Universally Sparse Hypergraphs with Applications to Coding Theory

Chong Shangguan*, Itzhak Tamo†

*Department of Electrical Engineering-Systems, Tel Aviv University, Tel Aviv 6997801, Israel, theoreming@163.com
†Department of Electrical Engineering-Systems, Tel Aviv University, Tel Aviv 6997801, Israel, zactamo@gmail.com

Abstract—For fixed integers $r \geq 2, e \geq 2, v \geq r + 1$, an $r$-uniform hypergraph is called $\mathcal{G}_r(v, e)$-free if the union of any $e$ distinct edges contains at least $v+1$ vertices. Let $f_r(n, v, e)$ denote the maximum number of edges in a $\mathcal{G}_r(v, e)$-free $r$-uniform hypergraph on $n$ vertices. Brown, Erdős and Sós showed in 1973 that there exist constants $c_1, c_2$ depending only on $r, v, e$ such that $c_1 n^{\frac{r-1}{r}} \leq f_r(n, v, e) \leq c_2 n^{\frac{r-1}{r}}$. For $e - 1 \mid er - v$, the lower bound matches the upper bound up to a constant factor; whereas for $e - 1 \nmid er - v$, it is a notoriously hard problem to determine the correct exponent of $n$. Our main result is an improvement $f_r(n, v, e) = \Omega(n^{\frac{r-1}{r}} \log n)^{\frac{r-1}{r+1}}$ for any $r, e, v$ satisfying $\gcd(e - 1, er - v) = 1$. Moreover, the hypergraph we constructed is not only $\mathcal{G}_r(v, e)$-free but also universally $\mathcal{G}_r(ir - \lfloor \frac{(r-1)(er-v)}{r} \rfloor, i)$-free for every $2 \leq i \leq e$. Interestingly, our new lower bound provides improved constructions for several seemingly unrelated topics in Coding Theory, namely, Parent-Identifying Set Systems, uniform Combinatorial Batch Codes and optimal Locally Recoverable Codes.

For a full version [1], see: https://arxiv.org/abs/1902.05903

I. Introduction

For an integer $r \geq 2$, an $r$-uniform hypergraph (henceforth an $r$-graph) $\mathcal{H} := (V(\mathcal{H}), E(\mathcal{H}))$ is a pair of vertices and edges, where the vertex set $V(\mathcal{H})$ is a finite set and the edge set $E(\mathcal{H})$ is a collection of $r$-subsets of $V(\mathcal{H})$. An $r$-graph is called $\mathcal{H}$-free if it contains no subgraph which is a copy of $\mathcal{H}$. For a family $\mathcal{H}$ of $r$-graphs, the Turán number, $\text{ex}_r(n, \mathcal{H})$, is the maximum number of edges in an $r$-graph on $n$ vertices which is $\mathcal{H}$-free for every $\mathcal{H} \in \mathcal{H}$. Since the pioneering work of Turán [2], the study of Turán-type problems has been playing a central role in the field of extremal graph theory. In this work, we present an improved probabilistic lower bound for an interesting hypergraph Turán-type problem introduced by Brown, Erdős and Sós [3] in 1973. Moreover, we show that the new bound provides improved constructions for several topics in Coding Theory, as mentioned in the abstract.

Notations. An $r$-graph $\mathcal{H}$ always stands for its edge set $E(\mathcal{H})$; the vertex set $V(\mathcal{H})$ can be simply viewed as a subset of the first $n$ integers, $[n] := \{1, \ldots, n\}$; given a finite set $X$, the family of $\binom{|X|}{r}$ distinct $r$-subsets of $X$ is denoted by $\binom{X}{r}$. Hence, $\mathcal{H} = E(\mathcal{H}) \subseteq \binom{[n]}{r}$. We frequently use the standard Bachmann-Landau notations $\Omega(\cdot), \Theta(\cdot), O(\cdot)$ and $o(\cdot)$.

A. Sparse hypergraphs

For integers $r \geq 2, e \geq 2, v \geq r + 1$, let $\mathcal{G}_r(v, e)$ be the family of all $r$-graphs formed by $e$ edges and at most $v$ vertices; that is, $\mathcal{G}_r(v, e) = \{ \mathcal{H} \subseteq \binom{[n]}{r} : |\mathcal{H}| = e, |V(\mathcal{H})| \leq v \}$. An $r$-graph $\mathcal{H}$ is called $\mathcal{G}_r(v, e)$-free if it does not contain any member of $\mathcal{G}_r(v, e)$, namely, the union of any $e$ distinct edges in $\mathcal{H}$ contains at least $v + 1$ vertices. In the literature, such $r$-graphs are termed sparse [4]. As in the previous papers, we use the notation $f_r(n, v, e) := \text{ex}_r(n, \mathcal{G}_r(v, e))$.

