Combinatorics, Probability and Computing (2006) 00, 000–000.
DOI: 10.1017/S0000000000000000 Printed in the United Kingdom

On Cancellative Set Families

JÁNOS KÖRNER and BLERINA SINAIMERI

Department of Computer Science, Università "La Sapienza",
via Salaria 113, 00198 Roma, Italy
korner@di.uniroma1.it and sinaimeri@di.uniroma1.it

A family of subsets of an n-set is 2-cancellative if for every four-tuple $\{A, B, C, D\}$ of its members $A \cup B \cup C = A \cup B \cup D$ implies $C = D$. This generalizes the concept of cancellative set families, defined by the property that $A \cup B \neq A \cup C$ for A, B, C all different. The asymptotics of the maximum size of cancellative families of subsets of an n-set is known, (Tolhuizen [7]). We provide a new upper bound on the size of 2-cancellative families improving the previous bound of $2^{0.458n}$ to $2^{0.42n}$.

1. Introduction

Extremal set theory or extremal hypergraph theory is an important and highly applicable part of combinatorics, [6]. Although problems of this kind have been studied systematically, almost none of them has been solved even in an asymptotic sense. The only notable exception is Tolhuizen's solution of an old problem of Erdős and Katona regarding the maximum size of cancellative set families. A family $\mathcal{F}$ of subsets of a ground set of n elements is *cancellative* if for any three of its elements $A \cup B = A \cup C$ implies $B = C$. Frankl and Füredi [4] have introduced an upper bound for the maximum cardinality of cancellative set families as a function of n. Tolhuizen, disproving the original conjecture of Erdős and Katona, showed in [7] that the upper bound of [4] is tight in the sense of exponential asymptotics. Given the exceptional status of this problem among similar ones about families with excluded triples of subsets, inasmuch as no other problem of a similar kind has been solved, it is reasonable to look for an analogous property concerning excluded four-tuples, in the hope that this might give rise to the "easiest" of all such problems for four-tuples. In what follows, we define what we claim is the needed natural generalization of the cancellative property. The corresponding problem has not been treated in the literature neither explicitly nor implicitly, even though corresponding bounds can be derived from those for similar properties. In particular, we will improve on the upper bound obtainable from the one in [3] for a similar but weaker property.

Let $\mathcal{F}$ be a family of subsets of an n-set with the property that $A \cup B \cup C = A \cup B \cup D$ implies $C = D$ for any four of its sets A, B, C, D. We call such a family 2-*cancellative*. Let

$M(n)$ be the maximum size of such a family. We would like to determine the exponential asymptotics

$$t(4) = \limsup_{n\to+\infty} \frac{1}{n} \log M(n).$$

As usual all logarithms have base two. We define the weight of a binary vector $\mathbf{x}$ of length n by $w(\mathbf{x}) = \sum_{i=0}^{n} x_i$, and denote by h the binary entropy function

$$h(p) = -p \log p - (1-p) \log (1-p), \text{ where } 0 \leq p \leq 1.$$

Without loss of generality we can suppose that the n-set underlying $\mathcal{F}$ is $[n] = \{1, \ldots, n\}$. We associate to every subset A in the family, its characteristic binary vector, $\mathbf{x} = x_1 \ldots x_n$, with $x_i = 1$ if $i \in A$ and $x_i = 0$ else. One can immediately see that requiring the family $\mathcal{F}$ to have the desired property is equivalent for its representation set of binary vectors, F, to satisfy the following: for every four-tuple $\{\mathbf{w}, \mathbf{x}, \mathbf{y}, \mathbf{z}\}$ of distinct vectors in the set, considered in an arbitrary but fixed order $(\mathbf{w}, \mathbf{x}, \mathbf{y}, \mathbf{z})$ there exist at least three different values of $k \in [n]$, such that the corresponding ordered quadruples (w_k, x_k, y_k, z_k) are all different while for each of them we have that $w_k + x_k + y_k + z_k = 1$ (we will sometimes refer to $w_k x_k y_k z_k$ as the k'th column of the ordered four-tuple $(\mathbf{w}, \mathbf{x}, \mathbf{y}, \mathbf{z})$).

