Distributed Storage Allocations and a Hypergraph Conjecture of Erdős

Yi-Hsuan Kao*, Alexandros G. Dimakis†, Derek Leong‡§, and Tracey Ho‡
*Department of Electrical Engineering, University of Southern California, Los Angeles, California 90089, USA
†Department of Electrical & Computer Engineering, The University of Texas at Austin, Austin, Texas 78712, USA
‡Department of Electrical Engineering, California Institute of Technology, Pasadena, California 91125, USA
§Institute for Infocomm Research, Singapore 138632, Singapore
yihsuank@usc.edu, dimakis@austin.utexas.edu, dleong@i2r.a-star.edu.sg, tho@caltech.edu

***Abstract*—We study two variations of the distributed storage allocation problem. The goal is to allocate a given storage budget in a distributed storage system for maximum reliability. It was recently discovered that this problem is related to an old conjecture in extremal combinatorics, on the maximum number of edges in a hypergraph subject to a constraint on its maximum matching number. The conjecture was recently verified in some regimes. In this paper we assume that the conjecture is true and establish new results for the optimal allocation for a variety of parameter values. We also derive new performance bounds that are independent of the conjecture, and compare them to the best previously known bounds.**

I. Introduction

In *distributed storage allocation* problems, a source has a single data object of unit size that has to be coded and stored in n storage nodes. Let x_i be the amount of coded data stored in node $i \in \{1 \dots, n\}$. The total amount of introduced redundancy is subject to a given storage budget T, i.e.,

$$\sum_{i=1}^{n} x_i < T.$$

Each of the storage nodes fails with a probabilistic model and the question is to choose the allocation values x_i to maximize the probability that the data can be recovered. In [1], Leong *et al.* formulated three allocation models and derived performance bounds and optimal solutions for some cases. Allocation problems are very simple to state but surprisingly difficult to optimize. A few recent papers [1]–[3] have introduced methods to obtain approximate solutions with provable approximation guarantees.

In 1965, Erdős [4] gave a conjecture on the largest number of edges in a uniform hypergraph with bounded matching number. In a recent breakthrough paper, Alon *et al.* [5] introduced a fractional generalization of the Erdős hypergraph conjecture and showed it is asymptotically true. It further showed the fractional conjecture is connected to the distributed storage allocation problem for fixed-size subset access. Our contributions are summarized at the end of Section I, after we have completed the definitions.

The first allocation problem assumes access to a random fixed-size subset of nodes. In this model, a data collector attempts to reconstruct the data object by accessing an r-subset of the n storage nodes. The r nodes are randomly selected from the collection of all $\binom{n}{r}$ possible r-subsets. With an appropriate code (e.g., a maximum distance separable (MDS) code), the data object can be recovered if the total amount of coded bits accessed is at least the data object size [1]. Hence, successful recovery occurs if the data collector gets at least 1 unit of coded bits. The goal is to allocate coded bits over n nodes from a given storage budget T such that the probability of successful recovery is maximized. When $T > \frac{n}{r}$, simply allocating $\frac{1}{r}$ to each node guarantees recovery with probability 1. On the other hand, if $T \leq 1$, there is no way to recover. Therefore, we are interested in cases where $1 < T \leq \frac{n}{r}$. This problem can be formulated as follows:

Problem 1 (Fixed-size subset):

$$\Pi_F^*(n, r, T) = \frac{1}{\binom{n}{r}} \max_{x_1,\dots,x_n} \sum_{\substack{\mathbf{s}\subseteq[n]:\\ |\mathbf{s}|=r}} \mathbf{1}\left[\sum_{i\in\mathbf{s}} x_i \geq 1\right],$$
$$\text{subject to } \sum_{i=1}^{n} x_i < T, \quad x_i \geq 0 \ \ \forall\, i \in [n].$$

Here, $\mathbf{1}[\cdot]$ is the indicator function, and $[n]$ denotes the set $\{1, \dots, n\}$. Leong *et al.* [1] examined this problem for the regime of high recovery probability.

