MATHEMATICS OF COMPUTATION
VOLUME 40, NUMBER 162
APRIL 1983, PAGES 667–676

Irregular Sets of Integers Generated by the Greedy Algorithm

By Joseph L. Gerver

Abstract. The greedy algorithm was used to generate sets of positive integers containing no subset of the form $\{x, x+y, x+2y\}$, $\{x, x+y, x+3y\}$, $\{x, x+2y, x+3y\}$, $\{x, x+3y, x+4y\}$, $\{x, x+3y, x+5y\}$, and $\{x, x+y, x+2y, x+3y\}$, respectively. All of these sets have peaks of density in roughly geometric progression.

In 1936, Erdös and Turan [1], in connection with the question of whether there exist arbitrarily long sequences of primes in arithmetic progression, conjectured that if the sum of the reciprocals of an infinite set of positive integers diverges, then the set must contain arbitrarily long sequences of elements in arithmetic progression. This conjecture, still unsettled, led to attempts over the years to construct denser and denser sets of integers containing no k elements in arithmetic progression, for various fixed values of k.

The earliest such construction, due to Szekeres, was actually mentioned by Erdos and Turan [1] in their original paper. Szekeres considered the case where k is prime; his set S_k consisted of all nonnegative integers which do not include the digit $k-1$ when written in the base k. Although Rankin [6] later constructed examples with greater asymptotic density, Szekeres's sets (when each element is increased by one, to avoid dividing by zero) still hold the record for the sum of the reciprocals of a set of positive integers with no arithmetic progression of k terms; see Gerver [2].

Szekeres's sets are generated by the greedy algorithm. That is, $n \in S_k$ if and only if $(S_k \cap [0, n-1]) \cup \{n\}$ contains no k elements in arithmetic progression (where $[0, n-1]$ is the set of integers from 0 through $n-1$). In 1979, Gerver and Ramsey [3] considered the sets S_k generated by the greedy algorithm in this manner, in the case where k is composite. We computed thousands of elements of S_4 and S_6 and found them to be distributed quite randomly, on a local scale, in contrast to the very regular pattern of Szekeres's sets. We also presented a heuristic argument, supported by the computations, that the number of elements of S_k less than n (k composite) should be asymptotically proportional to $n^{(k-2)/(k-1)}(\log n)^{1/(k-1)}$. This argument hinged on the assumption that the elements of S_k were truly random, in some sense, and that the density of S_k in the vicinity of x could be approximated by a function $\varphi(x)$ with the following property: There exists a constant p, $0 \leq p \leq 1$, such that, for every $c > 1$, $\lim_{x\to\infty} \varphi(cx)/\varphi(x) = c^{-p}$. (It was erroneously stated in [3] that this follows from the weaker property that $\lim_{x\to\infty} \log \varphi(x)/\log x = -p$.)

Received December 17, 1980; revised June 21, 1982.
1980 *Mathematics Subject Classification.* Primary 10L10, 10–04.
Current address: Rutgers University, Camden, New Jersey 08102.

0025-5718/82/0000-0892/$02.25

At about the same time, Odlyzko and Stanley [4] considered the case where $k = 3$, but the greedy algorithm applies only starting with the third element of the set, the first element being 0, and the second element, a_2, being arbitrary. They observed that when a_2 was a power of three or twice a power of three, the set closely resembled Szekeres's set S_3; these sets they called *regular*. The remaining sets, called *irregular*, appeared to be quite random for the first few hundred elements, like the sets S_4 and S_6. Odlyzko and Stanley conjectured that, for irregular sets, the number of elements less than n should be asymptotically proportional to $n^{1/2}(\log n)^{1/2}$, using essentially the same heuristic argument as in [3].

Later Odlyzko [5] computed the first irregular set, with $a_2 = 4$, up to several thousand elements. Far from decreasing smoothly, the density of this set oscillated up and down, with a sequence of peaks and valleys in roughly geometric progression. Figure 1 shows the number of elements of this set in consecutive intervals of 10000, up to 2.55×10^6 (i.e. 255 intervals of 10000; actually the first interval, here and in the other figures, has length 9999).