This problem can be seen in a more general context. We can require that for every ordered four-tuple of vectors in the set, there exists at least t different columns, $k_1, \ldots, k_t$ which sum up to 1 (obviously $t \in [4]$). In the case $t = 4$ we get back a well-known problem, the one about the largest cardinality of 3-*cover-free* families (see, for example [5], [6]). Instead, the case of $t = 1$ corresponds to 4-locally thin sets originally introduced by Alon *et al.* [1]. The best upper bound known for this problem comes from [3]. Unfortunately nothing more is known for the case $t = 2$ or $t = 3$ (which is the case we are considering). The best upper bound known for both these problems remains the bound on 4-locally thin sets. So, it follows from the main result of [3], that $t(4) < 0.4561$. Here we will improve this result by showing that $t(4) < 0.42$.

2. The Main Result

The following theorem is the main result of this article.

Theorem 1.

$$0.11 < t(4) \leq 0.42$$

Proof: In order to prove the lower bound, we first reformulate the 2-cancellative property. Consider the ordered four-tuple of different binary vectors $(\mathbf{w}, \mathbf{x}, \mathbf{y}, \mathbf{z})$ of length n. We require that the following holds

$$\exists i \in [n] \quad \text{such that} \quad (\{w_i, x_i\}, \{y_i, z_i\}) = (\{0,0\}, \{0,1\}). \tag{2.1}$$

For the sake of brevity we say that the underlying set $\{\mathbf{w}, \mathbf{x}, \mathbf{y}, \mathbf{z}\}$ has the *critical* property if the corresponding relation (2.1) holds for any ordering of its vectors. Observe

that once the set of four distinct vectors is fixed, the configuration of ordered pairs of unordered couples that we introduced is uniquely determined by the first couple of the pair. Therefore, given an arbitrary set $T = \{\mathbf{w}, \mathbf{x}, \mathbf{y}, \mathbf{z}\}$ of distinct binary vectors, we will refer to each of the configurations it generates, by the the first couple $A \in \binom{T}{2}$ of the ordered pair. It can be seen that T has the critical property if and only if it is 2-cancellative. The "if" part is obvious. If we have three different columns of length 4 and weight 1, then, necessarily, for every couple $A \in \binom{T}{2}$, one of these vectors has a 0 in both positions defined by the vectors of A. On the other hand, if we have a set of columns such that for every A at least one of them satisfies (2.1), then they must contain at least three different columns of length 4 and weight 1; in fact, if there were only two of them, then their positions of 1's would define a couple A for which the relation (2.1) is not satisfied. Thus, we can say that a set F of binary vectors is 2-cancellative if every four-tuple of its vectors $\{\mathbf{w}, \mathbf{x}, \mathbf{y}, \mathbf{z}\}$ has the critical property.

The lower bound is obtained by a standard random choice argument. Let F be a random collection of N binary vectors of length n, constructed by letting each coordinate of each member of F be, randomly and independently 1 with probability p and 0 with probability $1-p$ for a p that will be chosen later. Now, the expected number of configurations in F which don't have the critical property is $\binom{N}{4}\binom{4}{2}[1-2p(1-p)^3]^n$. By deleting an arbitrarily chosen string from each of these forbidden configurations we obtain a set F of vectors having the 2-cancellative property, of cardinality

$$|F| \geq N - 6\binom{N}{4}[1-2p(1-p)^3]^n.$$

Choosing $p = 1/4$, $N = \lfloor (101/128)^{n/3} \rfloor$ and recalling that $M(n) \geq |F|$, it follows that

$$t(4) = \limsup_{n \to +\infty} \frac{1}{n} \log M(n) \geq \frac{1}{3}(7 - \log 101) > 0.11$$

as claimed.

We next give the proof of the upper bound.

