New Bounds for Union-free Families of Sets

Don Coppersmith (email: copper@watson.ibm.com)
James B. Shearer (email: jbs@watson.ibm.com)
Mathematical Sciences Department
IBM Research Division
T. J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598 U.S.A.

Submitted: April 11, 1997; Accepted: July 24, 1998

Abstract: Following Frankl and Füredi [1] we say a family, F, of subsets of an n-set is weakly union-free if F does not contain four distinct sets A, B, C, D with $A \cup B = C \cup D$. If in addition $A \cup B = A \cup C$ implies $B = C$ we say F is strongly union-free. Let $f(n)$ ($g(n)$) be the maximum size of strongly (weakly) union-free families. In this paper we prove the following new bounds on f and g: $2^{[0.31349+o(1)]n} \leq f(n) \leq 2^{[0.4998+o(1)]n}$ and $g(n) \leq 2^{[0.5+o(1)]n}$.

AMS Subject Classification. 05B10

1 Introduction

Let F be a family of subsets of an n-set. Suppose F does not contain four distinct sets A, B, C, D such that $A \cup B = C \cup D$. Then following Frankl and Füredi [1] we say F is *weakly union-free*. If $A \cup B = A \cup C$ implies $B = C$ then we say F is *cancellative*. If F is both weakly union-free and cancellative we say F is *strongly union-free*. Let $f(n)$ (respectively $g(n)$) be the maximum size of a strongly (respectively weakly) union-free family of subsets of an n-set. In this paper we prove new bounds on $f(n)$ and $g(n)$. We show $2^{[0.31349+o(1)]n} \leq f(n) \leq 2^{[0.4998+o(1)]n}$ and $g(n) \leq 2^{[0.5+o(1)]n}$. The best bounds previously known were $2^{[0.2534+o(1)]n} \leq f(n) \leq 2^{[0.5+o(1)]n}$ and $2^{[0.3333+o(1)]n} \leq g(n) \leq 2^{[0.75+o(1)]n}$ (see Frankl and Füredi [1]). We were unable to improve the lower bound for $g(n)$.

We will need the following result of Fredman and Komlós ([3], see also [2]). Consider an alphabet consisting of k ordinary symbols $a_1, \ldots, a_k$ and one special symbol $*$ ($*$ can be thought of as a "don't-care" indicator). Following Fredman and Komlós we will say two vectors $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$ with elements chosen from this alphabet are *strongly different* if there exists a j $(1 \leq j \leq n)$ such that $x_j \neq y_j$ and $x_j \neq *, y_j \neq *$. Suppose we have m pairwise strongly different vectors (with elements $\{x_{ij} | 1 \leq i \leq m, 1 \leq j \leq n\}$). Let $h_{j\ell}$ be the number of vectors with jth element a_ℓ. Let h_{j*} be the number of vectors with jth element $*$. Note $m = h_{j1} + \cdots + h_{jk} + h_{j*}$ for $1 \leq j \leq n$. Let $p_{j\ell} = \frac{h_{j\ell}}{m}$. Let $p_{j*} = \frac{h_{j*}}{m}$. Let $q_{j\ell} = \frac{h_{j\ell}}{h_{j1} + \cdots + h_{jk}}$. Then we need the following bound on m which is a special case of Theorem 1 in ([3]). We include a proof.

Theorem 1 $m \ln(m) \leq \sum_{j=1}^{n} \left(\sum_{\ell=1}^{k} h_{j\ell} \right) \left(\sum_{\ell=1}^{k} -q_{j\ell} \ln q_{j\ell} \right).$

Proof: Intuitively this bound arises as follows. Let R be a random variable which selects one of the m pairwise strongly different vectors (with equal probability). Since there are m choices for R it has entropy $m \ln(m)$. Suppose we can ask about any position of R. If the symbol in that position is ordinary we are told its value. If the symbol in that position is $*$ we are randomly told it is an ordinary symbol with random distribution chosen to match the distribution of ordinary symbols in that position of R. (If R is always $*$ in that position the reply can be a_1 always.) Replying in this way conveys no information about R when the symbol is $*$. So the information about R conveyed is the probability the symbol is ordinary multiplied by the entropy of the distribution of ordinary symbols in that position of R. This is

$$\left(\sum_{\ell=1}^{k} p_{j\ell} \right) \sum_{\ell=1}^{k} - \left(\frac{p_{j\ell}}{\sum_{r=1}^{k} p_{jr}} \right) \ln \left(\frac{p_{j\ell}}{\sum_{r=1}^{k} p_{jr}} \right)$$

Clearly asking about every position of R determines its value (since the possibilities strongly differ). So the entropy of R must not exceed the sum of the information about R conveyed by each of the position queries. Thus

$$\ln m \leq \sum_{j=1}^{n} \left(\sum_{\ell=1}^{k} p_{j\ell} \right) \sum_{\ell=1}^{k} - \left(\frac{p_{j\ell}}{\sum_{r=1}^{k} p_{jr}} \right) \ln \left(\frac{p_{j\ell}}{\sum_{r=1}^{k} p_{jr}} \right)$$

or

$$\ln m \leq \sum_{j=1}^{n} \left(\frac{\sum_{\ell=1}^{k} h_{j\ell}}{m} \right) \sum_{\ell=1}^{k} -q_{j\ell} \ln q_{j\ell},$$

which can be rewritten as

$$m \ln m \leq \sum_{j=1}^{n} \left(\sum_{\ell=1}^{k} h_{j\ell} \right) \sum_{\ell=1}^{k} -q_{j\ell} \ln q_{j\ell},$$

which is the bound we wish to prove.

