Optimal Monotone Encodings

Noga Alon * Rani Hod †

Abstract

Moran, Naor and Segev have asked what is the minimal $r = r(n,k)$ for which there exists an (n,k)-*monotone encoding of length* r, i.e., a monotone injective function from subsets of size up to k of $\{1,2,\ldots,n\}$ to r bits. Monotone encodings are relevant to the study of tamper-proof data structures and arise also in the design of broadcast schemes in certain communication networks. To answer this question, we develop a relaxation of k-superimposed families, which we call α-fraction k-multi-user tracing ((k,α)-FUT families). We show that $r(n,k) = \Theta(k\log(n/k))$ by proving tight asymptotic lower and upper bounds on the size of (k,α)-FUT families and by constructing an (n,k)-monotone encoding of length $O(k\log(n/k))$. We also present an explicit construction of an $(n,2)$-monotone encoding of length $2\log n + O(1)$, which is optimal up to an additive constant.

1 Introduction

In their pursuit of history-independent schemes that use a write-once memory, motivated by cryptographic applications, Moran et al. [14] have considered monotone injective functions that map subsets of size up to k of $[n]$ into $2^{[r]}$ (all subsets of $[r]$), henceforth called (n,k)-monotone encodings of length r, or $\mathrm{ME}(n,k,r)$. They have shown the existence of an (n,k)-monotone encoding of length $O(k\log n\log(n/k))$ and raised the question of determining the minimal $r = r(n,k)$ for which an $\mathrm{ME}(n,k,r)$ exists.

A quick counting argument shows that $r(n,k) \geq \log\left(\sum_{i=0}^{k}\binom{n}{i}\right) = \Omega\left(k\log\frac{n}{k}\right)$ is required for any injective encoding, without even considering monotonicity. In this paper, we show that a monotone encoding of length $O\left(k\log\frac{n}{k}\right)$ exists, establishing that $r(n,k) = \Theta\left(k\log\frac{n}{k}\right)$, thus settling the open problem raised in [14]. We limit ourselves to $k \leq \frac{n}{2}$ since the trivial identity encoding is optimal for $k > \frac{n}{2}$.

Throughout the paper we use $[n]$ to denote $\{1,2,\ldots,n\}$. We denote subsets of $[n]$ of size k and up to k by $\binom{[n]}{k}$ and $\binom{[n]}{\leq k}$, respectively. All logarithms are binary unless stated otherwise. Floor and ceiling signs are omitted whenever these are not crucial.

*Schools of Mathematics and Computer Science, Raymond and Beverly Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv, Israel. Research supported in part by the Israel Science Foundation, and by a USA-Israeli BSF grant. Email: nogaa@post.tau.ac.il.

†School of Computer Science, Raymond and Beverly Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv, Israel. Email: ranihod@post.tau.ac.il.

1.1 A First Attempt: Superimposed Families

A general representation of a monotone function f is $f(S) = \bigcup_{S' \subseteq S} g(S')$ for some function $g(S)$. For f to be injective as well, we need all the relevant unions to be distinct.

A family of subsets of $[r]$ is called k-superimposed if all the unions of up to k sets from it are distinct. Clearly, a k-superimposed family $\mathcal{F} = \{A_i\}_{i=1}^n$ of cardinality n translates to an $\mathrm{ME}(n, k, r)$ f defined by $f(S) = \bigcup_{i \in S} A_i$.

Probabilistic and explicit constructions of k-superimposed families of cardinality n are known for $r = O\left(k^2 \log \frac{n}{k}\right)$ (see, for example, [7, 8, 1, 16]), yielding the same upper bound on the length of (n, k)-monotone encodings. However, [7, 17, 9] showed that for $n > k^2$, k-superimposed families require $r = \Omega\left(\frac{k^2}{\log k} \log n\right)$; Thus, an approach based solely on k-superimposed families will not achieve optimal monotone encodings.

Inspecting the monotone encoding induced by a k-superimposed family, we observe that only the "linear" terms $g(\{i\}) = A_i$ are non-empty. In a way, using "higher-order" terms can be regarded as a form of adaptive encoding (obtained in a non-adaptive fashion) since collisions in the unions of lower-order terms can be resolved by a higher-order term.

1.2 Our Contribution

A general monotone encoding does not need the strict distinct-unions requirement of superimposed families. We consider the following relaxation of superimposed families.

Definition 1. *Let $\mathcal{F} = \{A_i\}_{i=1}^n$ be a family of subsets of $[r]$ and let $S \subseteq [n]$. We denote $\bigcup_{i \in S} A_i$ by A_S. An element $j \in S$ is said to be $\mathcal{F}$-identifiable (with respect to S) if A_j has a unique element not present in any other subset $A_i \in \mathcal{F}$ that is covered by A_S, that is, if $A_j \not\subseteq \bigcup \{A_i \in \mathcal{F} : i \neq j, A_i \subseteq A_S\}$. An element $j \in S$ is said to be $\mathcal{F}$-obscured (with respect to S) if it is not $\mathcal{F}$-identifiable.*

Definition 2. *Let $k \geq 2$, $n \geq 2k$ and $0 < \alpha < 1$. A family $\mathcal{F} = \{A_i\}_{i=1}^n$ of subsets of $[r]$ is called α-fraction k-multi-user tracing, or (k, α)-FUT, if for any $S \in \binom{[n]}{\leq k}$, more than $\alpha|S|$ of its elements are $\mathcal{F}$-identifiable.*[1]

We prove almost tight upper and lower bounds on $(k, 1-\epsilon)$-FUT families.

Theorem 1. *There exists a constant $c_1 > 0$ such that for all $k \geq 2$, $n \geq 2k$ and $\frac{1}{k} \leq \epsilon \leq \frac{1}{2}$, there exists a $(k, 1-\epsilon)$-FUT of cardinality n where $r = c_1 \frac{k}{\epsilon} \log \frac{n}{k}$.*

Theorem 2. *There exists a constant $c_2 > 0$ such that for all $k \geq 2$, $\frac{1}{k} \leq \epsilon \leq \frac{1}{2}$, any $(k, 1-\epsilon)$-FUT of cardinality $n \geq \frac{k}{\epsilon}$ must have $r \geq c_2 \frac{k/\epsilon}{\log k/\epsilon} \log n$.*

Note that for $0 < \epsilon \leq \frac{1}{k}$, any $(k, 1-\epsilon)$-FUT family is actually k-superimposed since the number of obscured elements is strictly less than one. Substituting $\epsilon = \frac{1}{k}$ in Theorems 1 and 2 yields the known asymptotic upper and lower bounds for k-superimposed families.

[1] Or all of them, if $|S| \leq \frac{1}{\alpha}$.

Back to monotone encodings, we form an optimal $\mathrm{ME}\left(n, k, O\left(k \log \frac{n}{k}\right)\right)$ by chaining $\left(\frac{k}{2^t}, \frac{1}{2}\right)$-FUT families of cardinality n for $t = 0, 1, \ldots, \log k$. This yields the following theorem.

Theorem 3. *There exists a constant $c_3 > 0$ such that for all integers $n \geq 4$ and $2 \leq k \leq \frac{n}{2}$, there exists an (n, k)-monotone encoding of length $r = c_3 k \log \frac{n}{k}$.*

Definition 3. *For integers $0 \leq k \leq n$ we denote $\left\lceil \log \left(\sum_{i=0}^{k} \binom{n}{i} \right) \right\rceil$ by $\rho(n, k)$.*

We also present a lower bound on the length of monotone encodings.

Theorem 4. *There exists a constant $c_4 > 0$ such that $r(n, k) > (1 + c_4)\rho(n, k)$ for sufficiently large n and some $k = k(n)$.*

When k is small, constant factors may have a significant impact. In section 4 we present an explicit construction for $k = 2$ which is optimal up to an additive constant, yielding the following theorem.

Theorem 5. *There exists a constant $c_5 > 0$ such that for all integers $n \geq 4$, there exists an explicit $(n, 2)$-monotone encoding of length $\rho(n, 2) + c_5$.*

1.3 Related Work

1.3.1 Single-user and Multi-user Tracing Families

Although we described (k, α)-FUT families as a relaxation of superimposed families, they can also be seen as an extension of single-user tracing (SUT) families, an even simpler relaxation of superimposed families introduced by Csűrös and Ruszinkó [6]. Given the union of up to k subsets of a SUT family, we are able to identify at least one of them. While the lower bound remains $\Omega\left(k \log \frac{n}{k}\right)$, SUT families of cardinality n were shown by Alon and Asodi [2] to exist for $r = O\left(k \log \frac{n}{k}\right)$.