Consider a 2-cancellative set of binary vectors of length n, with the additional property that all its members have the same weight. Let $N(n)$ be the maximum cardinality of such a set. It is clear that $N(n)$ and $M(n)$ have the same exponential asymptotics, as we have that $N(n) \leq M(n) \leq (n+1)N(n)$. As we are only interested in the asymptotic behavior of $M(n)$, we can restrict ourselves to the case of sets in which each of the member vectors has the same weight. Let $F = F_n$ be such a set, which achieves maximum cardinality. In other words, the vectors in F have the same weight equal to np for some p, $0 \leq p \leq 1$. In order to proceed with the proof we next give some definitions:

Given a vector $\mathbf{x}$, its *projection* onto the set of coordinates $I = \{i_1, ...i_m\}$, (with the natural ordering $i_1 <, \ldots, < i_m$), is the vector $\mathbf{x}|_I = x_{i_1}...x_{i_m}$.

For every $\mathbf{x} \in F$ define the set of all the coordinates where $\mathbf{x}$ has a zero

$$I_{\mathbf{x}} = \{i : i \in [n]; x_i = 0\}.$$

Obviously, $|I_{\mathbf{x}}| = (1-p)n$.
For all $\mathbf{x} \in F$ define the projection of F onto the set of coordinates $I_{\mathbf{x}}$ as follows:

$$F^{\mathbf{x}} = \{\mathbf{y} : \mathbf{y} \in \{0,1\}^{(1-p)n}; \exists \mathbf{z} \in F \setminus \{\mathbf{x}\}; \mathbf{y} = \mathbf{z}|_{I_{\mathbf{x}}}\}.$$

Now let us fix $\mathbf{x} \in F$ and consider $F^{\mathbf{x}}$. The next observations follow directly from the 2-cancellative property:

1 $|F| = |F^{\mathbf{x}}| + 1$.
2 For every three vectors $\mathbf{w}, \mathbf{y}, \mathbf{z}$ in the set $F^{\mathbf{x}}$ there exist at least two integer values $k \in [0, n(1-p)]$, such that the ordered triples (w_k, y_k, z_k) are all different and $w_k + y_k + z_k = 1$.

Indeed, let $(\mathbf{w}, \mathbf{x}, \mathbf{y}, \mathbf{z})$ be an arbitrary four-tuple of distinct vectors in the set F, where $\mathbf{x}$ is the fixed element of F. We know that F is a 2-cancellative family. Hence there exist at least three different values of k, such that the ordered quadruples (w_k, x_k, y_k, z_k) are all different and $w_k + x_k + y_k + z_k = 1$. This implies that $x_k = 0$ for at least two different values of $k \in [n]$; observe that this forces k to belong to $I_{\mathbf{x}}$. Furthermore, in correspondence to these values of k we have at least two different triples (w_k, y_k, z_k) such that $w_k + y_k + z_k = 1$. Now, recalling that $k \in I_{\mathbf{x}}$ and considering the projections $w|_{I_{\mathbf{x}}}, y|_{I_{\mathbf{x}}}, z|_{I_{\mathbf{x}}}$, it is easy to verify the properties we claimed.
We can view $F^{\mathbf{x}}$ as the set of characteristic vectors of a family $\mathcal{F}^{\mathbf{x}}$ of subsets of an $n(1-p)$-set. In the terminology of Frankl and Füredi [4], $\mathcal{F}^{\mathbf{x}}$ is a *cancellative* family. In [4], they estimated the size of the largest cancellative family $\mathcal{F}_k$, consisting of k-element subsets of $[n]$. They proved that if $k \leq n/2$ then $|\mathcal{F}_k| \leq \binom{n}{k} 2^k / \binom{2k}{k}$.
Thus,

$$|F^{\mathbf{x}}| = |\mathcal{F}^{\mathbf{x}}| \leq \sum_{k=0}^{\hat{k}} \binom{\hat{n}}{k} 2^k / \binom{2k}{k}, \tag{2.2}$$

where $\hat{n} = n(1-p)$ is the length of the binary vectors of $F^{\mathbf{x}}$ and $\hat{k}$ is their maximum weight. The following claim is at the base of our proof.