A more rigorous proof follows. Note we have $\sum_{\ell=1}^{k} q_{j\ell} = 1$. So we have $\prod_{j=1}^{n} \left(\sum_{\ell=1}^{k} q_{j\ell} \right) = 1$. Now let $\sigma_j(x_{ij}) = q_{j\ell}$ if $x_{ij} = a_\ell$ and let $\sigma_j(x_{ij}) = \sum_{\ell=1}^{k} q_{j\ell} = 1$ if $x_{ij} = *$. Associate the ith vector $\{x_{ij} | j = 1, \ldots, n\}$ with the product $\prod_{j=1}^{n} \sigma_j(x_{ij})$. Since the m vectors are strongly different the products associated with the different vectors must consist of non-overlapping groups of terms of the product $\prod_{j=1}^{n} \left(\sum_{\ell=1}^{k} q_{j\ell} \right)$. Hence we have

$$\sum_{i=1}^{m} \prod_{j=1}^{n} \sigma_j(x_{ij}) \leq \prod_{j=1}^{n} \left(\sum_{\ell=1}^{k} q_{j\ell} \right) = 1.$$

The rest follows from the arithmetic-geometric mean:

$$\begin{array}{rcl} 1 & \geq & \sum_{i=1}^{m} \prod_{j=1}^{n} \sigma_j(x_{ij}) \\ & = & \sum_{i=1}^{m} \exp\left(\ln \prod_{j=1}^{n} \sigma_j(x_{ij}) \right) \\ & = & m \sum_{i=1}^{m} \frac{1}{m} \exp\left(\sum_{j=1}^{n} \ln \sigma_j(x_{ij}) \right) \\ & \geq & m \exp\left(\frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{n} \ln \sigma_j(x_{ij}) \right) \\ & = & m \exp\left(\frac{1}{m} \sum_{j=1}^{n} \sum_{\ell=1}^{k} h_{j\ell} \ln q_{j\ell} \right) \end{array}$$

This is readily seen to be equivalent to the inequality in the statement of theorem 1. ■

As noted above $\left(\sum_{\ell=1}^{k} p_{j\ell} \right) \sum_{\ell=1}^{k} -\left(\frac{p_{j\ell}}{\sum_{r=1}^{k} p_{jr}} \right) \ln \left(\frac{p_{j\ell}}{\sum_{r=1}^{k} p_{jr}} \right)$ can be thought of as a kind of generalized entropy of column j when the rows are chosen with equal probability. We will need the following lemma about this generalized entropy function.

Lemma 1 *Let* $J(x_1, \ldots, x_n) = (x_1 + \cdots + x_n) H(\frac{x_1}{(x_1+\cdots+x_n)}, \ldots, \frac{x_n}{(x_1+\cdots+x_n)})$ *where* H *is the ordinary entropy function,* $0 < x_1, \ldots, x_n < 1$ *and* $0 < x_1 + \cdots + x_n \leq 1$. *Then* J*, like* H*, is a convex cap function.*

Proof: Let $a = \frac{\lambda(x_1+\cdots+x_n)}{\lambda(x_1+\cdots+x_n)+(1-\lambda)(y_1+\cdots+y_n)}$. Then since H is convex cap

$$aH\left(\frac{x_1}{(x_1+\cdots+x_n)},\ldots,\frac{x_n}{(x_1+\cdots+x_n)}\right)$$

$$+(1-a)H\left(\frac{y_1}{(y_1+\cdots+y_n)},\ldots,\frac{y_n}{(y_1+\cdots+y_n)}\right)\le$$

$$H\left(\frac{ax_1}{(x_1+\cdots+x_n)}+\frac{(1-a)y_1}{(y_1+\cdots+y_n)},\ldots,\frac{ax_n}{(x_1+\cdots+x_n)}+\frac{(1-a)y_n}{(y_1+\cdots+y_n)}\right)$$

or multiplying through by $\lambda(x_1+\cdots+x_n)+(1-\lambda)(y_1+\cdots+y_n)$ and simplifying:

$$\lambda J(x_1,\ldots,x_n)+(1-\lambda)J(y_1,\ldots,y_n)\le J(\lambda x_1+(1-\lambda)y_1,\ldots,\lambda x_n+(1-\lambda)y_n)$$

which shows J is convex cap. ■

Identify subsets of an n-set with 0-1 vectors of length n in the usual way. Define $\ominus$ as follows:

$$1\ominus 0=1,\quad 0\ominus 0=0,\quad 0\ominus 1=-1,\quad 1\ominus 1=*.$$

Let $\ominus$ operate on vectors componentwise. The definition of $\ominus$ is motivated by the following lemma.