Laczay and Ruszinkó [12] extended SUT families in another direction, considering j-out-of-k multi-user tracing (MUT_j) families, ensuring that given the union of up to k subsets we are able to identify at least j of them.[2] By definition, a MUT_1 family is equivalent to a SUT family; Alon and Asodi [3] proved that MUT_j families exist for $r = O\left((k + j^2) \log \frac{n}{k}\right)$, effectively creating $\mathrm{MUT}_{\sqrt{k}}$ families for the same cost as SUT families. Nevertheless, MUT_j families are also j-superimposed, hence we cannot use them for $j = \omega(\sqrt{k \log k})$ while maintaining linear dependence in k.

1.3.2 Non-adaptive Conflict Resolution

Komlós and Greenberg [11] solved a problem similar to monotone encoding using techniques very similar to ours. They considered non-adaptive conflict resolution (NACR) in a multiple access OR channel. Here is a quick description of NACR.

> A communication channel is shared by n stations, some of which may want to broadcast a message. Multiple concurrent broadcasts cancel out and the stations involved are notified of this.

[2] Or all of them, if the given union is a union of less than j subsets.

Each station has a *scheme*, or a list of available time slots. Each active station will try to broadcast its message on every available time slot; it will deactivate if it succeeds (i.e., if it was the only station broadcasting on this time slot).

An *epoch*, or a sequence of time-slots, is called successful if no active stations remain (due to successful broadcasting) until the epoch ends.

A scheme set is valid if for all choices of initial k active stations, the epoch succeeds. What is the minimal length (in time slots) needed for a valid scheme set to exist?

Although the problem of monotone encodings can be reformulated in a similar language, two major differences exist between ME and NACR.

1. In NACR, stations are aware of their success/failure, i.e., they know whether there were 0, 1 or ≥ 2 concurrent broadcasts. In ME, an outside observer is required to identify active stations seeing only the channel activity indicator (0 or ≥ 1 broadcasts).

2. In NACR, an active station will stop once it has successfully broadcast its message. In ME, the situation is analogous to stations that remain active and cannot change their schemes. However, stations in ME are aware of each other, and are allowed to broadcast *more* if other stations are active.

For instance, the following valid NACR scheme set for $n = 3, k = 2$ uses three time slots: $S_1 = \{2, 3\}, S_2 = \{1, 3\}, S_3 = \{1, 2\}$. Nevertheless, the activity indicator of the channel gives no hint of *which* stations are active when any two of them are active!

Assume that the message each station broadcasts specifies its identifying number and consider the actual channel data rather than the channel activity indicator. This allows a successful broadcast to identify[3] the transmitting station. Thus, we may convert[4] an NACR solution to an ME at the cost of a factor of $\log n$. The $\mathrm{ME}(n, k, O(k \log n \log(n/k)))$ presented in [14] proceeds essentially along these lines.

1.3.3 Cryptographic Applications

In [14], monotone encodings are used to maintain a tamper-proof deterministic data structure that represents a subset of size up to k of $[n]$.

Instead of relying on cryptographic assumptions, the data structure is made tamper-proof by storing it on a write-only memory, i.e., all bits are initially 0 and can only be turned to 1. This imposes the monotonicity requirement.

Since elements are inserted one by one, another security-motivated requirement is that the representation of the data structure is independent of the order in which elements are inserted (for example, to ensure privacy in voting schemes). This rules out "adaptive" solutions like writing down the elements sequentially using $\log n$ bits per element. This requirement is expressed in ME by taking $\binom{[n]}{\leq k}$ (that is, *unordered* subsets) as the domain of the encoding.

[3] Some action is needed to ensure that multiple concurrent broadcasts are not misinterpreted as valid messages. For instance, encode 0 as '01' and 1 as '10', doubling the length of the data.

[4] Further modifications are necessary to work out the second difference as well, but the length of the data remains unaffected.

2 The Construction

2.1 FUT Families

The upper bound stated in Theorem 1 is implied by the following probabilistic construction.

Let $k \geq 2$, $\frac{1}{k} \leq \epsilon \leq \frac{1}{2}$. Let $d = \frac{2}{\epsilon} \log \frac{2en}{k}$ and $r = 16kd = O\left(\frac{k}{\epsilon} \log \frac{n}{k}\right)$. Let $h_1, \ldots, h_n$ be n random functions from $[d]$ to $[16k]$, i.e., the values $h_j(i)$ for $i \in [d]$ and $j \in [n]$ are chosen independently and equiprobably from $[16k]$ and let $\mathcal{F} = \{A_1, \ldots, A_n\} \subset 2^{[r]}$ be their representations as sets, that is, $A_j = \{16ki - h_j(i) + 1 : i \in [d]\} \subset [r]$.

Definition 4. *A family $\mathcal{F}$ of sets is said to have property A if for $t \in [2k]$ and for all distinct $A_1, \ldots, A_t \in \mathcal{F}$, A_j is covered by the union of $\{A_i : i \in [t], i \neq j\}$ for less than ϵt values of $j \in [t]$. In other words, more than $(1-\epsilon)t$ of the sets have a unique element in $\bigcup_{i=1}^{t} A_i$.*

Proposition 2.1. *With positive probability, property A holds for $\mathcal{F}$ as selected above.*

Proof. We use the union bound over all choices of such $A_1, \ldots, A_t \in \mathcal{F}$ to bound the probability that property A does not hold. Fix $t \in [2k]$ and let $\mu = \lfloor \log(2k/t) \rfloor \geq 0$. Then, $2^{-\mu}k < t \leq 2^{1-\mu}k$. Fix distinct $A_1, \ldots, A_t \in \mathcal{F}$ and assume that at least $m = \epsilon t$ of them are covered by the union of the others. Without loss of generality, assume that these are $A_1, \ldots, A_m$ (and maybe others).

Consider the random functions $\{h_j\}_{j=1}^{t}$ represented by $\{A_j\}_{j=1}^{t}$. Fix a coordinate $i \in [d]$. Assume that $h_j(i)$ are already determined for $j > m$ and now select $h_j(i)$ sequentially for $j = 1, \ldots, m$. At least $\frac{m}{2}$ of them collide with some previously determined $h_{j'}(i), j' < j$ as each of these must be covered by the union of all others. Since $\bigcup_{j=1}^{t} A_j$ covers at most $t \leq 2^{1-\mu}k$ elements of $[16k]$, the probability of this event is at most

$$\binom{m}{m/2} \left(\frac{t}{16k}\right)^{m/2} < 2^m \left(\frac{1}{4}\right)^{m/2} \left(\frac{t}{4k}\right)^{m/2} \leq \left(\frac{1}{2}\right)^{(1+\mu)m/2}$$

and the probability of it happening simultaneously at all d coordinates is at most

$$\left(\frac{1}{2}\right)^{(1+\mu)md/2} = \left(\frac{1}{2}\right)^{(1+\mu)\epsilon td/2} = \left(\frac{k}{2en}\right)^{(1+\mu)t}.$$

The number of choices for $A_1, \ldots, A_t \in \mathcal{F}$ and for the $m = \epsilon t$ covered sets among them is $\binom{n}{t}\binom{t}{m} \leq \left(\frac{en}{t}\right)^t 2^t$. Therefore, the probability that property A does not hold for this value of t is at most

$$\left(\frac{2en}{t}\right)^t \left(\frac{k}{2en}\right)^{(1+\mu)t} = \left(\frac{k}{2en}\right)^{\mu t} \left(\frac{k}{t}\right)^t \leq \left(\frac{1}{4e}\right)^{\mu t} \left(\frac{k}{t}\right)^t \leq \begin{cases} (4e)^{-\mu t} 2^{\mu t} = (2e)^{-\mu t} < \frac{1}{5^t} & ,\mu \geq 1 \\ \left(1 - \frac{t-k}{t}\right)^t < e^{k-t} & ,\mu = 0 \end{cases}.$$

Summing over $t \in [2k]$, the probability that property A does not hold is at most

$$\sum_{t=1}^{k} 5^{-t} + \sum_{t=k+1}^{2k} e^{k-t} < \sum_{t=1}^{\infty} 5^{-t} + \sum_{t'=1}^{\infty} e^{-t'} = \frac{1/5}{1-1/5} + \frac{1/e}{1-1/e} = \frac{1}{4} + \frac{1}{e-1} < 0.9\,.$$

□

Proposition 2.2. *Any family for which property A holds is $(k, 1-\epsilon)$-FUT; Thus, with positive probability, $\mathcal{F}$ as selected above is $(k, 1-\epsilon)$-FUT.*

Proof. Let $S \subset [n], |S| \leq k$ and consider $I = \{i \in [n] : A_i \subseteq A_S\}$. Obviously, $S \subseteq I$. By the definition of I, all $i \in I \setminus S$ are $\mathcal{F}$-obscured since A_i is covered by A_S. Assume that $|I| \geq 2k$. By property A, applied to some subset $S \subset I' \subseteq I$ of cardinality $2k$, more than $(1-\epsilon)2k \geq k \geq |S|$ elements of I' are $\mathcal{F}$-identifiable, which is absurd as they all reside in S.[5] Thus, $|I| < 2k$. By property A, more than $(1-\epsilon)|I| \geq (1-\epsilon)|S|$ elements of I are $\mathcal{F}$-identifiable. Again, they all reside in S. □

2.2 Monotone Encodings

Equipped with the tool we have just developed, we move on to describing a function as stated by Theorem 3.