Claim 1. *Let F be the set of characteristic vectors of a 2-cancellative family $\mathcal{F}$ of subsets of cardinality np, (for some $0 < p \leq 1$), of an n-set. Let ε be an arbitrary nonnegative constant, to be specified later. There is a binary vector $\boldsymbol{x} \in F$ and a constant $\gamma = \gamma_{(\varepsilon,p)}$, $(0 < \gamma_{(\varepsilon,p)} \leq 1)$, such that at least $\gamma|F^{\boldsymbol{x}}| - 1$ vectors in $F^{\boldsymbol{x}}$ have at most a weight $n(1-p)(p+\varepsilon)$.*

Proof: For each coordinate j, $(1 \leq j \leq n)$, denote by $c_j = \sum_{\mathbf{y} \in F} y_j$, the sum of the projections of all the vectors of F, to the j'th coordinate. With this notation, the overall weight of the projections of the vectors of F onto every possible set of coordinates $I_{\mathbf{x}}$, $(\mathbf{x} \in F)$, can be expressed as

$$\sum_{\mathbf{x} \in F} \sum_{\mathbf{y} \in F^{\mathbf{x}}} w(\mathbf{y}) = \sum_{\mathbf{x} \in F} \sum_{j \in I_{\mathbf{x}}} c_j = \sum_{j=1}^{n} c_j(|F| - c_j). \tag{2.3}$$

The first equality is simply obtained by "double counting" the number of coordinates equal to one in the set F (observe that x_j is also considered when calculating c_j, but its contribution to the sum is zero, as $x_j = 0$ when $j \in I_{\mathbf{x}}$). The second equality follows by observing that each c_j appears in the precedent sum in correspondence with all the vectors $\mathbf{x} \in F$ for which $j \in I_{\mathbf{x}}$, in other words, for all the vectors $\mathbf{x}$ such that $x_j = 0$. Obviously, the number of vectors having zero at their j'th coordinate is $|F| - c_j$. Now, by a simple application of Jensen's inequality to the function x^2, we can upper bound the right-most end of (2.3) by

$$\sum_{j=1}^{n} c_j|F| - \sum_{j=1}^{n} {c_j}^2 \leq \sum_{j=1}^{n} c_j|F| - \frac{1}{n}\Big(\sum_{j=1}^{n} c_j\Big)^2. \tag{2.4}$$

Again, a simple "double counting" argument shows that

$$\sum_{j=1}^{n} c_j = \sum_{\mathbf{y}\in F} w(\mathbf{y}) = |F|pn,$$

whence we can rewrite the last expression of (2.4) as

$$|F|^2np - \frac{1}{n}|F|^2n^2p^2 = |F|^2n(1-p)p. \tag{2.5}$$

Hence, recalling that $|F| = |F^{\mathbf{x}}| + 1$, the relations (2.3)-(2.5) give the upper bound

$$\frac{1}{|F|}\sum_{\mathbf{x}\in F}\sum_{\mathbf{y}\in F^{\mathbf{x}}} w(\mathbf{y}) \leq (|F^{\mathbf{x}}| + 1)n(1-p)p.$$

This immediately shows the existence of an $\mathbf{x} \in F$ such that

$$\sum_{\mathbf{y}\in F^{\mathbf{x}}} w(\mathbf{y}) \leq (|F^{\mathbf{x}}| + 1)n(1-p)p. \tag{2.6}$$

Given such an $\mathbf{x}$, we will show that a constant fraction of the vectors of $F^{\mathbf{x}}$ has a weight smaller than $(1-p)n(p+\varepsilon)$. Observing that a vector $\mathbf{y}$ in $F^{\mathbf{x}}$ has a length equal to $n(1-p)$, we can define its density $p_{\mathbf{y}}, (0 < p_{\mathbf{y}} \leq 1)$ as

$$p_{\mathbf{y}} = \frac{w(\mathbf{y})}{n(1-p)}.$$

Denote by $F_1^{\mathbf{x}}$ the set of all the vectors $\mathbf{y}$ of $F^{\mathbf{x}}$ for which we have $p_{\mathbf{y}} < p + \varepsilon$ and by $F_2^{\mathbf{x}}$ its complement in $F^{\mathbf{x}}$, i.e., $\mathbf{y} \in F_2^{\mathbf{x}}$ whenever $p_{\mathbf{y}} \geq p + \varepsilon$. Thus,

$$\sum_{\mathbf{y}\in F^{\mathbf{x}}} w(\mathbf{y}) \geq |F_2^{\mathbf{x}}|n(1-p)(p+\varepsilon).$$

