Journal of Combinatorial Theory, Series A **95**, 209–218 (2001)
doi:10.1006/jcta.2000.3162, available online at http://www.idealibrary.com on IDEAL®

A Better Bound for Locally Thin Set Families

Emanuela Fachini, János Körner, and Angelo Monti

Department of Computer Science, Università "La Sapienza,"
via Salaria 113, 00198 Rome, Italy
E-mail: fachini@dsi.uniroma1.it, korner@dsi.uniroma1.it, monti@dsi.uniroma1.it

Communicated by the Managing Editors

Received June 12, 2000; published online May 10, 2001

A family of subsets of an n-set is 4-locally thin if for every quadruple of its members the ground set has at least one element contained in exactly 1 of them. We show that such a family has at most $2^{0.4561n}$ members. This improves on our previous results with Noga Alon. The new proof is based on a more careful analysis of the self-similarity of the graph associated with such set families by the graph entropy bounding technique. © 2001 Academic Press

1. INTRODUCTION

Let $\mathscr{F}$ be a family of subsets of a ground set of n elements. We can suppose w.l.o.g. that our ground set is $[n]=\{1, 2, ..., n\}$. Following [1] we say that the family is 4-locally thin if for any quadruple of its distinct members at least one point $i \in [n]$ of the ground set is contained in exactly one of them. Let $M(n)$ denote the maximum cardinality of a 4-locally thin family of subsets of a ground set of n elements. We are interested in the exponential asymptotics of $M(n)$. More precisely, our aim is to sharpen the existing upper bounds on

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

(All the exp's and log's are binary.)

In two previous papers the present authors and Noga Alon [1, 2] have applied various versions of an information-theoretic bounding technique to obtain increasingly sharp upper bounds on $t(4)$. To make this paper self-contained, we recall some introductory material available in [2] and [1].

0097-3165/01 $35.00

As mentioned in [2], it follows from an old result of Lindström's [9] and is quite simple to see by elementary combinatorics that

$$t(4) \leqslant \tfrac{1}{2}.$$

To verify this, it is sufficient to note that if two couples of sets in a 4-locally thin set family have no member in common, then at least one point of the ground set belongs, in one of the couples, to exactly one member set, while the same point belongs to no member of the other couple. Likewise, if two couples from this family have one set in common, then there must be at least one point in the ground set belonging to exactly one of the two remaining members in the two couples, so that in conclusion, also in this case, the point in question belongs to an odd number of members of one of the couples and to an even number of members of the other couple. If we now associate with every couple of member sets the binary vector whose coordinate corresponding to point i of the ground set is 0 if i belongs to an even number of member sets of the couple and 1 otherwise, then we can see that the resulting binary strings associated with different couples of member sets must be different. This then implies that

$$\binom{M(n)}{2} \leqslant 2^n$$

which in turn gives the asserted inequality. It is not hard to see (cf. [2]) that this bound could be asymptotically tight only if the average cardinality of the member sets in the family were approximately 1/2. However, it was shown in [2] that such a choice cannot yield the asymptotic optimum, with the consequence that

$$t(4) < \tfrac{1}{2}.$$

The last bound was quantified in [1] to yield

$$t(4) < 0.4968.$$

While the second mentioned paper contains more general results of which the previous bound is just a special case, at present we are only concerned with set families with excluded 4-tuples. It is interesting to compare the last inequality with the Coppersmith–Shearer upper bound [3] on the maximum size of weakly union-free set families from the same ground set. Since a 4-locally thin set family is also weakly union-free, one would expect a better upper bound for our present problem. Yet the above upper bound from [1] is only very slightly less than the corresponding 0.5 in [3].

Our aim here is to prove the significantly better new bound

$$t(4) < 0.4561. \tag{2}$$

Other than yielding the above sharpening of the previous upper bound [1], our new proof has the additional advantage of not using any deeper results from extremal combinatorics.

To make the present paper self-contained we copy from [1] 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 by Körner [6]. It is 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. The random variable X takes its values in the vertex set of G, while the random variable Y is ranging over the stable sets of G. The condition $X \in Y \in S(G)$ means that the value of the variable X, a vertex, is an element of the value of the variable Y, a set of vertices. This condition is a restriction on the possible joint distributions of the random variables X and Y appearing in the above minimization. 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. Further, we will sometimes rewrite mutual information into the form

$$I(X \wedge Y) = H(X) - H(X \mid Y), \tag{3}$$

where $H(X \mid Y) = H(X, Y) - H(Y)$ is the conditional entropy of X given Y and can be expressed as the expected entropy of the conditional distributions of X given Y. For the basics in information theory the reader is referred to the book [4].)