We construct $f : \binom{[n]}{\leq k} \to 2^{[r]}$ inductively. Initialize the construction[6] with the trivial case $k = 1$. Let $f' : \binom{[n]}{\leq k/2} \to 2^{[r']}$ be a monotone encoding for subsets of size up to $\frac{k}{2}$ and let $\mathcal{F} = \{A_i\}_{i=1}^{n} \subset 2^{[r'']}$ be a $(k, \frac{1}{2})$-FUT family. Shift $\mathcal{F}$ by r' to make its support disjoint from $[r']$. All involved sets are now subsets of the ground set $[r]$ where $r = r' + r''$. We define $f(S) = A_S \cup f'(S')$ where S' consists of all $\mathcal{F}$-obscured elements of S (note that S' is well-defined given $\mathcal{F}$).

Since a $(k, \frac{1}{2})$-FUT family exists for $r = 2c_1 k \log \frac{n}{k}$, the size of the ground set for the entire construction is

$$\sum_{t=0}^{\log k} 2c_1 2^{-t} k \log\left(\frac{n}{2^{-t}k}\right) \leq 2c_1 k \sum_{t=0}^{\infty} 2^{-t}\left(t + \log\frac{n}{k}\right) = 2c_1 k\left(2 + 2\log\frac{n}{k}\right) = O\left(k \log\frac{n}{k}\right).$$

Proposition 2.3. *The function f is injective, i.e., $f(S) \neq f(T)$ for $S \neq T$.*

Proof. It is sufficient to see that given $f(S)$ we can unambiguously determine S. First we use A_S to determine all $\mathcal{F}$-identifiable elements of S. This is possible since if $j \in S$ is $\mathcal{F}$-identifiable, some element of A_S is present only in A_j of all $A_i \subseteq A_S$. Let $S' \subseteq S$ be the set of $\mathcal{F}$-obscured elements of S. By the definition of $\mathcal{F}$, $|S'| \leq \frac{1}{2}|S| \leq \frac{k}{2}$. By the induction hypothesis, we can determine S' from $f'(S')$. Putting both parts together, S is fully determined. □

Claim 2.4. *If $S \subseteq T$ then $S' \subseteq T'$.*

Proof. Let $j \in S'$ be an $\mathcal{F}$-obscured element with respect to S, that is, A_j is covered by the union of $\{A_i \subseteq A_S, i \neq j\}$. Clearly, $A_S \subseteq A_T$. Thus, A_j is also covered by the union of $\{A_i \subseteq A_T, i \neq j\}$. Hence, j is $\mathcal{F}$-obscured with respect to T as well, so $j \in T'$. □

Proposition 2.5. *The function f is monotone, i.e., $f(S) \subseteq f(T)$ for $S \subseteq T$.*

Proof. Fix some $S \subseteq T$. By Claim 2.4, $S' \subseteq T'$. By the induction hypothesis, f' is monotone, hence $f'(S') \subseteq f'(T')$. Now, $f(S) = A_S \cup f'(S') \subseteq A_T \cup f'(T') = f(T)$. □

[5] To be exact, these elements are $\mathcal{F}$-identifiable with respect to I' and not I, but it is all the same since $A_{I'} = A_S = A_I$.

[6] Another way to initialize the construction is with an ME$(n, \sqrt{k}, O(k \log \frac{n}{\sqrt{k}}))$ induced by a $\sqrt{k}$-superimposed family as described in section 1.1.

3 Lower Bounds

3.1 FUT Families

First we show a lower bound of $\Omega(k \log(n/k))$ on the length of *constant* fraction user-tracing families.

Proposition 3.1. *For all $k \geq 2$, any $(k, \frac{1}{2})$-FUT family of cardinality $n \geq k$ must have $r \geq \frac{1}{4} k \log \frac{n}{k}$.*

Proof. Let $k \geq 2$ and let $\mathcal{F} = \{A_i\}_{i=1}^{n} \subseteq 2^{[r]}$ be a $(k, \frac{1}{2})$-FUT family, where $n \geq k$. Without loss of generality, $r \geq \frac{k}{2}$ since otherwise we could cover the entire support of $\mathcal{F}$ by the union of only $\frac{k}{2}$ sets despite $\mathcal{F}$ being a $(k, \frac{1}{2})$-FUT family of cardinality $n \geq k$.

Assume for the sake of contradiction that $n > k 2^{4r/k}$. Consider all unions of $\frac{k}{2}$ sets from $\mathcal{F}$. By the pigeonhole principle, at least

$$\frac{1}{2^r}\binom{n}{k/2} \geq \frac{1}{2^r}\left(\frac{n}{k/2}\right)^{k/2} > \frac{1}{2^r}\left(2^{1+4r/k}\right)^{k/2} = 2^{k/2+r} \geq 2^k > \binom{k}{k/2}$$

of these unions are identical. In other words, there exist distinct $S_1, \ldots, S_m \in \binom{[n]}{k/2}, m > \binom{k}{k/2}$ such that $A_{S_1} = \cdots = A_{S_m}$. Therefore, the set $S = \bigcup_{i=1}^{m} S_i$ is of cardinality at least k. All elements of $S \setminus S_1$ are $\mathcal{F}$-obscured since $A_S = A_{S_1}$. We have reached a contradiction as $\mathcal{F}$ is a $(k, \frac{1}{2})$-FUT family but $|S \setminus S_1| \geq \frac{k}{2}$.[7] □

Next, we establish the lower bound stated in Theorem 2 by using a modified version of a technique from [17]. As the bound of Proposition 3.1 holds for any $(k, 1-\epsilon)$-FUT family, $\epsilon \leq \frac{1}{2}$, we henceforth assume $\epsilon < \frac{1}{4}$. Let $k \geq 2, \frac{1}{k} \leq \epsilon < \frac{1}{4}$ and let $\mathcal{F} = \{A_i\}_{i=1}^{n} \subseteq 2^{[r]}$ be a $(k, 1-\epsilon)$-FUT family, where $n \geq \frac{k}{\epsilon}$. We modify $\mathcal{F}$ in two phases, as follows.

1. As long as $\mathcal{F}$ contains a set F of cardinality at least $\beta = \frac{4r}{k}$, remove F from $\mathcal{F}$ and remove its elements from all other sets of $\mathcal{F}$. Call the resulting family $\mathcal{F}'$.

2. As long as $\mathcal{F}'$ contains a set F such that any 4ϵ-fraction of it is covered by some other set from $\mathcal{F}'$, and in particular, it is covered by the union of $t \leq \frac{1}{4\epsilon}$ other sets $A_1, \ldots, A_t \in \mathcal{F}'$, remove F and $\{A_i\}_{i=1}^{t}$ from $\mathcal{F}'$. Call the resulting family $\mathcal{F}''$.

Claim 3.2. *Phase 1 stops after at most $\frac{k}{4}$ iterations.*

Proof. Every iteration discards at least β elements from the ground set. We begin with r elements, so we stop after at most $\frac{r}{\beta} = \frac{k}{4}$ iterations. □

Claim 3.3. *Phase 2 stops after less than ϵk iterations.*

Proof. Assume for the sake of contradiction that Phase 2 continued for at least ϵk iterations. So, ϵk sets of $\mathcal{F}'$ are covered by at most $\frac{\epsilon k}{4\epsilon} = \frac{k}{4}$ other sets of $\mathcal{F}'$. Considering once again sets dropped by Phase 1, ϵk sets from $\mathcal{F}$ are covered by at most $\frac{k}{4} + \frac{k}{4} = \frac{k}{2}$ other sets from $\mathcal{F}$. In other words, at most $\frac{k}{2}$ sets are identifiable from the union of these $\left(\frac{1}{2} + \epsilon\right) k \leq k$ sets, contradicting the definition of $\mathcal{F}$. Hence, our assumption must be wrong. □

[7] More rigorously, apply this reasoning to some S' of cardinality exactly k such that $S_1 \subset S' \subseteq S$.