The second model assumes independent probabilistic access to each node. Specifically, the data collector accesses each of the n storage nodes independently with probability p. Similarly, our goal is to find an optimal allocation such that the recovery probability is maximized. This problem can be formulated as follows:

Problem 2 (Probabilistic access):

$$\Pi_I^*(n, p, T) = \max_{x_1,\dots,x_n} \sum_{\mathbf{r}\subseteq[n]} p^{|\mathbf{r}|}(1-p)^{n-|\mathbf{r}|}\mathbf{1}\left[\sum_{i\in\mathbf{r}} x_i \geq 1\right],$$
$$\text{subject to } \sum_{i=1}^{n} x_i < T, \quad x_i \geq 0 \ \ \forall\, i \in [n].$$

The work of Y.-H. Kao and A. G. Dimakis was supported in part by NSF Awards 1055099 and 1218235, and research gifts by Google, Intel, and Microsoft. The work of D. Leong and T. Ho was supported in part by the Air Force Office of Scientific Research under Grant FA9550-10-1-0166.

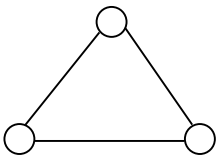

Fig. 1. An instance of the minimum vertex cover and the maximum matching.

Leong *et al.* [1] proved that for $pT > 1$, the symmetric allocation $x_1 = \cdots = x_n = \frac{T}{n}$ is optimal as n goes to infinity. (A *symmetric* allocation is one in which all nonempty nodes store the same amount of data.)

Observe that in Problem 2, the number of nodes accessed by the data collector is a binomial random variable with parameters n and p. We denote this random variable by X. If we know $X = r$, we can find the optimal allocation for accessing r nodes by solving Problem 1. In this paper, we use our new insights from Problem 1 to obtain stronger bounds for Problem 2.

We begin by explaining how Problem 1 can be interpreted as an extremal graph problem with appropriately defined constraints following the work of Alon *et al.* [5].

Definition 1 (Minimum Vertex Cover Problem)**.** An r-uniform hypergraph $H(r,n)$ is a pair (V,E), where $V = [n]$ is an index set of vertices, and $E = [m]$ is an index set of edges such that each edge $j \in E$ corresponds to an r-subset of V. Let V_j denote the r-subset of vertices that are incident to edge $j \in E$. The minimum vertex cover problem for H can be stated as follows:

$$\tau(H) = \min_{x_1,\ldots,x_n} \sum_{i\in V} x_i,$$
$$\text{subject to } \sum_{i\in V_j} x_i \geq 1 \ \ \forall\, j \in E, \quad x_i \in \{0,1\} \ \ \forall\, i \in V.$$

The optimal solution $\tau(H)$ is known as the minimum vertex cover. One can think of this problem as finding the smallest set of vertices that covers all the edges of H. Consider a relaxation of the minimum vertex cover problem in which the variables x_i's are allowed to be real numbers in the interval $[0,1]$. The resulting problem is a linear program, and the corresponding optimal solution $\tau^*(H)$ is known as the minimum fractional vertex cover. Trivially, we have $\tau^*(H) \leq \tau(H)$ for all H. Fig. 1 shows a graph with $n = 3$ and $r = 2$. We have to pick at least two vertices so that each edge is incident to at least one vertex. Thus, $\tau(H) = 2$ and is achieved by assigning $(x_1, x_2, x_3) = (1,1,0)$. On the other hand, $\tau^*(H) = 1.5$ is achieved by assigning $x_i = 0.5$ for each vertex $i \in V$.

In terms of Problem 1, each feasible allocation $(x_1,\ldots,x_n)$ can be mapped to an r-uniform hypergraph with vertex set $V = [n]$ and edge set E equal to the collection of successful r-subsets of nodes (i.e., with each subset storing at least a unit amount of data), and with $\tau^*(H) \leq \sum_{i=1}^n x_i < T$. Hence, Problem 1 can be equivalently formulated as follows:

$$\Pi_F^*(n,r,T) = \frac{1}{\binom{n}{r}} \max_{H(r,n)} \left\{|E(H)| : \tau^*(H) < T\right\}. \tag{1}$$

Therefore, we want to maximize the number of edges over all n-node r-uniform hypergraphs subject to $\tau^*(H) < T$.