Since the study of $f_r(n, v, e)$ for $e = 2$ or $r = 2$ has been quite extensive (see, e.g. [1]), here we focus on the asymptotic behavior of $f_r(n, v, e)$ for fixed integers $r \geq 3, e \geq 3, v \geq r + 1$ as $n \to \infty$. It was shown in [3] that in general

$$\Omega\left(n^{\frac{r-1}{r}} \log n\right) = f_r(n, v, e) = O\left(n^{\frac{r-1}{r}} \log n\right).$$

(1)

Observe that the exponent of $n$ in (1) is tight for $e - 1 \mid er - v$; however, for $e - 1 \nmid er - v$, it is a notoriously hard problem to determine the correct order of the exponent. In particular, for $r > k \geq 2, e \geq 3$ and $v = e(r - k) + k + 1$, the determination of $f_r(n, e(r - k) + k + 1, e)$ has attracted a lot of attention since 1973, and has been studied in depth for more than forty years. Improvements of (1) on sporadic or less general parameters have been obtained in a line of other works. We refer the reader to the full version of this paper [1] for a comprehensive review on the previous results. Despite the efforts of many researchers, the lower bound $f_r(n, e(r - k) + k + 1, e) = \Omega\left(n^{\frac{r-1}{r-1}} \log n\right)$ implied by (1) remains the best possible for $e \geq 4, r > k \geq 3$ and $e \notin \{3, 4, 5, 7, 8\}, r > k = 2$. We slightly improve the lower bound above by a $(\log n)^{\frac{r-1}{r+1}}$ factor.

Proposition I.1. For fixed integers $r > k \geq 2, e \geq 3$ and $n \to \infty$, $f_r(n, e(r - k) + k + 1, e) = \Omega\left(n^{\frac{r-1}{r-1}} \log n\right)$. Indeed, Proposition I.1 is an easy consequence of the following more general lower bound on $f_r(n, v, e)$.

Theorem I.2. For fixed integers $r \geq 2, e \geq 3, v \geq r + 1$ satisfying $\gcd(e - 1, er - v) = 1$ and sufficiently large $n$, there exists an $r$-graph with $\Omega\left(n^{\frac{r-1}{r}} \log n\right)$ edges, which is simultaneously $\mathcal{G}_r(ir - \lfloor \frac{(i-1)(er-v)}{r} \rfloor, i)$-free for every $2 \leq i \leq e$; in particular, for $n \to \infty$, $f_r(n, v, e) = \Omega\left(n^{\frac{r-1}{r-1}} \log n\right)$.

The proof of this theorem is presented in Section III. To see that Proposition I.1 indeed follows from Theorem I.2, one can write $er - v = k(e - 1) + q$ for some $1 \leq q \leq e - 2$, then $\gcd(e - 1, er - v) = 1$ for $q = 1$, implying Proposition I.1.

B. Universally sparse hypergraphs

Theorem I.2 provides a construction of an $r$-graph $\mathcal{H}$ that is not only sparse, but also universally sparse (for the union
Bujtás and Tuza [5] studied the following extremal problem which is related to the construction of uniform Combinatorial Batch Codes (see Section II below). An r-graph is said to be \( \mathcal{H}(q,e) \)-free if it is simultaneously \( \mathcal{H}_r(i-q-i,1) \)-free for every \( 1 \leq i \leq e \). In [5] it was shown that for fixed integers \( r \geq 2, e > q + r \geq 1 \), \( \exp_r(n, \mathcal{H}(q,e)) = \Omega(n^{r^{-1} + \frac{1}{2} + \frac{1}{e}}) \). The following proposition, whose proof is postponed to Section II, is a direct consequence of Theorem I.2.

**Proposition I.3.** For fixed integers \( r \geq 2, e > q + r \geq 1 \) with \( \gcd(e-1,r+q) = 1 \) and \( n \to \infty \), \( \exp_r(n, \mathcal{H}(q,e)) = \Omega(n^{r^{-1} + \frac{1}{2} + \frac{1}{e}} (\log n)^{\frac{1}{e}}) \).

For integers \( t \geq 2 \) and \( r \geq 2 \), a Berge t-cycle in an r-graph is a set of \( t \) distinct vertices \( v_1, \ldots, v_t \) associated with \( t \) distinct edges \( A_1, \ldots, A_t \) such that \( \{v_{i-1}, v_i\} \subseteq A_i \) for \( 2 \leq i \leq t \) and \( \{v_1, v_t\} \subseteq A_t \). An r-graph is said to be \( \mathcal{B}_t \)-free if it contains no Berge cycles of length at most \( t \). Note that for \( r = 2 \), Berge cycles coincide with the usual graph cycles, and hence a 2-graph is \( \mathcal{B}_t \)-free if and only if it has a girth at least \( t + 1 \), where the girth of a 2-graph is the length of a shortest cycle contained in it. The reader is referred to [1], [6] for known results on \( \exp_r(n, \mathcal{B}_t) \) for \( r = 2 \) or \( t \leq 4 \).