A crucial property of $H(G, P)$ needed in our proof is its sub-additivity with respect to graph union [7]. Given two arbitrary graphs, F and G, their union $F \cup G$ is defined by setting

$$V(F \cup G) = V(F) \cup V(G) \qquad \text{and} \qquad E(F \cup G) = E(F) \cup E(G).$$

In these terms, the above sub-additivity means that if F and G have the same vertex set, then for every P we have

$$H(F \cup G, P) \leqslant H(F, P) + H(G, P). \tag{4}$$

We shall use the notation $h(t) = -t \log t - (1-t) \log t$ for the binary entropy function. The interested reader can find a more complete introduction to graph entropy in the survey of Simonyi [10].

2. THE MAIN RESULT

Our goal in this section is to prove

THEOREM 2.1.

$$t(4) \leqslant \frac{1}{2} \max_{p \in [0,1]} \left[(1-p^2)\, h\left(\frac{1-p}{1+p}\right)\right]$$

Proof. Let us fix an n and let $N(n)$ stand for the maximum cardinality of a 4-locally thin family of subsets of $[n]$, satisfying the additional constraint that all the member sets have the same cardinality. Then

$$N(n) \leqslant M(n) \leqslant (n+1)\, N(n),$$

whence the functions $N(n)$ and $M(n)$ have the same exponential asymptotics, and thus it is sufficient to establish our claim for set families whose members have the same cardinality. Let therefore $\mathcal{G} = \mathcal{G}_n$ be an arbitrary family achieving $N(n)$ and let $np = np_n$ be the common cardinality of its member sets. We consider the graph $G = G_n$ whose vertices are all the unordered couples of distinct sets from the family $\mathcal{G}$, i.e., we set $V(G) = \binom{\mathcal{G}}{2}$. The vertices $A \in \binom{\mathcal{G}}{2}$, $B \in \binom{\mathcal{G}}{2}$ are defined to be adjacent in G if there is a point $i \in [n]$ which is contained in exactly one of the four (not necessarily distinct) sets belonging to A and/or B. Further, let $P = P_n$ be the uniform probability distribution on $V(G_n)$.

We will derive appropriate lower and upper bounds on $H(G_n, P_n)$, the entropy of the graph G_n with respect to the distribution P_n. We begin with an upper bound on $H(G, P)$, taken from [2] and whose proof we reproduce here for completeness. Given any $i \in [n]$ let G^i be the graph having the same vertex set as G and an edge set $E(G^i) \subseteq E(G)$ defined by making the vertices $A \in \binom{\mathcal{G}}{2}$, $B \in \binom{\mathcal{G}}{2}$ adjacent in G^i if exactly one of the four (not necessarily distinct) sets belonging to either or both of A and B contains i. Since every couple of sets $\{A, B\} \in E(G)$ must satisfy this for at least one $i \in [n]$, we immediately see that

$$G \subseteq \bigcup_{i=1}^{n} G^i.$$

But then by the sub-additivity of graph entropy (4) we get

$$H(G, P) \leqslant \sum_{i=1}^{n} H(G^i, P). \tag{5}$$

Next, for every $i \in [n]$, let us denote by $p_i = p_i(n)$ the fraction of those elements of $\mathcal{G}$ which contain the point i. We claim to have

$$H(G^i, P) = (1 - p_i^2)\, h\left(\frac{1-p_i}{1+p_i}\right). \tag{6}$$

To see this, consider the graph F with vertex set $V(F) = \{0, 1, 2\}$ and the single edge $\{0, 1\}$. Observe that the function g_i: $\binom{\mathcal{G}}{2} \to \{0, 1, 2\}$ defined by setting $g_i(\{A, B\})$ equal to the number of sets among the members of A and B that contain i is acting on the vertices of G^i in an edge-preserving manner. Next consider the probability distribution P_i on $\{0, 1, 2\}$ defined by

$$P_i(t) = P(g_i^{-1}(t)) = \sum_{\{A, B\} \in \binom{\mathcal{G}}{2},\, g_i(\{A, B\}) = t} P(\{A, B\}).$$