The dual of the minimum vertex cover problem is the maximum matching problem, which is defined as follows:

Definition 2 (Maximum Matching Problem)**.** Consider an r-uniform hypergraph $H(r,n) = (V,E)$. Let E_i denote the set of edges that are incident to vertex $i \in V$. The maximum matching problem for H can be stated as follows:

$$\nu(H) = \max \sum_{j\in E} y_j,$$
$$\text{subject to } \sum_{j\in E_i} y_j \leq 1 \ \ \forall\, i \in V, \quad y_j \in \{0,1\} \ \ \forall\, j \in E.$$

The optimal solution $\nu(H)$ is known as the maximum matching number. Similarly, if we allow y_j's to be real numbers in the interval $[0,1]$, then the resulting optimal solution $\nu^*(H)$ is known as the maximum fractional matching number. By strong duality, we have

$$\nu(H) \leq \nu^*(H) = \tau^*(H) \leq \tau(H) \quad \forall\, H. \tag{2}$$

In the example of Fig. 1, $\nu(H) = 1$ is achieved by picking only one edge. However, $\nu^*(H) = 1.5$ is achieved by assigning $y_j = 0.5$ for each edge $j \in E$. This example also illustrates the relationship expressed in (2). Since $\nu^*(H) = \tau^*(H)$ for any H, Problem 1 is therefore also equivalent to the following:

$$\Pi_F^*(n,r,T) = \frac{1}{\binom{n}{r}} \max_{H(r,n)} \left\{|E(H)| : \nu^*(H) < T\right\}. \tag{3}$$

The Erdős hypergraph conjecture provides an upper bound for the problem in (3), as will be shown in Section III. Specifically, Erdős [4] conjectured that for any integer $t \leq \frac{n}{r}$, the number of edges in an r-uniform hypergraph H is at most $\max\left\{\binom{rt-1}{r}, \binom{n}{r} - \binom{n-t+1}{r}\right\}$ if $\nu(H) < t$.

We refer to this as the (integral) Erdős hypergraph conjecture, which can be written as follows:

Conjecture 1 (Erdős Hypergraph Conjecture)**.** *If integers n, r and t satisfy $2 \leq r \leq n$ and $1 \leq t \leq \frac{n}{r}$, and*

$$M^*(n,r,t) = \max_{H(r,n)} \left\{|E(H)| : \nu(H) < t\right\}, \tag{4}$$

$$M(n,r,t) = \max\left\{\binom{rt-1}{r}, \binom{n}{r} - \binom{n-t+1}{r}\right\}, \tag{5}$$

then $M^*(n,r,t) = M(n,r,t)$.

As a consequence of the Erdős-Gaillai Theorem [6], the graph case (i.e., $r = 2$) is true. Fig. 2 shows that the extremal graph can either contain an $(rt-1)$-clique or contain all the edges intersecting a fixed set of $t-1$ vertices. Both of them clearly do not have t disjoint edges. In recent work [7], the Erdős hypergraph conjecture was verified for all $t < \frac{n}{3r^2}$.

If we replace $\nu(H)$ with $\nu^*(H)$ in (4), then the resulting problem becomes the same as (3). Alon *et al.* [5] formulated the fractional Erdős hypergraph conjecture that addresses $\Pi_F^*(n,r,T)$ and the corresponding optimal allocation. Further they showed the asymptotic validity of their conjecture. The fractional Erdős hypergraph conjecture can be written as follows:

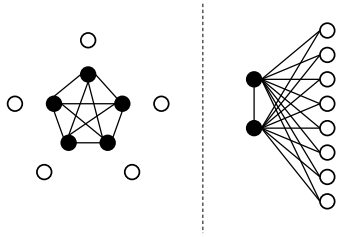

Fig. 2. Two extremal graphs with $n = 10,\ r = 2,\ t = 3$. In this example, the graph on the right-hand side achieves the maximum number of edges. In the context of distributed storage allocations, the left-hand side graph corresponds to maximal spreading, while the right-hand side graph corresponds to minimal spreading.

Conjecture 2 (Fractional Erdős Hypergraph Conjecture)**.** *If integers n, r and real number T satisfy $2 \le r \le n$ and $0 \le T \le \frac{n}{r}$, and*

$$S^*(n,r,T) = \max_{H(r,n)} \{|E(H)| : \nu^*(H) < T\},$$

$$S(n,r,T) = \max\left\{\binom{\lceil rT\rceil - 1}{r}, \binom{n}{r} - \binom{n-\lceil T\rceil + 1}{r}\right\}, \quad (6)$$

then $S^*(n,r,T) = S(n,r,T)$.