The ratio between two consecutive peaks in the density of this set is approximately 2.5. Odlyzko conjectured that this ratio should tend to $\phi^2 = 2.618\ldots$, where $\phi = (1 + \sqrt{5})/2$ is the golden ratio. Indeed, let r^2 be the ratio between successive peaks, and assume that each valley is at the geometric mean of the peaks on either side. Then two adjacent peaks and the following valley will be in the ratio $1 : r^2 : r^3$. If these three numbers are in arithmetic progression, then the two peaks will tend to reinforce the valley; likewise, to a lesser extent, two adjacent valleys will tend to reinforce the following peak. But $r > 1$ and $r^3 - r^2 = r^2 - 1$ if and only if $r = \phi$.

An arithmetic progression of three terms is a set of the form $\{x, x + y, x + 2y\}$. In order to test Odlyzko's conjecture, the greedy algorithm was used to generate sets of positive integers containing no subset of the form $\{x, x + y, x + 3y\}$, $\{x, x + 2y, x + 3y\}$, $\{x, x + 3y, x + 4y\}$, and $\{x, x + 3y, x + 5y\}$, respectively. The distribution of the elements of these four sets (that is, the number of elements in consecutive intervals of 10000) is shown in Figures 2, 3, 4, and 5, respectively. All of these sets appear to be irregular, in the sense that they have no simple nonrecursive definition, in terms of the digits of their elements in some base. It would be interesting to know whether any of these sets can be made regular by delaying the application of the greedy algorithm.

Table 1 compares some predicted and observed numbers associated with these sets. Equation 1 is derived by letting two successive peaks and the following valley (assumed to be in the ratio $1 : r^2 : r^3$) form the pattern of the avoided subset. This equation always has the root 1. Of the two remaining roots, one must be positive and one negative; the former is used to compute the ratio r^2 which follows equation 1. In the cases where the avoided subset is $\{x, x + 2y, x + 3y\}$ and $\{x, x + 3y, x + 4y\}$, we have $r^2 \leqslant 1$, which is meaningless when interpreted as the ratio of successive peaks. If instead of taking two successive peaks, we skip a peak, so the two peaks and the following valley are in the ratio $1 : r^4 : r^5$, and we let these peaks and valley form the pattern of the avoided subset, we then obtain equation 2. This equation also has the root $r = 1$, and one other positive root from which we can compute r^2.

The observed ratio between successive peaks was computed not from the positions of the peaks themselves, but from the positions of the ascending midpoints, defined

TABLE 1

Avoided subset:	$x, x+y, x+2y$ (Figure 1)						
Equation 1:	$r^3 - r^2 = r^2 - 1$, $r^2 = 2.62$						
Ascending midpoints:	12	31	79	195			
Ratios:	2.58	2.55	2.47	(2.53 ± .03)			
Number of elements:	1179	1991	3266	5405			
Ratios:	1.69	1.64	.1.65	(1.66 ± .01)			
Exponent:	.546						
Avoided subset:	$x, x+y, x+3y$ (Figure 2)						
Equation 1:	$r^3 - r^2 = 2(r^2 - 1)$, $r^2 = 7.46$						
Ascending midpoints:	16	67	260				
Ratios:	4.19	3.88	(4.03 ± .16)				
Number of elements	1652	3352	7036				
Ratios:	2.03	2.10	(2.06 ± .04)				
Exponent:	.519						
Avoided subset:	$x, x+2y, x+3y$ (Figure 3)						
Equation 1:	$2(r^3 - r^2) = r^2 - 1$, $r^2 = 1.00$						
Equation 2:	$2(r^5 - r^4) = r^4 - 1$, $r^2 = 1.82$						
Ascending midpoints:	11	19	32	56	95	164	279
Ratios:	1.73	1.68	1.75	1.70	1.73	1.70	(1.71 ± .01)
Number of elements	1526	2055	2766	3734	5048	6823	9174
Ratios:	1.347	1.346	1.350	1.352	1.352	1.345	(1.349 ± .001)
Exponent:	.555						
Avoided subset:	$x, x+3y, x+4y$ (Figure 4)						
Equation 1:	$3(r^3 - r^2) = r^2 - 1$, $r^2 = 0.59$						
Equation 2:	$3(r^5 - r^4) = r^4 - 1$, $r^2 = 1.27$						
Ascending midpoints:	9	13.5	20	29	42	62	91
Ratios:	1.50	1.48	1.45	1.45	1.48	1.47	(1.47 ± .01)
Number of elements:	1631	2015	2500	3117	3840	4743	5913
Ratios:	1.235	1.241	1.247	1.232	1.235	1.247	(1.239 ± .003)
Exponent:	.556						
Avoided subset:	$x, x+3y, x+5y$ (Figure 5)						
Equation 1:	$3(r^3 - r^2) = 2(r^2 - 1)$, $r^2 = 1.48$						
Ascending midpoints:	10	14	21	29	41	59	82
Ratios:	1.40	1.50	1.38	1.41	1.44	1.39	(1.42 ± .02)
Number of elements:	1890	2306	2881	3434	4187	5038	6119
Ratios:	1.22	1.25	1.19	1.22	1.20	1.22	(1.22 ± .01)
Exponent:	.558						