Recently, Xing and Yuan [7] used \( \mathcal{B}_t \)-free r-graphs to construct optimal Locally Recoverable Codes (see Section II below) and they showed that for any \( r \geq 3 \) and \( t \geq 5 \), \( \exp_r(n, \mathcal{B}_t) = \Omega(n^{r^{-1} + \frac{1}{t}}) \). It is not hard to verify (see, e.g. Theorem 5.1 in [7]) that an r-graph \( H \) is \( \mathcal{B}_t \)-free if and only if it is simultaneously \( \mathcal{H}_r(i-r,1) \)-free for every \( 1 \leq i \leq t \). By applying Theorem I.2 with \( v = tv-t \) and \( e = t \), it is easy to prove the following result.

**Proposition I.4.** For fixed integers \( r \geq 3, t \geq 5 \), and \( n \to \infty \), \( \exp_r(n, \mathcal{B}_t) = \Omega(n^{r^{-1} + \frac{1}{t}} (\log n)^{\frac{1}{t}}) \); or equivalently, there exists an r-graph with such number of edges, which is simultaneously \( \mathcal{H}_r(i-r,1) \)-free for every \( 1 \leq i \leq t \).

Finally we remark that in [7] the authors stated that in a private communication, Jacques Verstraëte suggested that a lower bound on \( \exp_r(n, \mathcal{B}_t) \), which is exactly the same with Proposition I.4, can also be proved by using the method of [8], [9] (known as the differential equations method for random graph processes, which is rather involved). Nevertheless, since [7] stated this result (see also Proposition II.7 below) without a proof, we present it here as an easy consequence of Theorem I.2.

**C. Applications**

In Section II we will discuss the applications of Theorem I.2 in Parent-Identifying Set Systems (IPPSs for short\(^1\)), uniform Combinatorial Batch Codes (uniform CBCs for short) and optimal Locally Recoverable Codes (optimal LRCs for short). For an introduction to more applications of sparse and universally sparse hypergraphs, see [1].

\(^1\)Since Parent-Identifying Set Systems are also known as Set Systems with the Identifiable Parent Property, the abbreviation IPPS was used, e.g. in [10].

## II. Applications

### A. Parent-Identifying Set Systems

An r-graph \( H \subseteq \binom{[n]}{r} \) is said to be a t-Parent-Identifying Set System (t-IPPS), denoted as t-IPPS\((r, |H|, n)\), if for any r-subset \( X \subseteq [n] \) which is contained in the union of at most \( t \) edges of \( H \), it holds that \( \cap_{P \in P_t(X)} P \neq \emptyset \), where \( P_t(X) = \{ P \subseteq H : |P| \leq t, X \subseteq \cup_{A \in P} A \} \). Any \( P \in P_t(X) \) is called a parent of \( X \).

Based on the combinatorial model introduced in [11], [12], IPPSs were introduced by Collins [13] as a technique to trace traitors in a secret sharing scheme. Generally speaking, an \((n, r)\)-threshold secret sharing scheme (see, e.g. [14], [15]) has one message and \( n \) keys such that any set of at least \( r \) keys can be used to decrypt this message but no set of fewer than \( r \) keys can. Consider the following model in [11]–[13]. Assume that there is a data supplier who wants to distribute a message to \( m \) users through an \((n, r)\)-threshold secret sharing scheme such that every user gets exactly \( r \) keys. Suppose there are at most \( t \) illegal users and a coalition of at most \( t \) users may collude by combining some of their keys to produce a new, unauthorized set \( T \) of \( r \) keys to decrypt this message.

Let \( H \) be a t-IPPS\((r, m, n)\) whose \( n \) vertices and \( m \) edges are indexed by the \( n \) keys and the \( m \) users, respectively. If the data supplier distributes the keys to the users according to \( H \) such that for \( 1 \leq i \leq m \), the \( i \)th user gets the \( r \) keys which form the \( i \)th edge of \( H \), then upon capturing an unauthorized set \( T \), the data supplier is able to identify at least one illegal user who contributed to \( T \). Indeed, by the property of a t-IPPS\((r, m, n)\), we have that \( \cap_{P \in P_t(T)} P \neq \emptyset \). By applying a brute-force search to all possible coalitions of at most \( t \) users, the data supplier can find at least one colluder (i.e., illegal user) who is guaranteed to be contained in the non-empty intersection of all possible coalitions of at most \( t \) users who can produce the unauthorized key \( T \).

For a t-IPPS\((r, m, n)\) with given \( t, r \) and \( n \), it was shown in [13] that \( m = O(n^{\frac{1}{r(r+1)}}) \). A better upper bound on \( m \) was obtained by Gu and Miao [10], showing that \( m = O(n^{\frac{1}{r^2}}) \). Recently, Gu et al. [16] showed that for fixed integers \( t, r \geq 2 \), there exists a t-IPPS\((r, m, n)\) with \( m = \Omega(n^{\frac{1}{r(r+1)}}) \), which implies that for \( (\lfloor t^2/4 \rfloor + t) r \) the upper bound in [10] is tight up to a constant factor.

The following proposition slightly improves the lower bound obtained in [16] for some pairs of \((r, t)\).

