

# A Survey on Network Codes for Distributed Storage

*In distributed storage systems where reliability is maintained using erasure coding, network codes can be designed to meet specific requirements.*

By ALEXANDROS G. DIMAKIS, *Member IEEE*, KANNAN RAMCHANDRAN, *Fellow IEEE*, YUNNAN WU, *Member IEEE*, AND CHANGHO SUH, *Student Member IEEE*

**ABSTRACT** | Distributed storage systems often introduce redundancy to increase reliability. When coding is used, the *repair problem* arises: if a node storing encoded information fails, in order to maintain the same level of reliability we need to create encoded information at a new node. This amounts to a partial recovery of the code, whereas conventional erasure coding focuses on the complete recovery of the information from a subset of encoded packets. The consideration of the repair network traffic gives rise to new design challenges. Recently, network coding techniques have been instrumental in addressing these challenges, establishing that maintenance bandwidth can be reduced by orders of magnitude compared to standard erasure codes. This paper provides an overview of the research results on this topic.

**KEYWORDS** | Distributed storage; erasure coding; interference alignment; multicast; network coding

## I. INTRODUCTION

In recent years, the demand for large-scale data storage has increased significantly, with applications like social networks, file, and video sharing demanding seamless storage, access and security for massive amounts of data. When the deployed storage nodes are individually unreliable, as is the case in modern data centers and peer-to-peer networks, redundancy must be introduced into the system to



**Fig. 1.** A  $(4, 2)$  MDS binary erasure code (evenodd code [10]). Each storage node (box) is storing two blocks that are linear binary combinations of the original data blocks  $A_1, A_2, B_1, B_2$ . In this example, the total stored size is  $M = 4$  blocks. Observe that any  $k = 2$  out of the  $n = 4$  storage nodes contain enough information to recover all the data.

improve reliability against node failures. The simplest and most commonly used form of redundancy is straightforward replication of the data in multiple storage nodes. However, erasure coding techniques can potentially achieve orders of magnitude more reliability for the same redundancy compared to replication (see, e.g., [2]). To realize the increased reliability of coding however, one has to address the challenge of maintaining an erasure encoded representation.

Given two positive integers  $k$  and  $n > k$ , an  $(n, k)$  maximum distance separable (MDS) code can be used for reliability: initially the data to be stored are separated into  $k$  information packets. Subsequently, using the MDS code, these are encoded into  $n$  packets (of the same size) such that any  $k$  out of these  $n$  suffice to recover the original data (see Fig. 1 for an example).

MDS codes are optimal in terms of the redundancy-reliability tradeoff because  $k$  packets contain the minimum amount of information required to recover the original data. In a distributed storage system, the  $n$  encoded packets are stored at different storage nodes (e.g., disks, servers, or peers) spread over a network, and the system can tolerate any  $(n - k)$  node failures without data loss. Note that throughout this paper we will assume a storage system of  $n$  storage nodes that can tolerate  $(n - k)$  node failures and use the idea of subpacketization: each storage node can

Manuscript received November 21, 2009; revised April 14, 2010; accepted October 17, 2010. Date of current version February 18, 2011.

A. G. Dimakis is with the Department of Electrical Engineering, University of Southern California, Los Angeles, CA 90089-2560 USA (e-mail: dimakis@usc.edu).

K. Ramchandran and C. Suh are with the Department of Electrical Engineering and Computer Science, University of California Berkeley, Berkeley, CA 94704 USA (e-mail: kannanr@eecs.berkeley.edu; chsuh@eecs.berkeley.edu).

Y. Wu is with Microsoft Research, Redmond, WA 98052 USA (e-mail: yunnanwu@microsoft.com).

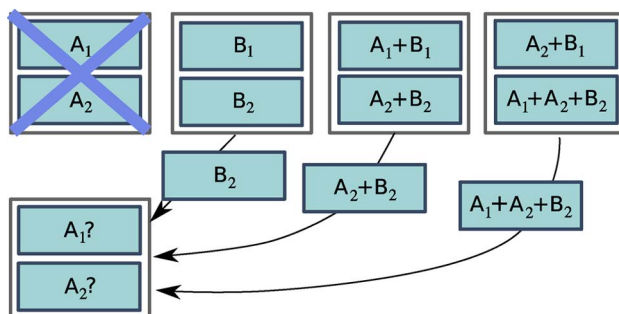
Digital Object Identifier: 10.1109/JPROC.2010.2096170

store multiple subpackets that will be referred to as blocks (essentially using the idea of array codes [10], [11]).

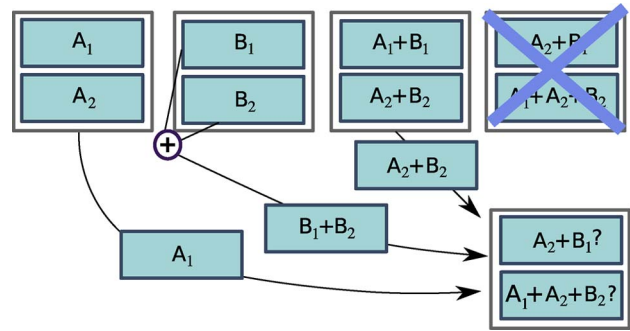
The benefits of coding for storage are well known and there has been a substantial amount of work in the area. Reed–Solomon codes [6] are perhaps the most popular MDS codes and together with the very similar information dispersal algorithm (IDA) [7] have been investigated in distributed storage applications (e.g., [3] and [5]). Fountain codes [8] and low-density parity-check (LDPC) codes [9] are recent code designs that offer approximate MDS properties and fast encoding-and-decoding complexity. Finally, there has been a large body of related work on codes for RAID systems and magnetic recording (e.g., see [10]–[13] and references therein).

In this tutorial, we focus on a new problem that arises when storage nodes are distributed and connected in a network. The issue of repairing a code arises when a storage node of the system fails. The problem is best illustrated through the example of Fig. 2. Assume a file of total size  $\mathcal{M} = 4$  blocks is stored using the  $(4, 2)$  evenodd code of the previous example and the first node fails. A new node (to be called the newcomer) needs to construct and store two new blocks so that the three existing nodes combined with the newcomer still form a  $(4, 2)$  MDS code. We call this the *repair problem* and focus on the required repair bandwidth. Clearly, repairing a single failure is easier than reconstructing all the data: since by assumption any two nodes contain enough information to recover all the data, the newcomer could download four blocks (from any two surviving nodes), reconstruct all four blocks, and store  $A_1, A_2$ . However, as the example shows, it is possible to repair the failure by communicating only three blocks  $B_2, A_2 + B_2, A_1 + A_2 + B_2$ , which can be used to solve for  $A_1, A_2$ .

Fig. 3 shows the repair of the fourth storage node. This can be achieved by using only three blocks [14] but one key difference is that the second node needs to compute a



**Fig. 2. Example of an (exact) repair.** Assume that the first node in the previous storage system failed. The issue is to repair the failure by creating a new node (the newcomer) that still forms a  $(4, 2)$  MDS code. In this example, it is possible to obtain exact repair by communicating three blocks, which is the information-theoretic minimum cutset bound.



**Fig. 3. Repairing the last node:** in some cases, it is necessary for storage nodes to compute functions of their stored data before communicating, as shown in the second node.

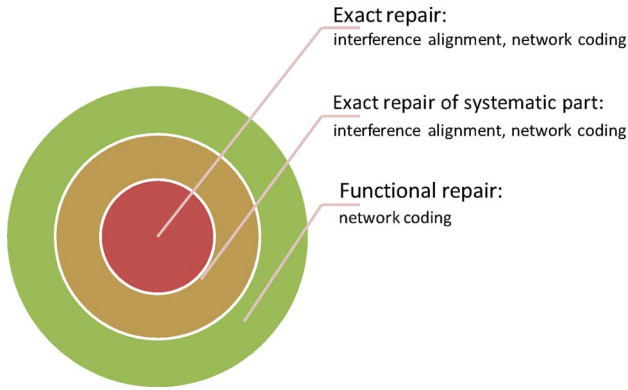
linear combination of the stored packets  $B_1, B_2$  and the actual communicated block is  $B_1 + B_2$ . This shows clearly the necessity of *network coding*, creating linear combinations in intermediate nodes during the repair process. If the network bandwidth is more critical resource compared to disk access, as is often the case, an important consideration is to find what is the minimum required bandwidth and which codes can achieve it.

The repair problem and the corresponding regenerating codes were introduced in [24] and received some attention in the recent literature [25]–[27], [31]–[38]. Somehow surprisingly these new code constructions can achieve a rather significant reduction in repair network bandwidth, compared with the straightforward application of Reed–Solomon or other existing codes. In this paper, we provide an overview of this recent work and discuss several related research problems that remain open.

### A. Various Repair Models

In the repair examples shown in Figs. 2 and 3, the newcomer constructs exactly the two blocks that were in failed node. Note, however, that our definition of repair only requires that the new node forms an  $(n, k)$  MDS code property (that any  $k$  nodes out of  $n$  suffice to recover the original whole data), when combined with existing nodes. In other words, the new node could be forming new linear combinations that were different from the ones in the lost node; a requirement that is strictly easier to satisfy.

Three versions of repair have been considered in the literature: *exact repair*, *functional repair*, and *exact repair of systematic parts*. In exact repair, the failed blocks are exactly regenerated, thus restoring exactly the lost encoded blocks with their exact replicas. In functional repair, the requirement is relaxed: the newly generated blocks can contain different data from that of the failed node as long as the repaired system maintains the MDS-code property. The exact repair of the systematic part is a hybrid repair model lying between exact repair and functional repair. In this hybrid model, the storage code is always a systematic



**Fig. 4. Various repair models and the key constructive techniques.**

code (meaning that one copy of the data exists in uncoded form). The systematic part is exactly repaired upon failures and the nonsystematic part follows a functional repair model where the repaired version may be different from the original copy. See Fig. 4 for an illustration.