We now have a family $\mathcal{F}''$ of cardinality greater than $n-k$ with the following property: every $F \in \mathcal{F}''$ has a subset of $4\epsilon|F|$ elements unique to F. Let $\gamma = \frac{16\epsilon}{k}$. Every such unique subset is of cardinality $4\epsilon|F| \leq 4\epsilon\beta = \frac{16\epsilon}{k}r = \gamma r$; Thus,

$$n - k + 1 \leq |\mathcal{F}''| \leq \left|\binom{[r]}{\leq \gamma r}\right| = \sum_{t=0}^{\gamma r}\binom{r}{t} \leq 1 + \gamma r\binom{r}{\gamma r} \leq 1 + \gamma r\left(\frac{re}{\gamma r}\right)^{\gamma r}.$$

Taking logarithms we get

$$\Omega(\log n) \leq (\log n) - 1 \leq \log(n-k) \leq \gamma r \log\frac{e}{\gamma} + o(r) = O\left(r\frac{\epsilon}{k}\log\frac{k}{\epsilon}\right).$$

Therefore, $r = \Omega\left(\frac{k/\epsilon}{\log k/\epsilon}\log n\right)$.

3.2 Monotone Encodings

As we already stated, $r(n,k) \geq \rho(n,k)$ by a counting argument. The trivial identity encoding implies that $r(n,k) \leq n$. Obviously, $r(n,1) = \rho(n,1) = \lceil\log(1+n)\rceil$. Theorem 3 states that $r(n,k) = \Theta(\rho(n,k))$. In Section 4 we will prove that $r(n,2) \leq \rho(n,2) + O(1)$. The following simple proposition shows that sometimes $r(n,2) \geq \rho(n,2) + 1$.

Proposition 3.4. *$r(5,2) = 5 > 4 = \rho(5,2)$.*

Proof. Assume for the sake of contradiction that $r(5,2) \leq \rho(5,2) = \lceil\log(1+5+10)\rceil = 4$ and let $f : \binom{[5]}{2} \to 2^{[4]}$ be an ME$(5,2,4)$. Without loss of generality, $f(\emptyset) = \emptyset$. There must be some $i \in [5]$ such that $|f(\{i\})| \geq 2$, since $\{f(\{i\})\}_{i=1}^{5}$ cannot all reside in $\binom{[4]}{1}$. Without loss of generality, assume that $\{1,2\} \subseteq f(\{1\})$.

We have reached a contradiction, as $f(\{1,2\}), f(\{1,3\}), f(\{1,4\}), f(\{1,5\})$ are 4 distinct subsets of $2^{[4]}$ that properly contain $\{1,2\}$. Thus, $r(5,2) \geq 5$. But obviously $r(5,2) \leq 5$, hence $r(5,2) = 5$. □

The proof of Proposition 3.4 extends to show that for some choices of n and k, $r(n,k)$ exceeds $\rho(n,k)$ by more than an additive constant.

Lemma 3.5. *Let $f : \binom{[n]}{\leq k} \to 2^{[n-1]}$ be an ME$(n,k,n-1)$ and let $0 \leq m \leq k$. Then, there exists $S \in \binom{[n]}{\leq m}$ such that $|f(S)| \geq 2m$.*

Proof. By induction on m. The claim trivially holds for $m = 0$; Assuming it holds for $m-1$, we prove it for m. Let S be a subset of $[n]$ of cardinality at most $m-1$ such that $|f(S)| \geq 2m-2$. Consider $\{S \cup \{j\} : j \notin S\}$. These are at least $n - m + 1$ sets of cardinality at most m. The images under f of all of these sets must be distinct and strictly contain $f(S)$. Since there are only $\leq n - 2m + 1 < n - m + 1$ such sets of cardinality $2m-1$, one of these has cardinality at least $2m$. □

Using this lemma, we can prove the following.

Corollary 3.6. *For all odd integers $n \geq 5$, $r(n, \frac{n-1}{2}) \geq 1 + \rho(n, \frac{n-1}{2})$.*

We omit the simple proof since we proceed to establish a stronger result. We need the following corollary of Stirling's approximation formula, where $H(\alpha) = \alpha \log \frac{1}{\alpha} + (1-\alpha) \log \frac{1}{1-\alpha}$ is the binary entropy function.

Claim 3.7. $\log \binom{t}{\alpha t} = tH(\alpha) - \frac{1}{2} \log (2\pi\alpha(1-\alpha)t) + o(1)$.

Proposition 3.8. *There exists a constant $\delta > 0$ such that $r\left(n, \frac{1-\delta}{2}n\right) = n$ for sufficiently large n.*

Proof. Let $k = \frac{1-\delta}{2}n$ and let $m = \left(\frac{1}{2} - \delta\right) n = k - \frac{\delta}{2}n$. Assume for the sake of contradiction that $r(n,k) \leq n-1$ and let $f : \binom{[n]}{\leq k} \to 2^{[n-1]}$ be an $\mathrm{ME}(n,k,n-1)$.

By Lemma 3.5, there exists some S of cardinality at most m such that $|f(S)| \geq 2m = (1-2\delta)n$. Consider $\{S \cup T : T \in \binom{[n]\setminus S}{\frac{\delta}{2}n}\}$. These are at least $\binom{n-m}{\frac{\delta}{2}n}$ sets in $\binom{[n]}{\leq k}$ whose images under f are all distinct and (properly) contain $f(S)$. Therefore, for sufficiently small δ and sufficiently large n,

$$\log \binom{n-m}{\frac{\delta}{2}n} = \log \binom{\left(\frac{1}{2}+\delta\right)n}{\frac{\delta}{2}n} > \left(\frac{1}{2}+\delta\right) nH\left(\frac{\delta}{1+2\delta}\right) - \frac{1}{2}\log n > 2\delta n = n - 2m > |[n-1]\setminus f(S)|,$$

which is a contradiction. Thus, $r(n,k) \geq n$ and hence $r(n,k) = n$. □

Another useful entropy-related estimation we need is

Claim 3.9. $H\left(\frac{1-\delta}{2}\right) = 1 - \frac{\delta^2}{2\ln 2} + O(\delta^4)$.

Proof. We use the fact that $\ln(1 \pm \delta) = \pm\delta - \frac{\delta^2}{2} + o(\delta^2)$.

$$\begin{aligned} 1 - H\left(\frac{1-\delta}{2}\right) &= 1 + \frac{1+\delta}{2}\log\left(\frac{1+\delta}{2}\right) + \frac{1-\delta}{2}\log\left(\frac{1-\delta}{2}\right) = \frac{1+\delta}{2}\log(1+\delta) + \frac{1-\delta}{2}\log(1-\delta) \\ &= \frac{1}{2\ln 2}\left((1+\delta)\left(\delta - \frac{\delta^2}{2}\right) - (1-\delta)\left(\delta + \frac{\delta^2}{2}\right)\right) + o(\delta^2) \\ &= \frac{\delta}{4\ln 2}\left((1+\delta)(2-\delta) - (1-\delta)(2+\delta)\right) + o(\delta^2) = \frac{\delta^2}{2\ln 2} + o(\delta^2)\,. \end{aligned}$$

Note that $H\left(\frac{1-\delta}{2}\right)$ is symmetric around $\delta = 0$ and has continuous derivatives of all orders, hence the $o(\delta^2)$ term is actually $O(\delta^4)$. □

We are now ready to prove the lower bound on $r(n,k)$ of Theorem 4.

Proof of Theorem 4. Choose $k = k(n) = \frac{1-\delta}{2}n$ where δ is the constant from Proposition 3.8. By Claim 3.9,

$$\rho(n,k) = \log\left(\sum_{i=0}^{k}\binom{n}{i}\right) < nH\left(\frac{k}{n}\right) = nH\left(\frac{1-\delta}{2}\right) = n\left(1 - \frac{\delta^2}{2\ln 2} + O(\delta^4)\right).$$

Hence, $r(n,k) = n > (1+c_4)\rho(n,k)$ for large enough n and sufficiently small $c_4 > 0$. □

In Proposition 3.8 we needed δ to satisfy $\left(\frac{1}{2}+\delta\right) H(\frac{\delta}{1+2\delta}) > 2\delta$, e.g., $\delta = 0.2276$. Thus, the largest value of c_4 in Theorem 4 that follows from the proof is roughly 0.038.

4 Tighter Bounds for $k = 2$

The ME construction presented in Section 2 is optimal up to a constant factor. Yet, it is interesting to see how small this constant can get and whether we can beat MEs induced by superimposed families, even for small values of k where asymptotic superiority still does not apply.