**Proposition II.1.** For fixed integers \( t, r \geq 2 \) satisfying \( \gcd(\lfloor t^2/4 \rfloor + t, r) = 1 \) and \( n \to \infty \), there exists a t-IPPS\((r, m, n)\) with \( m = \Omega(n^{\frac{1}{r(r+1)}} (\log n)^{\frac{1}{r(r+1)}}) \).

Proposition II.1 is proved by establishing a connection between IPPSs and sparse hypergraphs, as stated below.

**Lemma II.2.** Assume that \( H \subseteq \binom{[n]}{r} \) is a \( \mathcal{B}_r(\lfloor tr-r, e \rfloor) \)-free r-graph with \( e = \lfloor (t/2+1)^2 \rfloor \). Then it is also a t-IPPS\((r, |H|, n)\).
Proof of Proposition II.1. To prove the lower bound of \( m \), it suffices to apply Lemma II.2 and Theorem I.2 with \( v = er - r \) and \( e = \lceil (t/2 + 1)^2 \rceil \).

Since the proof of Lemma II.2 is very much in the spirit of Theorem 2.8 in [17], we omit it here. For the proof, see [1].

B. Uniform Combinatorial Batch Codes

An \( r \)-uniform CBC with parameters \( m, e, n \), denoted as \( r-(m, e, n)\)-CBC, is an \( r \)-uniform multihypergraph (i.e., hypergraphs allowing repeated edges) \( \mathcal{H} \) with \( n \) vertices and \( m \) edges, such that for every \( 1 \leq i \leq e \), the union of any \( i \) edges contains at least \( i \) vertices. For integers \( e > r \geq 2 \), let \( m(n, r, e) \) denote the maximum \( m \) such that an \( r-(m, e, n)\)-CBC exists.

Batch codes were introduced by Ishai et al. [18] as a solution to several practical problems in Computer Science. As a special family of batch codes, uniform-CBCs were introduced by Paterson and Stinson [19] to model the following scenario in a distributed database system. Loosely speaking, there are \( m \) data items which are stored in \( n \) servers so that any data item is replicated across \( r \) servers such that any \( e \) of the \( m \) data items can be retrieved by accessing \( e \) servers and reading exactly one data item from each. Given \( r, e \) and \( n \), Balachandran and Bhattacharya [20] formulated the problem of determining the maximum possible \( m \) as the aforementioned Turán-type problem of multihypergraphs, as explained below.

Definition II.3 (Hypergraph-based replication system). Let \( \mathcal{H} \subseteq \binom{[n]}{r} \) be an \( r \)-uniform multihypergraph whose \( n \) vertices and \( m \) edges are indexed by the servers and the data items, respectively. An \( \mathcal{H} \)-based replication system stores \( m \) data items among \( n \) servers as follows: for \( 1 \leq i \leq m \), the \( i \)-th data item is stored in the \( r \) servers which form the \( i \)-th edge of \( \mathcal{H} \).

Given a hypergraph-based replication system, the required retrieval condition on the servers and the data items can be expressed in a purely combinatorial way: every collection of at most \( e \) distinct edges of \( \mathcal{H} \) has a system of distinct representatives (SDR for short) from the \( n \) vertices, where for any \( e \) edges \( A = \{A_1, \ldots, A_e\} \subseteq \mathcal{H} \), an SDR of \( A \) is a set of \( e \) distinct elements \( \{x_1, \ldots, x_e\} \subseteq [n] \) such that \( x_i \in A_i \) for each \( 1 \leq i \leq e \). Applying Hall’s theorem [21] one can infer that this holds if and only if \( \mathcal{H} \) is \( \mathcal{G}_e(i-1, i) \)-free for every \( 1 \leq i \leq e \). We refer the uninformed reader to Section 5 of [22] for background on SDRs and Hall’s theorem.

Bujtás and Tuza [5] studied the above extremal problem in the following more general setting. Recall that an \( r \)-graph is said to be \( \mathcal{G}_e(q, e) \)-free if it is simultaneously \( \mathcal{G}_e(i-q-1, i) \)-free for every \( 1 \leq i \leq e \). Clearly, an \( r-(m, e, n)\)-CBC is equivalent to an \( \mathcal{G}_e(0, e) \)-free \( r \)-uniform multihypergraph with \( n \) vertices and \( m \) edges; consequently,

\[
m(n, r, e) \geq \text{ex}_e(n, \mathcal{G}_e(0, e)).
\]

For fixed integers \( r \geq 2, e \geq r + 1 \), it was shown in [19] that

\[
m(n, r, e) = \Omega(n^{r-1+\frac{e}{r(r+1)}}).
\]

Improved lower bounds for \( r = 2 \) or \( e = 6, e - \lceil \log e \rceil \leq r = e - 1 \) were obtained in [20]. The following result, which is an easy consequence of (2) and Proposition I.3, slightly improves the known lower bounds for some of the remaining pairs of \((r, e)\).