Clearly, (in the limit of n going to infinity, we can neglect the effect of not allowing repetitions and thus) we are allowed to suppose that $P_i(0) = (1 - p_i)^2$, $P^i(2) = p_i^2$ and $P^i(1) = 2p_i(1 - p_i)$. Therefore, as an easy consequence of the above definition of graph entropy, one sees that

$$H(G^i, P) = H(F, P^i) = (1 - p_i^2)\, h\left(\frac{1-p_i}{1+p_i}\right).$$

Now, from (5) and (6) we get

$$H(G_n, P_n) \leqslant \sum_{i=1}^{n} (1 - p_i^2)\, h\left(\frac{1-p_i}{1+p_i}\right). \tag{7}$$

In order to lower bound $H(G_n, P_n)$, we first take a closer look at the structure of the stable sets in the graph G_n. We start by some heuristic discussion put in parenthesis to indicate that it is not part of the proof. (If the graph G_n were complete, this would immediately imply that its entropy is $\log\binom{|\mathcal{G}|}{2}$, and this, combined with the upper bound (7) on the entropy of G_n would imply the upper bound on $t(4)$ stated by our theorem. On the other hand, the completeness of the graph in question is equivalent to require that for every pair of couples of member sets in our family sharing a member, there be a point $i \in [n]$ contained in exactly one of the four members of the two couples, which, since the common member of the couples appears twice, means, in particular, that i does not belong to the common member. It is easy to realize that this is precisely the cancellative property for set families introduced by Frankl and Füredi in [5]. In what follows we shall prove what amounts to say, that "our optimal set family almost has the cancellative property". More precisely, we shall show that the entropy of the graph G_n is close enough to that of the complete graph so as to enable us to get the same bound under our true–and weaker-condition. It is worth mentioning that a similar phenomenon is at the basis of a recent, powerful upper bound on the size of weakly union–free set families by Coppersmith and Shearer [3].)

Let y be a stable set of vertices in the graph G_n. Since we know that such a set must be a family of pairwise intersecting couples of member sets of $\mathcal{G}$, therefore, as soon as it has more than 3 elements, there exists a set $A \in \mathcal{G}$ such that every vertex in the stable set y is a couple of distinct sets from $\mathcal{G}$ one of whose members is the fixed set A. On the other hand, consider the family of all the couples $\{A, B\}$ as B is running over all the member sets, excepting A of the family $\mathcal{G}$. We shall call B and C A-twins if the couples $\{A, B\}$ and $\{A, C\}$ are non-adjacent vertices in G_n. Let now B and C be arbitrary A-twins. Clearly,

$$\bar{A} \cap B = \bar{A} \cap C, \tag{8}$$

for else $\{A, B\} \in V(G)$ and $\{A, C\} \in V(G)$ would be adjacent vertices in G, contrary to our hypothesis. Since the relation between A-twin sets is an equivalence relation, and as B is running through any its classes, the corresponding pairs $\{A, B\}$ form a maximal stable set of G_n. For later reference, we shall call each of these stable sets an *A-substar* and denote the family they form by $\mathcal{P}(A)$.

Consider now two pairs, $\{B, C\}$ and $\{B', C'\}$ of A-twin sets. Then also

$$\bar{A} \cap B' = \bar{A} \cap C'.$$

We have to distinguish two cases. Either these 4 sets are all different, meaning that some element $i \in [n]$ must belong to exactly one of them. Then the last relation and (8) imply that such an i must necessarily be in A. In the remaining case, when only 3 of these 4 sets are distinct, and, say, $C = C'$, then, obviously, all the 3 couples induced by B, B' and C are A-twins, and thus the intersections $A \cap B$, $A \cap B'$ and $A \cap C$ must themselves be 3 different sets. In either case we conclude that

$$(A \cap B) \triangle (A \cap C) \neq (A \cap B') \triangle (A \cap C'). \tag{9}$$

(Here $D \triangle E = (D - E) \cup (E - D)$ denotes symmetric difference of the sets involved.)

