COMBINATORICA 1 (4) (1981) 381—384

INTERSECTIONS OF k-ELEMENT SETS

D. J. KLEITMAN, J. SHEARER*

and

D. STURTEVANT

Massachusetts Institute of Technology
Cambridge, MA 02 139 U.S.A.

Received 15 June 1981

Let F be a collection of k-element sets with the property that the intersection of no two should be included in a third. We show that such a collection of maximum size satisfies $.2715k+o(k) \leqq \leqq \log_2 |F| \leqq .7549k+o(k)$ settling a question raised by Erdős. The lower bound is probabilistic, the upper bound is deduced via an entropy argument. Some open questions are posed.

0. Introduction

Let $\{A_1, \ldots, A_m\}$ be a collection of k-element sets with the following property:

(P1) *If i, j, l are unequal, then* $A_i \cap A_j \not\subseteq A_l$.

In this paper we consider the question: How large can m be? If you only take into account the fact that the $m-1$ intersections with A_1 are distinct, you find $m=O(2^k)$. Erdős [1] conjectured that $m=O((2-\varepsilon)^k)$ where ε is a positive real number. We settle this conjecture in the affirmative, showing that

$$\log m \leqq k \log\left(\frac{27}{16}\right)+o(k) \doteq .7549k+o(k).$$

(All logarithms in this paper have base 2.)

To get this bound, we use information theory to derive a basic lemma. We then apply this lemma to collections of sets satisfying the weaker property

(P2) *If* $\{i, j\} \neq \{p, q\}$, *then* $A_i \cap A_j \neq A_p \cap A_q$

to get a bound $f_1(\lambda)$, where λ is the average proportion (in some sense) of sets containing a particular element. Again using the lemma, we derive (under (P1)) another bound $f_2(\lambda)$. The maximum over λ of the minimum of the two functions yields the upper bound stated.

This research has been supported in part by the Office of Naval Research under Contract N00014-76-C-0366.

* Supported in part by a NSF postdoctoral Fellowship.

AMS subject classification (1980): 05 C 65; 05 C 35

In the final section we derive a probabilistic lower bound $\log m \geqq .2715k + o(k)$.

Before proceeding with the proofs, we pose some open questions.

(1) What, in fact, is $\lim_{k\to\infty} m^{1/k}$ (or does this limit even exist)?

(2) λ is roughly k/n, where n is the size of the base set. The upper bound given can only be attained when $\lambda \sim \frac{1}{3}$. The authors feel that a largest collection will have λ near $\frac{1}{2}$.

(3) Jim Shearer has given a heuristic argument which suggests that the probabilistic methods used for the lower bound can do no better than the square root of the actual answer. If this is true, that together with an affirmative answer to (2) would give $.54k \leqq \log m \leqq .62k$.

(4) No explicit construction of size exponential in k is known.

1. The basic lemma

The information theory in this section may be found in [2].

Lemma. *Let F be a collection of distinct subsets of $\{1, \ldots, n\}$ where i occurs in a proportion α_i of the sets in F. Then $\log |F| \leqq \sum H(\alpha_i)$ where $H(\alpha) = -\alpha \log \alpha - (1-\alpha) \log (1-\alpha)$.*

Proof. Let $F = \{S_1, \ldots, S_r\}$. Let S be a uniformly distributed random variable taking values in F, that is, $\Pr(S = S_i) = 1/r$ $(i = 1, \ldots, r)$. Let X_j $(j = 1, \ldots, n)$ be random variables taking values in $\{0, 1\}$ defined by $X_j = 1$ iff $j \in S$. Notice that $\Pr(X_j = 1) = \alpha_j$.

The information theoretic entropy of S is

$$H(S) = -\sum_{i=1}^{r} \Pr(S = S_i) \log \Pr(S = S_i) = \log r.$$

The proof of the following proposition may be found in [2], p. 33.