Proposition II.4. For fixed integers \( r \geq 3, e \geq r + 1 \) satisfying \( \gcd(e, r - 1) = 1 \) and \( n \to \infty \), \( m(n, r, e) \geq \text{ex}_e(n, \mathcal{G}_e(0, e)) = \Omega(n^{r-1+\frac{e}{r(r+1)}}) \).

It remains to prove Proposition I.3.

Proof of Proposition I.3. Apply Theorem I.2 with \( v = e - q - 1 \). Since for every \( 1 \leq i \leq e \),

\[
ir - \left(\frac{(i-1)(er - e + q + 1)}{e - 1}\right) = ir - (i - 1)(r - 1) - \frac{(i - 1)(r + q)}{e - 1} \geq i + r - 1 - (r + q) = i - q - 1.
\]

Therefore there exists an \( r \)-graph with the desired number of edges, which is \( \mathcal{G}_e(i - q - 1, i) \)-free for every \( 1 \leq i \leq e \).

C. Optimal Locally Recoverable Codes

An \( [n, k, d] \) linear code \( C \) over \( \mathbb{F}_q \) is a subspace of dimension \( k \) of \( \mathbb{F}_q^n \), such that \( d = \min\{\text{wt}(\mathbf{x}) : \mathbf{x} \in C \setminus \{\mathbf{0}\}\} \) and \( \text{wt}(\mathbf{x}) \) is the number of nonzero coordinates of \( \mathbf{x} \). A linear code is naturally associated with two matrices, the generator and parity check matrices. A generator matrix of \( C \) is a \( k \times n \) matrix \( G \) whose rows form a basis of the code. A parity check matrix of \( C \) is an \((n - k) \times n\) matrix \( H \) such that \( \mathbf{x} \in C \iff H \cdot \mathbf{x}^T = \mathbf{0} \). The following observation is well-known and it follows easily from the above discussion.

Observation II.5. Let \( C \subseteq \mathbb{F}_q^n \) be an \([n, k, d]\)-linear code, then

(i) the generator matrix \( G \) defines an encoding mapping: \( \mathbb{F}_q^k \to \mathbb{F}_q^n \), such that \( y = y \cdot G \);

(ii) any \( d - 1 \) distinct columns of the parity check matrix \( H \) are linearly independent, hence any codeword \( \mathbf{x} \in C \) can be uniquely reconstructed from any \( n - d + 1 \) of its coordinates.

A linear code \( C \subseteq \mathbb{F}_q^n \) of dimension \( k \) is called Locally Recoverable Code (LRC) with locality \( r \) if for any \( i \in [n] \) there exists \( r \) other coordinates \( i_1, \ldots, i_r \) such that for any codeword \( \mathbf{x} = (x_1, \ldots, x_n) \in C \), \( x_i \) can be recovered from \( x_{i_1}, \ldots, x_{i_r} \). We denote such a code by \((n, k, r)\)-LRC.

In [23] it was shown that the minimum distance \( d \) of an \((n, k, r)\)-LRC satisfies \( d \leq n - k - \lceil \frac{k}{r} \rceil + 2 \), and the code is called optimal if the bound is achieved with equality. The reader is referred to [24] and the references therein for constructions of optimal linear LRCs.

LRCs have found applications in modern distributed storage systems, as we explain next. In a distributed storage system a

---

2In most cases, the generator and parity check matrices of a given linear code are not unique.
data of $k$ symbols of $\mathbb{F}_q$ is encoded to a codeword $(x_1, \ldots, x_n)$ of an $[n, k, d]$-linear code $C \subseteq \mathbb{F}_q^n$ (see, e.g. Observation II.5 (i)), and is stored on $n$ storage nodes, where the $i$th node stores the symbol $x_i \in \mathbb{F}_q$. Such encoding scheme provides resiliency to any $d - 1$ concurrent node failures by accessing the remaining $n - (d - 1)$ nodes (see e.g. Observation II.5 (ii)). However, by far the most common scenario is the failure of a single storage node. The efficiency of the repair in such a scenario is measured by the repair locality, i.e., the number of nodes that participate in the repair process. Clearly, by definition an $(n, k, r)$-LRC can repair any single failed node by accessing $r$ other nodes. This property makes LRCs attractive for modern distributed storage systems (e.g. Windows Azure [25] and Facebook [26]).

In order to reduce the complexity of the operations in the finite field, it is desired to define the LRCs over small enough fields. In other words, given the size of the underlying field, our goal is to construct the longest possible optimal-LRC. The following theorem was proved in [7]. Note that for $r \geq d - 2$ and $(r + 1)! n$, an $(n, k, r)$-LRC is optimal if and only if $n - k - \frac{n}{r+1} - 2$ (see Corollary 2.3 [7]).

**Lemma II.6** (see Theorem 3.1, [7]). For $d \geq 11, r \geq d - 2$ and $(r + 1)! n$, if there exists a family $A \subseteq \binom{\mathbb{F}_q}{\frac{r+1}{}}$ which is $\mathcal{G}_{r+1}(i\epsilon, i\epsilon)$-free for each $1 \leq i \leq \lfloor \frac{d-1}{r+1} \rfloor$, then one can construct an optimal $(n, k, r)$-LRC with $n = (r + 1)|A|$.