Trivially, $r(n, 1) = \lceil \log(n + 1) \rceil$, as we just need n different non-empty subsets of $[r]$. Hence, the first interesting case is $k = 2$.

The obvious lower bound[8] is $\rho(n, 2) = \left\lceil \log\left(1 + n + \binom{n}{2}\right) \right\rceil = \lceil 2 \log n \rceil - 1 + o(1)$. We prove Theorem 5 by explicitly constructing an $(n, 2)$-monotone encoding of length $\rho(n, 2) + O(1)$.

In contrast to the $r(n, 2) \leq 2 \log n + O(1)$ bound of Theorem 5, Coppersmith and Shearer [5] have shown that for $r < (2.0008 - o(1)) \log n$, no family $\{A_i\}_{i=1}^{n}$ of n subsets of $[r]$ exists for which $\{A_i \cup A_j\}_{1 \leq i < j \leq n}$ are all different.

4.1 Construction Time Again

A monotone function f over the domain $\binom{[n]}{\leq 2}$ can be defined by[9]

$$A_i = f(\{i\}) \text{ for } i \in [n] \quad \text{and} \quad A_{ij} = A_{ji} = f(\{i, j\}) \setminus (A_i \cup A_j) \text{ for } \{i, j\} \in \binom{[n]}{2}.$$

Assuming that $\{A_i : i \in [n]\}$ and $\{A_{ij} : \{i, j\} \in \binom{[n]}{2}\}$ have disjoint supports, it is sufficient to require three conditions for the function to be injective:

(*i*) $A_i \neq A_j$ for $i \neq j$.

(*ii*) $A_{ij} \neq \emptyset$ for $i \neq j$ if $\exists i'$ such that $A_i \cup A_j = A_{i'}$.

(*iii*) $A_{ij} \neq A_{i'j'}$ for $\{i, j\} \neq \{i', j'\}$ satisfying $A_i \cup A_j = A_{i'} \cup A_{j'}$.

As we now strive for a result optimal up to an additive constant, we cannot continue neglecting the effects of rounding.

Definition 5. *For $x \in \mathbb{R}$, define $\lfloor x \rceil = \{\lfloor x \rfloor, \lceil x \rceil\}$.*

Let $p = 1 - \frac{1}{\sqrt{2}} \approx 0.29$ and select minimal a and $b \in \lfloor pa \rceil$ subject to $\binom{a}{b} \geq n$. Select distinct $A_1, \ldots, A_n \in \binom{[a]}{b}$. Obviously, these satisfy condition (*i*). In addition, $|A_i \cup A_j| > b$ for $i \neq j$, so condition (*ii*) is satisfied as well, albeit in the null sense. Condition (*iii*) will be satisfied by an appropriate selection of $A_{ij} \in 2^{[s]}$ where

$$s = \lceil \log B_{max} \rceil, \quad B_{max} = \max_{A \subseteq [a]} B(A) \quad \text{and} \quad B(A) = \left| \left\{ \{i, j\} \in \binom{[n]}{2} : A_i \cup A_j = A \right\} \right|.$$

In simple words, we differentiate between the $B(A)$ pairs colliding at A by labeling each one with a unique number between 1 and $B(A) \leq B_{max} \leq 2^s$.

[8] As Proposition 3.4 demonstrated, $\rho(n, 2)$ is not tight for some values of n.

[9] Without loss of generality we may assume $f(\emptyset) = \emptyset$.

Thus, we only need to determine B_{max} to get an $(n,2)$-monotone encoding of length $r = a + s$. By symmetry,[10] $B(A)$ depends only on $|A|$. For $0 \le m \le a$, denote the value of $B([m])$ by $B(m)$. Clearly,

$$B(m) = \begin{cases} \frac{1}{2}\binom{m}{b}\binom{b}{m-b} = \frac{m!}{2(m-b)!^2(2b-m)!}, & b < m \le 2b, \\ 0, & \text{otherwise.} \end{cases}$$

Proposition 4.1. *$B(m)$ has a single maximum, achieved at $m^* \in \left\lfloor \frac{b}{2p} \right\rceil = \lfloor (2-p)b \rceil$.*

Proof. Let $C(m) = \frac{B(m)}{B(m-1)} = \frac{m(2b-m+1)}{(m-b)^2} = \frac{m+b^2}{(m-b)^2} - 1$. $C(m)$ is decreasing for $b+1 \le m \le 2b+1$, from $C(b+1) = b + b^2 \ge 2$ to $C(2b+1) = 0$. Therefore, $B(m)$ has a single maximum, achieved close to $x \in \mathbb{R}$ satisfying $C(x) = 1$. Let $\underline{m} = \lfloor (2-p)b \rfloor$ and $\overline{m} = \lceil (2-p)b \rceil$.

$$1 + C(1+\overline{m}) = \frac{1 + \overline{m} + b^2}{(1+\overline{m}-b)^2} < \frac{2 + (2-p)b + b^2}{((2-p)b + 1 - b)^2} < \frac{2 + 2\sqrt{2}b + b^2}{((1-p)b+1)^2} = \frac{(b+\sqrt{2})^2}{(\frac{b}{\sqrt{2}}+1)^2} = 2,$$

$$1 + C(\underline{m}) = \frac{\underline{m} + b^2}{(\underline{m}-b)^2} > \frac{b^2}{(\underline{m}-b)^2} > \frac{b^2}{((2-p)b-b)^2} = \frac{b^2}{(1-p)^2 b^2} = \frac{1}{(1-p)^2} = 2\,.$$

We've shown that $C(\underline{m}) > 1 > C(1+\overline{m})$, hence $B(m)$ achieves its maximum at either $\underline{m}$ or $\overline{m}$. □

A rough estimate. We now bound the difference between r and the lower bound $\rho(n,2)$. First, we use Claim 3.7 to get a quick estimate, neglecting $o(1)$ terms and the effects of rounding. Assume that $n = \binom{a}{b}$ and recall that $m^* \approx (2-p)b = \frac{b}{2p} \approx \frac{pa}{2p} = \frac{a}{2}$. Hence,

$$\begin{aligned} r - \rho(n,2) &\approx (a + \log B_{max}) - (2\log n - 1) = a + \log\binom{m^*}{b} + \log\binom{b}{m^*-b} - 1 - 2\log\binom{a}{b} + 1 \\ &\approx a + \log\binom{\frac{a}{2}}{pa} + \log\binom{b}{(1-p)b} - 2\log\binom{a}{pa} \\ &\approx a + \left(\frac{a}{2}H(2p) - \frac{1}{2}\log\left(2\pi 2p(1-2p)\frac{a}{2}\right)\right) + \left(paH(1-p) - \frac{1}{2}\log\left(2\pi(1-p)p^2a\right)\right) \\ &\quad - (2aH(p) - \log(2\pi p(1-p)a)) \\ &= a\left(1 + \frac{1}{2}H(2p) + pH(p) - 2H(p)\right) + \frac{1}{2}\log\frac{(2\pi p(1-p)a)^2}{(2\pi p(1-2p)a)(2\pi(1-p)p^2a)} \\ &= \frac{1}{2}\log\frac{1-p}{p(1-2p)} = \log(1+\sqrt{2}) \approx 1.272\,. \end{aligned}$$

During the computation we used the following identity.

Claim 4.2. *$1 + \frac{1}{2}H(2p) = (2-p)H(p)$ for our choice of p.*

[10] For this analysis we assume that $n = \binom{a}{b}$. If $n < \binom{a}{b}$, $B(A)$ will decrease for some values of A, but the maximum B_{max} should remain unaffected.

Proof.

$$\begin{aligned}(2-p)H(p) = \frac{H(p)}{2p} &= -\frac{p\log p + (1-p)\log(1-p)}{2p} = -\frac{\log p - \frac{1-p}{2p}}{2} = \frac{1-\log p}{2} + \frac{1-2p}{4} \\ &= 1 - \frac{1}{2}\log(2p) - \frac{1-2p}{2}\log(1-p) = 1 - \frac{1}{2}\log(2p) - \frac{1-2p}{2}\log\frac{1-2p}{2p} \\ &= 1 - \frac{\log(2p) - (1-2p)\log(2p) + (1-2p)\log(1-2p)}{2} \\ &= 1 - \frac{2p\log(2p) + (1-2p)\log(1-2p)}{2} = 1 + \frac{1}{2}H(2p)\,.\end{aligned}$$

□

Next we delve into details to check where the estimation above is inaccurate. We lose a little due to the following reasons: (1) While $p \approx \frac{b}{a}$ is irrational, a and b must be integers; (2) If n is just a little bigger than $\binom{a}{b}$, we are forced to increase either a or b.

It can be verified for small values of a and b that the $o(1)$ terms cause no further loss (see Table 1).

