LOCALLY THIN SET FAMILIES

Noga Alon *
noga@math.tau.ac.il
Tel–Aviv University
ISRAEL

Emanuela Fachini
fachini@dsi.uniroma1.it
"La Sapienza" University of Rome
ITALY

János Körner
korner@dsi.uniroma1.it
"La Sapienza" University of Rome
ITALY

February 22, 2002

Abstract

A family of subsets of an n–set is k–locally thin if for every k of its member sets the ground set has at least one element contained in exactly 1 of them. We derive new asymptotic upper bounds for the maximum cardinality of locally thin set families for every even k. This improves on previous results of two of us with Monti.

*Research supported in part by a USA Israeli BSF grant and by a grant from the Israel Science Foundation.

1 Introduction

Let $\mathcal{F}$ be a family of subsets of a ground set of n elements. As usual, we can suppose without loss of generality that our ground set is $[n] = \{1, 2, \ldots, n\}$. We say that the family is k– locally thin if for any k of its distinct member sets at least one point $i \in [n]$ of the ground set is contained in exactly one of them. We say further that a set family is thin if it is k–locally thin for every $k \leq |\mathcal{F}|$. We are interested in finding out how large these families can get. It is quite obvious at the outset that a thin family of subsets of $[n]$ can have at most n members and this is achieved by the family of all the subsets of n having a single element. Let $M(n, k)$ denote the maximum cardinality of a k–locally thin family of subsets of a ground set of n elements. We are interested in the exponential asymptotics of $M(n, k)$ for arbitrary but fixed k. More precisely, we would like to determine the sequence

$$t(k) = \limsup_{n\to\infty} \frac{1}{n} \log M(n, k) \tag{1}$$

(All the exp's and log's are binary.) The above is a notoriously hard task. In particular, one doesn't even know whether $t(3) < 1$. This is one of the famous open questions about strong Δ–systems, (cf. [3] and the recent survey [5].) Another startling lack of our knowledge is not to be able to decide whether $t(k)$ is monotonic in k. In this paper we will present improvements over previously known bounds for every even value of k.

Our starting point is a recent paper of Alon, Körner and Monti in which the authors prove that

Theorem AKM([1], Theorem 2)

$$\frac{1}{3}(6 - \log 37) \leq t(4) < \frac{1}{2}$$

while in general,

Theorem AKM([1], Theorem 3)

$$t(k) \leq \frac{\log 3}{2} \quad for\ every \quad k \geq 5$$

Our main objective here is to improve these bounds for even values of k. The two prerequisites for our proof are the information–theoretic bounding technique of [7] based on graph entropies and a theorem of Frankl and Füredi [4]. To make the present paper self–contained we recall the basic information–theoretic definitions needed. Graph entropy $H(G, P)$ is an information–theoretic functional of a graph G with a probability distribution P on its vertex set, introduced in Körner [6]. It is usually defined as

$$H(G, P) = \min_{X \in Y \in S(G),\, P_X = P} I(X \wedge Y),$$

where $S(G)$ denotes the family of the stable sets of vertices in G. (A subset of the vertex set is called stable if it does not contain any edge. For the basics in information theory

the reader is referred to the book [2]. We recall that the mutual information $I(X \wedge Y)$ of the random variables X and Y equals $H(X) + H(Y) - H(X,Y)$, where e. g. $H(X,Y)$ is the entropy of the random variable (X,Y). Notice that the entropy of a random variable is the entropy of its distribution.) A crucial property of $H(G,P)$ needed in our proof is its sub–additivity with respect to graph union [7]; if F and G are two graphs on the same vertex set V and $F \cup G$ denotes the graph on V with edge set $E(F \cup G) = E(F) \cup E(G)$, then for every P we have

$$H(F \cup G, P) \leq H(F,P) + H(G,P). \tag{2}$$

A straightforward consequence of the definition is the lower bound (cf. [7])

$$H(P) - \log \alpha(G) \leq H(G,P), \tag{3}$$

where $\alpha(G)$, the stability number of the graph G, is the maximum cardinality of a stable set in the graph. We shall use the notation $h(t) = -t \log t - (1-t) \log t$ for the binary entropy function.