Combining this with the inequality (2.6), we get

$$|F_2^{\mathbf{x}}| \leq \frac{p}{p+\varepsilon}(|F^{\mathbf{x}}| + 1).$$

Recalling that $|F_2^{\mathbf{x}}| = |F^{\mathbf{x}}| - |F_1^{\mathbf{x}}|$ we obtain

$$|F_1^{\mathbf{x}}| \geq \frac{\varepsilon}{p+\varepsilon}|F^{\mathbf{x}}| - \frac{p}{p+\varepsilon}. \tag{2.7}$$

This bound shows that there is a constant $\gamma = \dfrac{\varepsilon}{p+\varepsilon}$, $(0 < \gamma_{(\varepsilon,p)} \leq 1)$, such that at least $\gamma|F^{\mathbf{x}}| - 1$ vectors in $F^{\mathbf{x}}$ have a weight at most $n(1-p)(p+\varepsilon)$. This completes the proof of the claim.

Now we are ready to prove the theorem. Applying the previous result to our set F, for every fixed $\varepsilon > 0$ we can find an $\mathbf{x} \in F$ and a γ, such that at least a γ-fraction of the vectors of $F^{\mathbf{x}}$, has a weight smaller than $\widehat{k} = n(1-p)(p+\varepsilon)$. Combining this with the Frankl-Füredi bound [4] we have for $0 < p \leq 1/2$ that

$$\gamma|F^{\mathbf{x}}| - 1 \leq |F_1^{\mathbf{x}}| \leq \sum_{k=0}^{\hat{k}} \binom{\hat{n}}{k} 2^k / \binom{2k}{k}, \tag{2.8}$$

where $\gamma = \dfrac{\varepsilon}{p+\varepsilon}$, and since we are considering vectors in $F^{\mathbf{x}}$, $\hat{n} = (1-p)n$, therefore $\hat{k} = \hat{n}(p+\varepsilon) = n(1-p)(p+\varepsilon)$. We need the following exponential bound for binomial coefficients, (see, *e.g*, Section 1.2 in Csiszár and Körner [2])

$$\binom{\hat{n}}{k} \leq \exp_2\left(\hat{n}h\left(\frac{k}{\hat{n}}\right)\right) \quad , \quad \binom{2k}{k} \geq \frac{1}{2k+1}\exp_2(2k) \tag{2.9}$$

Combining (2.8) and (2.9) we have

$$\gamma|F^{\mathbf{x}}| \leq (\hat{k}+1)\max_{k\leq\hat{k}} \binom{\hat{n}}{k} \exp_2(k) / \binom{2k}{k} \tag{2.10}$$

$$\leq \hat{n}\exp_2\left(\hat{n}\left(\max_{k\leq\hat{k}}\left(h\left(\frac{k}{\hat{n}}\right) - \frac{k}{\hat{n}} + \frac{\log\hat{n}}{\hat{n}}\right)\right)\right). \tag{2.11}$$

Thus,

$$\begin{aligned}\frac{1}{n}\log|F^{\mathbf{x}}| \leq \frac{1}{n}\log\left(\frac{n(1-p)}{\gamma}\right) \\ + (1-p)\max_{k\leq\hat{k}}\left(h\left(\frac{k}{\hat{n}}\right) - \frac{k}{\hat{n}} + \frac{\log\hat{n}}{\hat{n}}\right).\end{aligned} \tag{2.12}$$

Now, set $q = \dfrac{k}{\hat{n}}$ and rewrite the right hand side of (2.12) as

$$= \frac{1}{n}\log\left(\frac{n(1-p)}{\gamma}\right) + (1-p)\max_{q\leq p+\varepsilon}\left(h(q) - q + \frac{\log\hat{n}}{\hat{n}}\right). \tag{2.13}$$