The following result (also stated in [7], but without a proof) follows by combining Lemma II.6 and Proposition I.4.

**Proposition II.7** (see also Theorem 1.1, [7]). Suppose that $d \geq 11, r \geq d - 2$ and $(r + 1)! n$, then there exists an optimal $(n, k, r)$-LRC with minimum distance $d$ and length $n = \Omega(q(q \log q)^{(r+1)/r}}$.

**III. PROOF OF THEOREM I.2**

We view the parameters $v, e$, and $r$ as constants, whereas $n$ tends to infinity. Since we are only interested in the asymptotic behavior we do not make an attempt to optimize the constants.

To prove Theorem I.2 we need the following result of Duke et al. [27] on the lower bound of hypergraph independence number. Note that an $r$-graph is said to be linear if any two distinct edges share at most one vertex.

**Lemma III.1** (see Theorem 2, [27]). For all fixed $r \geq 3$ there exists a constant $c > 0$ depending on $r$ such that every linear $r$-graph on $n$ vertices with average degree at most $d$ has an independent set of size at least $cn(\log \frac{d}{n})^{r/4}$. 

**Proof of Theorem I.2.** Set $p := p(n) = \Theta(n^{\frac{r-1}{r}})$ for some $\epsilon > 0$, which will be made explicit later. Generate an $r$-graph $\mathcal{H}_0 \subseteq \binom{[n]}{r}$ by picking each element of $\binom{[n]}{r}$ with probability $p$ uniformly and independently at random. Let $X$ denote the number of edges in $\mathcal{H}_0$. Clearly,

$$E[X] = p \binom{n}{r} = \Theta(n^{\frac{r-1}{r}}).$$

For $2 \leq i \leq e - 1$, let $\mathcal{B}_i$ be the collection of all $i$ distinct edges of $\mathcal{H}_0$ whose union contains at most $i r - f(i)$ vertices, where $f(i)$ will be determined later. Let $Y_i$ denote the size of $\mathcal{B}_i$.

$$E[Y_i] = O(p^n n^{i r - f(i)}) = O(n^{\frac{(r-1)(e-i)}{e-i} + f(i)}),$$

where the first estimate follows from the fact that there are at most $O(n^2 r - f(i))$ ways to choose at most $i r - f(i)$ vertices, and that for every fixed choice the probability of having exactly $i$ edges is $O(p^2)$. We say that $e$ distinct edges of $\mathcal{H}_0$ form a bad $e$-system if their union contains at most $v$ vertices. Clearly, two distinct bad $e$-systems can share at most $e - 1$ edges. For each $2 \leq i \leq e - 1$, let $\mathcal{Z}_i$ be the collection of the unordered pairs of bad $e$-systems which share precisely $i$ edges, such that the union of those $i$ edges contains at least $i r - f(i) + 1$ vertices. For $\mathcal{Z}, \mathcal{Z}' \in \mathcal{Z}_i$, it is easy to verify that $|E(\mathcal{Z} \cup E(\mathcal{Z}'))| = 2e - i$ and $|V(\mathcal{Z} \cup V(\mathcal{Z}'))| \leq 2v - (i r - f(i) + 1)$. Let $Z_i$ denote the size of $\mathcal{Z}_i$. Then

$$E[Z_i] = O(p^{2e-i} n^{2e-(i r - f(i) + 1)}) = O(n^{f(i) + O(e-i - 1)}),$$

where the first estimate follows from the fact that there are at most $O(n^{2} r - f(i) + 1)$ ways to choose $V(\mathcal{Z}) \cup V(\mathcal{Z}'')$, and that for every fixed choice the probability of having exactly $2e - i$ edges is $O(p^{2e-i})$. Finally, let $W$ denote the number of bad $e$-systems in $\mathcal{H}_0$. By a similar reasoning as before,

$$E[W] = O(p^n v) = O(n^{\frac{r-1}{r} + \epsilon}).$$

Next, we will bound from above the number of pairs of bad $e$-systems which share at least two edges, by picking $\epsilon$ and $f(i)$’s for which $E[Y_i] = o(E[X])$ and $E[Z_i] = o(E[X])$ as $n \to \infty$ and $2 \leq i \leq e - 1$. From (3), (4) and (5), it is not hard to see that $E[Y_i] = o(E[X])$ holds if and only if

$$f(i) > \frac{(i - 1)(e r - v)}{e - 1},$$

and $E[Z_i] = o(E[X])$ holds if and only if

$$f(i) < \frac{(i - 1)(e r - v)}{e - 1} - \epsilon(2e - i - 1) + 1.$$

Let

$$a = \min_{2 \leq i \leq e - 1} \left\{ \frac{1}{i - 1} \left( f(i) - \frac{(i - 1)(e r - v)}{e - 1} \right), \frac{1}{2e - i - 1} \left( \frac{(i - 1)(e r - v)}{e - 1} + 1 - f(i) \right) \right\}.$$ 