There is one important benefit in keeping the code in systematic form: as shown in Fig. 1, if the code contains the original data as a subset, reading parts of the data can be performed very quickly by just accessing the corresponding storage node without requiring decoding. Interestingly, as we will see, exact repair, which is the most interesting problem in practice, is also the most challenging one and determining a large part of the achievable region remains open.

The functional repair problem is completely understood because, as shown in [24], it can be reduced to a multicasting problem on an appropriately constructed graph called the information flow graph. The pioneering work of Ahlswede *et al.* [15] characterized the multicasting rates by showing that cutset bounds are achievable. Further work showed that linear network coding suffices [16], [18] and random linear combinations construct good network codes with high probability [19]. See also the survey [21] and references therein. Since functional repair is reduced to multicasting, we can completely characterize the minimum repair bandwidth by evaluating the min-cut bounds and network coding provides effective and constructive solutions. In Section II, we present the results that characterize the achievable functional repair region and show a tradeoff between storage and repair bandwidth.

The exact repair problem is strictly harder than functional repair. In exact repair, the new node accesses some existing storage nodes and exactly reproduces the lost coded blocks. As will be described subsequently, repair codes come with fundamental tradeoffs between storage cost and repair bandwidth. The two important special cases involve operating points corresponding to maximal storage and minimal bandwidth versus minimal storage with maximal bandwidth point. Exact repair for the minimal bandwidth operating point is described in Section II-B) and describes the recent work of

[33], which develops optimal exact repair codes for this operating point without any loss of optimality with respect to only functional repair.

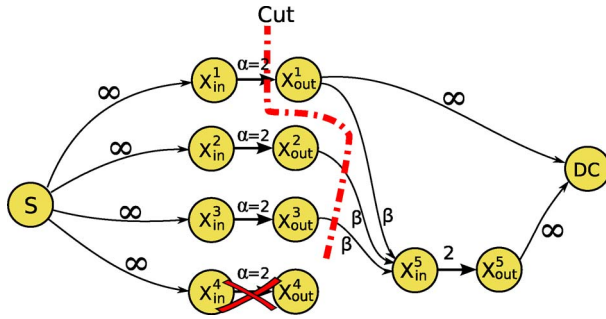
The special case of the operating point that corresponds to minimal storage, which also corresponds to minimizing the repair bandwidth while keeping the same storage cost of MDS codes, turns out to be more challenging. It turns out that, in this case, the new node needs to recover part of the data that are interfering with other data packets. When an information sink receives a set of linear equations and tries to decode for some variables, we call undesired variables, mixed into these equations, interference. It is the need to carefully handle interference that makes the problem difficult. The constructive techniques perform algebraic alignment so that the effective dimension of unwanted information is reduced, thus reducing the repair traffic. These constructive techniques achieve perfect alignment and characterize the repair bandwidth for low-rate MDS codes ( $k/n \leq 1/2$ ). Achieving the cut-set bound for high-rate MDS codes is only known to be achievable by asymptotic nonpractical techniques as we discuss subsequently.

The exact repair of systematic parts model is a relaxation of the exact repair model. As in the exact repair model, the core constructive techniques are interference alignment and network coding. In Section IV, we will see that this relaxation addresses some problem space not covered by exact repair.

## II. MODEL I: FUNCTIONAL REPAIR

As shown in [24], the functional repair problem can be represented as multicasting over an *information flow graph*. The *information flow graph* represents the evolution of information flow as nodes join and leave the storage network (see also [23] for a similar construction). Fig. 5 gives an example of an information flow graph. In this graph, each storage node is represented by a pair of nodes  $x_{in}^i$  and  $x_{out}^i$  connected by an edge whose capacity is the storage capacity of the node. There is a virtual source node  $s$  corresponding to the origin of the data object. Suppose initially we store a file of size  $M = 4$  blocks at four nodes, where each node stores  $\alpha = 2$  blocks and the file can be reconstructed from any two nodes. Virtual sink nodes called *data collectors* connect to any  $k$  node subsets and ensure that the code has the MDS property (that any  $k$  out of  $n$  suffices to recover). Suppose storage node 4 fails, then the goal is to create a new storage node, node 5, which communicates the minimum amount of information and then stores  $\alpha = 2$  blocks. This is represented in Fig. 5 by the unit-capacity edges  $x_{out}^1 x_{in}^5$ ,  $x_{out}^2 x_{in}^5$ , and  $x_{out}^3 x_{in}^5$  that enter node  $x_{in}^5$ .

The functional repair problem for distributed storage can be interpreted as a multicast communication problem defined over the information flow graph, where the source  $s$  wants to multicast the file to the set of all possible data collectors. For multicasting, it is known that the maximum multicast rate is equal to the minimum-cut capacity



**Fig. 5. Illustration of the information flow graph  $\mathcal{G}$  corresponding to the (4, 2) code of Fig. 1. A distributed storage scheme uses an (4, 2) erasure code in which any two nodes suffice to recover the original data. If node  $x^4$  becomes unavailable and a new node joins the system, we need to construct new encoded blocks in  $x^5$ . To do so, node  $x_{in}^5$  is connected to the  $d = 3$  active storage nodes. Assuming  $\beta$  bits communicated from each active storage node, of interest is the minimum  $\beta$  required. The min-cut separating the source and the data collector must be larger than  $\mathcal{M} = 4$  blocks for regeneration to be possible. For this graph, the min-cut value is given by  $\alpha + 2\beta$ , implying that communicating  $\beta \geq 1$  block is sufficient and necessary. The total repair bandwidth to repair one failure is therefore  $\gamma = d\beta = 3$  blocks.**

separating the source from a receiver and it can be achieved using linear network coding [16]. Since the current problem can be viewed as a multicast problem, the fundamental limit can be characterized by the min-cuts in the information flow graph and network coding provides effective constructive solutions. One complication is that since the number of failures/repairs is unbounded, the resulting information flow graph can grow unbounded in size. Hence, we have to deal with cuts, flows, and network codes in graphs that are potentially infinite.

In Section II-A, we present the cut analysis of information flow graphs [24], [25]. In Section II-B, we discuss the two extreme points corresponding to minimum repair bandwidth and minimum storage cost.

### A. Cut Analysis of Information Flow Graphs

By analyzing the connectivity in the information flow graph, we can derive fundamental performance bounds about codes. In particular, if the minimum cut between  $s$  and a data collector is less than the size of original file, then we can conclude that it is impossible for the data collector to reconstruct the original file. In this section, we review the cut analysis of [24] and [25]. The setup is as follows: there are always  $n$  active storage nodes. Each node can store  $\alpha$  bits. An information flow graph (as illustrated by Fig. 5) corresponds to a particular evolution of the storage system after a certain number of failures/repairs. We call each failure/repair a “stage”; in each stage, a single storage node fails and the code gets repaired by downloading  $\beta$  bits each from any  $d$  surviving nodes. Therefore, the total repair bandwidth is  $\gamma = d\beta$ .

See Fig. 5 for an example. In the initial stage, the system consists of nodes 1, 2, 3, and 4; in the second stage, the system consists of nodes 2, 3, 4, and 5. For each set of parameters  $(n, d, \alpha, \gamma = d\beta)$ , there is a family of finite or infinite information flow graphs, each of which corresponds to a particular evolution of node failures/repairs. We denote this family of directed acyclic graphs by  $\mathcal{G}(n, d, \alpha, \gamma)$ . We restrict our attention to the symmetric setup where it is required that any  $k$  storage nodes can recover the original file, and a newcomer receives the same amount of information from each of the existing nodes. An  $(n, k, d, \alpha, \gamma)$  tuple will be feasible, if a code with storage  $\alpha$  and repair bandwidth  $\gamma$  exists. For the example in Fig. 2, the total file has size  $\mathcal{M} = 4$  blocks and the point  $(n = 4, k = 2, d = 3, \alpha = 2$  blocks,  $\gamma = 3$  blocks) is feasible. On the contrary, a standard erasure code that communicates the whole data object would correspond to  $\gamma = 4$  blocks instead. Note that  $n, k, d$  must be integers. If there is one failure, the newcomer can connect to at most all the  $n - 1$  surviving nodes, so  $d \leq n - 1$  and  $\alpha, \beta, \gamma = d\beta$  are the nonnegative real-valued parameters of the repair process.

*Theorem 1:* For any  $\alpha \geq \alpha^*(n, k, d, \gamma)$ , the points  $(n, k, d, \alpha, \gamma)$  are feasible and linear network codes suffice to achieve them. It is information theoretically impossible to achieve points with  $\alpha < \alpha^*(n, k, d, \gamma)$ . The threshold function  $\alpha^*(n, k, d, \gamma)$  is the following:

$$\alpha^*(n, k, d, \gamma) = \begin{cases} \frac{\mathcal{M}}{k}, & \gamma \in [f(0), +\infty) \\ \frac{\mathcal{M} - g(i)\gamma}{k - i}, & \gamma \in [f(i), f(i - 1)) \end{cases} \quad (1)$$

where

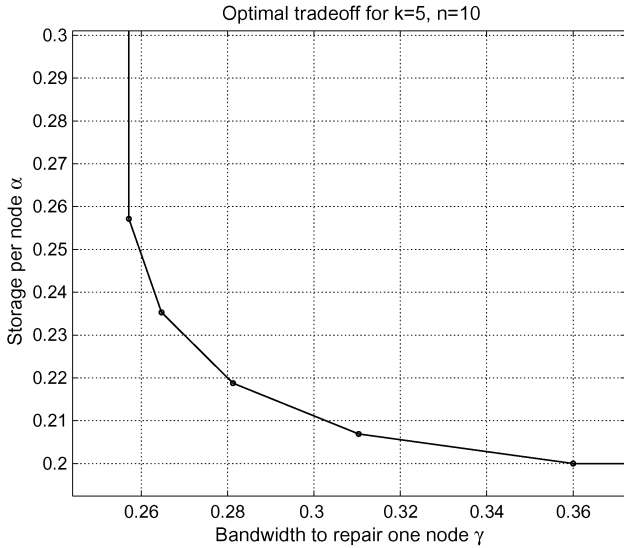
$$f(i) \triangleq \frac{2\mathcal{M}d}{(2k - i - 1)i + 2k(d - k + 1)} \quad (2)$$

$$g(i) \triangleq \frac{(2d - 2k + i + 1)i}{2d} \quad (3)$$

where  $d \leq n - 1$ . Given  $(n, k, d)$ , the minimum repair bandwidth  $\gamma$  is

$$\gamma_{\min} = f(k - 1) = \frac{2\mathcal{M}d}{2kd - k^2 + k} \quad (4)$$

One important observation is that the minimum repair bandwidth  $\gamma = d\beta$  is a decreasing function of the number  $d$  of nodes that participate in the repair. While the newcomer communicates with more nodes, the size of each communicated packet  $\beta$  becomes smaller fast enough to



**Fig. 6. Optimal tradeoff curve between storage  $\alpha$  and repair bandwidth  $\gamma$ , for  $k = 5$  and  $n = 10$ . Here  $\mathcal{M} = 1$  and  $d = n - 1$ . Note that traditional erasure coding corresponds to the point  $(\gamma = 1, \alpha = 0.2)$ .**

make the product  $d\beta$  decrease. Therefore, the minimum repair bandwidth can be achieved when  $d = n - 1$ .