Thus,

$$t(4) = \limsup_{n\to+\infty} \frac{1}{n} \log M(n) \tag{2.14}$$

$$\leq \limsup_{n\to+\infty} \frac{1}{n} \log\big((n+1)N(n)\big) \tag{2.15}$$

$$\leq \limsup_{n\to+\infty} \Bigg(\frac{1}{n}\log(n+1) + \frac{1}{n}\log\left(\frac{n(1-p)}{\gamma}\right) + (1-p)\max_{q\leq p+\varepsilon}\left(h(q) - q + \frac{\log \hat{n}}{\hat{n}}\right)\Bigg). \tag{2.16}$$

Whence it follows that

$$t(4) \leq \max_{0\leq p\leq \frac{1}{2}} (1-p) \max_{q\leq p+\varepsilon} \Big(h(q) - q\Big). \tag{2.17}$$

Now, since $h(q) - q$ is monotonically increasing in $[0, 1/3]$ and monotonically decreasing elsewhere, we have for every $\varepsilon > 0$

$$\max_{0\leq p\leq \frac{1}{3}-\varepsilon} (1-p) \max_{q\leq p+\varepsilon} \Big(h(q) - q\Big) \leq \max_{p\leq \frac{1}{3}-\varepsilon} (1-p)\Big(h(p) - p\Big) \tag{2.18}$$

and

$$\max_{\frac{1}{3}-\varepsilon<p\leq \frac{1}{2}} (1-p) \max_{q\leq p+\varepsilon} \Big(h(q) - q\Big) \leq \max_{\frac{1}{3}-\varepsilon<p\leq \frac{1}{2}} (1-p)\left(h\left(\frac{1}{3}\right) - \frac{1}{3}\right). \tag{2.19}$$

In conclusion, (2.17), (2.18) and (2.19) can be summarized as

$$\begin{aligned} t(4) &\leq \max\left\{ \max_{p\leq \frac{1}{3}-\varepsilon} (1-p)(h(p)-p), \max_{\frac{1}{3}-\varepsilon<p\leq 1/2} (1-p)\left(h\left(\frac{1}{3}\right) - \frac{1}{3}\right)\right\} \\ &= \max\left\{ \max_{p\leq \frac{1}{3}-\varepsilon} (1-p)(h(p)-p), \max_{\frac{1}{3}-\varepsilon<p\leq 1/2} (1-p)(\log 3 - 1)\right\}. \end{aligned}$$

Choosing $\varepsilon = 0.01$ we obtain

$$t(4) \leq \max\big\{0.42, 0.4\big\} = 0.42$$

as claimed.

□

As already mentioned in the introduction, the asymptotic version of the Frankl–Füredi upper bound [4] for cancellative set families is tight, as proved by Tolhuizen [7]. This is the only non–trivial problem in extremal set theory where for an excluded configuration of size greater than two, the exact exponential asymptotics is known. Unfortunately, in our generalization there remains a huge gap between the two bounds on the exponent.

References

[1] Alon, N., Fachini, E. and Körner, J. (2000) Locally thin set families *Combinatorics, Prob. Computing* **6** 481–488.

[2] Csiszár, I. and Körner, J. (1982) *Information Theory : Coding Theorems for Discrete Memoryless Systems*, Academic Press Inc, Orlando, FL, USA.
[3] Fachini, E., Körner, J. and Monti, A. (2001) A Better Bound for Locally Thin Set Families *J. Combin. Theory Ser. A* **95** 209–218.
[4] Frankl, P. and Füredi, Z. (1984) Union-free Hypergraphs and Probability Theory *Europ. J. Combinatorics* **5** 127–131.
[5] Füredi, Z. (1996) On r-cover-free families *J. Combin. Theory Ser. A* **73** 172–173.
[6] Jukna, S. (2000) *Extremal Combinatorics With Applications in Computer Science*, Springer, Berlin.
[7] Tolhuizen, L. (2000) New rate pairs in the zero-error capacity region of the binary multiplying channel without feedback *IEEE Trans. Inform. Theory* **46** 1043–1046.