A rigorous proof. Although empirical results show that the first loss is always close to $\log(1+\sqrt{2}) \approx 1.272$, so after rounding up only 2 bits are lost, we will rigorously prove only a weaker bound.

We need the two following technical lemmata.

Lemma 4.3. *Let $\tau \geq 1$ be fixed. Then, $B(m^* \pm u) > (1-o(1))e^{-2\sqrt{2}\tau}B_{max}$ for $0 < u \leq \sqrt{\tau b}$.*

Proof. Fix $0 < u \leq \sqrt{\tau b}$ and let $0 \leq v < u$.

$$\begin{aligned}1 + C(m^* - v) = \frac{m^* - v + b^2}{(m^* - v - b)^2} &< \frac{(2-p)b + 1 - v + b^2}{((1-p)b - v - 1)^2} = \frac{b^2}{((1-p)b - v)^2} + O(1/b) \\ &< \frac{(b - \sqrt{2}v)^2 + 2\sqrt{2}bv}{((1-p)b - v)^2} + O(1/b) = 2 + \frac{4\sqrt{2}v}{b} + O(1/b)\,.\end{aligned}$$

Recall that $C(m) = \frac{B(m)}{B(m-1)}$, hence

$$\begin{aligned}\ln\frac{B(m^*)}{B(m^* - u)} = \ln\prod_{v=0}^{u-1} C(m^* - v) &< \sum_{v=0}^{u-1}\ln\left(1 + \frac{4\sqrt{2}v + O(1)}{b}\right) < \sum_{v=0}^{u-1}\frac{4\sqrt{2}v + O(1)}{b} \\ &= \frac{4\sqrt{2}}{b}\binom{u}{2} + O(u/b) < 2\sqrt{2}\frac{u^2}{b} + o(1) \geq 2\sqrt{2}\tau + o(1)\,.\end{aligned}$$

The proof that $1 + C(m^* + v) > 2 - \frac{4\sqrt{2}v}{b} - \Omega(1/b)$ and thus $\ln\frac{B(m^*+u)}{B(m^*)} > -2\sqrt{2}\tau - o(1)$ is analogous. □

Lemma 4.4. *Let $\tau \geq 1$. Then, $2^{-a}\sum_{m=m^*-\sqrt{\tau b}}^{m^*+\sqrt{\tau b}}\binom{a}{m} > 1 - (2+o(1))e^{-2\tau p}$.*

Proof. It is sufficient to show that $\sum_{m=0}^{m^*-\sqrt{\tau b}} \binom{a}{m} \leq (1+o(1))2^a e^{-2\tau p}$. By claim 3.9,

$$\begin{aligned}\log \sum_{m=0}^{m^*-\sqrt{\tau b}} \binom{a}{m} &< aH\left(\frac{m^*-\sqrt{\tau b}}{a}\right) = aH\left(\frac{1}{2} - \frac{\sqrt{\tau b}}{a} + O(1/a)\right) \\ &= a\left(1 - \frac{1}{2\ln 2}\frac{4\tau b}{a^2} + O\left(\frac{\sqrt{b}}{a^2} + \frac{b^2}{a^4}\right)\right) = a - 2\tau\frac{b}{a}\log e + O\left(\frac{\sqrt{b}}{a} + \frac{b^2}{a^3}\right) \\ &< a - 2\tau p \log e + O(1/\sqrt{a})\,.\end{aligned}$$

□

Proposition 4.5. *Assume that $n = \binom{a}{b}$. Then, the first loss is bounded by 10 bits.*

Proof. We will show that the image of f as constructed above covers a fraction of at least $\frac{1}{1000}$ of its range $2^{[r]}$. Select $\tau = 1.5$, satisfying $e^{-2\sqrt{2}\tau} > \frac{1}{70}$ and $1-2e^{-2\tau p} > \frac{1}{6}$; Recall that $B_{max} \leq 2^s < 2B_{max}$. Now,

$$\begin{aligned}\frac{1}{2^r}|Im(f)| &\geq \frac{1}{2^r}\sum_{m=b+1}^{2b}\binom{a}{m}B(m) \geq \frac{1}{2^r}\sum_{m=m^*-\sqrt{\tau b}}^{m^*+\sqrt{\tau b}}\binom{a}{m}B(m) \\ &\geq \frac{B_{max}}{2^s}(1-o(1))e^{-2\sqrt{2}\tau}2^{-a}\sum_{m=m^*-\sqrt{\tau b}}^{m^*+\sqrt{\tau b}}\binom{a}{m} \geq \frac{1}{2}(1-o(1))e^{-2\sqrt{2}\tau}\left(1-(2+o(1))e^{-2\tau p}\right) \\ &> (1-o(1))\frac{1}{2\cdot 70\cdot 6} = (1-o(1))\frac{1}{840} > \frac{1}{1000}\,.\end{aligned}$$

Therefore, $r \leq \lceil \log(1000|Im(f)|)\rceil \leq \lceil \log 1000\rceil + \left\lceil \log\left(1+n+\binom{n}{2}\right)\right\rceil = 10 + \rho(n,2)$. □

Proposition 4.6. *The second loss is bounded by one bit.*

Proof. Fix n. Let a and $b \in \lfloor pa\rceil$ be maximal[11] such that $\binom{a}{b} < n$. If $b \in \lfloor p(a+1)\rceil$, then we are able to use $\binom{[a+1]}{b}$ since by our choice of a, $\binom{a+1}{b} \geq n$. Here we lose exactly one bit as we increased a by 1 while s did not change.

Otherwise, $\lfloor p(a+1)\rfloor - 1 \leq \lfloor pa+1\rfloor - 1 = \lfloor pa\rfloor \leq b < \lfloor p(a+1)\rfloor$, so $b+1 = \lfloor p(a+1)\rfloor$.

Claim 4.7. $b+1 = \lceil p(a-2)\rceil$.

Proof. We use the fact that two distinct integers must differ by at least 1.

$$\begin{aligned}\lfloor p(a+1)\rfloor - \lceil p(a-2)\rceil &< p(a+1) - p(a-2) = 3p < 1 \\ \lfloor p(a+1)\rfloor - \lceil p(a-2)\rceil &\geq \lfloor pa\rfloor + 1 - \lceil p(a-2)\rceil > pa - p(a-2) - 1 = 2p - 1 > -1\end{aligned}$$

Hence $\lfloor p(a+1)\rfloor - \lceil p(a-2)\rceil = 0$ and $b+1 = \lfloor p(a+1)\rfloor = \lceil p(a-2)\rceil$. □

Corollary 4.8. *By our choice of b, $n \leq \binom{a-2}{b+1}$.*

In the following claim, s', $B'(m)$ and B'_{max} are used to indicate new values of s, $B(m)$ and B_{max} (resp.) when $\binom{[a-2]}{b+1}$ is used instead of $\binom{[a]}{b}$.

[11] To be exact, first select maximal b such that there exists a for which $b \in \lfloor pa\rceil$ and $\binom{a}{b} < n$ and then select maximal a such that $b \in \lfloor pa\rceil$ and $\binom{a}{b} < n$.

Claim 4.9. $s' \le s+3$.

Proof. B'_{max} is achieved at either m^*+1, m^*+2, or m^*+3 since $(2-p)(b+1) = 2+(2-p)b$.

$$\begin{aligned}
\frac{B'(m^*+1)}{B(m^*)} &= \frac{\frac{1}{2}\binom{m^*+1}{b+1}\binom{b+1}{m^*-b}}{\frac{1}{2}\binom{m^*}{b}\binom{b}{m^*-b}} = \frac{m^*+1}{2b-m^*+1} < \frac{(2-p)b+2}{2b-(2-p)b} = \frac{2-p}{p} + o(1) = 3+2\sqrt{2}+o(1)\,,\\
\frac{B'(m^*+2)}{B(m^*)} &= \frac{\frac{1}{2}\binom{m^*+2}{b+1}\binom{b+1}{m^*+1-b}}{\frac{1}{2}\binom{m^*}{b}\binom{b}{m^*-b}} = \frac{(m^*+1)(m^*+2)}{(m^*+1-b)^2} < \frac{(m^*+2)^2}{(m^*+1-b)^2}\\
&< \frac{((2-p)b+3)^2}{((2-p)b-b)^2} = \left(\frac{(2-p)b+3}{(1-p)b}\right)^2 = 2(2-p)^2 + o(1) = 3+2\sqrt{2}+o(1)\,,\\
\frac{B'(m^*+3)}{B(m^*)} &= \frac{\frac{1}{2}\binom{m^*+3}{b+1}\binom{b+1}{m^*+2-b}}{\frac{1}{2}\binom{m^*}{b}\binom{b}{m^*-b}} = \frac{(m^*+1)(m^*+2)(m^*+3)(2b-m^*)}{(m^*+1-b)^2(m^*+2-b)^2}\\
&< \frac{(m^*+2)^3(2b-m^*)}{(m^*+1-b)^4} < \frac{((2-p)b+3)^3(2b-(2-p)b+1)}{((2-p)b-b)^4}\\
&= \frac{((2-p)b+3)^3(pb+1)}{(1-p)^4b^4} = 4p(2-p)^3 + o(1) = 3+2\sqrt{2}+o(1)\,.
\end{aligned}$$

Therefore, $\dfrac{B'_{max}}{B_{max}} \le 3+2\sqrt{2}+o(1) < 5.83+o(1) < 8$ and $s'-s = \lceil \log B'_{max} \rceil - \lceil \log B_{max} \rceil \le \lceil \log 8 \rceil = 3$.