We will use the following beautiful theorem.

Theorem FF([4])

Let X be an n-set and let $\mathcal{G}$ be a family of k-element subsets of X. Given nonnegative integers l and l' such that $l + l' < k$, we say that $\mathcal{G}$ is an (n,k,l,l')-system if, for all distinct pairs $F, F' \in \mathcal{G}$, either $|F \cap F'| < l$ or $|F \cap F'| \geq k - l'$. Let $m(n,k,l,l')$ be the maximum cardinality of such a system. There exists a positive constant d_k such that $m(n,k,l,l') < d_k n^{\max\{l,l'\}}$ holds.

If k is even then, taking $l = k/2$, $l' = k/2 - 1$ we conclude:

Corollary FF

For every even k there exists a positive constant d_k such that the maximum cardinality of a family of k-subsets of an m-element set in which no two members intersect in precisely $k/2$ elements is at most $d_k m^{k/2}$.

These are the prerequisites to our proof. The interested reader can find a gentler introduction to graph entropy in the survey of Simonyi [9].

2 Locally thin families

Our goal in this section is to prove

Theorem 1 *For every even $k > 2$ we have*

$$t(k) \leq \frac{2}{k} \max_{p \in [0,1]} \left[1 - \sum_{j=k/2+2}^{k} \binom{k}{j} p^j (1-p)^{k-j} \right] h\left(\frac{\sum_{t=0}^{\lceil k/4 \rceil} \binom{k}{2t} p^{2t} (1-p)^{k-2t}}{1 - \sum_{j=k/2+2}^{k} \binom{k}{j} p^j (1-p)^{k-j}} \right) \tag{4}$$

Proof.

Let us fix a (large) n and an arbitrary family $\mathcal{G}$ achieving $M(n,k)$ with $k=2l$. We consider the graph G whose vertices are all the k–tuples of different sets from the family $\mathcal{G}$, i. e., we define $V(G)=\binom{\mathcal{G}}{k}$. Given an element $i\in[n]$ and a family $A\subseteq 2^{[n]}$ of subsets of $[n]$ we denote by $\chi(i,A)$ the number of those sets in the family A that contain i. We draw an edge between two vertices, A and B in G if

$$|A\cap B|=l \tag{5}$$

By hypothesis, since the symmetric difference of such sets, $A\triangle B$, has cardinality k, there exists at least one point $i\in[n]$ for which

$$\chi(i,A\triangle B)=1 \tag{6}$$

We define, for every $i\in[n]$ the graph G_i by setting $V(G_i)=V(G)$ and by drawing an edge between two vertices A and B of G_i if they are adjacent in G and furthermore, the point $i\in[n]$ satisfies (6). This implies that each of the n graphs G_i is a subgraph of G. Further, more importantly, we have

$$G=\bigcup_{i=1}^{n}G_i \tag{7}$$

Then, by virtue of the sub–additivity (2) of graph entropy for the equidistribution P on $V(G)$ the graph relation (7) implies

$$H(G,P)\leq\sum_{i=1}^{n}H(G_i,P). \tag{8}$$

Let k_i be the number of different sets in $\mathcal{G}$ that contain $i\in[n]$. We set

$$p_i=\frac{k_i}{n}$$

Let us consider the graph F with vertex set $\{0,1,\ldots,k\}$ and edge set consisting of the unordered pairs $\{j,j+1\}$ of vertices for every j with $0\leq j\leq\frac{k}{2}$. Consider further the function $f_i:V(G)\to V(F)$ defined by setting

$$f_i(A)=\chi(i,A)$$

for every $A\in V(G)$. By our hypothesis, if $\{A,B\}\in E(G_i)$, we must have

$$|\chi(i,A)-\chi(i,B)|=1$$

and

$$\chi(i,A)\leq\frac{k}{2}+1,\qquad\chi(i,B)\leq\frac{k}{2}+1$$

implying that f_i acts on $\binom{V(G)}{2}$ in an edge–preserving manner. Let Q_i denote the probability distribution generated on $V(F)$ by f_i and P. Clearly, the $\frac{k}{2}+2$ non–isolated vertices of the graph F form a path and the rest consists of $\frac{k}{2}-1$ isolated points. We have, (up to an asymptotically negligible correcting factor compensating the effect of not allowing for repetitions in forming the k–sets),