Proposition. If $S = (X_1, \ldots, X_n)$ is a random vector, then $H(S) \leqq \sum H(X_i)$.

The lemma is now immediate, as $H(X_i) = H(\alpha_i)$. ∎

2. $f_1(\lambda)$

Suppose we are given a collection $G = \{A_1, \ldots, A_m\}$ of k-element subsets of $\{1, \ldots, n\}$ satisfying (P2), where i occurs in a proportion λ_i of the sets in G. Out of the $\binom{m}{2}$ pairwise intersections, i appears in $\binom{\lambda_i m}{2}$, that is, in a proportion $\lambda_i^2 - \lambda_i(1-\lambda_i)/(m-1)$. We are working toward an exponential upper bound on m, so the second term may be neglected.

By the lemma, $\log\binom{m}{2} \leqq \sum H(\lambda_i^2) = k \sum_i \left(\frac{\lambda_i}{k}\right) \frac{H(\lambda_i^2)}{\lambda_i}$. As $\sum \left(\frac{\lambda_i}{k}\right) = 1$ and $H(\lambda^2)/\lambda$ is concave for $\lambda \in [0, 1]$, Jensen's inequality (cf. [2], p. 277) gives

$$\log\binom{m}{2} \leqq k \frac{H(\lambda^2)}{\lambda} \quad \text{where} \quad \lambda = \sum \left(\frac{\lambda_i}{k}\right) \lambda_i$$

or, setting $f_1(\lambda) = H(\lambda^2)/2\lambda$,

$$\log m \leqq k f_1(\lambda) + o(k).$$

Note that this bound already settles Erdős's conjecture, for $\max_\lambda f_1(\lambda) \doteq .8114$, attained when $\lambda \doteq .4914$.

3. $f_2(\lambda)$

Now let $G = \{A_1, \ldots, A_m\}$ satisfy (P1), with λ_i as above. In this section we will work with the complements of the sets in G. We claim that for any i, and any $S \subseteq \bar{A}_i$, either $(\forall j \neq i)\, S \nsubseteq \bar{A}_j$ or $(\forall j \neq i)\ \bar{A}_i \setminus S \nsubseteq \bar{A}_j$. Otherwise there exist p and q such that $S \subseteq \bar{A}_p$ and $\bar{A}_i \setminus S \subseteq \bar{A}_q$, so $\bar{A}_i \subseteq \bar{A}_p \cup \bar{A}_q$, violating (P1). Hence for each i we may associate a collection C_i of 2^{n-k-1} subsets of $\bar{A}_i$ included in no other $\bar{A}_j$.

For $j \in \bar{A}_i$, let j be an element of a proportion $\frac{1}{2} + \varepsilon_{ij}$ of the sets in C_i. By the lemma,

$$n-k-1 \leqq \sum_{j \in \bar{A}_i} H\left(\frac{1}{2} + \varepsilon_{ij}\right) \leqq n-k-c \sum_{j \in \bar{A}_i} \varepsilon_{ij}^2$$

where c is a positive constant, which implies $\sum_j \varepsilon_{ij}^2 = O(1)$ for any i.

Setting $\varepsilon_{ij} = 0$ if $j \notin \bar{A}_i$, let $\varepsilon_j = \frac{1}{m} \sum_i \varepsilon_{ij}$. Then $\sum_j \varepsilon_j^2 = \sum_j \left(\frac{1}{m} \sum_i \varepsilon_{ij}\right)^2 \leqq$

$$\leqq \sum_j \left(\frac{1}{m} \sum_i \varepsilon_{ij}^2\right) = O(1).$$

Let $C = \cup C_i$, so $|C| = m 2^{n-k-1}$. Notice that j appears in a proportion $(1-\lambda_j)/2 + \varepsilon_j$ of the sets in C. Using the lemma yet again,

$$\log m + n - k - 1 \leqq \sum_j H\left(\frac{1-\lambda_j}{2} + \varepsilon_j\right)$$