The solution in (6), which corresponds to an optimal allocation, can also be interpreted in terms of extremal graphs, as shown in Fig. 2. One graph corresponds to maximal spreading, which allocates $\frac{1}{r}$ equally over $\lceil rT\rceil - 1$ nodes such that any r-subset of them sums up to 1; this is maximal in the sense that there is no way to recover if we allocate less than $\frac{1}{r}$ symmetrically over nodes. The other graph corresponds to minimal spreading, which allocates 1 over $\lceil T\rceil - 1$ nodes such that any r-subset containing one of these nodes can recover. Conjecture 2 simply says that one of these two graphs corresponds to an optimal allocation. This is consistent with the heuristic principle of allocation proposed in [1]: when the budget is small, spread it minimally; when the budget is large, spread it maximally. However, there exist choices of (n,r,T) for which the optimal allocation is neither maximal spreading nor minimal spreading [1]. Thus, Conjecture 2 is not true in general.

Our Contributions: For Problem 1, we give an upper bound for $\Pi_F^*(n,r,T)$ that is independent of any conjecture. Assuming that the Erdős hypergraph conjecture is true, we derive an additional bound that is tight enough to settle $\Pi_F^*(n,r,T)$ in some cases. For Problem 2, by applying an appropriate Chernoff bound with the assumption that the original Erdős hypergraph conjecture is true, we verify $\Pi_I^*(n,p,T)$ for fixed (p,T) satisfying $p\lceil T\rceil < 1$ as n goes to infinity.

In the following section, we derive an upper bound for $\Pi_F^*(n,r,T)$ that is valid in general. The relationship between the Erdős hypergraph conjecture and $\Pi_F^*(n,r,T)$ is addressed in Section III. Section IV contains an asymptotic analysis of $\Pi_I^*(n,p,T)$. We compare these upper bounds and present some numerical results in Section V. Finally, we conclude in Section VI. Detailed results and proofs can be found in the extended paper [8].

II. An Upper Bound for $\Pi_F^*(n,r,T)$

The following lemma presents several upper bounds for $\Pi_F^*(n,r,T)$ that are independent of any conjecture:

Lemma 1. *For given n, r and T satisfying $1 < T \le \frac{n}{r}$, $\Pi_F^*(n,r,T)$ has the following upper bounds:*

$$\Pi_F^*(n,r,T) \le 1 - \frac{\gcd(r,r')}{\alpha\gcd(r,r') + r'} \qquad \text{if } T < \frac{n}{r}, \quad (7)$$

$$\Pi_F^*(n,r,T) \le \lfloor T\rfloor \frac{r}{n} \qquad \text{if } T < \left\lfloor \frac{n}{r}\right\rfloor, \quad (8)$$

$$\Pi_F^*(n,r,T) \le \frac{\lfloor \beta T\rfloor}{\beta}\frac{r}{n}, \quad (9)$$

where α and r' are uniquely defined integers satisfying $n = \alpha r + r'$, $\alpha \in \mathbb{Z}_0^+$, and $r' \in \{r, \cdots, 2r-1\}$, and $\beta \equiv \frac{\mathrm{lcm}(n,r)}{n}$.

Upper bound (7) is a corollary of [1, Theorems 5 and 6], while upper bounds (8) and (9) are derived using permutation counting arguments similar to Katona's proof of the Erdős-Ko-Rado theorem [9], [10]. A combined upper bound for $\Pi_F^*(n,r,T)$ can be obtained by picking the tightest of these bounds and applying the fact that $\Pi_F^*(n,r,T)$ is at most 1:

$$\Pi_F^*(n,r,T) \le \min\left\{1 - \mathbf{1}\left[T < \frac{n}{r}\right]\cdot\frac{\gcd(r,r')}{\alpha\gcd(r,r')+r'},\right.$$
$$\left. 1 - \mathbf{1}\left[T < \left\lfloor\frac{n}{r}\right\rfloor\right]\cdot\left(1 - \lfloor T\rfloor\frac{r}{n}\right), \frac{\lfloor\beta T\rfloor}{\beta}\frac{r}{n}\right\}. \quad (10)$$

III. An Upper Bound for $\Pi_F^*(n,r,T)$ based on the Erdős hypergraph conjecture

The Erdős hypergraph conjecture considers integer t values only. However, when T is not an integer, say $t-1 < T < t$, $M^*(n,r,t)$ and $M^*(n,r,T)$ are the same due to the fact that both of them are maximized over the same set $\{H : \nu(H) < t\}$ since $\nu(H)$ is an integer. The following lemma enables us to get a stronger upper bound:

Lemma 2. *If integers n, r and real number T satisfy $0 < r < n$ and $1 < T \le \frac{n}{r}$, then $S^*(n,r,T) \le M^*(n,r,\lceil T\rceil)$.*

Proof: We compare the set $\mathcal{S}_1$ and $\mathcal{S}_2$, where

$$\mathcal{S}_1 = \{H(r,n) : \nu^*(H) < T\}, \quad \mathcal{S}_2 = \{H(r,n) : \nu(H) < \lceil T\rceil\}.$$

It suffices to show that $\mathcal{S}_1 \subseteq \mathcal{S}_2$. Let $\mathcal{S}_2' = \{H(r,n) : \nu(H) < T\}$. Since $\nu(H)$ is a nonnegative integer for any graph H, we have $\mathcal{S}_2' = \mathcal{S}_2$. From (2), we know that $\nu(H) \le \nu^*(H)$ for any graph H. Hence, $\mathcal{S}_1 \subseteq \mathcal{S}_2'$, which implies $\mathcal{S}_1 \subseteq \mathcal{S}_2$. ∎

Assuming that the conjecture is true, we have the following bounds as a consequence of Lemma 2:

Theorem 1. *If Conjecture 1 is true, then for $1 < T \le \frac{n}{r}$,*

$$\frac{1}{\binom{n}{r}} S(n,r,T) \le \Pi_F^*(n,r,T) \le \frac{1}{\binom{n}{r}} M(n,r,\lceil T\rceil).$$

Here, $S(n,r,T)$ and $M(n,r,\lceil T\rceil)$ are as defined in (5) and (6). The upper bound follows from Lemma 2 and (3). The lower bound follows from the two feasible allocations minimal spreading and maximal spreading. This theorem also implies we can find $\Pi_F^*(n,r,T)$ when the upper bound matches

the lower bound. Furthermore, the optimal allocation is consistent with Conjecture 2. That is, in the regimes where $S(n,r,T) = M(n,r,\lceil T\rceil)$, we can verify Conjecture 2 by assuming Conjecture 1 is true. We summarize these results in the following corollaries:

Corollary 1. *If Conjecture 1 is true and $T \in \mathcal{R}_1(n,r) \cup \mathcal{R}_2(n,r)$, where*

$$\mathcal{R}_1(n,r) = \left\{T : \binom{r\lceil T\rceil - 1}{r} \le \binom{n}{r} - \binom{n-\lceil T\rceil+1}{r}\right\},$$
$$\mathcal{R}_2(n,r) = \{T : r\lceil T\rceil = \lceil rT\rceil\},$$

then Conjecture 2 is true.

Corollary 1 specifies the regimes where $M(n,r,\lceil T\rceil) = S(n,r,T)$ in Theorem 1. Hence, $\Pi_F^*(n,r,T)$ can be verified to be $\frac{1}{\binom{n}{r}}S(n,r,T)$ as claimed in Conjecture 2. We study some cases for fixed (r,T) and growing n. First, $\mathcal{R}_2(n,r)$ does not depend on n. Therefore, if $T \in \mathcal{R}_2(n,r)$, the two bounds merge for all n and the optimal solution is either maximal spreading or minimal spreading. On the other hand, if $T \in \mathcal{R}_1(n,r)$, i.e., $\binom{r\lceil T\rceil-1}{r} \le \binom{n}{r} - \binom{n-\lceil T\rceil+1}{r}$, then $\binom{\lceil rT\rceil-1}{r} \le \binom{n}{r} - \binom{n-\lceil T\rceil+1}{r}$ since $r\lceil T\rceil \ge \lceil rT\rceil$. Hence, the two bounds merge and the optimal solution is minimal spreading. For the latter case, we are interested in how the upper bound matches the lower bound as n grows; this behavior is expressed in the following corollary:

Corollary 2. *If $T > 1$, $\lceil T\rceil < \frac{n+1}{r}$, and $n \ge (r+1)\lceil T\rceil - 2$, then*

$$\binom{n}{r} - \binom{n-\lceil T\rceil+1}{r} \ge \binom{r\lceil T\rceil - 1}{r}.$$

Together with Theorem 1, Corollary 2 specifies a sufficient condition for minimal spreading to be optimal.