Note: this method of proof does not hold for very small values of b, but it can be verified that $s' \le s+3$ remains true for these (see Table 1). □

Together, these claims state that $\binom{[a-2]}{b+1}$ satisfies $b+1 \in \lfloor p(a-2) \rceil$, has at least n sets and that the obtained length is at most $(a-2)+s' \le a-2+(s+3) \le r+1$. □

Thus, we have proved that our construction is optimal up to an additive constant $c = 11$. Again, empirical evidence shows that the correct value of this constant is 3, but to avoid further complication in the proof we settled for the above estimate.

$n = \binom{a}{b}$	a	b	B_{max}	r	$\rho(n,2)$	loss (1): best n	losses (1)+(2): worst n
6	6	1	1	6	5	1	
15	6	2	3	8	7	1	3
21	7	2	3	9	8	1	2
28	8	2	3	10	9	1	2
35	7	3	15	11	10	1	2
56	8	3	15	12	11	1	2
84	9	3	15	13	12	1	2
120	10	3	15	14	13	1	2
165	11	3	15	15	14	1	2
220	12	3	15	16	15	1	2
286	13	3	15	17	16	1	2
330	11	4	70	18	16	2	2
495	12	4	70	19	17	2	3
715	13	4	70	20	18	2	3
1001	14	4	70	21	19	2	3
1365	15	4	70	22	20	2	3
1820	16	4	70	23	21	2	3
2002	14	5	315	23	21	2	2
3003	15	5	315	24	23	1	3
4368	16	5	315	25	24	1	2
6188	17	5	315	26	25	1	2
8568	18	5	315	27	26	1	2
11628	19	5	315	28	27	1	2
15504	20	5	315	29	27	2	2
18564	18	6	1575	29	28	1	2
27132	19	6	1575	30	29	1	2
38760	20	6	1575	31	30	1	2
54264	21	6	1575	32	31	1	2
74613	22	6	1575	33	32	1	2
100947	23	6	1575	34	33	1	2
116280	21	7	8316	35	33	2	2
170544	22	7	8316	36	34	2	3
245157	23	7	8316	37	35	2	3
346104	24	7	8316	38	36	2	3
480700	25	7	8316	39	37	2	3
657800	26	7	8316	40	38	2	3
735471	24	8	42042	40	38	2	2
888030	27	7	8316	41	39	2	3
1081575	25	8	42042	41	40	1	2
1562275	26	8	42042	42	41	1	2

Table 1: Parameters of the $k = 2$ construction for small values of a and b.

5 Concluding Remarks and Open Problems

5.1 Encoding and Decoding Algorithms

Proposition 2.3 fuels a recursive algorithm to decode S from $f(S)$:

1. Separate $f(S)$ to A_S and $f'(S')$.
2. Determine S' by running the algorithm recursively on $f'(S')$.
3. Find all sets $A_i \subseteq A_S$.
4. Add j to S'' if some $x \in A_S$ is present solely in A_j.
5. Return $S' \cup S''$.

A quick calculation shows that the running time of the whole decoding algorithm is $O\left(nk \log \frac{n}{k}\right)$. This is rather expensive as it is exponential in $r = O\left(k \log \frac{n}{k}\right)$ for $k = \text{poly}(\log n)$. The encoding algorithm suffers from the same behaviour, since basically it determines S' in the same way and encodes it recursively.

Our explicit construction for $k = 2$, however, has polynomial-time encoding and decoding algorithms. We will use the following algorithms as subroutines.

Claim 5.1. *Fix integers $a \geq b \geq 0$. Let φ_b^a be the lexicographic isomorphism from $\binom{[a]}{b}$ to $[\binom{a}{b}]$ and let $\psi_b^a : [\binom{a}{b}] \to \binom{[a]}{b}$ be its inverse. There exist $\text{poly}(a, b)$-time algorithms computing $\varphi_b^a(S)$ given S and $\psi_b^a(m)$ given m.*

Proof. Recursive algorithms based on the identity $\binom{a}{b} = \binom{a-1}{b-1} + \binom{a-1}{b}$ will do the job. Initialize $\varphi_a^a([a]) = \varphi_0^a(\emptyset) = 1$, $\psi_a^a(1) = [a]$ and $\psi_0^a(1) = \emptyset$. Now define recursively

$$\varphi_b^a(S) = \begin{cases} \binom{a-1}{b} + \varphi_{b-1}^{a-1}(S \setminus \{a\}), & a \in S, \\ \varphi_b^{a-1}(S), & \text{otherwise} \end{cases}$$

and

$$\psi_b^a(m) = \begin{cases} \{a\} \cup \psi_{b-1}^{a-1}\left(m - \binom{a-1}{b}\right), & m > \binom{a-1}{b}, \\ \psi_b^{a-1}(m), & \text{otherwise.} \end{cases}$$

Notice that all numbers $\binom{a'}{b'}$ for $a' \leq a, b' \leq b$ can be computed using dynamic programming based on the above-mentioned identity. This requires ab additions of $(b \log(a/b))$-bit integers, that is, $\text{poly}(a, b)$-time as well. □

Proposition 5.2. *The $(n, 2)$-monotone encoding presented in section 4 and its inverse can be computed in $\text{poly}(\log n)$-time (per input).*

Proof. Let a, b be the parameters of the construction. Encoding and decoding the empty set is trivial. By Claim 5.1, encoding and decoding a singleton takes $\text{poly}(a, b)$ time as well.[12] We are left with the interesting case - a set of cardinality 2.

[12] When decoding, we can detect a singleton by checking that the cardinality of the input is exactly b.

Encoding: Let $S = \{i, j\}$ be the input for encoding. Calculate A_i and A_j using Claim 5.1. Recall that S should be encoded by $X \cup A_{ij}$ where $X = A_i \cup A_j$, $m = |X|$ and A_{ij} encodes an index between 1 and $B(m) = \binom{m}{2(m-b)}\binom{2(m-b)-1}{m-b}$. Using Claim 5.1, convert $Y = A_i \cap A_j$ to a number $u \in [\binom{m}{2(m-b)}]$ and convert $A_i \setminus Y$ to a number $v \in [\binom{2(m-b)-1}{m-b}]$.

It might appear as though one needs a range of $\binom{2(m-b)}{m-b}$ to properly split $X \setminus Y$ to $A_i \setminus Y$ and $A_j \setminus Y$, but since the order of i and j does not matter, we may pick i such that $A_i \setminus Y$ does not contain the maximal element of $X \setminus Y$, saving a bit.

Combine u and v to a single number $w = u + (v-1)\binom{m}{2(m-b)} \in [B(m)]$ and encode it as a set $A_{ij} \in 2^{[s]}$.

Decoding: Let $X \cup A_{ij}$ be the input for decoding. We kept the support of $\{A_i\}_{i=1}^n$ separate from $\{A_{ij}\}$'s, so we have $X = A_i \cup A_j$ and $A_{ij} \in 2^{[s]}$. A_{ij} simply encodes w, from which we can recover u and v as the remainder and the quotient of w divided by $\binom{m}{2(m-b)}$ (we know $m = |X|$). Using Claim 5.1, convert u back to Y and v back to $A_i \setminus Y$. Now it is easy to determine $A_i = Y \cup (A_i \setminus Y)$ and $A_j = Y \cup (X \setminus A_i)$ and to use Claim 5.1 once more to determine i and j.