As we mentioned, code repair can be achieved if and only if the underlying information flow graph has sufficiently large min-cuts. This condition leads to the repair rates computed in Theorem 1, and when these conditions are met, simple random linear combinations will suffice with high probability as the field size over which coding is performed grows, as shown by Ho et al. [19]. The optimal tradeoff curve for  $k = 5$ ,  $n = 10$ , and  $d = 9$  is shown in Fig. 6.

### B. Two Special Cases

It is of interest to study the two extremal points on the optimal tradeoff curve, which correspond to the best storage efficiency and the minimum repair bandwidth, respectively. We call codes that attain these points minimum-storage regenerating (MSR) codes and minimum-bandwidth regenerating (MBR) codes, respectively.

From Theorem 1, it can be verified that the minimum storage point is achieved

$$(\alpha_{\text{MSR}}, \gamma_{\text{MSR}}) = \left( \frac{\mathcal{M}}{k}, \frac{\mathcal{M}d}{k(d-k+1)} \right). \quad (5)$$

As discussed, the repair bandwidth  $\gamma_{\text{MSR}} = d\beta_{\text{MSR}}$  is a decreasing function of the number of nodes  $d$  that participate in the repair. Since the MSR codes store  $\mathcal{M}/k$  bits at each node while ensuring the MDS-code property, they are equivalent to standard MDS codes. Observe that when

$d = k$ , the total communication for repair is  $\mathcal{M}$  (the size of the original file). Therefore, if a newcomer is allowed to contact only  $k$  nodes, it is inevitable to download the whole data object to repair one new failure and this is the naive repair method that can be performed for any MDS codes.

However, allowing a newcomer to contact more than  $k$  nodes, MSR codes can reduce the repair bandwidth  $\gamma_{\text{MSR}}$ , which is minimized when  $d = n - 1$

$$(\alpha_{\text{MSR}}, \gamma_{\text{MSR}}^{\min}) = \left( \frac{\mathcal{M}}{k}, \frac{\mathcal{M}}{k} \cdot \frac{n-1}{n-k} \right). \quad (6)$$

We have separated the  $\mathcal{M}/k$  factor in  $\gamma_{\text{MSR}}^{\min}$  to illustrate that MSR codes communicate an  $(n-1)/(n-k)$  factor more than what they store. This represents a fundamental expansion necessary for MDS constructions that are optimal on the reliability–redundancy tradeoff. For example, consider a  $(n, k) = (14, 7)$  code. In this case, the newcomer needs to download only  $\mathcal{M}/49$  bits from each of the  $d = n - 1 = 13$  active storage nodes, making the repair bandwidth equal to  $(\mathcal{M}/7) \cdot (13/7)$ . Notice that we need only an expansion factor of  $13/7$ , while a factor of 7 is required for the naive repair method.

At the other end of the tradeoff are MBR codes, which have minimum repair bandwidth. It can be verified that the minimum repair bandwidth point is achieved by

$$(\alpha_{\text{MBR}}, \gamma_{\text{MBR}}) = \left( \frac{2\mathcal{M}d}{2kd - k^2 + k}, \frac{2\mathcal{M}d}{2kd - k^2 + k} \right). \quad (7)$$

Note that in the minimum bandwidth regenerating codes, the storage size  $\alpha$  is equal to  $\gamma$ , the total number of bits communicated during repair. If we set the optimal value  $d = n - 1$ , we obtain

$$(\alpha_{\text{MBR}}^{\min}, \gamma_{\text{MBR}}^{\min}) = \left( \frac{\mathcal{M}}{k} \cdot \frac{2n-2}{2n-k-1}, \frac{\mathcal{M}}{k} \cdot \frac{2n-2}{2n-k-1} \right). \quad (8)$$

Notice that  $\alpha_{\text{MBR}}^{\min} = \gamma_{\text{MBR}}^{\min}$ : MBR codes incur no repair bandwidth expansion at all, just like a replication system does, downloading exactly the amount of information stored during a repair. However, MBR codes require an expansion factor of  $(2n-2)/(2n-k-1)$  in the amount of stored information and are no longer optimal in terms of their reliability for the given redundancy.

### III. MODEL II: EXACT REPAIR

As we discussed, the repair–storage tradeoff for functional repair can be completely characterized by analyzing the cutset of the information flow graphs. However, as

mentioned earlier, functional repair is of limited practical interest since there is a need to maintain the code in systematic form. Also, under functional repair, significant system overhead is incurred in order to continually update repairing-and-decoding rules whenever a failure occurs. Moreover, the random-network-coding-based solution for the function repair can require a huge finite-field size to support a dynamically expanding graph size (due to continual repair). This can significantly increase the computational complexity of encoding and decoding. Furthermore, functional repair is undesirable in storage security applications in the face of eavesdroppers. In this case, information leakage occurs continually due to the dynamics of repairing-and-decoding rules that can be potentially observed by eavesdroppers [40]. These drawbacks motivate the need for *exact* repair of failed nodes. This leads to the following question: Is it possible to achieve the cutset lower bound region presented, with the extra constraint of exact repair?

Recently, significant progress has been made on the two extreme points of the family of regenerating codes (and arguably most interesting): the MBR point [33] and the MSR point [31], [34], [35]. Rashmi *et al.* [33] showed that for  $d = n - 1$ , the optimal MBR point can be achieved with a deterministic scheme requiring a small finite-field size and repair bandwidth matching the cutset bound of (8).

For the MSR point, Wu and Dimakis [31] showed that it can be attained for the cases of  $k = 2$  and  $k = n - 1$  when  $d = n - 1$ . Subsequently, Shah *et al.* [34] established that, for  $(k/n) > (1/2) + (2/n)$ , cutset bounds cannot be achieved for exact repair under *scalar linear* codes (i.e.,  $\beta = 1$ ) where symbols are not allowed to be split into arbitrarily small subsymbols as with vector linear codes.<sup>1</sup> For large  $n$ , this case boils down to  $(k/n) > (1/2)$ . Suh and Ramchandran [35] showed that exact-MSR codes can match the cutset bound of (5) for the case of  $(k/n) \leq (1/2)$  and  $d \geq 2k - 1$ .<sup>2</sup> For the in-between regime  $(k/n) \in (1/2, (1/2) + (2/n)]$ , Cullina *et al.* [32] and Suh and Ramchandran [35] showed that cutset bounds are achievable for the case of  $k = 3$ . A construction that can match the cutset bound for the MBR point for all  $n, k, d$  and for MSR codes if the rate  $(k/n) \leq (1/2)$  was presented by Rashmi *et al.* [46].

Finally, it was very recently established that MSR codes that can match the repair communication cutset bound for all  $n, k$  exist asymptotically. This surprising result was independently obtained in [35] and [45] by using the breakthrough technique of symbol extension introduced by

<sup>1</sup>This is equivalent to having large block lengths in the classical setting. Under nonlinear and vector linear codes, tightness of cutset bounds remains open.

<sup>2</sup>The idea was inspired by the code structure in [34] where exact repair is guaranteed for the systematic part only. Indeed, it is shown in [35] that the code introduced in [34] for exact repair of only the systematic nodes can also be used to repair the nonsystematic (parity) node failures exactly provided repair construction schemes are appropriately designed.

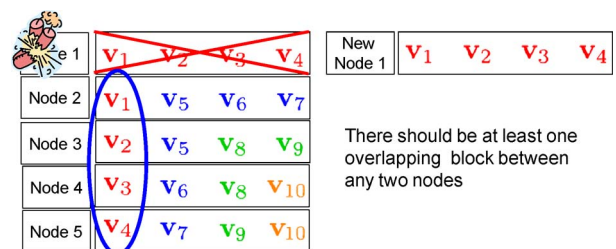
Cadambe and Jafar [29]. It is surprising how symbol extension, a technique developed to exploit independent fading of wireless channels, maps exactly to problem of exact repair at the MSR point. Recent work of Papailiopoulos *et al.* [43], [44] explores this connection further. We note that while this work shows that high-rate exact MSR codes exist, the constructions of [35] and [45] are not practical since they require exponential field size and subpacketization. Concerning the intermediate points beyond MSR and MBR, finding the fundamental limits of storage and repair communication remains a challenging open problem. We now briefly summarize some of these recent results.

### A. Exact-MBR Codes

*Theorem 2 (Exact-MBR Codes [33]):* For  $d = n - 1$ , the cutset lower bound of (8) can be achieved with a deterministic scheme that requires a finite-field alphabet size of at most  $(n - 1)n/2$ .