to be the point before each peak at which the density has a value midway between that of the peak and that of the previous valley. The positions of the ascending midpoints are given in units of 10000, and are estimated by eye in those cases where there is some ambiguity. The ascending midpoint was chosen, rather than the descending midpoint, because of the empirical fact that, for all five sets, the density function rises faster than it falls during each cycle. This means that the valleys are not located at the geometric mean of the two adjacent peaks, but closer to the following peak. Below the position of each ascending midpoint is listed the ratio of the next position to that position;-these ratios are followed, in parentheses, by their mean μ_1 and the standard deviation of their mean. Next is listed the number of elements in the set up to the valley immediately before each ascending midpoint, followed by the ratios of these numbers, and the mean μ_2 and standard deviation of the mean of these ratios. Finally, we compute the exponent $\log \mu_2 / \log \mu_1$; this is, roughly speaking, the power of n which approximates the number of elements less than n.

FIGURE 1

FIGURE 2

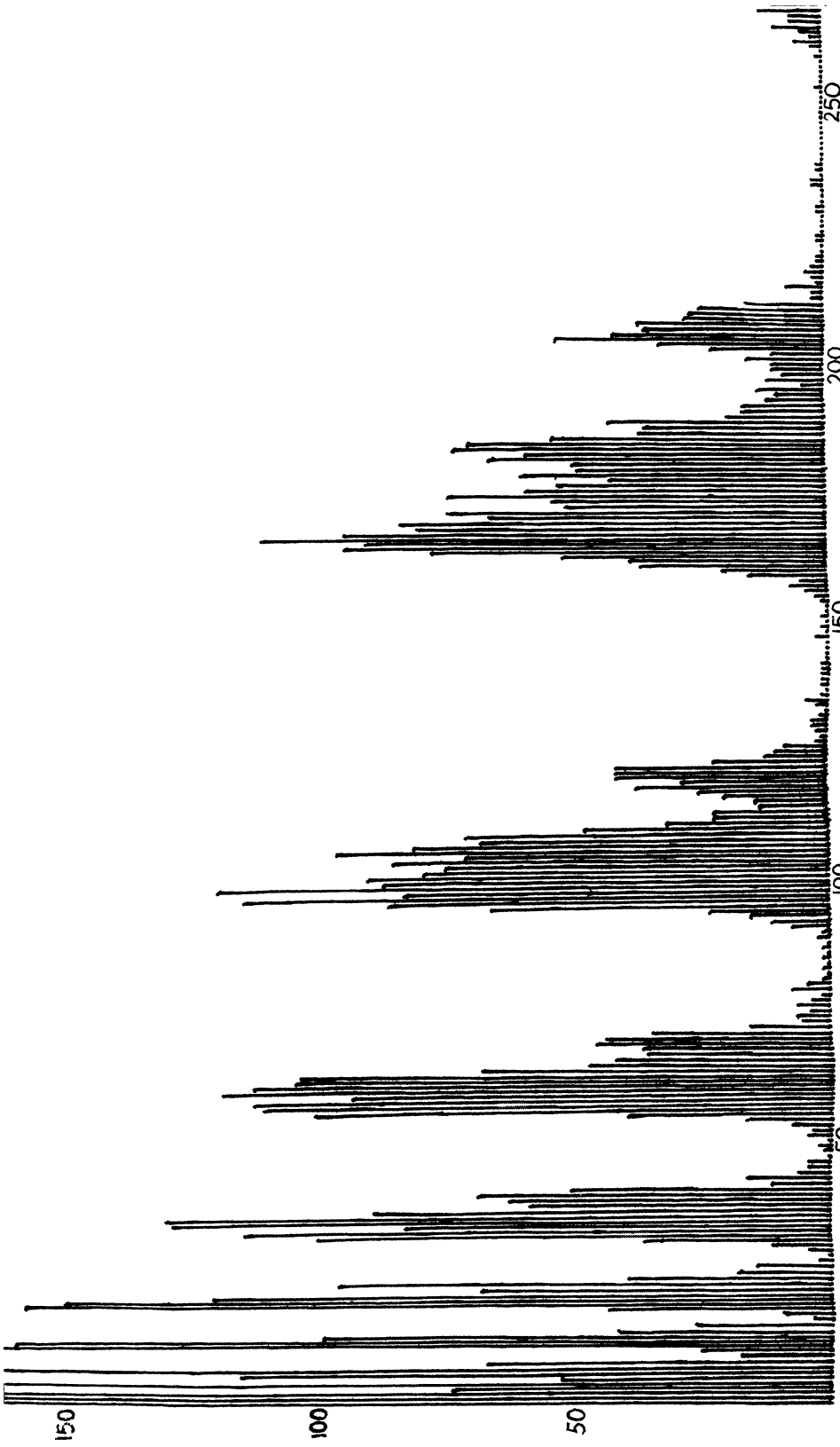

FIGURE 3

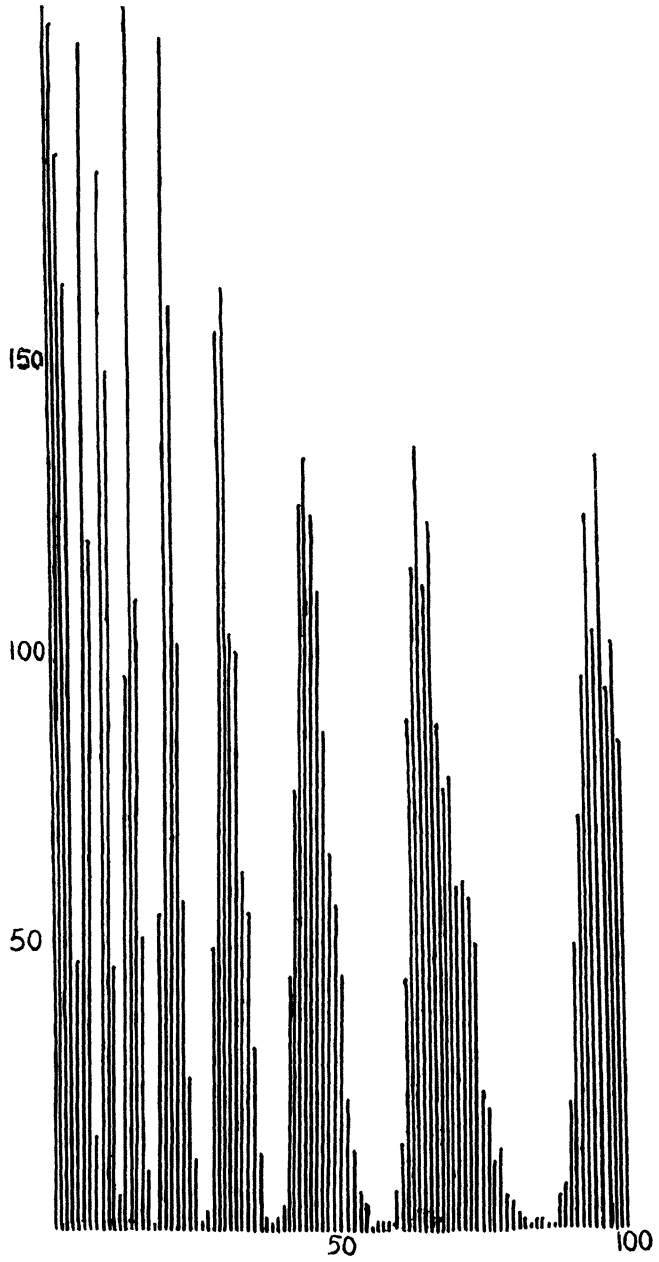

FIGURE 4

In all cases, the observed exponent agrees well with the conjecture that the number of elements less than n is roughly proportional to $n^{1/2}(\log n)^{1/2}$. As n tends to infinity, the exponent should approach $1/2$, but in this region, where $\log n$ is on the order of 12.5, the exponent should be about $(1/2)[1 + (1/12.5)] = .54$.

On the other hand, it is evident that equations 1 and 2 are not of much help in predicting the ratio of successive density peaks.