Lemma 2 *Suppose $A\ominus B$ is not strongly different from $C\ominus D$. Then $A\cup D = C\cup B$.*

Proof: For each $x\in\{1,\ldots,n\}$, if $(A\ominus B)_x=*$, then $x\in A$ and $x\in B$, so that $x\in A\cup D$ and $x\in C\cup B$. Similarly if $(C\ominus D)_x=*$, then $x\in A\cup D$ and $x\in C\cup B$. Otherwise $(A\ominus B)_x=(C\ominus D)_x\in\{-1,0,1\}$, so that $x\in A\Leftrightarrow x\in C$, and $x\in B\Leftrightarrow x\in D$. In either case, $x\in A\cup D\Leftrightarrow x\in C\cup B$. This holds for all x, and we have $A\cup D=C\cup B$. ■

Lemma 2 combined with theorem 1 will allow us to bound the number of "nearby" (in the Hamming sense) pairs $\{A,B\}$ in an union-free family. This in turn will yield bounds on the size of union-free families.

2 Weakly Union-Free Upper Bound

We consider first weakly union-free families. We need the following lemma.

Lemma 3 *Let F be a weakly union-free family of subsets of an n-set. Suppose we have four or more pairs (A_i, B_i) such that $A_i \cup B_i = X$. Then some $A \in F$ is a member of every pair (A_i, B_i).*

Proof: It is easy to see the only way to avoid having a common member of every pair is if we have three pairs (A, B), (A, C) and (B, C) with $A \cup B = A \cup C = B \cup C$. This is impossible if we have more than three pairs. ■

We are now ready to prove an upper bound on $g(n)$ which improves the bound $g(n) \leq 2^{[0.75+o(1)]n}$ of Frankl and Füredi [1]. We will use lg for $\log_2$.

Theorem 2

$$g(n) \leq 2^{[0.5+o(1)]n}.$$

Proof: We now use H to denote the binary entropy function. Let F be a weakly union-free family of subsets of an n-set. Suppose F contains $2^{\alpha n}$ subsets and that each subset in F contains pn elements. Let $\phi(p)$ be the convex hull of the function $H(2p - p^2)(0 \leq p \leq 1)$. (i.e. $\phi(p) = \max \{\lambda H(2p_1 - p_1^2) + (1-\lambda)H(2p_2 - p_2^2) \mid \lambda p_1 + (1-\lambda)p_2 = p,\ 0 \leq \lambda \leq 1,\ 0 \leq p_1 \leq p_2 \leq 1\}$) Note $\phi(p) \leq 1$. Let $\beta(p) = \max[0, \frac{8}{11}p - \frac{3}{11}\phi(p)]$. Note $\beta(p) \leq p$. Consider unions $X = A \cup B$ of sets $A, B \in F$. Say a union X is *good* if there are at most $2n2^{n\beta(p)}$ ways of expressing it as $X = A_i \cup B_i$ $(A_i, B_i \in F)$. Otherwise say the union is *bad*.

Suppose first $A \cup B$ is bad for at most a fraction $\frac{1}{n}$ of the ordered pairs (A, B) $(A, B \in F)$. Consider the random variable $X = A \cup B$. It has entropy at least $(1 - \frac{1}{n}) \lg \left(\frac{2^{2\alpha n}}{2n2^{n\beta(p)}} \right)$ or $(2\alpha - \beta(p))n + o(n)$ as $n \to \infty$. Consider X to be a 0-1 vector $(x_1, \ldots, x_n)$. Let p_i be the fraction of the sets of F which contain element i. Let $h(x_i)$ be the entropy of the ith component of X. Clearly as $n \to \infty$, $h(x_i) \to H(2p_i - p_i^2)$. Therefore we have

$$[2\alpha - \beta(p) + o(1)]n \leq \sum_{i=1}^{n} H(2p_i - p_i^2) \leq n\phi(p)$$

(since $pn = \sum p_i$). Therefore

$$\alpha n \leq \frac{1}{2}[\beta(p) + \phi(p) + o(1)]n.$$

Now $\beta(p) + \phi(p) = \max[\phi(p), \frac{8}{11}(p + \phi(p))]$. A calculation shows $p + \phi(p) < 1.35$ and $(\frac{8}{11})(1.35) < 0.982$. Therefore $\beta(p) + \phi(p) \leq 1$. Hence $\alpha \leq \frac{1}{2}[1 + o(1)]$.

Suppose next $A \cup B$ is bad for at least $\frac{1}{n}$ of the ordered pairs (A, B) $(A, B \in F)$. By Lemma 3 every bad union X has associated with it some set, A_X, which is involved in every expression of X. It follows that there is some fixed set A so that at least $\frac{1}{2n}$ of the $2^{\alpha n}$ (unordered) unions X involving A are bad (with $A_X = A$). Fix some $\beta(p)n$ of the pn elements of A. Let A' be the remaining $(p - \beta(p))n$ elements of A.