Fig. 7 illustrates an idea through the example of  $(n, k, d, \alpha, \gamma) = (5, 3, 4, 4, 4)$  where the maximum file size of  $\mathcal{M} = 9$  (matching the cutset bound) can be stored. Let  $\mathbf{a}$  be nine-dimensional data file. Each node stores four blocks with the form of  $\mathbf{a}^t \mathbf{v}_i$ , where  $\mathbf{v}_i$  can be interpreted as a 1-D subspace of data file. We simply write only subspace vector to represent an actually stored block. Notice that the degree  $d$  is equal to the number of storage blocks to be repaired, i.e., the number of available equations matches the number of desired variables for exact repair of a single node. Hence, for exact repair, there must be at least one duplicated block between node 1 and node  $i$  for all  $i \neq 1$ .

This observation motivates the following idea. The idea is to have other nodes  $i$  ( $i \neq 1$ ) store each block of node 1, respectively: nodes 2, 3, 4, and 5 store  $\mathbf{a}^t \mathbf{v}_1$ ,  $\mathbf{a}^t \mathbf{v}_2$ ,  $\mathbf{a}^t \mathbf{v}_3$ , and  $\mathbf{a}^t \mathbf{v}_4$  in its own place, respectively. Notice that for ensuring repair, it suffices to have only one duplicated block between any two storage nodes. Hence, node 2 can store another new three blocks of  $\mathbf{a}^t \mathbf{v}_5$ ,  $\mathbf{a}^t \mathbf{v}_6$ , and  $\mathbf{a}^t \mathbf{v}_7$  in



**Fig. 7. Repairing node 1 for a (5, 3)-MBR code. Note that the number of desired blocks (that need to be repaired) is equal to the number of available equations (that can be downloaded). Hence, the code should be designed such that undesired blocks (interference) are totally avoided.**

the remaining other places. In accordance with the above procedure, nodes 3, 4, and 5 then copy each of three blocks in their space, respectively. We repeat this procedure until  $10 (= 4 + 3 + 2 + 1)$  blocks are stored in total. One can see that this construction guarantees exact repair of any failed node, since at least one block is duplicated between any two storage nodes and also the duplicated block is *distinct*. See the example in Fig. 7.

The remaining issue is now to design these ten subspace vectors  $\mathbf{v}_i, i = 1, \dots, 10$ . The detailed construction comes from the MDS-code property that any three nodes out of five need to recover the whole data file. Observe in Fig. 7 that nine distinct vectors can be downloaded from any three nodes. Hence, any  $(10, 9)$  MDS code can construct these  $\mathbf{v}_i$ 's. In this example, using the parity-check code defined over  $\mathbf{GF}(2)$ , we can design the  $\mathbf{v}_i$ 's as follows:  $\mathbf{v}_i = \mathbf{e}_i, \forall i = 1, \dots, 9$  and  $\mathbf{v}_{10} = [1, \dots, 1]^t$ . It has been shown in [33] that this idea can be extended to an arbitrary  $(n, k)$  case.

This construction can be interpreted as an optimal *interference avoidance* technique. To see this, observe in the figure that the number of desired blocks for exact repair matches the number of available equations that can be downloaded. Hence, the involvement of any undesired blocks (interference) precludes exact repair. A natural question arises: Can this interference-avoidance technique provide solutions to the other extreme MSR point? It turns out that a new idea is needed to cover this point.

### B. Exact-MSR Codes

The new idea is *interference alignment* [28], [29]. The idea of interference alignment is to align multiple interference signals in a signal subspace whose dimension is smaller than the number of interferers. Specifically, consider the following setup where a decoder has to decode one desired signal that is linearly interfered with by two separate undesired signals. How many linear equations (relating to the number of channel uses) does the decoder need to recover its desired input signal? As the aggregate signal dimension spanned by desired and undesired signals

is at most three, the decoder can naively recover its signal of interest with access to three linearly independent equations in the three unknown signals. However, as the decoder is interested in only one of the three signals, it can decode its desired unknown signal even if it has access to only two equations, provided the two undesired signals are judiciously aligned in a 1-D subspace. See [28]–[30] for details.

This concept relates intimately to our repair problem that involves recovery of a subset (related to the subspace spanned by a failed node) of the overall aggregate signal space (related to the entire user data dimension). This attribute was first observed in [31], where it was shown that interference alignment could be exploited for exact-MSR codes.

Fig. 8 illustrates interference alignment for exact repair of failed node 1 for  $(n, k, d, \alpha, \gamma) = (4, 2, 3, 2, 2)$  where the maximum file size of  $\mathcal{M} = 4$  can be stored. We introduce matrix notation for illustration purposes. Let  $\mathbf{a} = (a_1, a_2)^t$  and  $\mathbf{b} = (b_1, b_2)^t$  be 2-D information-unit vectors. Let  $\mathbf{A}_i$  and  $\mathbf{B}_i$  be 2-by-2 encoding matrices for parity node  $i$  ( $i = 1, 2$ ), which contain encoding coefficients for the linear combination of  $(a_1, a_2)$  and  $(b_1, b_2)$ , respectively. For example, parity node 1 stores blocks in the form of  $\mathbf{a}^t \mathbf{A}_1 + \mathbf{b}^t \mathbf{B}_1$ , as shown in Fig. 8. The encoding matrices for systematic nodes are not explicitly defined since those are trivially inferred. Finally, we define 2-D projection vectors  $\mathbf{v}_{\alpha i}$ 's ( $i = 1, 2, 3$ ) because of  $\beta = 1$ .

Let us explain the interference-alignment scheme. First, two blocks in each storage node are projected into a *scalar* with projection vectors  $\mathbf{v}_{\alpha i}$ 's. By connecting to three nodes, we get:  $\mathbf{v}_{\alpha 1}^t \mathbf{b}; (\mathbf{A}_1 \mathbf{v}_{\alpha 2})^t \mathbf{a} + (\mathbf{B}_1 \mathbf{v}_{\alpha 2})^t \mathbf{b}; (\mathbf{A}_2 \mathbf{v}_{\alpha 3})^t \mathbf{a} + (\mathbf{B}_2 \mathbf{v}_{\alpha 3})^t \mathbf{b}$ . Here the goal is to decode two desired unknowns out of three equations including four unknowns. To achieve this goal, we need

$$\text{rank} \left( \begin{bmatrix} (\mathbf{A}_1 \mathbf{v}_{\alpha 2})^t \\ (\mathbf{A}_2 \mathbf{v}_{\alpha 3})^t \end{bmatrix} \right) = 2 \quad \text{rank} \left( \begin{bmatrix} \mathbf{v}_{\alpha 1}^t \\ (\mathbf{B}_1 \mathbf{v}_{\alpha 2})^t \\ (\mathbf{B}_2 \mathbf{v}_{\alpha 3})^t \end{bmatrix} \right) = 1.$$

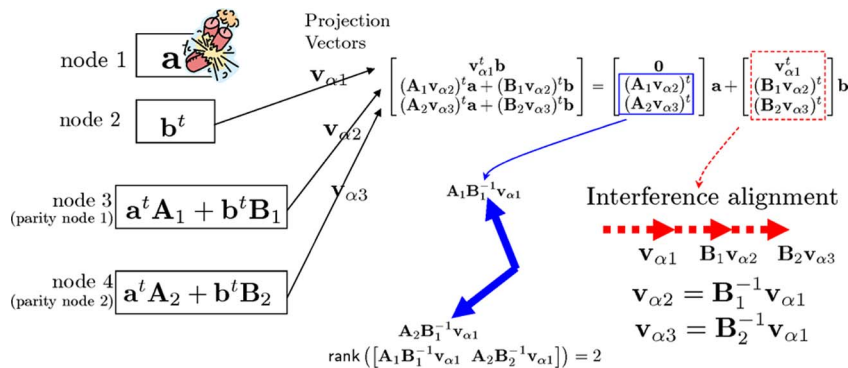
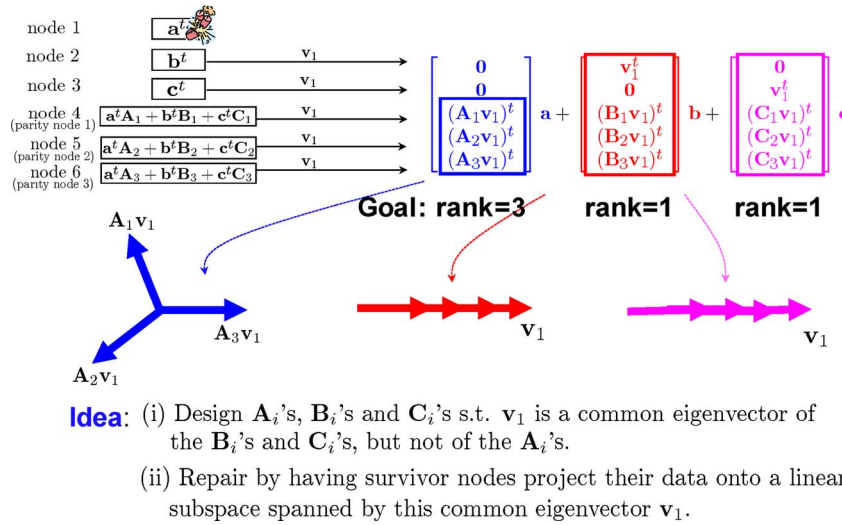


Fig. 8. Repairing a  $(4, 2)$ -MSR code, when node 1 fails [31].



**Fig. 9. Repairing the (6, 3)-MSR code when a systematic node fails. A common eigenvector concept is employed to achieve interference alignment simultaneously.**

The second condition can be met by setting  $\mathbf{v}_{\alpha 2} = \mathbf{B}_1^{-1} \mathbf{v}_{\alpha 1}$  and  $\mathbf{v}_{\alpha 3} = \mathbf{B}_2^{-1} \mathbf{v}_{\alpha 1}$ . This choice forces the interference space to be collapsed into a 1-D linear subspace, thereby achieving interference alignment. On the other hand, we can satisfy the first condition as well by carefully choosing the  $\mathbf{A}_i$ 's and  $\mathbf{B}_i$ 's. For exact repair of node 2, we can apply the same idea. For parity node repair, we can remap parity node information and then apply the same technique.