Now we are ready to develop our lower bound for the entropy of our graph. To this end, consider the random variables X and Y attaining the minimum in the definition of the entropy of the graph $G = G_n$. Clearly, by (3), we have

$$H(G_n, P_n) = I(X \wedge Y) = H(X) - H(X \mid Y) \tag{10}$$

$$= H(P_n) - \sum_{y \in S(G)} \Pr\{Y = y\}\ H(X \mid Y = y) \tag{11}$$

$$\geqslant H(P_n) - \sum_{y \in S(G)} \Pr\{Y = y\}\ \log |y|, \tag{12}$$

where the right-most inequality follows from the fact that the conditional distribution of the random variable X given that the random variable Y takes its value y is concentrated in the stable set y so that its entropy is upper bounded by $\log |y|$. In order to lower bound this graph entropy, we shall upper bound

$$\sum_{y \in S(G)} \Pr\{Y = y\}\ \log |y| \tag{13}$$

$$= \sum_{x \in V(G)} \sum_{y;\, x \in y \in S(G)} \Pr\{X = x\}\ \Pr\{Y = y \mid X = x\}\ \log |y| \tag{14}$$

$$\leqslant \log \left(\sum_{x \in V(G)} \sum_{y;\, x \in y \in S(G)} \Pr\{X = x\}\ \Pr\{Y = y \mid X = x\}\ |y| \right), \tag{15}$$

where the last inequality follows by the cap-convexity of the binary logarithm. Recalling that P_n is the uniform distribution over $\binom{\mathscr{G}}{2}$, we can rewrite the right-most end of (15) as

$$\log \left(\sum_{x \in V(G)} \sum_{y;\, x \in y \in S(G)} \Pr\{X = x\}\ \Pr\{Y = y \mid X = x\}\ |y| \right) \tag{16}$$

$$\leqslant \log \left(\frac{1}{\binom{|\mathscr{G}|}{2}} \sum_{x \in V(G)} \max_{y;\, x \in y \in S(G)} |y| \right). \tag{17}$$

To continue with the upper bound in (16), we observe that no A-substar can appear in the last sum in (16) for more than two values of A. (This takes care also of the case when a stable set of the covering has 3 vertices. As observed before, such a set is not necessarily a star, nevertheless its size can be upper bounded by twice that of the corresponding star.) This consideration gives

$$\sum_{x\in V(G)} \max_{y;\, x\in y\in S(G)} |y| \leqslant 2 \sum_{A\in\mathscr{G}} \sum_{S\in\mathscr{P}(A)} \sum_{B\in S} |S| \tag{18}$$

$$\leqslant 2 \sum_{A\in\mathscr{G}} \sum_{S\in\mathscr{P}(A)} |S|^2 \tag{19}$$

$$= 2 \sum_{A\in\mathscr{G}} \sum_{S\in\mathscr{P}(A)} |S| + 4 \sum_{A\in\mathscr{G}} \sum_{S\in\mathscr{P}(A)} \frac{|S|\,(|S|-1)}{2}. \tag{20}$$

Next we observe that for each fixed $A\in\mathscr{G}$ the inner sum $\sum_{S\in\mathscr{P}(A)} \frac{|S|(|S|-1)}{2}$ is exactly the number of all the A-twin pairs, whence, in virtue of (9) we get

$$\sum_{S\in\mathscr{P}(A)} \left|\binom{S}{2}\right| \leqslant 2^{np}.$$

(We recall that $np = np_n$ is the common cardinality of the sets in $\mathscr{G}$.) More trivially, we also have

$$\sum_{S\in\mathscr{P}(A)} |S| \leqslant |\mathscr{G}|$$

and thus in conclusion the last two equations yield

$$2 \sum_{S\in\mathscr{P}(A)} |S| + 4 \sum_{S\in\mathscr{P}(A)} \frac{|S|\,(|S|-1)}{2} \leqslant 2 \max\{2^{np+2}, 2\,|\mathscr{G}|\}. \tag{21}$$

Substituting this inequality into (20) and comparing all the inequalities from (10) to (20) we obtain

$$H(G_n, P_n) \geqslant H(P_n) - \log\left(\frac{1}{\binom{|\mathscr{G}|}{2}} \sum_{x\in V(G)} \max_{y;\, x\in y\in S(G)} |y|\right) \tag{22}$$

$$\geqslant \log\binom{|\mathscr{G}|}{2} - \log\left(\frac{1}{\binom{|\mathscr{G}|}{2}} \sum_{A\in\mathscr{G}} \max\{2^{np+3}, 4\,|\mathscr{G}|\}\right) \tag{23}$$

$$\geqslant \log\binom{|\mathscr{G}|}{2} + \log\left(\frac{\binom{|\mathscr{G}|}{2}}{|\mathscr{G}|} \min\left\{2^{-np-3}, \frac{1}{4\,|\mathscr{G}|}\right\}\right) \tag{24}$$

$$\geqslant \log\frac{|\mathscr{G}|^2}{3} + \log\frac{|\mathscr{G}|}{3}$$

$$+ \min\{-np-3, -\log|\mathscr{G}|-2\} \tag{25}$$

$$\geqslant \min\{3\log|\mathscr{G}| - np - \log 72, 2\log|\mathscr{G}| - \log 36\}. \tag{26}$$