Consider the partition, P, of the elements of F into groups depending on the value of $A \cup B, B \in F$. By the choice of A at least $\frac{1}{2n}2^{\alpha n}$ elements of F lie in groups of size at least $2n2^{n\beta(p)}$. Now consider the refined partition P' formed by using the value of $A' \cup B$ rather than the value of $A \cup B$. Clearly each group of P will be divided into at most $2^{n\beta(p)}$ parts in P' (since $|A - A'| = \beta(p)n$). Hence any group, G, of size at least $2n2^{n\beta(p)}$ in P will be divided into at most $2^{n\beta(p)}$ subgroups of average size at least $2n$. Say a subgroup is large iff it has size at least n. It is easy to see this means at least half the sets in the group G will lie in large subgroups in P' (since $2^{n\beta(p)}$ subgroups of size less than n can account for at most $n2^{n\beta(p)}$ elements of G).

Thus we have that at least $\frac{1}{4n}2^{\alpha n}$ sets of F lie in subgroups G' of size at least n in P'. Divide each such large subgroup G' into pairs of elements (uniformly) at random. (If the size of G' is odd leave one element unpaired.) Let m be the total number of pairs. Then we have $m \geq \frac{(1-\frac{1}{n})}{8n}2^{\alpha n}$. (The $\frac{1}{n}$ term is due to the possibly unpaired elements.) Let $\{(B_i, C_i)\}$ be the collection of these pairs.

Consider the collection of vectors D_i where $D_i = B_i \ominus C_i$. Suppose $D_i \sim D_j$ (where $\sim$ means "not strongly different from"). Then $B_i \ominus C_i \sim B_j \ominus C_j$. Then by Lemma 2, $B_i \cup C_j = B_j \cup C_i$. Since we are assuming F is weakly union-free, and B_i, C_i, B_j, C_j are all distinct, this cannot occur. Therefore all the vectors $\{D_i\}$ must strongly differ. Note since $A' \cup B_i = A' \cup C_i$, $B_i \ominus C_i$ will be 0 or $*$ on the complement of A'. So in fact the restrictions of the $\{D_i\}$ to A' all strongly differ. Fix $x \in A'$. Let random variables n_1, n_2, n_3, n_4 be the number of times position x of D_i is equal to $*$, 1, -1, 0 respectively for our random pairing (B_i, C_i). We are interested in bounding the expected value, $\bar{S}_x$, of the generalized column entropy S_x. Now $n_1 + n_2 + n_3 + n_4 = m$.

Set $p_i = \frac{n_i}{n_2+n_3+n_4}$, $i = 2, 3, 4$. Then

$$S_x = \tfrac{n_2+n_3+n_4}{m} \times$$
$$\left[-\tfrac{n_2}{n_2+n_3+n_4} \lg \tfrac{n_2}{n_2+n_3+n_4} - \tfrac{n_3}{n_2+n_3+n_4} \lg \tfrac{n_3}{n_2+n_3+n_4} - \tfrac{n_4}{n_2+n_3+n_4} \lg \tfrac{n_4}{n_2+n_3+n_4}\right]$$
$$= \frac{1}{m}\left[(n_2+n_3+n_4)\lg(n_2+n_3+n_4) - n_2 \lg n_2 - n_3 \lg n_3 - n_4 \lg n_4\right].$$

It follows from Lemma 1 that S_x is a convex cap function of n_2, n_3 and n_4.

Therefore the expected value, $\bar{S}_x$, of S_x is less than or equal to this function of the expected values of n_2, n_3 and n_4. Let $\bar{n}_i$ be the expected value of n_i $(i = 1, \ldots, 4)$. So we have

$$\bar{S}_x \leq \frac{1}{m}\left[(\bar{n}_2+\bar{n}_3+\bar{n}_4)\lg(\bar{n}_2+\bar{n}_3+\bar{n}_4) - \bar{n}_2 \lg \bar{n}_2 - \bar{n}_3 \lg \bar{n}_3 - \bar{n}_4 \lg \bar{n}_4\right].$$

The expected values $\bar{n}_1$, $\bar{n}_2$, $\bar{n}_3$ and $\bar{n}_4$ are the sums of the corresponding expected values of these quantities for pairs in each large subgroup G' of P'. The values in each subgroup depend on how many sets in the group contain x. Let the fraction of sets in G' which contain x be $p(G')$. Let $\bar{n}_1(G'), \ldots, \bar{n}_4(G')$ be the expected counts for pairs in G'. Let G' have $m(G')$ pairs. Then

$$\begin{aligned}
\bar{n}_1(G') &= p(G')^2 m(G') + O(1) \\
\bar{n}_2(G') &= \bar{n}_3(G') = [p(G') - p(G')^2] m(G') + O(1) \\
\bar{n}_4(G') &= [1 - 2p(G') + p(G')^2] m(G') + O(1)
\end{aligned}$$

The $O(1)$ terms arise because we are considering pairs of distinct terms. Since we are considering large subgroups with $m(G') \geq n$ they will become negligible as n goes to infinity.