It turned out that this idea cannot be generalized to arbitrary  $(n, k)$  case: it provides the optimal codes only for the case of  $k = 2$ . Recently, significant progress has been made: for the case of  $(k/n) \leq (1/2)$ , it has been shown that there is no price with exact repair for attaining the cutset lower bound of (5).

**Theorem 3 (Exact-MSR Codes [35]):** Suppose the MDS code rate is at most  $1/2$ , i.e.,  $(k/n) \leq (1/2)$  and the degree  $d \geq 2k - 1$ . Then, the cutset bound of (5) can be achieved with interference alignment. The achievable scheme is deterministic and requires a finite-field alphabet size of at most  $2(n - k)$ .

A more sophisticated idea arises to cover this case: *simultaneous interference alignment*. Fig. 9 illustrates the interference-alignment technique through the example of  $(n, k, d, \alpha, \gamma) = (6, 3, 5, 3, 3)$  where  $\mathcal{M} = 9$ . Let  $\mathbf{a} = (a_1, a_2, a_3)^t$ ,  $\mathbf{b} = (b_1, b_2, b_3)^t$ , and  $\mathbf{c} = (c_1, c_2, c_3)^t$  be 3-D information-unit vectors. Let  $\mathbf{A}_i$ ,  $\mathbf{B}_i$ , and  $\mathbf{C}_i$  be 3-by-3 encoding matrices for parity node  $i$  ( $i = 1, 2, 3$ ). We define 3-D projection vectors  $\mathbf{v}_{\alpha i}$ 's ( $i = 1, \dots, 5$ ).

By connecting to five nodes, we get five equations shown in the figure. In order to successfully recover the desired signal components of  $\mathbf{a}$ , the matrix associated with  $\mathbf{a}$  should have full rank of 3, while the other matrices corresponding to  $\mathbf{b}$  and  $\mathbf{c}$  should have rank 1, respectively.

In accordance with the (4, 2) code example in Fig. 8, if one were to set  $\mathbf{v}_{\alpha 3} = \mathbf{B}_1^{-1} \mathbf{v}_{\alpha 1}$ ,  $\mathbf{v}_{\alpha 4} = \mathbf{B}_2^{-1} \mathbf{v}_{\alpha 1}$ , and  $\mathbf{v}_{\alpha 5} = \mathbf{B}_3^{-1} \mathbf{v}_{\alpha 1}$ , then it is possible to achieve interference alignment with respect to  $\mathbf{b}$ . However, this choice also specifies the interference space of  $\mathbf{c}$ . If the  $\mathbf{B}_i$ 's and  $\mathbf{C}_i$ 's are not designed judiciously, interference alignment is not guaranteed for  $\mathbf{c}$ . Hence, it is not evident how to achieve interference alignment at the same time.

In order to address the challenge of simultaneous interference alignment, a *common eigenvector* concept is invoked. The idea consists of two parts: 1) designing the  $(\mathbf{A}_i, \mathbf{B}_i, \mathbf{C}_i)$ 's such that  $\mathbf{v}_1$  is a common eigenvector of the  $\mathbf{B}_i$ 's and  $\mathbf{C}_i$ 's, but not of  $\mathbf{A}_i$ 's<sup>3</sup>; and 2) repairing by having survivor nodes *project* their data onto a linear subspace spanned by this common eigenvector  $\mathbf{v}_1$ . We can then achieve interference alignment for  $\mathbf{b}$  and  $\mathbf{c}$  at the same time, by setting  $\mathbf{v}_{\alpha i} = \mathbf{v}_1, \forall i$ . As long as  $[\mathbf{A}_1 \mathbf{v}_1, \mathbf{A}_2 \mathbf{v}_1, \mathbf{A}_3 \mathbf{v}_1]$  is invertible, we can also guarantee the decodability of  $\mathbf{a}$ . See Fig. 9.

The challenge is now to design encoding matrices to guarantee the existence of a common eigenvector while also satisfying the decodability of desired signals. The difficulty comes from the fact that in the (6, 3, 5) code example, these constraints need to be satisfied for *all* six possible failure configurations. The structure of *elementary matrices* (generalized matrices of Householder and Gauss matrices) gives insights into this. To see this, consider a 3-by-3 elementary matrix  $\mathbf{A}$

$$\mathbf{A} = \mathbf{u}\mathbf{v}^t + \alpha \mathbf{I} \quad (9)$$

<sup>3</sup>Of course, five additional constraints also need to be satisfied for the other five failure configurations for this (6, 3, 5) code example.

where  $\mathbf{u}$  and  $\mathbf{v}$  are 3-D vectors. Note that the dimension of the null space of  $\mathbf{v}$  is 2 and the null vector  $\mathbf{v}^\perp$  is an eigenvector of  $\mathbf{A}$ , i.e.,  $\mathbf{A}\mathbf{v}^\perp = \alpha\mathbf{v}^\perp$ . This motivates the following structure:

$$\begin{aligned} \mathbf{A}_1 &= \mathbf{u}_1\mathbf{v}_1^t + \alpha_1\mathbf{I} \\ \mathbf{B}_1 &= \mathbf{u}_1\mathbf{v}_2^t + \beta_1\mathbf{I} \\ \mathbf{C}_1 &= \mathbf{u}_1\mathbf{v}_3^t + \gamma_1\mathbf{I} \\ \mathbf{A}_2 &= \mathbf{u}_2\mathbf{v}_1^t + \alpha_2\mathbf{I} \\ \mathbf{B}_2 &= \mathbf{u}_2\mathbf{v}_2^t + \beta_2\mathbf{I} \\ \mathbf{C}_2 &= \mathbf{u}_2\mathbf{v}_3^t + \gamma_2\mathbf{I} \\ \mathbf{A}_3 &= \mathbf{u}_3\mathbf{v}_1^t + \alpha_3\mathbf{I} \\ \mathbf{B}_3 &= \mathbf{u}_3\mathbf{v}_2^t + \beta_3\mathbf{I} \\ \mathbf{C}_3 &= \mathbf{u}_3\mathbf{v}_3^t + \gamma_3\mathbf{I} \end{aligned} \quad (10)$$

where  $\mathbf{v}_i$ 's are 3-D linearly independent vectors and so are  $\mathbf{u}_i$ 's. The values of the  $\alpha_i$ 's,  $\beta_i$ 's, and  $\gamma_i$ 's can be arbitrary nonzero values. For simplicity, we consider the simple case where the  $\mathbf{v}_i$ 's are *orthonormal*, although these need not be orthogonal, but only linearly independent. We then see that  $\forall i = 1, 2, 3$

$$\begin{aligned} \mathbf{A}_i\mathbf{v}_1 &= \alpha_i\mathbf{v}_1 + \mathbf{u}_i \\ \mathbf{B}_i\mathbf{v}_1 &= \beta_i\mathbf{v}_1 \\ \mathbf{C}_i\mathbf{v}_1 &= \gamma_i\mathbf{v}_1. \end{aligned} \quad (11)$$

Importantly, notice that  $\mathbf{v}_1$  is a common eigenvector of the  $\mathbf{B}_i$ 's and  $\mathbf{C}_i$ 's, while simultaneously ensuring that the vectors of  $\mathbf{A}_i\mathbf{v}_1$  are linearly independent. Hence, setting  $\mathbf{v}_{\alpha i} = \mathbf{v}_1$  for all  $i$ , it is possible to achieve simultaneous interference alignment while also guaranteeing the decodability of the desired signals. On the other hand, this structure also guarantees exact repair for  $\mathbf{b}$  and  $\mathbf{c}$ . We use  $\mathbf{v}_2$  for exact repair of  $\mathbf{b}$ . It is a common eigenvector of the  $\mathbf{C}_i$ 's and  $\mathbf{A}_i$ 's, while ensuring  $[\mathbf{B}_1\mathbf{v}_2, \mathbf{B}_2\mathbf{v}_2, \mathbf{B}_3\mathbf{v}_2]$  invertible. Similarly,  $\mathbf{v}_3$  is used for  $\mathbf{c}$ .

Parity nodes can be repaired by drawing a *dual* relationship with systematic nodes. The procedure has two steps. The first is to remap parity nodes with  $\mathbf{a}'$ ,  $\mathbf{b}'$ , and  $\mathbf{c}'$ , respectively. Systematic nodes can then be rewritten in terms of the prime notations

$$\begin{aligned} \mathbf{a}^t &= \mathbf{a}^t\mathbf{A}'_1 + \mathbf{b}^t\mathbf{B}'_1 + \mathbf{c}^t\mathbf{C}'_1 \\ \mathbf{b}^t &= \mathbf{a}^t\mathbf{A}'_2 + \mathbf{b}^t\mathbf{B}'_2 + \mathbf{c}^t\mathbf{C}'_2 \\ \mathbf{c}^t &= \mathbf{a}^t\mathbf{A}'_3 + \mathbf{b}^t\mathbf{B}'_3 + \mathbf{c}^t\mathbf{C}'_3 \end{aligned} \quad (12)$$

where the newly mapped encoding matrices  $(\mathbf{A}'_i, \mathbf{B}'_i, \mathbf{C}'_i)$ 's are defined as

$$\begin{bmatrix} \mathbf{A}'_1 & \mathbf{A}'_2 & \mathbf{A}'_3 \\ \mathbf{B}'_1 & \mathbf{B}'_2 & \mathbf{B}'_3 \\ \mathbf{C}'_1 & \mathbf{C}'_2 & \mathbf{C}'_3 \end{bmatrix} := \begin{bmatrix} \mathbf{A}_1 & \mathbf{A}_2 & \mathbf{A}_3 \\ \mathbf{B}_1 & \mathbf{B}_2 & \mathbf{B}_3 \\ \mathbf{C}_1 & \mathbf{C}_2 & \mathbf{C}_3 \end{bmatrix}^{-1}. \quad (13)$$

With this remapping, one can dualize the relationship between systematic and parity node repair. Specifically, if all of the  $\mathbf{A}'_i$ 's,  $\mathbf{B}'_i$ 's, and  $\mathbf{C}'_i$ 's are *elementary matrices* and form a similar code structure as in (10), exact repair of the parity nodes becomes transparent. It was shown that a special relationship between  $[\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3]$  and  $[\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3]$  through the correct choice of  $(\alpha_i, \beta_i, \gamma_i)$ 's can also guarantee the *dual* structure of (10) [35].

