JOURNAL OF COMBINATORIAL THEORY, Series A 54, 296–303 (1990)

The Largest Set Partitioned by a Subfamily of a Cover

KEVIN J. COMPTON*

EECS Department, University of Michigan, Ann Arbor, Michigan 48109

AND

CARLOS H. MONTENEGRO

Department of Mathematics, University of Michigan, Ann Arbor, Michigan 48109

Communicated by the Managing Editors

Received July 29, 1988

Define $\lambda(n)$ to be the largest integer such that for each set A of size n and cover $\mathscr{F}$ of A, there exist $B \subseteq A$ and $\mathscr{G} \subseteq \mathscr{F}$ such that $|B| = \lambda(n)$ and the restriction of $\mathscr{G}$ to B is a partition of B. It is shown that when $n \geqslant 3$

$$\frac{n}{(1+\ln n)} \leqslant \lambda(n) \leqslant \frac{2(n-1)}{(1+\lg(n-1)-\lg\lg(n-1))}.$$

The lower bound is proved by a probabilistic method. A related probabilistic algorithm for finding large sets partitioned by a subfamily of a cover is presented.

1. INTRODUCTION

The exact cover problem asks whether, for a given set A and a cover $\mathscr{F}$ of A, there is a subcover $\mathscr{G} \subseteq \mathscr{F}$ that partitions A. When no such subcover exists, we may consider a related problem: is there a "large" set $B \subseteq A$ which is partitioned by some $\mathscr{G}$, a subfamily of $\mathscr{F}$ (but perhaps not a subcover)? In this paper we investigate the problem of how large B can be in general.

* Research partially supported by NSF Grant DCR 86-05358.

0097-3165/90 $3.00

For $n > 0$ fix a set A of size n. Let $\lambda(n)$ be the largest integer k such that if $\mathscr{F} \subseteq 2^A$ is a cover of A, then there exist $B \subseteq A$ and $\mathscr{G} \subseteq \mathscr{F}$ such that $|B| = k$ and $\mathscr{G} \upharpoonright \mathscr{B} = \{B \cap C \mid C \in \mathscr{G}\}$ is a partition of B; i.e., each element of B is contained in precisely one set in $\mathscr{G}$. Let $\ln n$ denote $\log_e n$ and $\lg n$ denote $\log_2 n$. We show that when $n \geqslant 3$

$$\frac{n}{1 + \ln n} \leqslant \lambda(n) \leqslant \frac{2(n-1)}{1 + \lg(n-1) - \lg\lg(n-1)}.$$

The definition of $\lambda(n)$ may be formulated in the language of hypergraphs (see Berge [1]): $\lambda(n)$ is the largest integer k such that every hypergraph of size n has a partial subhypergraph of size k that is a matching.

The proof of the lower bound for $\lambda(n)$ is by a probabilistic argument. We assume that the reader is familiar with the basic concepts from probability theory found in introductory texts (see, e.g., Loéve [4]). We will present a related probabilistic algorithm for finding $B \subseteq A$ and $\mathscr{G} \subseteq \mathscr{F}$ partitioning B where $|B|$ approaches $\lambda(n)$.

We use the falling factorial notation $(n)_i = n(n-1)\cdots(n-i+1)$. Thus $\binom{n}{i} = (n)_i/i!$. By convention $(n)_0 = 1$. H_n will denote the nth harmonic number $1 + (1/2) + (1/3) + \cdots + (1/n)$.

2. Lower Bound for $\lambda(n)$

We first establish the following simple identity.

Lemma 1. *Let $0 \leqslant k \leqslant m$. Then*

$$\sum_{i=1}^{m-k+1} \frac{(m-k)_{i-1}}{(m)_i} = \frac{1}{k}.$$

Proof. Let $n = m - k$. Reversing the summation above, we see that we must show $\sum_{i=1}^{n+1} (n)_{i-1}/(m)_i = 1/(m-n)$, when $n \leqslant m$. We prove this by induction on n. It is clear lfor $n = 0$. If $n > 0$,

$$\sum_{i=1}^{n+1} \frac{(n)_{i-1}}{(m)_i} = \frac{1}{m} + \sum_{i=2}^{n+1} \frac{(n)_{i-1}}{(m)_i} = \frac{1}{m} + \frac{n}{m} \sum_{i=1}^{n} \frac{(n-1)_{i-2}}{(m-1)_{i-1}}$$