Now the values of $\bar{n}_1, \ldots, \bar{n}_4$ will be determined by the weighted average values of $p(G')$ and $p(G')^2$ (weighted by $m(G')$ for all large subgroups G' in P'). Let p be the weighted average value of $p(G')$ and $p^2 + \epsilon$ be the weighted average value of $p(G')^2$ ($\epsilon \geq 0$ because x^2 is convex cup). Then

$$\begin{aligned}
\bar{S}_x &\leq \tfrac{1}{m}[(1-p^2-\epsilon)m \lg(1-p^2-\epsilon)m - 2(p-p^2-\epsilon)m \lg(p-p^2-\epsilon)m \\
&\qquad - (1-2p+p^2+\epsilon)m \lg(1-2p+p^2+\epsilon)m + O(m/n)] \\
&= (1-p^2-\epsilon)\lg(1-p^2-\epsilon) - 2(p-p^2-\epsilon)\lg(p-p^2-\epsilon) \\
&\qquad - (1-2p+p^2+\epsilon)\lg(1-2p+p^2+\epsilon) + O(1/n).
\end{aligned}$$

The right hand side is maximized when $p = \frac{1}{3}$ and $\epsilon = 0$. Hence

$$\bar{S_x} \leq \frac{4}{3} + O(1/n)$$

This will be true for each $x \in A'$. Therefore by Theorem 1

$$\lg(m) \leq \frac{4}{3}(p - \beta(p))n + O(1)$$

Now $m \geq \left(\frac{1-\frac{1}{n}}{8n}\right) 2^{\alpha n}$ so as $n \to \infty$ we have

$$\alpha \leq \frac{4}{3}(p - \beta(p)) + o(1).$$

Now $\beta(p) = \max[0, \frac{8}{11}p - \frac{3}{11}\phi(p)]$ so $[p - \beta(p)] = \min[p, \frac{3}{11}(p + \phi(p)]$ and $\alpha \leq \min[\frac{4}{3}p, \frac{4}{11}(p + \phi(p))] \leq \frac{4}{11}(p + \phi(p))$.

As above $p + \phi(p) < 1.35$ and $\frac{4}{11}(p + \phi(p)) < 0.491$. Therefore $\alpha < 0.491 + o(1)$.

Hence, in either case, we have shown $\alpha \leq 0.5 + o(1)$ as $n \to \infty$.

We assumed that all members of F contained the same number of elements. However, removing this assumption will increase the size of F by at most a factor of $n + 1$. Thus

$$g(n) \leq (n+1)2^{\alpha n} \leq (n+1)2^{[0.5+o(1)]n} = 2^{[0.5+o(1)]n}$$

which completes the proof. ∎

3 Strongly Union-Free Upper Bound

We now consider strongly union-free families. Recall $f(n)$ is the maximum size of a strongly union-free family of subsets of an n-set. It is easy to see that $f(n) \leq 2^{[0.5+o(1)]n}$ (see Frankl and Füredi [1]). We show below how to improve this slightly to $f(n) \leq 2^{[0.4998+o(1)]n}$. We need the following lemma.

Lemma 4 *Let F be a strongly union-free family of subsets of an n-set. Suppose all members of F contain exactly pn elements and that there are $2^{\beta n}$ pairs $A, B \in F$ such that $\mid A \cap B \mid = tn$. Then $\beta \leq (1-t)H(\frac{p-t}{1-t}, \frac{p-t}{1-t}, \frac{1-2p+t}{1-t})$.*

Proof: Consider the $2^{\beta n}$ vectors $A \ominus B$ constructed from the $2^{\beta n}$ pairs with $\mid A \cap B \mid = tn$. Clearly each such vector will contain tn $*$'s, $(p-t)n$ 1's, $(p-t)n$ -1's and $(1-2p+t)n$ 0's. By Lemma 2 these vectors must be strongly different. So by Theorem 1, $\beta n \leq \sum_{j=1}^{n} J_i$ where J_i is the generalized entropy of column i (considering the $2^{\beta n}$ vectors as a $2^{\beta n}$ by n array). However by Lemma 1 the generalized column entropy function is convex. It follows that $\sum_{j=1}^{n} J_i \leq nJ(p-t, p-t, 1-2p+t) = (1-t)nH(\frac{p-t}{1-t}, \frac{p-t}{1-t}, \frac{1-2p+t}{1-t})$. The lemma follows. ∎

We can now prove our theorem.

Theorem 3

$$f(n) \leq 2^{[0.4998+o(1)]n}.$$

Proof: Let F be a strongly union-free family of subsets of an n-set. Suppose F contains $2^{(\alpha+o(1))n}$ subsets. We may neglect terms which can be buried in the $o(1)$ term. So we may assume each subset in F contains exactly pn elements. For each $i \in \{1, \ldots, n\}$ let p_i be the fraction of sets in F containing i, so that $p = \frac{1}{n} \sum p_i$.