We recall that $|\mathscr{G}| = N(n)$ and compare the lower bound (26) to the upper bound (7) arriving at either

$$3 \log N(n) - np - \log 72 \leqslant \sum_{i=1}^{n} (1 - p_i^2)\, h\left(\frac{1 - p_i}{1 + p_i}\right) \tag{27}$$

or

$$2 \log N(n) - \log 36 \leqslant \sum_{i=1}^{n} (1 - p_i^2)\, h\left(\frac{1 - p_i}{1 + p_i}\right) \tag{28}$$

Upon observing that

$$\frac{1}{n} \sum_{i=1}^{n} p_i = p$$

we divide both sides of the inequality (27) by $3n$ to get

$$\frac{1}{n} \log N(n) - \frac{\log 72}{3n} \leqslant \frac{1}{3n} \sum_{i=1}^{n} \left[(1 - p_i^2)\, h\left(\frac{1 - p_i}{1 + p_i}\right) + p_i \right] \tag{29}$$

$$\leqslant \frac{1}{3} \max_{p \in [0, 1]} \left[(1 - p^2)\, h\left(\frac{1 - p}{1 + p}\right) + p \right]. \tag{30}$$

Likewise, dividing both sides of (28) by $2n$ we obtain

$$\frac{1}{n} \log N(n) - \frac{\log 36}{2n} \leqslant \frac{1}{2n} \sum_{i=1}^{n} \left[(1 - p_i^2)\, h\left(\frac{1 - p_i}{1 + p_i}\right) \right] \tag{31}$$

$$\leqslant \frac{1}{2} \max_{p \in [0, 1]} \left[(1 - p^2)\, h\left(\frac{1 - p}{1 + p}\right) \right]. \tag{32}$$

The last chain of equations gives a weaker bound than the preceding one and this implies the theorem.

Computer calculations show that the upper bound in (30) is less than 0.4098 while that in (32) is less than 0.4561. We recall that the best lower bound for $t(4)$ is about 0.26, obtained by random choice, (cf. [1]). Thus we have somewhat reduced the previous gap between the best upper and lower bound for $t(4)$. However, all indications are that a simple random choice is not the proper way to get good constructions for this particular problem.

ACKNOWLEDGMENT

In the original version of this manuscript a stronger result was claimed. Unfortunately, the proof contained an error. We are extremely grateful to the anonymous referee whose careful reading saved us from publishing a false statement.

REFERENCES

1. N. Alon, E. Fachini, and J. Körner, Locally thin set families, *Combinatorics*, *Prob*. *Computing*, to appear.
2. N. Alon, J. Körner and A. Monti, String quartets in binary, *Combinatorics*, *Prob*. *Computing*, to appear.
3. D. Coppersmith and J. B. Shearer, New bounds for union-free families of sets, *Electronic J*. *Combin*. **5** (1998), R39.
4. I. Csiszár and J. Körner, "Information Theory: Coding Theorems for Discrete Memoryless Systems," Academic Press, New York, 1982; Akadémiai Kiadó, Budapest, 1981.
5. P. Frankl and Z. Füredi, Union-free hypergraphs and probability theory, *European J. Combin.* **5** (1984), 127–131.
6. J. Körner, Coding of an information source having ambiguous alphabet and the entropy of graphs, *in* "Transactions of the 6th Prague Conference on Information Theory, 1971," pp. 411–425, Academia, Prague, 1973.
7. J. Körner, Fredman–Komlós bounds and information theory, *SIAM J*. *Algebraic Discrete Meth.* **4** (1986), 560–570.
8. J. Körner and K. Marton, New bounds for perfect hashing via information theory, *European J*. *Combin*. **9** (1988), 523–530.
9. B. Lindström, Determination of two vectors from the sum, *J*. *Combin*. *Theory Ser*. *A* **6** (1969), 402–407.
10. G. Simonyi, Graph entropy: a survey, *in* "Combinatorial Optimization," DIMACS Series on Discrete Math. and Computer Science (W. Cook, L. Lovász, P. D. Seymour, Eds.), Vol. 20, pp. 399–441, 1995.