$$= \frac{1}{m} + \frac{n}{m} \frac{1}{m-n} = \frac{1}{m-n}$$

by the induction hypothesis. ∎

We thank Joel Spencer for suggesting the following alternate proof of Lemma 1. Consider an urn containing m marbles, k of which are red, the remainder being blue. Draw marbles from the urn (without replacement) until a red marble is found. Let us compute the probability that precisely i marbles will be drawn: Of the $(m)_i$ possible sequences of i marbles, $(m-k)_{i-1}k$ consist of $i-1$ blue marbles followed by a red one, so the probability is $(m-k)_{i-1}k/m_i$. Since a red marble will occur at the latest by the time $m-k+1$ marbles are drawn,

$$\sum_{i=1}^{m-k+1} \frac{(m-k)_{i-1}k}{(m)_i} = 1$$

We now prove the lower bound.

THEOREM 2. $n/(1+\ln n) \leqslant \lambda(n)$.

Proof. Let $|A| = n$ and $\mathscr{F} \subseteq 2^A$ be any cover of A. We will show that there are a set $B \subseteq A$ of size at least n/H_n and a subfamily $\mathscr{G} \subseteq \mathscr{F}$ such that $\mathscr{G} \restriction B$ is a partition of B. We may suppose that $\mathscr{F}$ is a minimal covering of A—i.e., that no proper subfamily of $\mathscr{F}$ covers A. Put $|\mathscr{F}| = m$. We know $m \leqslant n$ since every element of $\mathscr{F}$ covers some element of A which is covered by no other element of $\mathscr{F}$.

The proof proceeds as follows. We define a probability measure P on the set $\Omega = = \{\mathscr{G} \subseteq \mathscr{F} \mid \mathscr{G} \neq \varnothing\}$. For $\mathscr{G} \in \Omega$ let $B(\mathscr{G})$ be the set of elements in A covered by precisely one set in $\mathscr{G}$ and define a random variable $\mathbf{X}$ on Ω by $\mathbf{X}(\mathscr{G}) = |B(\mathscr{G})|$. We then show that $E(\mathbf{X})$, the expected value of $\mathbf{X}$, is n/H_m so there must be a subfamily $\mathscr{G} \subseteq \mathscr{F}$ such that $|B(\mathscr{G})| \geqslant n/H_m$. Clearly, if we take $B = B(\mathscr{G})$, $\mathscr{G} \restriction B$ is a partition of $\mathscr{B}$ with $|B| = n/H_m$.

We now define P. For $\mathscr{G} \in \Omega$, if $|\mathscr{G}| = i$ then set $P\{\mathscr{G}\} = (i\binom{m}{i}H_m)^{-1}$. To see that $P(\Omega) = 1$ note that there are $\binom{m}{i}$ elements $\mathscr{G} \in \Omega$ such that $|\mathscr{G}| = i$. Hence, $P(|\mathscr{G}| = i) = (iH_m)^{-1}$. But for every $\mathscr{G} \in \Omega$, $1 \leqslant |\mathscr{G}| \leqslant m$, so $P(\Omega) = \sum_{i=1}^{m} (iH_m)^{-1} = 1$.

Define a function $\mathbf{Y}: \Omega \times A \to \{0, 1\}$ as follows. $\mathbf{Y}(\mathscr{G}, a) = 1$ if and only if a is covered by precisely one element of $\mathscr{G}$. Thus $\mathbf{X}(\mathscr{G}) = \sum_{a \in A} \mathbf{Y}(\mathscr{G}, a)$. Also define for each $a \in A$ a random variable $\mathbf{Y}_a$ on Ω by $\mathbf{Y}_a(\mathscr{G}) = \mathbf{Y}(\mathscr{G}, a)$. We have

$$\begin{aligned} E(\mathbf{X}) &= \sum_{\mathscr{G} \in \Omega} \sum_{a \in A} \mathbf{Y}(\mathscr{G}, a)\, P\{\mathscr{G}\} \\ &= \sum_{a \in A} \sum_{\mathscr{G} \in \Omega} \mathbf{Y}(\mathscr{G}, a)\, P\{\mathscr{G}\} = \sum_{a \in A} E(\mathbf{Y}_a). \end{aligned}$$

We will show that $E(\mathbf{Y}_a) = 1/H_m$ for every $a \in A$, from which it follows that $E(\mathbf{X}) = n/H_m$.