As before, let $\phi(p)$ be the convex hull of the function $H(2p-p^2)$ $(0 \leq p \leq 1)$. Consider the random variable $X = A \cup B$ with A, B chosen uniformly and independently from F. Since F is strongly union-free X will take on $2^{[2\alpha+o(1)]n}$ distinct values and will have entropy $(2\alpha + o(1))n$. This entropy is upper-bounded by the sum of the entropies of the entries of the random vector X. Thus

$$(2\alpha + o(1))n \leq \sum_i H(2p_i - p_i^2) \leq \sum_i \phi(p_i) \leq n\phi(p).$$

Therefore

$$\alpha \leq .5\phi(p) + o(1) \text{ as } n \to \infty$$

We will show below how the above bound can be improved by using Lemma 4 for values of $p \leq .3014-$. Since $.5\phi(p)$ attains its maximum of .5 when $p = 1 - \sqrt{.5} = .2929-$ this yields a slight improvement in the overall bound. To apply Lemma 4 we need to show F must contain many pairs of subsets with some relatively large intersection tn. This can be done as follows. Choose the maximal s so that

$$\mid F \mid \binom{pn}{sn} > 2\binom{n}{sn}. \tag{1}$$

In the worst case (by which we mean the case for which we will prove the weakest bound), with $p = 0.3014-$ we would have $s = 0.2179+$. The left-hand side of (1) counts the tuples (B, S) where $B \in F$, $S \subseteq B$ and $|S| = sn$. The right-hand side of (1) is twice the number of sets $S \subseteq \{1, \ldots, n\}$ with $|S| = sn$. A counting argument shows that some tuples must share the same set S: There are at least $2^{[\alpha+o(1)]n}\binom{pn}{sn}$ triples (B, C, S) with $B, C \in F$; $S \subseteq B$; $S \subseteq C$; and $|S| = sn$, with the two triples (B, C, S) and (C, B, S) counting as one. So we have found pairs of subsets with large intersection. However such pairs may have intersection tn greater than sn. Every such pair will contribute $\binom{tn}{sn}$ triples. Fix a value of t which contributes at least $\frac{1}{n}$ of the triples. Let $2^{\beta n}$ be the number of pairs of subsets of F with intersection of size tn. Then we have

$$2^{\beta n}\binom{tn}{sn} > 2^{[\alpha+o(1)]n}\binom{pn}{sn} \tag{2}$$

where some terms have been incorporated in the $o(1)$. Taking logs and letting $n \to \infty$ equations (1) and (2) become

$$\alpha + pH\left(\frac{s}{p}\right) = H(s) \tag{3}$$

$$\beta + tH\left(\frac{s}{t}\right) \geq \alpha + pH\left(\frac{s}{p}\right) \tag{4}$$

Furthermore by Lemma 4 we have

$$\beta \leq (1-t)H\left(\frac{p-t}{1-t}, \frac{p-t}{1-t}, \frac{1-2p+t}{1-t}\right) \tag{5}$$

Calculations show that for $p \leq .3014-$ if we set $\alpha = .5\phi(p)$, the first bound obtained above, it is impossible to find a value of t, $s \leq t \leq p$ so that equations (3), (4) and (5) are satisfied. Let $\alpha = \psi(p)$ be defined as the maximum value of α (as a function of p) which allows equations (3), (4) and (5) to be satisfied. Further calculations show $\psi(p)$ is increasing for $p \leq .3014-$. Therefore the maximum of the combined bounds occurs at $p = .3014-$ at which point $\alpha = .4998- = .5\phi(p)$, $s = .2117+$ and $t = .2144-$. This suffices to prove the theorem. ∎

4 Lower bound on f

We give a construction of a (strongly) union-free family of subsets of an $n-$set $N = \{1, \ldots, n\}$, containing $2^{[\delta+o(1)]n}$ members, where $\delta > 0.31349$. This improves Frankl and Füredi's [1] bound with $\delta = \frac{1}{2}\lg(\frac{27}{19}) = .2534+$.

Theorem 4

$$f(n) \geq 2^{[0.31349+o(1)]n}.$$

Proof: The idea behind our construction is the following. Frankl and Füredi [1] use a simple random construction which shows $g(n) \geq 2^{[0.33333+o(1)]n}$ which is the best lower bound known for weakly union-free families. However this construction does not work so well for strongly union-free families yielding $f(n) \geq 2^{[0.2534+o(1)]n}$ as noted above. The problem seems to be the cancellative property. Cancellative families produced by the random construction are much smaller than those which can be explicitly constructed. This suggests trying a combined construction. By basing the random construction on explicitly constructed cancellative families we find it easier to ensure the cancellative property thereby bringing the lower bound for $f(n)$ closer to that for $g(n)$. The details are a little complicated because simpler versions of the idea do not seem to give the best results.