Observe that there exist $\epsilon \in (0, a)$ and $f(i) \in \mathbb{N}$ which satisfy (7) and (8) if and only if for each $2 \leq i \leq e - 1$, there exists an $f(i) \in \mathbb{N}$ such that

$$\frac{(i - 1)(e r - v)}{e - 1} - f(i) < \frac{(i - 1)(e r - v)}{e - 1} - \epsilon(2e - i - 1) + 1.$$ 

By the integrality of $f(i)$, (10) holds if and only if $e - 1 \nmid (i - 1)(e r - v)$ for each $2 \leq i \leq e - 1$. It is easy to verify that those $e - 2$ indistinguishability conditions hold simultaneously if and only if $\gcd(e - 1, e r - v) = 1$. Under this condition, it suffices to pick for each $2 \leq i \leq e - 1$, $f(i) = \lfloor \frac{(i - 1)(e r - v)}{e - 1} \rfloor$ and an
arbitrary $\epsilon \in (0, a)$ (note that the choices of $f(i)$’s guarantee that there must exist some constant $a > 0$ satisfying (9)).

We form a subgraph $H_1$ of $H_0$ as follows. For every $2 \leq i \leq e - 1$, delete one edge from each member of $\mathcal{F}_i$, and one edge from $E(Z) \cup E(Z')$ for each pair $\{Z, Z'\} \in \mathcal{Z}_{i}$. By linearity of expectation, $H_1$ has the following properties:

(i) $E[|E(H_1)|] = \Theta(n^{\frac{e - 1}{e + 1} + \epsilon})$;
(ii) the expected number of bad $e$-systems contained in $H_1$ is at most $O(p^r n^v) = O(n^{\frac{e - 2}{e + 1} + \epsilon r})$;
(iii) for each $2 \leq i \leq e - 1$, the union of any $i$ distinct edges in $H_1$ contains more than $\epsilon r - f(i)$ vertices;
(iv) any two bad $e$-systems in $H_1$ can share at most one edge.

Indeed, (i) is a consequence of the following calculation:

$$E[|E(H_1)|] \geq |E(H_0)| - \sum_{i=2}^{e-1} |\mathcal{F}_i| - \sum_{i=2}^{e-1} |\mathcal{Z}_i|$$

$$= E[X] - \sum_{i=2}^{e-1} E[Y_i] - \sum_{i=2}^{e-1} E[Z_i]$$

$$= E[X] - o(E[X]) = \Theta(E[X]);$$

(ii) follows from (6) and the observation that deleting edges from $H_0$ does not increase the number of bad $e$-systems; (iii) holds since $H_1$ does not contain any member of $\mathcal{F}_i$ for any $2 \leq i \leq e - 1$. It remains to verify (iv). Assume to the contrary that $H_1$ still contains two bad $e$-systems sharing $i$ edges for some $2 \leq i \leq e - 1$. On one hand, if those $i$ edges are spanned by at least $\epsilon r - f(i) + 1$ vertices, then the pair of those two bad $e$-systems must belong to $\mathcal{Z}_{i}$, which is a contradiction; on the other hand, if those $i$ edges are spanned by at most $\epsilon r - f(i)$ vertices, then they must form a member of $\mathcal{F}_i$, which is again a contradiction.

Next we construct an auxiliary $e$-graph $U \subseteq \left( E(H_1) \right)$:

- the vertex set of $U$ is formed by the edge set of $H_1$;
- $e$ vertices of $U$ form an edge if and only if the corresponding $e$ $r$-edges in $H_1$ form a bad $e$-system.

It is routine to check that the following hold:

- $U$ is linear (by (iv));
- $U$ has at least $\Theta(n^{\frac{e - 1}{e + 1} + \epsilon})$ vertices (by (i)) and at most $O(p^r n^v) = O(n^{\frac{e - 2}{e + 1} + \epsilon r})$ edges (by (ii));
- $d(U)$ the average degree of $U$ is at most $d(U) = \Theta(\frac{e \cdot |E(U)|}{|V(U)|}) = O(n^{(1-e)})$.

It follows from Lemma III.1 that $U$ has an independent set of size at least

$$\Omega\left(|V(U)|\left(\frac{\log d(U)}{d(U)}\right)^{\frac{1}{e - 1 + \epsilon}}\right) = \Omega\left(n^{\frac{e - 1}{e + 1}}(\log n)^{\frac{1}{e - 1 + \epsilon}}\right). \quad (11)$$

To conclude, by Markov’s inequality it is easy to see that with positive probability there exists an $H_1 \subseteq \left( E(H_1) \right)$ which satisfies (i)-(iv), implying that there also exists a $U \subseteq \left( E(H_1) \right)$ with independence number lower bounded by (11). Now the theorem follows from the following simple observation: every independent set $I \subseteq V(U)$ corresponds to a $\mathcal{G}_r(v, e)$-free subhypergraph $H_I \subseteq H_1$ with $|I|$ edges; moreover, by (iii) $H_I$ is also $\mathcal{G}_r(\epsilon r - f(i), i)$-free for each $2 \leq i \leq e - 1$. □