$$\leqq \sum_j \left[H\left(\frac{1-\lambda_j}{2}\right) + H'\left(\frac{1}{2} - \frac{\lambda_j}{2}\right)\varepsilon_j\right]$$

by the concavity of H. If $|x| < \frac{1}{2} - \varepsilon$, then $H'\left(\frac{1}{2} - x\right) = O(x)$. For our purposes this applies, for we are trying to bound the size of a collection of k-sets by r^k, and if $\lambda_i > 1/r$, then the collection of $(k-1)$-sets $G' = \{A - \{i\}\colon i \in A \in G\}$ satisfies $|G'| \leqq$ $\leqq r|G'|$. Hence by induction on k we may assume λ_i is bounded away from 1. It follows that

$$\log m \leqq \sum_j \left[H\left(\frac{1-\lambda_j}{2}\right) - 1 + \lambda_j\right] + O\left(\sum_j \lambda_j \varepsilon_j\right).$$

We may bound the error term by $\sqrt{k}$ by expanding $\sum_j (\lambda_j - \sqrt{k}\varepsilon_j)^2 \geqq 0$ and using $\sum \varepsilon_j^2 = O(1)$ and $\sum \lambda_j^2 \leqq k$ (as $\sum \lambda_j = k$ and $\lambda_j \leqq 1$). Thus

$$\log m \leqq k \sum_j \left(\frac{\lambda_j}{k}\right)\left[\frac{H\left(\frac{1-\lambda_j}{2}\right) - 1 + \lambda_j}{\lambda_j}\right] + o(k)$$

or, as $f_2(\lambda) = \frac{1}{\lambda}\left[H\left(\frac{1-\lambda}{2}\right) - 1 + \lambda\right]$ is concave on [0, 1], $\log m \leqq k f_2(\lambda) + o(k)$, where $\lambda = \frac{1}{k}\sum_j \lambda_j^2$ as before. Combining the two bounds, we have

$$\log m \leqq k \max_{\lambda} \min_{i} \{f_i(\lambda)\} + o(k)$$

which by direct calculation is

$$\log m \leqq k \log\left(\frac{27}{16}\right) + o(k)$$

attained when $\lambda = \frac{1}{3}$.

4. A Probabilistic Lower Bound

The idea is to find q as large as possible so that for some n and for a random collection Q of q k-element subsets of an n-set, the expected number of triples of sets in Q which violate (P1) is no more than $q/2$. From such a Q one may remove $q/2$ sets so that (P1) is satisfied.

Given n and k, the number of triples (A, B, C) of k-element subsets of an n-set N with $A \cap B \subseteqq C$ is

$$(*) \qquad b(n, k) = \binom{n}{k} \sum_x \binom{n-k}{k-x}\binom{n-x}{k-x}\binom{k}{x}.$$

The expected number of such triples in a random collection of q k-element subsets of N is thus $= O\left(q^3 b(n, k)\left(\frac{n}{k}\right)^{-3}\right)$.

If $x_0 = x_0(n, k)$ indexes the largest term in the sum $(*)$, then the q sought satisfies

$$\log q = \frac{1}{2}\max \log\left[\binom{n}{k}^2 \Big/ \binom{k}{x_0}\binom{n-k}{k-x_0}\binom{n-x_0}{k-x_0}\right] + o(k).$$

By using standard techniques, we find that the quantity on the right is maximized when $n \approx k(2-\sqrt{2})$ $\left(\text{for which } x_0 \approx \frac{1}{2}k\right)$ and

$$\log q = k \log\left(\frac{1+\sqrt{2}}{2}\right) + o(k)$$
$$= .2715k + o(k).$$

References

[1] P. Erdős, private communication.
[2] R. McEliece, *The Theory of Information and Coding*, Encyclopedia of Mathematics, Vol. 3, Addison-Wesley, Reading, Massachusetts, 1977.