$$Q_i(j) = \binom{k}{j} p_i^j (1-p_i)^{k-j} \tag{9}$$

In this graph the total probability of the non–isolated points is

$$1 - \sum_{j=k/2+2}^{k} \binom{k}{j} p_i^j (1-p_i)^{k-j}$$

These constitute a unique path along which they have an alternating parity and therefore can be partitioned into two stable sets, one consisting of odd vertices and the other consisting of the even ones. Hence, as an easy consequence of the very definition of graph entropy (or else cf. [7] for more detail), we obtain

$$H(G_i, P) \leq \left[1 - \sum_{j=k/2+2}^{k} \binom{k}{j} p_i^j (1-p_i)^{k-j}\right] h\left(\frac{\sum_{t=0}^{\lceil k/4 \rceil} \binom{k}{2t} p_i^{2t} (1-p_i)^{k-2t}}{1 - \sum_{j=k/2+2}^{k} \binom{k}{j} p_i^j (1-p_i)^{k-j}}\right) \tag{10}$$

Thus, (8) and (10) imply

$$H(G, P) \leq n \max_{p \in [0,1]} \left[1 - \sum_{j=k/2+2}^{k} \binom{k}{j} p^j (1-p)^{k-j}\right] h\left(\frac{\sum_{t=0}^{\lceil k/4 \rceil} \binom{k}{2t} p^{2t} (1-p)^{k-2t}}{1 - \sum_{j=k/2+2}^{k} \binom{k}{j} p^j (1-p)^{k-j}}\right) \tag{11}$$

We turn to lower bounding the entropy $H(G, P)$. In the previous development we have kept n fixed. Let $\mathcal{G}(n)$ now stand for the family hitherto denoted simply by $\mathcal{G}$, and let $G(n)$ stand for the corresponding graph, etc. Let $S(n)$ be a stable set of $G(n)$ having maximum size $\alpha(G(n))$. Clearly, the vertices of $S(n)$ form a family of k-element subfamilies of $\mathcal{G}$ satisfying the condition of Corollary FF, and thus we have

$$|S(n)| \leq d_k |\mathcal{G}|^{k/2} \tag{12}$$

for some positive constant d_k. Substituting the last inequality into (3) and observing that P is the equidistribution on $V(G)$, we obtain

$$H(G(n), P) \geq \log |V(G(n))| - \log |S(n)| =$$

$$= \log \binom{|\mathcal{G}|}{k} - \frac{k}{2} \log |\mathcal{G}| - \log d_k$$

implying that

$$\frac{1}{n}H(G,P) \geq \frac{k}{2n}\log|\mathcal{G}| - \epsilon_n(k) \tag{13}$$

for some $\epsilon_n(k) \to 0$. Comparing this with (11) results in

$$\frac{k}{2n}\log|\mathcal{G}| \leq \max_{p\in[0,1]}\left[1-\sum_{j=k/2+2}^{k}\binom{k}{j}p^j(1-p)^{k-j}\right] h\left(\frac{\sum_{t=0}^{\lceil k/4\rceil}\binom{k}{2t}p^{2t}(1-p)^{k-2t}}{1-\sum_{j=k/2+2}^{k}\binom{k}{j}p^j(1-p)^{k-j}}\right)+\epsilon_n(k)$$

whence we conclude that $\frac{1}{n}\log M(n,k)$ has the upper bound

$$\frac{2}{k}\max_{p\in[0,1]}\left[1-\sum_{j=k/2+2}^{k}\binom{k}{j}p^j(1-p)^{k-j}\right] h\left(\frac{\sum_{t=0}^{\lceil k/4\rceil}\binom{k}{2t}p^{2t}(1-p)^{k-2t}}{1-\sum_{j=k/2+2}^{k}\binom{k}{j}p^j(1-p)^{k-j}}\right) + \epsilon_n(k) \tag{14}$$

which, passing to the limit in n concludes the proof.