Fig. 10 shows a numerical example for exact repair of systematic node 1 [Fig. 10(a)] and parity node 1 [Fig. 10(b)] where  $[\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3] = [2, 2, 2; 2, 3, 1; 2, 1, 3]$ . This example illustrates the code structure that generalizes the code introduced in [34]. See [35] for details. This generalized code structure allows for a much larger design space for exact repair.

Notice that the projection vector solution for systematic node repair is simple:  $\mathbf{v}_{\alpha i} = 2^{-1}\mathbf{v}_1 = (1, 1, 1)^t, \forall i$ . Note that this choice enables simultaneous interference alignment, while guaranteeing the decodability of  $\mathbf{a}$ . Notice that  $(b_1, b_2, b_3)$  and  $(c_1, c_2, c_3)$  are aligned into  $b_1 + b_2 + b_3$  and  $c_1 + c_2 + c_3$ , respectively, while three equations associated with  $\mathbf{a}$  are linearly independent.

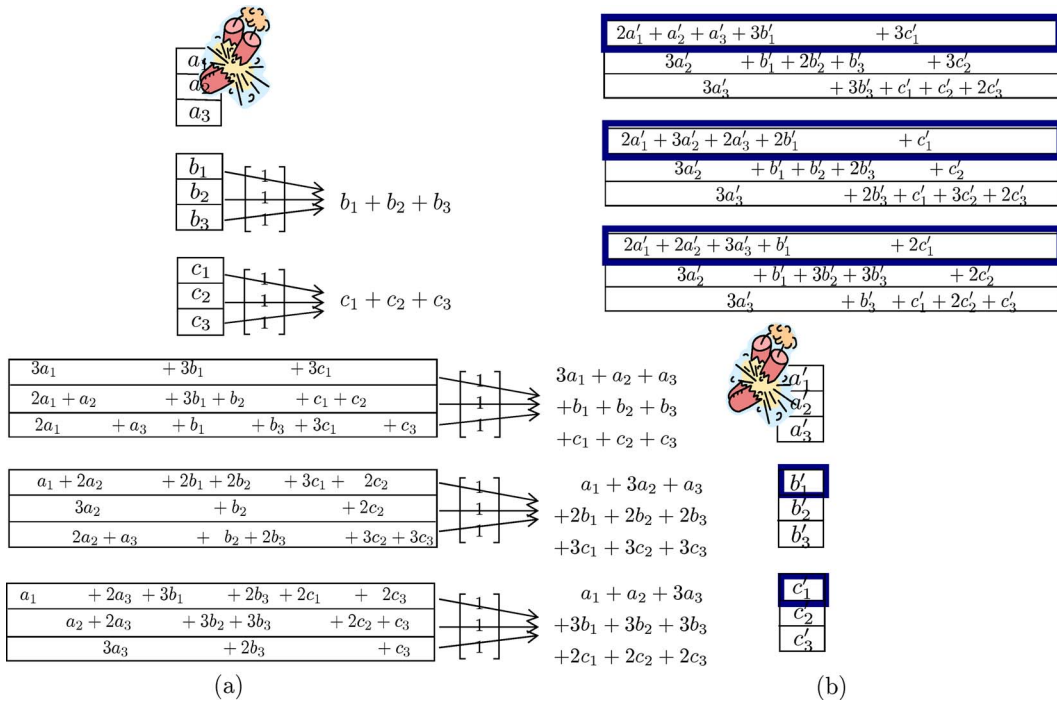
The dual structure also guarantees exact repair of parity nodes. Importantly, we have chosen code parameters from the generalized code structure of [35] such that parity node repair is quite simple. As shown in Fig. 10(b), downloading only the first equation from each survivor node ensures exact repair. Notice that the five downloaded equations contain only five unknown variables of  $(a'_1, a'_2, a'_3, b'_1, c'_1)$  and three equations associated with  $\mathbf{a}'$  are linearly independent. Hence, we can successfully recover  $\mathbf{a}'$ .

It has been shown in [35] that this alignment technique can be easily generalized to arbitrary  $(n, k, d)$  where  $n \geq 2k$  and  $d \geq 2k - 1$ .

#### IV. MODEL III: EXACT REPAIR OF THE SYSTEMATIC PART

In this section, we review the constructive scheme given in [36], which gives a construction of systematic  $(n, k)$ -MDS codes for  $2k \leq n$  that achieves the minimum repair bandwidth when repairing from  $k + 1$  nodes.

The scheme is illustrated in Fig. 11. Let  $\mathbb{F}$  denote the finite field where the code is defined in. In Fig. 11,  $\mathbf{x} \in \mathbb{F}^{2k}$  is a vector consisting of the  $2k$  original information symbols. Each node stores two symbols  $\mathbf{x}^t\mathbf{u}_i$  and  $\mathbf{x}^t\mathbf{v}_i$ . The vectors  $\{\mathbf{u}_i\}$  do not change over time but  $\{\mathbf{v}_i\}$  change as



**Fig. 10.** Illustration of exact repair for a  $(6, 3, 5)$  E-MSR code defined over  $\text{GF}(4)$  where a generator polynomial  $g(x) = x^2 + x + 1$ . The solution for systematic node repair is simple: setting all of the projection vectors as  $(1, 1, 1)^T$ . This enables simultaneous interference alignment, while guaranteeing the decodability of  $\mathbf{a}$ . For our carefully chosen parameters, parity node repair is much simpler. For the repair, we download only the first equation from each survivor node to solve five linear equations containing only five unknowns. (a) Exact repair of systematic node 1. (b) Exact repair of parity node 1.

the code repairs. We maintain the invariant property that the  $2n$  length- $2k$  vectors  $\{\mathbf{u}_i, \mathbf{v}_i\}$  form an  $(2n, 2k)$ -MDS code; that is, any  $2k$  vectors in the set  $\{\mathbf{u}_i, \mathbf{v}_i\}$  have full rank  $2k$ . This certainly implies that the  $n$  nodes form an  $(n, k)$ -MDS code. We initialize the code using any  $(2n, 2k)$  systematic MDS code over  $\mathbb{F}$ .

Now we consider the situation of a repair. Without loss of generality, suppose node  $n$  failed and is repaired by accessing nodes  $1, \dots, k + 1$ . As illustrated in Fig. 11, the replacement node downloads  $\alpha_i \mathbf{x}^T \mathbf{u}_i + \beta_i \mathbf{x}^T \mathbf{v}_i$  from each

node of  $\{1, \dots, k + 1\}$ . Using these  $k + 1$  downloaded symbols, the replacement node computes two symbols  $\mathbf{x}^T \mathbf{u}_n$  and  $\mathbf{x}^T \mathbf{v}'_n$  as follows:

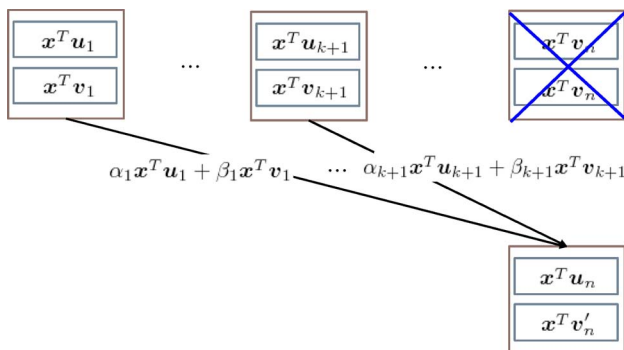
$$\sum_{i=1}^{k+1} (\alpha_i \mathbf{x}^T \mathbf{u}_i + \beta_i \mathbf{x}^T \mathbf{v}_i) = \mathbf{x}^T \mathbf{u}_n \quad (14)$$

$$\sum_{i=1}^{k+1} \rho_i (\alpha_i \mathbf{x}^T \mathbf{u}_i + \beta_i \mathbf{x}^T \mathbf{v}_i) = \mathbf{x}^T \mathbf{v}'_n. \quad (15)$$

Note that  $\mathbf{v}'_n$  is allowed to be different from  $\mathbf{v}_n$ ; the property that we maintain is that the repaired code continues to be an  $(2n, 2k)$ -MDS code. Here  $\{\alpha_i, \beta_i, \rho_i\}$  and  $\mathbf{v}'_n$  are the variables that we can control. The following theorem shows that we can choose these variables so that (14) and (15) are satisfied and the repaired code continues to be an  $(2n, 2k)$ -MDS code.

**Theorem 4 [36]:** Let  $\mathbb{F}$  be a finite field whose size is greater than

$$d_0 = 2 \binom{2n-1}{2k-1}. \quad (16)$$



**Fig. 11.** Illustration of the scheme in [36].

Suppose the old code specified by  $\{\mathbf{u}_i, \mathbf{v}_i\}$  is an  $(2n, 2k)$ -MDS code defined over  $\mathbb{F}$ . When node  $n$  fails, there exists an assignment of the variables  $\{\alpha_i, \beta_i, \rho_i\}$  such that (14) and (15) are satisfied and the repaired code continues to be an  $(2n, 2k)$ -MDS code.

*Corollary 1 [A Systematic  $(n, k)$ -MDS Code]:* The above scheme gives a construction of systematic  $(n, k)$ -MDS codes for  $2k \leq n$  that achieves the minimum repair bandwidth when repairing from  $k + 1$  nodes.