### References

1. C. Shangguan and I. Tamo, “Universally Sparse Hypergraphs with Applications to Coding Theory,” arXiv preprint arXiv:1902.05903, 2019.
2. P. Turán, “Eine Extremalaufgabe aus der Graphentheorie,” Mat. Fiz. Lapok, vol. 48, pp. 436–452, 1941.
3. W. G. Brown, P. Erdős, and V. T. Sós, “Some extremal problems on $r$-graphs,” in New directions in the theory of graphs. Academic Press, New York, 1973, pp. 53–63.
4. Z. Füredi and M. Ruszinkó, “Uniform hypergraphs containing no grids,” Adv. Math., vol. 240, pp. 302–324, 2013.
5. C. Bujtás and Z. Tuza, “Turán numbers and batch codes,” Discrete Appl. Math., vol. 186, pp. 45–55, 2015.
6. J. Verstraëte, “Extremal problems for cycles in graphs,” in Recent trends in combinatorics, ser. IMA Vol. Math. Appl. Springer, [Cham], 2016, vol. 159, pp. 83–116.
7. C. Xing and C. Yuan, “Construction of optimal locally recoverable codes and connection with hypergraph,” arXiv:1811.09142, 2018.
8. T. Bohman and P. Keevash, “The early evolution of the $H$-free process,” Invent. Math., vol. 181, no. 2, pp. 291–336, 2010.
9. P. Bennett and T. Bohman, “A note on the random greedy independent set algorithm,” Random Structures Algorithms, vol. 49, no. 3, pp. 479–502, 2016.

10. Y. Gu and Y. Miao, “Bounds on traceability schemes,” IEEE Trans. Inform. Theory, vol. 64, no. 5, pp. 3450–3460, 2018.
11. B. Chor, A. Fiat, M. Naor, and B. Pinkas, “Tracing traitors,” IEEE Trans. Inform. Theory, vol. 46, no. 3, pp. 893–910, 2000.
12. D. R. Stinson and R. Wei, “Combinatorial properties and constructions of traceability schemes and framework codes,” SIAM J. Discrete Math., vol. 11, no. 1, pp. 41–53, 1998.
13. M. J. Collins, “Upper bounds for parent-identifying set systems,” Des. Codes Cryptogr., vol. 51, no. 2, pp. 167–173, 2009.
14. A. Shamir, “How to share a secret,” Comm. ACM, vol. 22, no. 11, pp. 612–613, 1979.
15. G. R. Blakley, “Safeguarding cryptographic keys,” in Proceedings of the national computer conference, vol. 48, 1979, pp. 313–317.
16. Y. Gu, “Private Communication,” 2018.
17. J. Staddon, D. Stinson, and R. Wei, “Combinatorial properties of framework and traceability codes,” IEEE Trans. Inform. Theory, vol. 47, no. 3, pp. 1042–1049, 2001.
18. Y. Ishai et al., “Batch codes and their applications,” in Proceedings of the 36th Annual ACM Symposium on Theory of Computing. ACM, New York, 2004, pp. 262–271.
19. M. B. Paterson, D. R. Stinson, and R. Wei, “Combinatorial batch codes,” Adv. Math. Commun., vol. 3, no. 1, pp. 13–27, 2009.
20. N. Balachandran and S. Bhattacharya, “On an extremal hypergraph problem related to combinatorial batch codes,” Discrete Appl. Math., vol. 162, pp. 373–380, 2014.
21. P. Hall, “On representatives of subsets,” J. Lond. Math. Soc., vol. 1, no. 1, pp. 26–30, 1935.
22. S. Jukna, Extremal combinatorics, 2nd ed., ser. Texts in Theoretical Computer Science. An EATCS Series. Springer, Heidelberg, 2011, with applications in computer science.
23. P. Gopalan, C. Huang, H. Simonciti, and S. Yekhanin, “On the locality of codeword symbols,” IEEE Trans. Inform. Theory, vol. 58, no. 11, pp. 6925–6934, 2012.
24. I. Tamo and A. Barg, “A family of optimal locally recoverable codes,” IEEE Trans. Inform. Theory, vol. 60, no. 8, pp. 4661–4676, 2014.
25. C. Huang et al., “Erasure coding in windows azure storage,” in Presented as part of the 2012 USENIX Annual Technical Conference (USENIX ATC 12), Boston, MA, 2012, pp. 15–26.
26. M. Sathiamoorthy et al., “Xoring elephants: Novel erasure codes for big data,” in Proceedings of the VLDB Endowment, vol. 6, no. 5, VLDB Endowment, 2013, pp. 325–336.
27. R. A. Duke, H. Lefmann, and V. Rödl, “On uncrowded hypergraphs,” Random Structures Algorithms, vol. 6, no. 2-3, pp. 209–212, 1995.