□

3 Comments

Our theorem gives a complicated looking upper bound on $t(k)$. One immediately sees however that

Corollary 1 *For every even value of $k > 2$*

$$t(k) < \frac{2}{k}$$

In particular,

$$t(4) < 0.4968, \qquad t(6) < 0.3328$$

Proof.

Notice that $\left[1-\sum_{j=k/2+2}^{k}\binom{k}{j}p^j(1-p)^{k-j}\right] \leq 1$ with equality only if $p = 0$. In this case we have, however that

$$h\left(\frac{\sum_{t=0}^{\lceil k/4\rceil}\binom{k}{2t}p^{2t}(1-p)^{k-2t}}{1-\sum_{j=k/2+2}^{k}\binom{k}{j}p^j(1-p)^{k-j}}\right) = 0.$$

If $p > 0$, the latter expression, as a binary entropy, is still upper bounded by 1. For small values of k the maximum of the above function in p is easily calculated giving the claimed values.

□

Beyond that of determining the precise value of $t(k)$ many less specific open questions seem challenging. One of these is to decide whether $t(k)$ is monotonically decreasing in k. It is not hard to see that

Proposition 1

$$t(k+l) \leq \max\{t(k), t(l)\}$$

Proof.

Suppose, to the contrary, that $t(k+l) > \max\{t(k), t(l)\}$ and let $\mathcal{F}_n$ be a sequence of $k+l$–thin set families with $\mathcal{F} \subseteq 2^{[n]}$. Divide $\mathcal{F}_n$ into nearly equal parts, $\mathcal{F}_n(k)$ and $\mathcal{F}_n(l)$ so that both

$$\limsup_{n\to\infty} \frac{1}{n} \log |\mathcal{F}_n(k)| > \max\{t(k), t(l)\}$$

$$\limsup_{n\to\infty} \frac{1}{n} \log |\mathcal{F}_n(l)| > \max\{t(k), t(l)\}$$

This means that for n sufficiently large $\mathcal{F}_n(k)$ will contain k sets forming a subfamily which is not locally k–thin and likewise $\mathcal{F}_n(l)$ will have a subfamily of l sets which is not locally l–thin. Then, however, the union of these families will be a family of $k+l$ members which is not locally $k+l$–thin; a contradiction.

□

4 Further Bounds for $M(n,k)$

So far we have said nothing on $M(n,k)$ for odd values of k. We obtain the following asymptotic result on $t(k)$.

Theorem 2 *There are two absolute positive constants c_1, c_2 such that for every $k > 2$*

$$c_1 \frac{1}{k} \leq t(k) \leq c_2 \frac{\log k}{k}.$$

Proof. The lower bound is by a simple probabilistic construction, which is left to the reader. The upper bound for even values of k follows from the assertion of Theorem 1, and by choosing c_2 appropriately it is trivial for, say, $k < 9$. It thus remains to prove it for $k = 2s+3$, $s \geq 3$. It is convenient to define $M'(n,k)$ as the maximum length of a sequence of (not necessarily distinct) subsets of $[n]$, such that for every k members of the sequence there is an $i \in [n]$ that lies in exactly one of them, and to call a sequence of this type k-thin. Trivially $M'(n,k) \geq M(n,k)$. Therefore, to prove the desired result it suffices to prove the following.

Claim: Suppose $k = 2s+3$, $s \geq 3$, and let $\alpha \leq 1/2$ be a positive real satisfying

$$2^{-h(\alpha)s} < \frac{\alpha}{2}, \tag{15}$$

where $h(\alpha)$ is the binary entropy of α. Then, for all $n \geq 1$,

$$M'(n,k) < 2k \cdot 2^{h(\alpha)n}. \tag{16}$$

It is easy to check that there is an α satisfying (15) for which $\alpha = O(1/k)$, and hence the assertion of the claim implies the desired upper bound. We also note that the constant 2 in the right-hand-side of (15) can be improved, but we make no attempt to optimize the absolute constants here.