Express $E(\mathbf{Y}_a) = \sum_{i=1}^{m} E(\mathbf{Y}_a \mid |\mathscr{G}| = i)\, P(|\mathscr{G}| = i)$, where $E(\mathbf{Y}_a \mid |\mathscr{G}| = i)$ is the conditional expectation of $\mathbf{Y}_a$ given that $|\mathscr{G}| = i$. Suppose that precisely

k elements of $\mathscr{F}$ cover a. Then if $i > m - k + 1$, at least two elements of $\mathscr{G}$ cover a when $|\mathscr{G}| = i$, so $E(\mathbf{Y}_a \mid |\mathscr{G}| = i) = 0$. If $i \leqslant m - k + 1$, there are $\binom{m}{i}$ elements $\mathscr{G} \in \Omega$ with $|\mathscr{G}| = i$. Of these, $k\binom{m-k}{i-1}$ cover a precisely once. Form $\mathscr{G}$ by choosing one of the k elements of $\mathscr{F}$ covering a and $i-1$ of the $n-k$ elements of $\mathscr{F}$ not covering a. Hence,

$$E(\mathbf{Y}_a \mid |\mathscr{G}| = i) = \frac{k\binom{m-k}{i-1}}{\binom{m}{i}} = \frac{ik(m-k)_{i-1}}{(m)_i}.$$

We know that $P(|\mathscr{G}| = i) = (iH_m)^{-1}$ so

$$E(\mathbf{Y}_a) = \sum_{i=1}^{m-k+1} \frac{k(m-k)_{i-1}}{(m)_i H_m} = \frac{k}{H_m} \sum_{i=1}^{m-k+1} \frac{(m-k)_{i-1}}{(m)_i} = \frac{1}{H_m}$$

by Lemma 1. Thus, $E(\mathbf{X}) = n/H_m$ and there is a $\mathscr{G} \in \Omega$ such that $|B(\mathscr{G})| \geqslant n/H_m$.

Since $m \leqslant n$, $H_m - 1 \leqslant H_n - 1 \leqslant \ln n$, so $\lambda(n) \geqslant n/H_m \geqslant n/(1 + \ln n)$. ∎

We can improve this estimate slightly by observing that $H_n = \gamma + \ln n + O(1/n)$, where γ is Euler's constant (see Knuth [3]). Hence $\lambda(n) \geqslant n/(\gamma + \ln n) + O(1)$.

3. Upper Bound for $\lambda(n)$

The upper bound is obtained by construction. We will describe how to find, for a set A of size n, a cover $\mathscr{F} \subseteq 2^A$ such for all $\mathscr{G} \subseteq \mathscr{F}$

$$|B(\mathscr{G})| \leqslant \frac{2(n-1)}{1 + \lg(n-1) - \lg\lg(n-1)}.$$

Lemma 3. *Let $t_0, t_1, ..., t_k$ be a sequence of integers such that for all i with $1 \leqslant i \leqslant k$, $t_0 + t_1 + \cdots + t_{i-1} \leqslant t_i$. Let $n = \sum_{i=0}^{k} t_i 2^{k-i}$ and $m = \sum_{i=0}^{k} t_i$. Then there is a cover $\mathscr{F}$ of each A of size n such that whenever $\mathscr{G} \subseteq \mathscr{F}$, $|B(\mathscr{G})| \leqslant m$.*

Proof. By induction on k. The case $k = 0$ is obvious. Induction step: Assume the statement for k. Let

$$\tilde{n} = \sum_{i=0}^{k+1} t_i 2^{k+1-i} = t_{k+1} + 2\sum_{i=0}^{k} t_i 2^{k-i} = t_{k+1} + 2n$$

$$\tilde{m} = \sum_{i=0}^{k+1} t_i = t_{k+1} + \sum_{i=0}^{k} t_i = t_{k+1} + m.$$

By the induction hypotheses, for any set A of size n there is a cover $\mathscr{F}$ of A such that $|B(\mathscr{G})| \leqslant m$ for every $\mathscr{G} \subseteq \mathscr{F}$. Let $\mathscr{F}$ and $\mathscr{F}'$ be such covers for A and A', respectively, where $|A| = |A'| = n$ and $A \cap A' = \varnothing$. Also let C be any set of size t_{k+1} disjoint from A and from A'. Define a cover of $\tilde{A} = A \cup A' \cup C$: $\tilde{\mathscr{F}} = \{C \cup S \mid S \in \mathscr{F} \text{ or } S \in \mathscr{F}'\}$.