We start by defining constants

$$\alpha = \frac{1}{63}\lg(21 \times 3^{19}) \approx 0.5477238879$$

$$\beta = \frac{1}{63}\lg(861 \times 15^{19}) \approx 1.333028425$$

$$p = 0.28765$$

$$q = 1 - p = 0.71235$$

$$\nu = \text{ solution of } \left[(p-\nu)^4 = \nu^2(1-2p+\nu)(2p-\nu)\right] \approx 0.083426$$

$$\tau = 4H(p) - 2H(\nu, p-\nu, p-\nu, 1-2p+\nu) + H(2p-\nu) \approx 0.9992855$$

Find constants

$$\begin{aligned} \epsilon &\approx 0.14521 \\ \gamma &\approx 0.418076 \\ s &\approx 0.1106935 \\ \delta &\approx 0.31349 \end{aligned}$$

which solve the four equations

$$\epsilon = 2H(p) - H(s, s, p-s, 1-p-s)$$

$$\delta = \alpha\gamma + \epsilon(1-\gamma)$$
$$\delta = \beta\gamma - \tau(1-\gamma) + 4\epsilon(1-\gamma)$$
$$\delta = (1-\gamma)\left[H(p) - (1-2s)H((p-2s)/(1-2s))\right]$$

Let $k \approx n\gamma$ be a multiple of 63, and set $\ell = n-k$. Let K be a specific $k-$element subset of N, and L its complement. For instance $K = \{1, \cdots, k\}$ and $L = \{k+1, \cdots, n\}$.

We will use the fact that a violation $A \cup B = A \cup C$ of the cancellative property is equivalent to A containing the symmetric difference $B\Delta C = (B-C) \cup (C-B)$.

Following Shearer [4], we construct a cancellative family of subsets of K as follows: Break K into $k/63$ blocks of 63 elements, and further break each 63-block into 21 triplets. Within each triplet, assign the elements labels 0,1,2. For each subset in our family, for each block, select one of the triplets and take all three of its elements; select one element from each other triplet in the block, in such a way that a parity condition holds: the sum of the 20 labels is divisible by 3. The number of choices for each block is then 21×3^{19}, and the total number of subsets is

$$M_1 = \left(21 \times 3^{19}\right)^{k/63} = 2^{\alpha k}.$$

This family is cancellative because if $A \cup B = A \cup C$ with $B \neq C$, in each block on which B and C differ, there are at least two triplets where $(B\Delta C)$ contains two members each. (If B and C selected different triplets to take all three members, those two triplets suffice; otherwise the parity condition is used.) But then $(B\Delta C) \subset A$ is impossible.

Corresponding to each member A of this family, select

$$M_2 = 2^{\epsilon\ell}\ell^{-1}$$

different subsets $R_{A,i}$ of L, uniformly from those subsets of size

$$h = \lfloor p\ell \rfloor.$$

Remark: The factor ℓ^{-1} is chosen to help with the "deletion method" [5]. At each stage below, the expected number of deletions is quadratic or quartic in M_2, so that the fraction of elements deleted is linear or cubic in M_2. We want each fraction to be bounded by $1/10$, and we make M_2 small accordingly.

For each member A, the expected number of pairs $\{i,j\}$ such that

$$|R_{A,i} \cap R_{A,j}| \geq (p-s)\ell.$$

is

$$O\left(M_2 \frac{M_2}{2^{\epsilon\ell}\ell^{1/2}}\right),$$

and will be bounded by $M_2/10$ for n sufficiently large. Delete one member $R_{A,i}$ of each such pair. We retain at least $0.9M_2$ elements $R_{A,i}$ for each A, all enjoying the *small-intersection property*

$$|R_{A,i} \cap R_{A,j}| < (p-s)\ell.$$

This implies a large symmetric difference ($|R_{A,i}\Delta R_{A,j}| > 2s\ell$), unlikely to be contained in some third element $R_{B,k}$.

Define the family

$$S = \{A \cup R_{A,i} | \text{ all } A, i \text{ where } R_{A,i} \text{ retained}\}$$

It has M members, denoted A, B, C, D, where $0.9M_1M_2 \leq M \leq M_1M_2$.

We will delete some more elements. Whenever $A \cap K = C \cap K$ and $A \neq C$, we have $|A\Delta C| > 2s\ell$, by the small-intersection property. Given any triple (A, B, C) of distinct elements with $A \cap K = C \cap K$ and $(A\Delta C) \subset B$, we delete B. The expected number of such triples is bounded by

$$\sum_{m>s\ell} M_1 M_2^2 \frac{\binom{\ell}{m,m,h-m,\ell-h-m}}{\binom{\ell}{h}^2} M_1 M_2 \frac{\binom{\ell-2m}{h-2m}}{\binom{\ell}{h}}$$

where $m = |A-C|$. The summand is clearly maximized when m is minimized (near $s\ell$); for this value of m we have

$$M_2 \frac{\binom{\ell}{m,m,h-m,\ell-h-m}}{\binom{\ell}{h}^2} = O(\ell^{-3/2})$$

and

$$M_1 M_2 \frac{\binom{\ell-2m}{h-2m}}{\binom{\ell}{h}} = O(\ell^{-1}).$$

So the total number of deletions for $m = s\ell$ is $O(M_1 M_2 \ell^{-5/2})$, and this number decreases geometrically as m increases. For n sufficiently large, the total number of deletions for all m, namely $O(M_1 M_2 \ell^{-5/2})$, is bounded by $0.1 M_1 M_2$, leaving at least $0.8 M_1 M_2$ elements $R_{A \cap K, i}$.