*Proof:* Consider  $n \geq 2k$ . Note that in the above scheme, we can initialize the code  $\{\mathbf{u}_1, \dots, \mathbf{u}_n, \mathbf{v}_1, \dots, \mathbf{v}_n\}$  with any  $(2n, 2k)$ -MDS code. In particular, we can use a systematic code and assign the  $2k$  systematic code vectors to  $\{\mathbf{u}_1, \dots, \mathbf{u}_{2k}\}$ . Since  $\{\mathbf{u}_1, \dots, \mathbf{u}_n\}$  do not change over time, the code remains a systematic  $(2n, 2k)$ -MDS code. Thus, the  $n$  nodes form a systematic  $(n, k)$ -MDS code. The code repairs a failure by downloading  $k + 1$  blocks from  $d = k + 1$  nodes, with the total file size  $\mathcal{M} = 2k$ , achieving the cutset bounds derived in Section II. ■

## V. DISCUSSION AND OPEN PROBLEMS

We provided an overview of recent results about the problem of reducing repair traffic in distributed storage systems based on erasure coding. Three versions of the repair problems are considered: *exact repair*, *functional repair*, and *exact repair of systematic parts*. In the exact repair model, the lost content is exactly regenerated; in the functional repair model, only the same MDS-code property is maintained before and after repairing; in the exact repair of systematic parts, the systematic part is exactly reconstructed but the nonsystematic part follows a functional repair model.

The functional repair problem is in essence a problem of multicasting from a source to an unbounded number of receivers over an unbounded graph. As we showed there is a tradeoff between storage and repair bandwidth and the two extremal points are achieved by MBR and MSR codes. The repair bandwidth is characterized by the min-cut bounds and therefore the functional repair problem is well understood.

Problems that require exact repair correspond to network coding problems having sinks with overlapping subset demands. For such problems, cutset bounds are not tight in general and linear codes might not even suffice [22]. The recent work we discussed [33] showed that for MBR codes the repair bandwidth given by the cutset bound is achievable for the interesting case of  $d = n - 1$ . The minimum-storage point seems harder to understand. The best known constructions [35] we presented match the cutset bound for  $k/n \leq 1/2$  for the interesting regime of connectivity  $d \in [2k - 1, n - 1]$ . A corresponding negative result [34] established that, for  $(k/n) > (1/2) + (2/n)$ , the cutset bound cannot be achieved by scalar interference-alignment-based linear schemes. However, the symbol extension [35], [45] method showed that the cutset bound can be asymptotically approached for very large subpacketization  $\beta$ .

Table 1 summarizes what is known for the repair bandwidth region and an online editable bibliography (Wiki) can be found online [1]. All the cases marked correspond to regimes where the cutset bound is known to be achievable. To the best of our knowledge there are no information-theoretic upper bounds other than the cutset bound and it would be very interesting to see if the region could be universally achievable. A reasonable conjecture is that the whole tradeoff region can be asymptotically approached with sufficient subpacketization and field size.

In addition to the complete characterization of the repair rate region for storage, there are several other interesting open problems.

*OPI:* The first problem is to investigate the influence of network topology, as initiated recently [38] for trees. All the prior work so far has been assuming a complete connectivity topology for the storage network. However, most networks of interest will have different communication capacities and sparse topologies. For these cases, communication will have a different cost and it would be interesting to formulate this as an optimization problem.

*OP2:* While most of this work has focused on the size of communicated packets, to create these packets the amount of information that must be read from the storage nodes is

**Table 1** Known Results for Exact MBR and MSR Codes. All Points Correspond to Regimes Where the Cutset Bound Region Is Known to be Achievable

	MBR	MSR
Functional		[24]: $\forall n, k, d$
Hybrid	?	[34]: $\frac{k}{n} \leq \frac{1}{2}, d \geq 2k - 1$ . [36], [33]: $\frac{k}{n} \leq \frac{1}{2}, d = k + 1$
Exact	[33]: $d = n - 1$ [46]: $\forall n, k, d$	[35]: $\frac{k}{n} \leq \frac{1}{2}, d \geq 2k - 1$ [35], [45]: $\forall n, k, d$ (non-practical)

large. It would be very interesting to characterize the minimum communication that must be read to repair a code.

Most research on distributed storage has focused on designing MDS (or near-MDS) codes that are easily repairable. A different approach is to find ways to repair existing codes beyond the naive approach of reconstructing all the information. This is especially useful to leverage the benefits of known constructions such as reduced update complexity and efficient decoding under errors. The practical relevance of repairing a family of codes with a given structure depends on the applicability of this family in distributed storage problems. While the problem can be studied for any family of error correcting codes, two cases that are of special interest are array codes and Reed–Solomon codes.

*OP3 (Repairing array codes):* Array codes are widely used in data storage systems [11], [12], [42]. For the special case of evenodd codes [10], a repair method that improves on the naive method of reconstructing the whole data object by a factor of 0.75 was established in [14]. There is still a gap from the cutset lower bound and it remains open if the minimal repair communication can be achieved if we enforce the Evenodd code structure.

*OP4 (Repairing Reed–Solomon codes):* Another important family is Reed–Solomon codes [6]. A repair strategy that improves on the naive method of reconstructing the whole data object for each single failure would be directly applicable to storage systems that use Reed–Solomon codes. The repair of Reed–Solomon codes poses some challenges: since each encoded block corresponds to the

evaluation of a polynomial, during repair, a *partial evaluation* would have to be communicated from each surviving node. This step would require a nonlinear operation and it is unclear how to create the missing evaluation from partial evaluations at the other  $n - 1$  points. One way around this would be to use the idea of subpacketization to allow the communication of units smaller than the stored packet and stay within the framework of linear codes.

*OP5:* A coding theory problem that has been studied in depth is that of locally decodable codes [41]. The issue there is to recover a symbol by reading a small subset of other (noisy) symbols. This is very similar to exact repair with the important difference that during repair the newcomer is accessing many nodes and receives small parts of symbols. In addition, repair is assuming noiseless access to other encoded symbols. Connecting exact repair to the deep theory of locally decodable codes seems like an interesting research direction.

*OP6:* The issues of security and privacy are important for distributed storage. When coding is used, errors can be propagated in several mixed blocks through the repair process [39] and an error-control mechanism is required. A related issue is that of privacy of the data by information leakage to eavesdroppers during repairs [40]. ■

## Acknowledgment

The authors would like to thank Prof. P. V. Kumar (of IISs) and his students, N. B. Shah and K. V. Rashmi, for insightful discussions and fruitful collaboration.

## REFERENCES

- [1] *The Coding for Distributed Storage Wiki*. [Online]. Available: <http://tinyurl.com/storagecoding>
- [2] H. Weatherspoon and J. D. Kubiatowicz, "Erasure coding vs. replication: A quantitative comparison," in *Proc. Int. Workshop Peer-to-Peer Syst.*, 2002.
- [3] J. Kubiatowicz, D. Bindel, Y. Chen, S. Czerwinski, P. Eaton, D. Geels, R. Gummadi, S. Rhea, H. Weatherspoon, W. Weimer, C. Wells, and B. Zhao, "OceanStore: An architecture for global-scale persistent storage," in *Proc. 9th Int. Conf. Architectural Support Programm. Lang. Oper. Syst.*, Boston, MA, Nov. 2000, pp. 190–201.
- [4] S. Rhea, C. Wells, P. Eaton, D. Geels, B. Zhao, H. Weatherspoon, and J. Kubiatowicz, "Maintenance-free global data storage," *IEEE Internet Comput.*, pp. 40–49, Sep. 2001.
- [5] R. Bhagwan, K. Tati, Y.-C. Cheng, S. Savage, and G. M. Voelker, "Total recall: System support for automated availability management," in *Proc. Symp. Netw. Syst. Design Implementation*, 2004, pp. 337–350.
- [6] I. Reed and G. Solomon, "Polynomial codes over certain finite fields," *J. Soc. Ind. Appl. Math.*, vol. 8, no. 2, pp. 300–304, Jun. 1960.
- [7] M. O. Rabin, "Efficient dispersal of information for security, load balancing and fault tolerance," *J. Assoc. Comput. Mach.*, vol. 36, no. 2, pp. 335–348, 1989.
- [8] A. Shokrollahi, "Raptor codes," *IEEE Trans. Inf. Theory*, vol. 52, no. 6, pp. 2551–2567, Jun. 2006.
- [9] T. Richardson and R. Urbanke, *Modern Coding Theory*. Cambridge, U.K.: Cambridge Univ. Press, 2008.
- [10] M. Blaum, J. Brady, J. Bruck, and J. Menon, "EVENODD: An efficient scheme for tolerating double disk failures in raid architectures," *IEEE Trans. Comput.*, vol. 44, no. 2, pp. 192–202, Feb. 1995.
- [11] M. Blaum, J. Bruck, and A. Vardy, "MDS array codes with independent parity symbols," *IEEE Trans. Inf. Theory*, vol. 42, no. 2, pp. 529–542, Mar. 1996.
- [12] M. Blaum, P. G. Farrell, and H. van Tilborg, "Book chapter on array codes," in *Handbook of Coding Theory*, V. S. Pless and W. C. Huffman, Eds., 1998.
- [13] B. Marcus, R. M. Roth, and P. Siegel, "Constrained systems and coding for recording channels," in *Handbook of Coding Theory*, V. Pless and W. C. Huffman, Eds., 1998, pp. 1635–1764.
- [14] Z. Wang, A. G. Dimakis, and J. Bruck, "Rebuilding for array codes in distributed storage systems," in *Proc. Workshop Appl. Commun. Theory Emerging Memory Technol.*, 2010. [Online]. Available: <http://arxiv.org/abs/1009.3291>
- [15] L. Ahlswede, N. Cai, S.-Y. R. Li, and R. W. Yeung, "Network information flow," *IEEE Trans. Inf. Theory*, vol. 46, no. 4, pp. 1204–1216, Jul. 2000.
- [16] S.-Y. R. Li, R. W. Yeung, and N. Cai, "Linear network coding," *IEEE Trans. Inf. Theory*, vol. 49, no. 2, pp. 371–381, Feb. 2003.
- [17] L. Lima, J. Barros, M. Médard, and A. Toledo, "Protecting the code: Secure multiresolution network coding," in *Proc. IEEE Inf. Theory Workshop*, 2009.
- [18] R. Koetter and M. Médard, "An algebraic approach to network coding," *IEEE/ACM Trans. Netw.*, vol. 11, no. 5, pp. 782–795, Oct. 2003.
- [19] T. Ho, M. Médard, R. Koetter, D. Karger, M. Effros, J. Shi, and B. Leong, "A random linear network coding approach to multicast," *IEEE Trans. Inf. Theory*, vol. 52, no. 10, pp. 4413–4430, Oct. 2006.
- [20] S. Jaggi, P. Sanders, P. A. Chou, M. Effros, S. Egner, K. Jain, and L. Tolhuizen, "Polynomial time algorithms for network code construction," *IEEE Trans. Inf. Theory*, vol. 51, no. 6, pp. 1973–1982, Jun. 2005.
- [21] C. Fragouli, J. L. Boudec, and J. Widmer, "Network coding: An instant primer," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 36, no. 1, Jan. 2006, DOI:10.1145/1111322.1111337.
- [22] R. Dougherty, C. Freiling, and K. Zeger, "Insufficiency of linear coding in network information flow," *IEEE Trans. Inf. Theory*, vol. 51, no. 8, pp. 2745–2759, Aug. 2005.