Since A, A', and C are disjoint sets, $|\tilde{A}| = \tilde{n}$. We show that $\tilde{\mathscr{F}}$ is a cover of $\tilde{A}$ with the desired property. Let $\mathscr{G} \subseteq \tilde{\mathscr{F}}$ be any subset. If $|\mathscr{G}| = 1$, then $|B(\mathscr{G})| \leqslant t_{k+1} + m = \tilde{m}$. If $|\mathscr{G}| > 1$, then since each member of $\tilde{\mathscr{F}}$ contains C, $B(\mathscr{G}) \subseteq A \cup A'$ and so $|B(\mathscr{G})| \leqslant 2m \leqslant t_{k+1} + m = \tilde{m}$ (the last inequality holds by the assumption on the t_i's). This shows that for all $\mathscr{G} \subseteq \mathscr{F}$, $|B(\mathscr{G})| \leqslant \tilde{m}$ and so the lemma follows. ∎

Now for a given m, let $k = \lfloor \lg m \rfloor$, and let $t_i = \lfloor m/2^{k-i} \rfloor - \lfloor m/2^{k-i+1} \rfloor$. It is easy to see that the sequence $t_0, t_1, \ldots, t_k$ satisfies Lemma 3 and that $m = \sum_{i=0}^{k} t_i$. Let $v(m) = \sum_{i=0}^{k} t_i 2^{k-i}$.

LEMMA 4. *$2v(m) \geqslant (m+1)\lg(m+1)$ for all $m \geqslant 1$.*

Proof. By definition

$$v(m) = \sum_{i=0}^{k} \left(\left\lfloor \frac{m}{2^{k-i}} \right\rfloor - \left\lfloor \frac{m}{2^{k-i+1}} \right\rfloor \right) 2^{k-i},$$

where $k = \lfloor \lg m \rfloor$. Doubling and summing by parts, we have

$$2v(m) = m + \sum_{i=0}^{k} \left\lfloor \frac{m}{2^{k-i}} \right\rfloor 2^{k-i}.$$

We may suppose that this defines $v(m)$ for all positive real m, where k is an integer such that $2^k - 1 < m \leqslant 2^{k+1} - 1$. We prove by induction on k that $2v(m) \geqslant (m+1)\lg(m+1)$.

For the basis case $k = 0$ we must verify that $2m \geqslant (m+1)\lg(m+1)$ when $0 < m \leqslant 1$. The functions $2m$ and $(m+1)\lg(m+1)$ have the same values at $m = 0$ and 1. Also, $2m$ is linear while $(m+1)\lg(m+1)$ is convex since its second derivative is positive. Therefore, $2m$ dominates $(m+1)\lg(m+1)$ on the interval $0 < m \leqslant 1$.

Suppose that $k \geqslant 1$ and the result holds for smaller values. Then

$$2v(m) \geqslant 2m + \sum_{i=0}^{k-1} \left\lfloor \frac{m-1}{2^{k-i}} \right\rfloor 2^{k-i} = m + 1 + v\left(\frac{m-1}{2}\right).$$

Now $2^{k-1} - 1 < (m-1)/2 \leqslant 2^k - 1$, so by the induction hypothesis,

$$2v\left(\frac{m-1}{2}\right) \geqslant \frac{m-1}{2} \lg\left(\frac{m-1}{2}\right).$$

Combining inequalities and simplifying, we have $2\nu(m) \geqslant (m+1)\lg(m+1)$. ∎

We now prove the upper bound.

THEOREM 5. $\lambda(n) \leqslant 2(n-1)/(1+\lg(n-1)-\lg\lg(n-1))$ when $n \geqslant 3$.

Proof. Given n, let be m such that $\nu(m-1) < n \leqslant \nu(m)$. By Lemma 3, there is a cover $\mathscr{F}$ of each A of size $\nu(m)$ such that whenever $\mathscr{G} \subseteq \mathscr{F}$, $|B(\mathscr{G})| \leqslant m$. Since $n \leqslant \nu(m)$, the same statement holds for each A of size n.

By Lemma 4

$$\frac{m \lg m}{2} \leqslant \nu(m-1) \leqslant n-1.$$

Apply the function $f(x) = x/(\lg x - \lg\lg x)$ to this inequality to obtain

$$\frac{m}{2}\,\frac{\lg m}{2\lg((m\lg m)/2) - \lg\lg((m\lg m)/2)} \leqslant \frac{n-1}{\lg(n-1)-\lg\lg(n-1)}.$$

The inequality is preserved because f is monotonic. It is easy to check that the left side is at least $m/2$ so we have

$$\lambda(n) \leqslant m \leqslant \frac{2(n-1)}{\lg(n-1)-\lg\lg(n-1)}. \quad ∎$$

4. A Probabilistic Algorithm

Theorem 2, which gives the lower bound for $\lambda(n)$, is not constructive. However, it does provide a polynomial time probabilistic algorithm for finding a large set partitioned by a subfamily of a cover. We do not expect that there is a deterministic polynomial time algorithm for finding the largest set partitioned by a subfamily of a cover because the exact cover problem is a special case of this problem. (Recall that the exact cover problem asks whether there is a subcover $\mathscr{G} \subseteq \mathscr{F}$ that partitions A.) The exact subcover problem is NP-complete, even when the sets in $\mathscr{F}$ are restricted to be three element sets (see Garey and Johnson [2, p. 53]).