By now S is cancellative: for any instance of $A \cup B = C \cup B$, the cancellative property on K tells us $A \cap K = C \cap K$, and in that case we have deleted any instance where $(A \Delta C) \subset B$. This ensures $A \cup B \neq C \cup B$.

To make S weakly union-free (and thus strongly union-free), we first estimate the number of 4-tuples (A, B, C, D) of distinct members violating the weakly union-free property: $A \cup B = C \cup D$.

Let $A' = A \cap K$, $B' = B \cap K$, $C' = C \cap K$ and $D' = D \cap K$ be among the M_1 subsets of K being considered. The number of 4-tuples (A', B', C', D') (with repetition allowed) whose unions agree:

$$A' \cup B' = C' \cup D'$$

is upper-bounded as follows.

Consider a particular block of 63 elements. If there is only one triplet on which $A' \cup B'$ has all three elements, then the special triplet chosen by A' is the same one chosen by B', C' and D'. There are 21 choices of location for this triplet. For each of 19 other triplets, either the union has one element (in which case B', C' and D' agreed with the choice made by A'), or the union has two elements, in which case B' disagreed with A', C' agreed with either B' or A', and the choice of D' is forced. The total number of choices on this triplet is

$$3 \times 1 \times 1 \times 1 + 3 \times 2 \times 2 \times 1 = 15.$$

Values on the last triplet are forced. (Because we want an upper bound, we can ignore the chance that these forced values might cause a disagreement in the unions; taking this into consideration would improve our bound in the fifth decimal place.) The number of choices for one block, in this case, is then at most 21×15^{19}.

If $A' \cup B'$ contains all elements of two triplets, then B' made a different choice of triplet than did A', and also C' and D' made the same choices in some order; the number of such choices is

$$21 \times 20 \times 2 \times 1 = 840.$$

The other 19 triplets again allow 15^{19} choices. The triplet chosen by A' masks one of the triplets of B', and we let that triplet take care of the parity condition for B'. The number of choices for the block, in this case, is 840×15^{19}. Summing, the number of choices of (A', B', C', D') on one block is bounded by 861×15^{19}; and on all $k/63$ blocks, by

$$\left(861 \times 15^{19}\right)^{k/63} = 2^{\beta\gamma n}.$$

Given a 4-tuple (A, B, C, D) of *distinct* members whose unions agree on K, we evaluate the probability ρ that the unions agree on L. Consider the contribution to ρ due to instances (A, B, C, D) where $|A \cap B \cap L| = |C \cap D \cap L| = \nu'\ell$. Its logarithm, namely,

$$-\ell\left[4H(p) - 2H(\nu', p-\nu', p-\nu', 1-2p+\nu') + H(2p-\nu')\right] - \frac{1}{2}\log \ell + O(1),$$

is maximized when we select $\nu' = \nu$ to be the solution of

$$(p-\nu)^4 = \nu^2(1-2p+\nu)(2p-\nu)$$

lying between 0 and p. With ν and τ as defined above, this gives

$$\rho = O(2^{-\tau\ell}\ell^{-1/2}) = O(2^{-\tau(1-\gamma)n}n^{-1/2}).$$

So the total number of violations (4-tuples of distinct members satisfying $A \cup B = C \cup D$) is upper-bounded by

$$O\left(2^{\beta\gamma n}2^{-\tau(1-\gamma)n}n^{-1/2}2^{4\epsilon(1-\gamma)n}n^{-4}\right) = O\left(Mn^{-7/2}\right)$$

by our choice of parameters.

For n sufficiently large, this number is less than $M/10$. For each violation, discard one of the four sets A, B, C, D. Then we retain more than $M/2$ sets.

Our resulting family has size at least $M/2 = 2^{[\delta+o(1)]n}$ and is strongly union-free. This proves the lower bound. ∎

The authors acknowledge the efforts of the referee to improve the presentation of the results in this paper.

References

[1] P. Frankl and Z. Füredi, “Union-free Hypergraphs and Probability Theory”, *European Journal of Combinatorics,* **5** (1984), pp. 127-131.

[2] M. L. Fredman, “The Complexity of Maintaining an Array and Computing its Partial Sums”, *JACM,* **29**(1982), pp. 250-260.

[3] M. L. Fredman and J. Komlós, “On the Size of Separating Systems and Families of Perfect Hash Functions”, *SIAM J. Alg. Disc. Meth.,* **5** (1984), pp. 61-68.

[4] J. B. Shearer, “A New Construction for Cancellative Sets”, *The Electronic Journal of Combinatorics* **3**(1996), # R15.

[5] J. Spencer, *Ten Lectures on the Probabilistic Method,* SIAM CBMS **52** (1987).