IV. Asymptotic Analysis of $\Pi_I^*(n,p,T)$

In this section, we derive upper and lower bounds for $\Pi_I^*(n,p,T)$ using a Chernoff bound. Further, assuming Conjecture 1 is true, we analyze how the gap between upper and lower bounds decreases as n grows.

Let random variable X be the number of nodes accessed by the data collector in Problem 2; thus X is a binomial random variable with parameters (n,p). The probability that X deviates from its mean by a given distance can be bounded by the following Chernoff bound:

$$\mathbb{P}\{|X - np| \ge \delta np\} \le 2\mathrm{e}^{-\frac{np\delta^2}{3}}. \tag{11}$$

This inequality implies that the event that X is far from its mean is rare when n is large enough. Based on the formulation in Problem 2, the following theorem describes how $\Pi_I^*(n,p,T)$ can be bounded:

Theorem 2. *Given (n,p,T),*

$$\left(1-\frac{2}{n}\right)\Pi_F^*(n,r_{\min},T) \le \Pi_I^*(n,p,T) \le \left(1-\frac{2}{n}\right)\Pi_F^*(n,r_{\max},T)+\frac{2}{n},$$

where

$$r_{\min} = \left\lfloor np - \sqrt{3np\ln n}\right\rfloor + 1,\ r_{\max} = \left\lceil np + \sqrt{3np\ln n}\right\rceil - 1.$$

Proof: We first derive the upper bound. From Problem 2, $\Pi_I^*(n,p,T)$ can be equivalently written as

$$\max_{x_1,\ldots,x_n} \sum_{r=0}^{n} \mathbb{P}\{X=r\}\frac{1}{\binom{n}{r}} \sum_{\substack{\mathbf{s}\subseteq[n]:\\ |\mathbf{s}|=r}} \mathbf{1}\left[\sum_{i\in\mathbf{s}} x_i \ge 1\right]. \tag{12}$$

Hence, $\Pi_I^*(n,p,T)$ can be upper bounded by maximizing the recovery probability corresponding to individual events $X = r$. Thus,

$$\begin{aligned}
\Pi_I^*(n,p,T) &\le \sum_{r=0}^{n} \mathbb{P}\{X=r\}\frac{1}{\binom{n}{r}} \max_{x_1,\ldots,x_n} \sum_{\substack{\mathbf{s}\subseteq[n]:\\ |\mathbf{s}|=r}} \mathbf{1}\left[\sum_{i\in\mathbf{s}} x_i \ge 1\right] \\
&= \sum_{r=0}^{n} \mathbb{P}\{X=r\}\Pi_F^*(n,r,T) &(13)\\
&= \mathbb{P}\{X\in R\}\mathbb{E}\{\Pi_F^*(n,X,T) \mid X\in R\} \\
&\quad + \mathbb{P}\{X\notin R\}\mathbb{E}\{\Pi_F^*(n,X,T) \mid X\notin R\} &(14)\\
&\le (1-\epsilon)\mathbb{E}\{\Pi_F^*(n,X,T) \mid X\in R\} + \epsilon \\
&\le (1-\epsilon)\Pi_F^*(n,r_{\max},T) + \epsilon. &(15)
\end{aligned}$$

In (14), we define $R = \{r_{\min}, r_{\min}+1, \ldots, r_{\max}\}$. The optimal recovery probability $\Pi_F^*(n,X,T)$ is a function of random variable X. Further, let $\mathbb{P}\{X\in R\} \ge 1-\epsilon$, where $0 \le \epsilon \le 1$. The loosest upper bound given by (15) follows from the fact that $\Pi_F^*(n,r,T)$ is increasing with r.

A lower bound can also be derived from (12) in a similar manner:

$$\begin{aligned}
\Pi_I^*(n,p,T) &\ge \max_{x_1,\ldots,x_n} \sum_{r\in R} \mathbb{P}\{X=r\} \sum_{\substack{\mathbf{s}\subseteq[n]:\\ |\mathbf{s}|=r}} \frac{1}{\binom{n}{r}}\mathbf{1}\left[\sum_{i\in\mathbf{s}} x_i \ge 1\right] \\
&\ge \max_{x_1,\ldots,x_n} \sum_{r\in R} \mathbb{P}\{X=r\} \sum_{\substack{\mathbf{s}\subseteq[n]:\\ |\mathbf{s}|=r_{\min}}} \frac{1}{\binom{n}{r_{\min}}}\mathbf{1}\left[\sum_{i\in\mathbf{s}} x_i \ge 1\right] &(16)\\
&= \mathbb{P}\{X\in R\}\Pi_F^*(n,r_{\min},T) \\
&\ge (1-\epsilon)\Pi_F^*(n,r_{\min},T). &(17)
\end{aligned}$$

Inequality (16) follows from the fact that for a data collector that is allowed to access r nodes, the recovery probability of a given allocation $(x_1,\ldots,x_n)$ increases with r. The rest follows similarly as the derivation of the upper bound.