Running time is poly$(\log n)$-time for encoding or decoding as we only used Claim 5.1 and basic integer arithmetic for numbers of order $\log n$. □

5.2 Open Problems

Exact constructions. In spite of Proposition 3.4, we believe that usually $r(n, 2) = \rho(n, 2)$. For a fixed n, the following method can be used to check if $r(n, 2) = \rho(n, 2)$. First, we assign all singletons $\{f(\{i\})\}_{i=1}^n$ to small subsets of $2^{[\rho(n,2)]}$ (and obviously $f(\emptyset)$ to $\emptyset$). Next, we build the bipartite constraints graph:

- On one side $U = \binom{[n]}{2}$ we have all pairs,
- On the other side $V \subset 2^{[\rho(n,2)]}$ we have all unassigned targets;
- An edge connects $\{i, j\} \in U$ and $A \in V$ iff $f(\{i\}) \cup f(\{j\}) \subseteq A$.

A matching in this graph that saturates U translates into an $(n, 2)$-monotone encoding.[13]

Using Hopcroft-Karp's maximum-cardinality bipartite matching algorithm (see A.1), we verified that a saturating matching exists for $23 \leq n \leq 250$. Especially interesting is $n = 90$ since $1 + 90 + \binom{90}{2} = 2^{12}$, rendering the ME surjective as well. This suggests the following conjecture.

Conjecture 1. *For all $n \geq 23$, $r(n, 2) = \rho(n, 2)$.*

Maybe the following stronger version is true as well.

Conjecture 2. *For every fixed $k \geq 2$, $r(n, k) \neq \rho(n, k)$ for only a finite number of values of n.*

[13] It is possible that $r(n, 2) = \rho(n, 2)$ and still the graph does not contain a matching saturating U, as the values of $\{f(\{i\})\}_{i=1}^n$ we have chosen are not necessarily those leading to an optimal encoding.

Note that in the notation above, almost always $|U| < |V|$, i.e., there is some 'extra' space. Indeed, this is a simple consequence of the ABC conjecture, as we explain next.

Masser and Oesterlé conjectured in 1985 that for any $\epsilon > 0$ there exists a constant $K_\epsilon > 0$ such that for every triple of coprime positive integers a, b, c satisfying $a+b = c$ we have $c \leq K_\epsilon(\mathrm{rad}(abc))^{1+\epsilon}$ where $\mathrm{rad}(m)$ is defined as the product of all distinct prime divisors of m. This is known as the ABC Conjecture (see [13, 15]) and has numerous number-theoretic consequences including the following one.

Claim 5.3. *For any fixed M we have $2^{\rho(n,2)} = 1 + n + \binom{n}{2} + M$ for only a finite number of values of n, under the assumption that the ABC Conjecture holds.*

Proof. Fix M and let n be an integer such that $2^{\rho(n,2)} = 1 + n + \binom{n}{2} + M = \frac{(2n+1)^2+7}{8} + M$. Let $a = (2n+1)^2$, $b = 8M + 7$ and $c = 2^{3+\rho(n,2)}$. Note that $\gcd(a, b, c) = 1$ and that $\mathrm{rad}(abc) \leq 2\sqrt{a}b$. By the ABC Conjecture with $\epsilon = \frac{1}{2}$, we have $a \leq c \leq K_{1/2}(\mathrm{rad}(abc))^{3/2} = O(a^{3/4}M^{3/2})$. This is possible only for a finite number of values of n. □

Corollary 5.4. *For any fixed M we have $2^{\rho(n,2)} \geq 1 + n + \binom{n}{2} + M$, i.e., $|V| \geq |U| + M$ for all but a finite number of values of n, under the assumption that the ABC Conjecture holds.*

In other words, the matching is almost never required to be nearly perfect. It seems likely that the assertion of the last two claims can be proved without relying on any unproven conjectures, using the theory of imaginary quadratic fields, but as this is not very essential for our purpose in this paper, we include only the conditional simple proof above.

Explicit constructions, general case. Although the ME construction of Theorem 3 is explicit, it relies on using FUT families of various sizes as building blocks, for which we only presented a probabilistic construction. A bipartite graph $G = (U, V, E)$ in which the degree of every vertex $u \in U$ is s is called *a (k, δ)-expander* if any $U' \subset U$ of size at most k has at least $\delta s|U'|$ neighbors in V. $(k, 1-\epsilon)$-expanders for some small $\epsilon > 0$, called *lossless expanders*, may assist us in building FUT families as any $(k, 1-\epsilon)$-expander yields a $(k, 1-2\epsilon)$-FUT family; However, the best known explicit constructions of these (see [4, 10]) do not suffice for the recursive chaining procedure of Theorem 3.

A Computer Programs

A.1 Python code verifying the existence of an exact construction for $k = 2$

```
#! /usr/bin/python2.5

import math

# Hopcroft-Karp bipartite max-cardinality matching and max independent set
# David Eppstein, UC Irvine, 27 Apr 2002
# http://www.ics.uci.edu/~eppstein/PADS/BipartiteMatching.py
from BipartiteMatching import *

def powerSet(s):
  if not s: return [set()]
  res = powerSet(s[:-1])
  last = set(s[-1:])
  return res + map(last.union, res)

def generateGraph(n):
  rho2 = math.log(1 + n + n*(n-1)/2, 2)
  r = int(rho2 +.999999)
  R = powerSet(range(r))
  R.sort(key=len)
  S,R = R[1:1+n],R[1+n:] # drop empty set, separate singletons
  C = [set( tuple(A) for A in R if A.issuperset(S[i]) ) for i in xrange(n)]
  G = {}
  for i in xrange(n):
    for j in xrange(i):
      G[j,i] = list(C[i].intersection(C[j]))
  return G

if __name__ == '__main__':
  for n in xrange(4, 250):
    M = matching(generateGraph(n))[0] # take only the matching
    if len(M) < n*(n-1)/2:
      print "Saturating matching not found for n=%d" % n

#OUTPUT:
# Saturating matching not found for n=5
# Saturating matching not found for n=7
# Saturating matching not found for n=10
# Saturating matching not found for n=15
# Saturating matching not found for n=22
```

References

[1] *N. Alon*, Explicit construction of exponential sized families of k-independent sets, Discrete Mathematics 58(2), pp. 191–193 (1986).

[2] *N. Alon and V. Asodi*, Tracing a single user, European Journal of Combinatorics 27(8), pp. 1227–1234 (2006).

[3] *N. Alon and V. Asodi*, Tracing many users with almost no rate penalty, IEEE Transactions on Information Theory 53(1), pp. 437–439 (2007).

[4] *M. Capalbo, O. Reingold, S. Vadhan and A. Wigderson*, Randomness conductors and constant-degree lossless expanders, in proc. of the 34th STOC, pp. 659–668 (2002).

[5] *D. Coppersmith and J. Shearer*, New bounds for union-free families of sets, Electronic Journal of Combinatorics 5(1), #R39 (1998).

[6] *M. Csűrös and M. Ruszinkó*, Single user tracing and disjointly superimposed codes, IEEE Transactions on Information Theory 51(4), pp. 1606–1611 (2005).

[7] *A. G. Dyachkov and V. V. Rykov*, Bounds on the length of disjunctive codes, Problemy Peredachi Informatsii 18(3), pp. 158–166 (1982).

[8] *P. Erdős, P. Frankl and Z. Füredi*, Families of finite sets in which no set is covered by the union of r others, Israel Journal of Mathematics 51(1-2), pp. 79–89 (1985).

[9] *Z. Füredi*, A note on r-cover-free families, Journal of Combinatorial Theory Series A 73(1), pp. 172–173 (1996).

[10] *V. Guruswami, C. Umans and S. Vadhan*, Unbalanced Expanders and Randomness Extractors from Parvaresh-Vardy Codes, in proc. of the 22nd CCC, pp. 96–108 (2007).

[11] *J. Komlós and A. Greenberg*, An asymptotically fast nonadaptive algorithm for conflict resolution in multiple-access channels, IEEE Transactions on Information Theory 31(2), pp. 302–306 (1985).

[12] *B. Laczay and M. Ruszinkó*, Multiple user tracing codes, in proc. of ISIT 2006, pp. 1900–1904 (2006).

[13] *D. W. Masser*, Note on a conjecture of Szpiro, Astérisque 183, pp. 19–23 (1990).

[14] *T. Moran, M. Naor and G. Segev*, Deterministic history-independent strategies for storing information on write-once memories, in proc. of the 34th ICALP, pp. 303–315 (2007).

[15] *J. Oesterlé*, Nouvelles approches du "théorème" de Fermat, Astérisque 161/162, pp. 165–186 (1988).

[16] *E. Porat and A. Rothschild*, Better construction of error-correcting codes meeting the Gilbert-Varshamov bound and better non-adaptive combinatorial group testing schemes, submitted.

[17] *M. Ruszinkó*, On the upper bound of the size of the r-cover-free families, Journal of Combinatorial Theory Series A 66(2), pp. 302–310 (1994).