To prove the claim we apply induction on n. The result is trivial for $n = 1$, since every sequence of at least $2k$ subsets of $[1]$ contains either the empty set or the set $\{1\}$ itself k times, and hence cannot be k-thin. Let, now, $\mathcal{F}$ be a sequence of at least $2k \cdot 2^{h(\alpha)n}$ subsets of $[n]$, where $n > 1$, and suppose the assertion of the claim holds for $n - 1$. If there is some $i \in [n]$ that does not belong to at least $2^{-h(\alpha)}|\mathcal{F}|$ members of $\mathcal{F}$, then the collection of all these members is a sequence of at least $2k \cdot 2^{h(\alpha)(n-1)}$ subsets of an $(n-1)$-element set, and hence is not k-thin, by the induction hypothesis.

We can thus assume that for every $i \in [n]$ the number of members of $\mathcal{F}$ that do not contain i is smaller than $2^{-h(\alpha)}|\mathcal{F}|$. This implies that for every $i \in [n]$ the probability that the union of a set of s randomly chosen members of $\mathcal{F}$ does not contain i is at most $2^{-h(\alpha)s} < \alpha/2$. Let us choose, randomly, a collection of $m = 2\lceil 2^{h(\alpha)n} \rceil$ subsets $S_1, \ldots, S_m$ of $\mathcal{F}$, each consisting of precisely s members of $\mathcal{F}$, where the chosen collections are pairwise disjoint. (This can be done, for example, by randomly permuting all elements of $\mathcal{F}$, and by splitting the first sm elements in this permutation into m pairwise disjoint blocks, each consisting of s consecutive elements.) By linearity of expectation, for each fixed $1 \leq j \leq m$, the expected number of elements in $[n]$ that do not lie in the union $\cup_{F \in S_j} F$ is at most $\alpha n/2$. Therefore, using again linearity of expectation, there is a choice for the sets S_j such that at least $m/2$ of these unions are of cardinality at least $n - n\alpha$. Since the total number of subsets of cardinality at most αn in an n-element set is smaller than $m/2 = \lceil 2^{h(\alpha)n} \rceil$, this implies that there are two such unions that coincide. This gives a collection of $2s$ members of $\mathcal{F}$, such that each element in their union is covered at least twice, and their union is of size at least $n - \alpha n$. By the pigeonhole principle there are 3 additional members of $\mathcal{F}$ whose intersections with the complement of the above mentioned union are identical. These 3 members together with the previous $2s$ ones show that $\mathcal{F}$ is not k-thin and hence complete the proof.

□

5 Acknowledgement

Thanks are due to Ron Holzman for helpful criticism.

References

[1] N. Alon, J. Körner and A. Monti, String quartets in binary, *Combinatorics, Prob. Computing*, submitted,

[2] I. Csiszár, J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Academic Press, New York, 1982 and Akadémiai Kiadó, Budapest, 1981,

[3] P. Erdős, E. Szemerédi, Combinatorial properties of systems of sets, *JCT Ser. A*, **24**(1978), pp. 308–313,

[4] P. Frankl and Z. Füredi, Forbidding just one intersection, *JCT Ser. A*, **39**(1985), pp. 160–176,

[5] A. V. Kostochka (1998), Extremal problems on Δ-systems, manuscript

[6] J. Körner, Coding of an information source having ambiguous alphabet and the entropy of graphs. *Transactions of the 6th Prague conference on Information Theory, etc., 1971*, Academia, Prague, (1973), pp. 411–425,

[7] J. Körner, Fredman-Komlós bounds and information theory, *SIAM J. on Algebraic and Discrete Meth.*, 4(**7**), (1986), pp. 560–570,

[8] B. Lindström, Determination of two vectors from the sum, *JCT Ser. A*, **6**(1969), pp. 402–407,

[9] G. Simonyi, Graph entropy: a survey, in *Combinatorial Optimization*, DIMACS Series on Discrete Math. and Computer Science Vol. 20 (W. Cook, L. Lovász, P. D. Seymour eds.) pp. 399–441, (1995)