Combining (15) and (17) yields

$$(1-\epsilon)\Pi_F^*(n,r_{\min},T) \le \Pi_I^*(n,p,T) \le (1-\epsilon)\Pi_F^*(n,r_{\max},T)+\epsilon. \tag{18}$$

Finally, applying the Chernoff bound (11) with $\delta = \sqrt{\frac{3}{np}\ln n}$ and $\epsilon = \frac{2}{n}$ produces the required bounds. ∎

If Conjecture 1 is true, then the asymptotically optimal allocation for $p\lceil T\rceil < 1$ can be inferred from the limit of $\Pi_I^*(n,p,T)$ as n goes to infinity:

Theorem 3. *If Conjecture 1 is true, and p and T are fixed such that $T > 1$ and $p\lceil T\rceil < 1$, then*

$$\lim_{n\to\infty} \Pi_I^*(n,p,T) = 1 - (1-p)^{\lceil T\rceil - 1},$$

and

$$\Pi_F^*(n,r_{\max},T) - \Pi_F^*(n,r_{\min},T) < c\cdot\sqrt{\frac{p\ln n}{n}} \tag{19}$$

for sufficiently large n, where c is a constant.

As a corollary of this theorem, minimal spreading is asymptotically optimal for the specified regime since its recovery probability matches $\lim_{n\to\infty}\Pi_I^*(n,p,T)$.

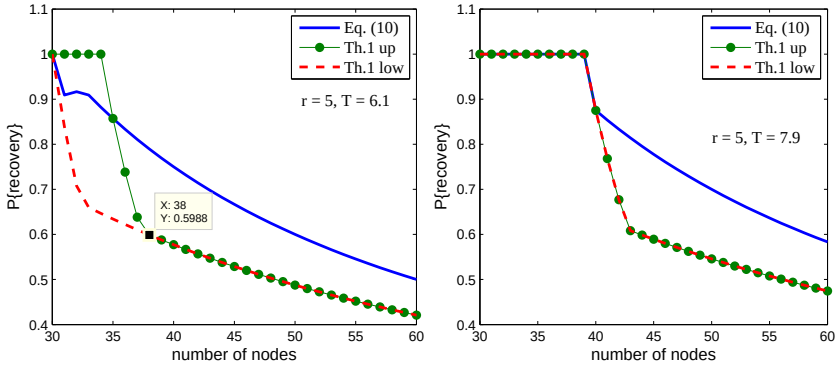

Fig. 3. Plots of the derived bounds against n, for fixed (r, T). In the left-hand side plot, the upper and lower bounds for $\Pi_F^*(n, r, T)$ from Theorem 1 (which are conditioned on Conjecture 1 being true) match for $n \geq 38$; in comparison, Corollary 2 provides a sufficient condition of $n \geq 40$. In the right-hand side plot, both bounds match for all n and so the corresponding optimal allocations can be easily inferred.

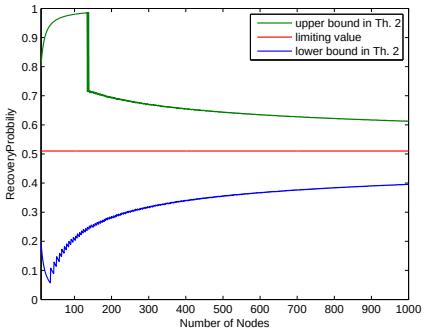

Fig. 4. Convergence of the bounds in Theorem 2 for $p = 0.3$, $T = 2.1$.