- [23] A. Jiang, "Network coding for joint storage and transmission with minimum cost," in *Proc. Int. Symp. Inf. Theory*, Jul. 2006, pp. 1359–1363.
- [24] A. G. Dimakis, P. G. 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.
- [25] Y. Wu, A. G. Dimakis, and K. Ramchandran, "Deterministic regenerating codes for distributed storage," in *Proc. Allerton Conf. Control Comput. Commun.*, Monticello, IL, Oct. 2007.
- [26] Y. Wu, "Existence and construction of capacity-achieving network codes for distributed storage," in *Proc. IEEE Int. Symp. Inf. Theory*, Seoul, Korea, Jun. 2009, pp. 1150–1154.
- [27] Y. Wu, "Existence and construction of capacity-achieving network codes for distributed storage," *IEEE J. Sel. Areas Commun.*, vol. 28, no. 2, pp. 277–288, Feb. 2010.
- [28] M. A. Maddah-Ali, S. A. Motahari, and A. K. Khandani, "Communication over MIMO X channels: Interference alignment, decomposition, and performance analysis," *IEEE Trans. Inf. Theory*, vol. 54, no. 8, pp. 3457–3470, Aug. 2008.
- [29] V. R. Cadambe and S. A. Jafar, "Interference alignment and the degrees of freedom for the K user interference channel," *IEEE Trans. Inf. Theory*, vol. 54, no. 8, pp. 3425–3441, Aug. 2008.
- [30] C. Suh and D. Tse, "Interference alignment for cellular networks," in *Proc. Allerton Conf. Control Comput. Commun.*, Urbana, IL, Sep. 2008.
- [31] Y. Wu and A. G. Dimakis, "Reducing repair traffic for erasure coding-based storage via interference alignment," in *Proc. IEEE Int. Symp. Inf. Theory*, Seoul, Korea, Jul. 2009, pp. 2276–2280.
- [32] D. Cullina, A. G. Dimakis, and T. Ho, "Searching for minimum storage regenerating codes," in *Proc. Allerton Conf. Control Comput. Commun.*, Urbana, IL, Sep. 2009.
- [33] K. V. Rashmi, N. B. Shah, P. V. Kumar, and K. Ramchandran, "Exact regenerating codes for distributed storage," in *Proc. Allerton Conf. Control Comput. Commun.*, Urbana, IL, Sep. 2009. [Online]. Available: <http://arxiv.org/abs/0906.4913>
- [34] N. B. Shah, K. V. Rashmi, P. V. Kumar, and K. Ramchandran, "Explicit codes minimizing repair bandwidth for distributed storage," in *Proc. IEEE Inf. Theory Workshop*, Jan. 2010, DOI: 10.1109/ITWIKSPS.2010.5503165. [Online]. Available: <http://arxiv.org/abs/0908.2984>
- [35] C. Suh and K. Ramchandran, "Exact regeneration codes for distributed storage repair using interference alignment," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun. 2010. [Online]. Available: <http://arxiv.org/abs/1001.0107v2>
- [36] Y. Wu. (2009, Aug.). A construction of systematic MDS codes with minimum repair bandwidth. *IEEE Trans. Inf. Theory*. [Online]. Available: <http://arxiv.org/abs/0910.2486>
- [37] A. Duminuco and E. Biersack, "A practical study of regenerating codes for peer-to-peer backup systems," in *Proc. Int. Conf. Distrib. Comput. Syst.*, 2009, pp. 376–384.
- [38] J. Li, S. Yang, X. Wang, and B. Li, "Tree-structured data regeneration in distributed storage systems with regenerating codes," in *Proc. IEEE INFOCOM*, 2010, DOI: 10.1109/INFOCOM.2010.5462122.
- [39] T. Dikaliotis, A. G. Dimakis, and T. Ho, "Security in distributed storage systems by communicating a logarithmic number of bits," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun. 2010, pp. 1948–1952.
- [40] S. Pawar, S. El Rouayheb, and K. Ramchandran, "On security for distributed storage systems," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun. 2010.
- [41] S. Yekhanin, "Locally decodable codes," *Found. Trends Theor. Comput. Sci.*, 2010.
- [42] C. Huang and L. Xu, "STAR: An efficient coding scheme for correcting triple storage node failures," in *Proc. 4th Usenix Conf. File Storage Technol.*, San Francisco, CA, Dec. 2005.
- [43] D. Papailiopoulos and A. G. Dimakis, "Interference alignment as a rank constrained rank minimization," in *Proc. IEEE Global Telecommun. Conf.*, 2010, vol. 2, pp. 895–900.
- [44] D. Papailiopoulos and A. G. Dimakis, "Connecting distributed storage codes and secure degrees-of-freedom of multiple access channels," in *Proc. Allerton Conf. Control Comput. Commun.*, 2010.
- [45] V. Cadambe, S. Jafar, and H. Maleki, "Distributed data storage with minimum storage regenerating codes—Exact and functional repair are asymptotically equally efficient," in *Proc. IEEE Int. Workshop Wireless Netw. Coding*, Apr. 2010. [Online]. Available: <http://arxiv.org/abs/1004.4299>
- [46] K. 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*. [Online]. Available: <http://arxiv.org/abs/1005.4178>

## ABOUT THE AUTHORS

**Alexandros G. Dimakis** (Member, IEEE) received the Diploma degree in electrical and computer engineering from the National Technical University of Athens, Athens, Greece, in 2003 and the M.S. and Ph.D. degrees in electrical engineering and computer science from the University of California Berkeley, Berkeley, in 2005 and 2008, respectively.

Currently, he is an Assistant Professor at the Viterbi School of Engineering, University of Southern California, Los Angeles. He has been a faculty member in the Department of Electrical Engineering—Systems since 2009. He was a Postdoctoral Scholar at the Center for the Mathematics of Information (CMI), California Institute of Technology (Caltech), Pasadena, in 2008. His research interests include communications, coding theory, signal processing, and networking, with a current focus on distributed storage, network coding, large-scale inference, and message passing algorithms.

Dr. Dimakis received the Eliahu Jury award in 2008 for his thesis work on codes for distributed storage, two outstanding paper awards, the University of California Berkeley Regents Fellowship and the Microsoft Research Fellowship.



**Kannan Ramchandran** (Fellow, IEEE) is currently a Professor of Electrical Engineering and Computer Science at the University of California at Berkeley, Berkeley, where he has been since 1999. Prior to that, he was with the University of Illinois at Urbana-Champaign, Urbana, from 1993 to 1999, and was at AT&T Bell Laboratories from 1984 to 1990. His current research interests include distributed signal processing algorithms for wireless sensor and *ad hoc* networks, multimedia and peer-to-peer networking, multiuser information and communication theory, and wavelets and multiresolution signal and image processing. He has published extensively in his field and holds eight patents.

Prof. Ramchandran received the Eliahu Jury award for the best doctoral thesis in the systems area at Columbia University, the National Science Foundation (NSF) CAREER award, the Office of Naval Research (ONR) and ARO Young Investigator Awards, two Best Paper awards from the IEEE Signal Processing Society, a Hank Magnuski Scholar award for excellence in junior faculty at the University of Illinois, and an Okawa Foundation Prize for excellence in research at Berkeley. He serves as an active consultant to industry, and has held various editorial and Technical Program Committee positions.



**Yunnan Wu** (Member, IEEE) received the Ph.D. degree from Princeton University, Princeton, NJ, in January 2006.

Since August 2005, he has been a Researcher at Microsoft Corporation, Redmond, WA. His research interests include networking, graph theory, information theory, game theory, wireless communications, and multimedia.

Dr. Wu was a recipient of the Best Student Paper Award at the 2000 SPIE and IS&T Visual Communication and Image Processing Conference, and a recipient of the Student Paper Award at the 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing. He was awarded a Microsoft Research Graduate Fellowship for 2003–2005.



**Changho Suh** (Student Member, IEEE) received the B.S. and M.S. degrees in electrical engineering from Korea Advanced Institute of Science and Technology, Daejeon, Korea, in 2000 and 2002, respectively.

Since 2006, he has been with the Department of Electrical Engineering and Computer Science, University of California at Berkeley, Berkeley. Prior to that, he was with the Department of Telecommunication R&D Center, Samsung Electronics, Korea. His research interests include information theory and wireless communications.



Mr. Suh is a recipient of the Best Student Paper Award from the IEEE International Symposium on Information Theory 2009 and the Outstanding Graduate Student Instructor Award in 2010. He was awarded several fellowships: the Vodafone U.S. Foundation Fellowship in 2006 and 2007; the Kwanjeong Educational Foundation Fellowship in 2009; and the Korea Government Fellowship from 1996 to 2002.