Let $|A| = n$ and $\mathscr{F} \subseteq 2^A$ be a cover of A. We may assume that $\mathscr{F} = m \leqslant n$. Consider the random variable $\mathbf{X}(\mathscr{G}) = |B(\mathscr{G})|$ defined in the proof of Theorem 2. It was shown there that $E(\mathbf{X})$, the expected value of $\mathbf{X}$ with respect to the probability measure P, is n/H_m (denote this value by M).

Take $\varepsilon > 0$ and let $p = P(\mathbf{X} \geqslant (1-\varepsilon)M)$. Now since $\mathbf{X}$ is bounded by n, we have

$$pn + (1-p)(1-\varepsilon)M \geqslant M$$

whence

$$p \geqslant \frac{\varepsilon M}{n - (1-\varepsilon)M} \geqslant \frac{\varepsilon M}{n} = \frac{\varepsilon}{H_m}.$$

That is, if a nonempty $\mathscr{G} \subseteq \mathscr{F}$ is selected according to the probability measure P, the probability that $\mathscr{G}$ partitions a set of size at least $(1-\varepsilon)M$ is at least ε/H_m. Suppose we independently repeat such a selection N times. The probability that we do not find a set of size $(1-\varepsilon)M$ partitioned by some $\mathscr{G}$ among the N choices is at most $(1-\varepsilon/H_m)^N$. Take $\varepsilon = \varepsilon(n)$ tending to 0 and a polynomial $N = N(n)$ such that $N\varepsilon/H_m$ tends to ∞. (For example, let $\varepsilon = 1/n$ and $N = n^2$.) Then $(2-\varepsilon/H_m)^N$ tends to 0 so the probability of finding $\mathscr{G}$ with $|B(\mathscr{G})|$ nearly as large as $\lambda(n)$ within N selections is nearly certain.

Our algorithm can now be simply stated for ε and N as above.

> **GIVEN**: A of size n; cover $\mathscr{F} \subseteq 2^A$ of size $m \leqslant n$.
> **REPEAT**
> Select $k \in \{1, ..., m\}$ according to the harmonic distribution;
> Select $\mathscr{G} \subseteq \mathscr{F}$ of size k according to the uniform distribution;
> N **TIMES OR UNTIL** $|B(\mathscr{G})| \geqslant (1-\varepsilon)\,\lambda(n)$.

5. Concluding Remarks

The lower bound for $\lambda(n)$ proved in Theorem 2 is asymptotic to $n/\ln n$. The upper bound proved in Theorem 5 is asmptotic to $(2\ln 2)\,n/\ln n = (1.386\cdots)\,n/\ln n$, which is surprisingly close to the lower bound. We are naturally led to conjecture that $\lambda(n) \sim Kn/\ln n$ for some constant K. Since the lower bound was obtained by probabilistic methods, we would expect K to correspond more closely to the upper bound value $2\ln 2$.

The algorithm in the previous section is quite modest. For a given cover $\mathscr{F} \subseteq 2^A$, the size k of the largest set partitioned by a subfamily of $\mathscr{F}$ may be much larger than $\lambda(n)$. However, the algorithm yields only a set of size $(1-\varepsilon)\,\lambda(n)$ with high probability. We would like to have an algorithm that yields a set of size $(1-\varepsilon)k$ in all cases, or an algorithm that yields a set of size k with high probability.

References

1. C. Berge, "Graphs and Hypergraphs," North-Holland, Amsterdam, 1973.
2. M. R. Garey and D. S. Johnson, "Computers and Intractability: A Guide to the Theory of NP-Completeness," Freeman, New York, 1979.
3. D. E. Knuth, "The Art of Computer Programming, Vol. 1," Addison–Wesley, Reading, MA, 1968.
4. M. Loéve, "Probability Theory I," 4th ed., Springer-Verlag, New York, 1977.