TABLE I
BOUNDS FOR REGIMES IN WHICH THE ERDŐS CONJECTURE IS VERIFIED

Regime	Upper bound for $\Pi_F^*(n, r, T)$
$r = 2,\ T \in \mathcal{R}_1$	$\frac{1}{\binom{n}{r}}\left(\binom{n}{r} - \binom{n-\lceil T\rceil+1}{r}\right)$
$T \in \mathcal{R}_1,\ T \leq \frac{n}{3r^2}$	
$r = 2,\ T \in \mathcal{R}_2$	$\frac{1}{\binom{n}{r}} \max\{\binom{\lceil rT\rceil-1}{r}, \binom{n}{r} - \binom{n-\lceil T\rceil+1}{r}\}$
$T \in \mathcal{R}_2,\ T \leq \frac{n}{3r^2}$	
$T \leq \frac{n}{3r^2}$	$\frac{1}{\binom{n}{r}} \max\{\binom{r\lceil T\rceil-1}{r}, \binom{n}{r} - \binom{n-\lceil T\rceil+1}{r}\}$

V. NUMERICAL RESULTS

Fig. 3 illustrates how the value of $\Pi_F^*(n, r, T)$ can be established using the bounds in Theorem 1. We observe that upper bound (10), which is independent of the conjecture, is tighter than the upper bound of Theorem 1 in the high recovery probability regime. Fig. 4 shows the convergence of the bounds in Theorem 3 when $p = 0.3$, $T = 2.1$. Both upper and lower bounds converge to the limit $1 - (1-p)^{\lceil T\rceil - 1}$ as n goes to infinity, at a rate of about $\frac{1}{\sqrt{n}}$ according to (19). Minimal spreading is asymptotically optimal in this regime. Table I presents verified bounds that are based on Conjecture 1.

VI. CONCLUSION

We examined two variations of the distributed storage allocation problem. For the case of access to a random fixed-size subset of nodes, two upper bounds for the optimal recovery probability were presented; the first is a general upper bound, while the second is based on the assumption that the Erdős hypergraph conjecture is true. Using the second bound, we determined the optimal allocation for a variety of cases. We also presented bounds for regimes in which the conjecture is verified; for these cases, the optimal allocation can be easily inferred. For the case of independent probabilistic access to each node, we determined an asymptotically optimal allocation in a specific regime by applying a Chernoff bound, assuming that the Erdős hypergraph conjecture is true.

ACKNOWLEDGMENT

The authors would like to thank Shaowei Lin, Liang Ze Wong, and Daniel Chen for the helpful discussions.

REFERENCES

[1] D. Leong, A. G. Dimakis, and T. Ho, "Distributed storage allocations," *IEEE Trans. Inf. Theory*, vol. 58, no. 7, pp. 4733–4752, Jul. 2012.
[2] M. Sardari, R. Restrepo, F. Fekri, and E. Soljanin, "Memory allocation in distributed storage networks," in *Information Theory Proceedings (ISIT), 2010 IEEE International Symposium on*. IEEE, 2010, pp. 1958–1962.
[3] V. Ntranos, G. Caire, and A. Dimakis, "Allocations for heterogenous distributed storage," in *Information Theory Proceedings (ISIT), 2012 IEEE International Symposium on*. IEEE, 2012, pp. 2761–2765.
[4] P. Erdős, "A problem on independent r-tuples," *Ann. Univ. Sci. Budapest. Eotvos Sect. Math.*, vol. 8, pp. 93–95, 1965.
[5] N. Alon, P. Frankl, H. Huang, V. Rödl, A. Ruciński, and B. Sudakov, "Large matchings in uniform hypergraphs and the conjectures of Erdős and Samuels," *Journal of Combinatorial Theory Series A*, vol. 119, pp. 1200–1215, 2012.
[6] P. Erdős and T. Gallai, "On maximal paths and circuits of graphs," *Acta Math. Acad. Sci. Hungar.*, vol. 10, pp. 337–356, 1959.
[7] H. Huang, P.-S. Loh, and B. Sudakov, "The size of a hypergraph and its matching number," *Cambridge Univ. Press, Combinatorics, Probability and Computing*, vol. 21, pp. 442–450, 2012.
[8] Y.-H. Kao, A. G. Dimakis, D. Leong, and T. Ho, "Distributed storage allocations and a hypergraph conjecture of Erdős." [Online]. Available: http://www.purl.org/net/ISIT2013
[9] J. H. van Lint and R. M. Wilson, *A Course in Combinatorics*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2001.
[10] G. O. H. Katona, "Extremal problems for hypergraphs," in *Combinatorics*, M. Hall, Jr. and J. H. van Lint, Eds. Dordrecht, Holland: D. Reidel, 1974.