11] 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, Oct. 2017.

[12] H. Hou, P. P. C. Lee, K. W. Shum, and Y. Hu, "Rack-aware regenerating codes for data centers," *IEEE Trans. Inf. Theory*, vol. 65, no. 8, pp. 4730–4745, Aug. 2019.

[13] J. Lu, X. Guang, and F.-W. Fu, "Distributed storage over unidirectional ring networks," in *Proc. IEEE Int. Symp. Inf. Theory Appl. (ISITA)*, Oct. 2014, pp. 368–372.

[14] A. Mazumdar, "Storage capacity of repairable networks," *IEEE Trans. Inf. Theory*, vol. 61, no. 11, pp. 5810–5821, Nov. 2015.

[15] A. Mazumdar, A. McGregor, and S. Vorotnikova, "Storage capacity as an information-theoretic vertex cover and the index coding rate," *IEEE Trans. Inf. Theory*, vol. 65, no. 9, pp. 5580–5591, Sep. 2019.

[16] V. Ramkumar, M. Vajha, S. B. Balaji, M. N. Krishnan, B. Sasidharan, and P. Kumar, "Codes for distributed storage," in *Concise Encyclopedia Coding Theory*, W. Huffman, J.-L. Kim, and P. Solé, Eds. Boca Raton, FL, USA: CRC Press, 2021, pp. 735–762, ch. 31.

[17] 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, Aug. 2011.

[18] 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, Apr. 2017.

[19] N. Raviv, I. Tamo, and E. Yaakobi, "Private information retrieval in graph-based replication systems," *IEEE Trans. Inf. Theory*, vol. 66, no. 6, pp. 3590–3602, Jun. 2020.

[20] 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, Mar. 2012.
[21] K. W. Shum and Y. Hu, "Cooperative regenerating codes," *IEEE Trans. Inf. Theory*, vol. 59, no. 11, pp. 7229–7258, Nov. 2013.
[22] 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, Apr. 2014.
[23] 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, Oct. 2017.
[24] M. Ye, "New constructions of cooperative MSR codes: Reducing node size to $exp(O(n))$," *IEEE Trans. Inf. Theory*, vol. 66, no. 12, pp. 7457–7464, Dec. 2020.
[25] 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, Apr. 2017.
[26] M. Ye and A. Barg, "Cooperative repair: Constructions of optimal MDS codes for all admissible parameters," *IEEE Trans. Inf. Theory*, vol. 65, no. 3, pp. 1639–1656, Mar. 2018.

Adway Patra (Graduate Student Member, IEEE) received the bachelor's degree in telecommunications from Jadavpur University, Kolkata, India, and the master's degree in electrical communications engineering from the Indian Institute of Science, Bangalore, India. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, University of Maryland, College Park, MD, USA. His research interests include coding theory, information theory, and their applications to real world problems.

Alexander Barg (Fellow, IEEE) is currently a Professor with the Department of Electrical and Computer Engineering and the Institute for Systems Research (ISR), University of Maryland, College Park, MD, USA. His research interests include information and coding theory, applied probability, and algebraic combinatorics. He received the 2015 Information Theory Society Paper Award. He was a Plenary Speaker at the 2016 IEEE International Symposium on Information Theory, Barcelona, Spain. From 2018 to 2019, he served as the Editor-in-Chief for this IEEE TRANSACTIONS. He is currently the Editor-in-Chief of *Foundations and Trends in Communications and Information Theory*.