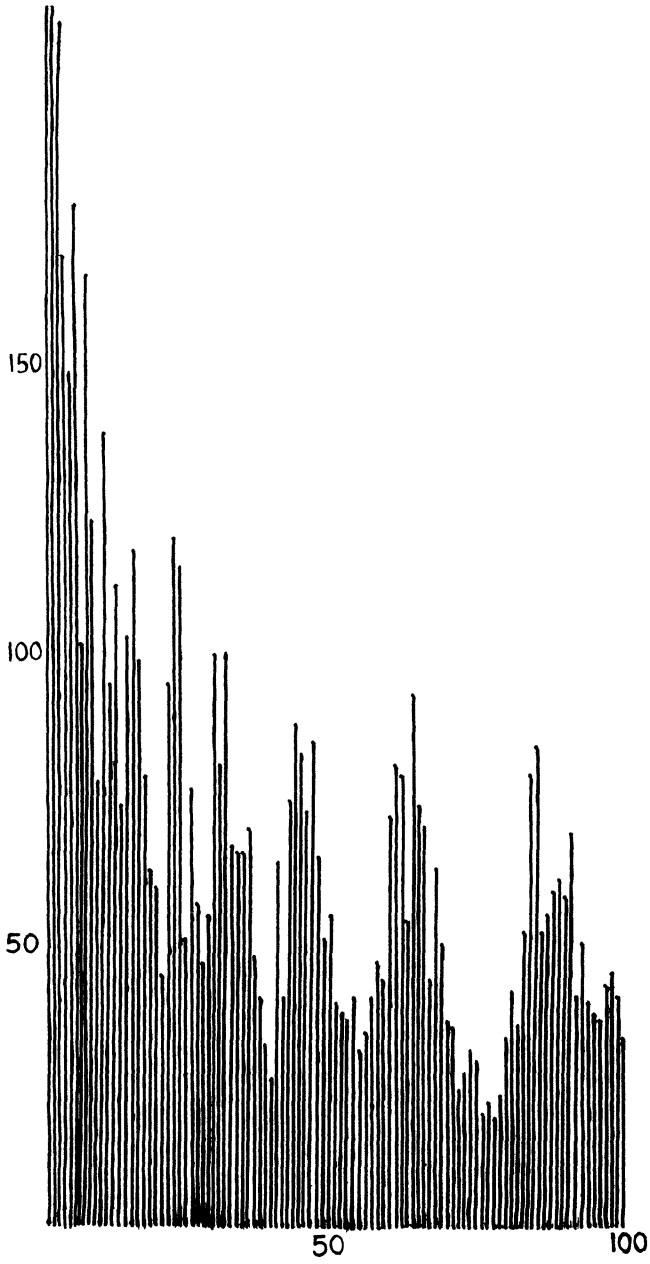

FIGURE 5

An alternative hypothesis is that each valley is caused by the previous peak acting in concert with the dense region at the beginning of the set. Thus if $\{x, x + by, x + cy\}$ is the avoided subset, a valley should appear at c/b times the position of each peak, and the next peak should occur shortly afterwards. For all the sets except the one in Figure 5, this model seems to fit, with the ratio between successive peaks approximately $1 + (3/2)[(c/b) - 1]$. In Figure 5, this number is very close to the square of the ratio between successive peaks, so it is possible that in this case the fundamental frequency is actually half of what is appears to be, and the first harmonic is unusually strong for some reason.

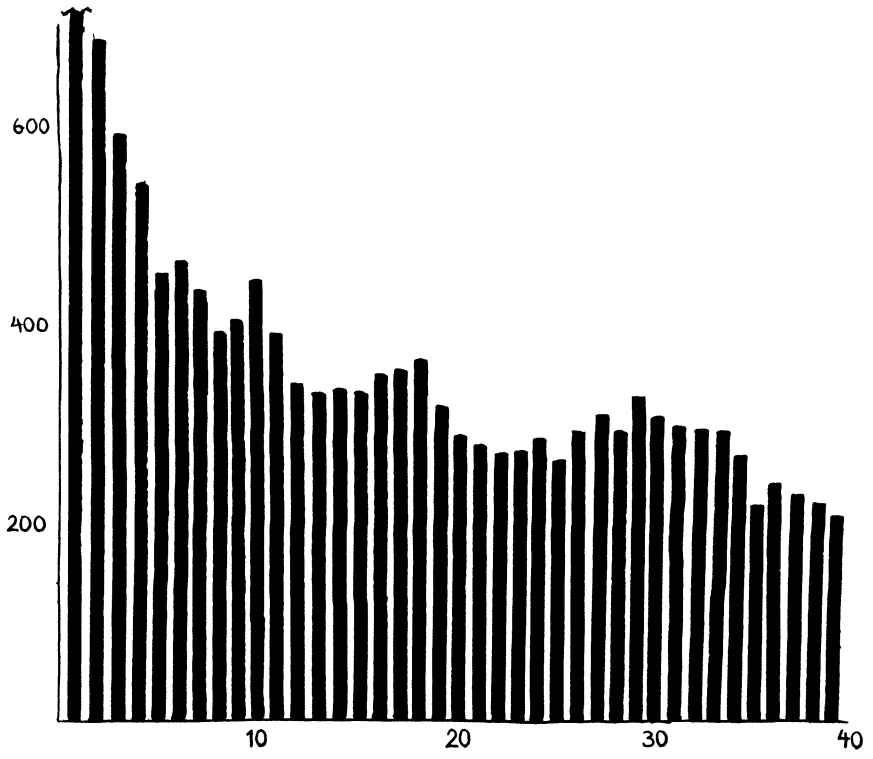

FIGURE 6

Whatever the precise mechanism may be which gives rise to the peaks and valleys, one might expect the following to be true for all sets S generated by the greedy algorithm in this manner.

CONJECTURE. For each set S, there exists a real number α, $0 \leqslant \alpha \leqslant 1$, and a function g, where $\lim_{u\to\infty} g(u) = \infty$, such that, for each real number ν, there exists

$$f_\nu = \lim_{u\to\infty} \left(\sum_{\substack{n\in S \\ u\leqslant n\leqslant u+g(u)}} n^{-\alpha} e^{i\nu \log n} \Big/ \sum_{\substack{n\in S \\ u\leqslant n\leqslant u+g(u)}} n^{-\alpha} \right).$$

One might strengthen this conjecture to state that $f_\nu = 0$ except for a discrete set of ν; one might even speculate that this discrete set always consists of all the integral multiplies of a single fundamental frequency, characteristic of S. Indeed, this last version of the conjecture is true for Szekeres's original sets, where $\alpha = \log(k-1)/\log k$, and the fundamental frequency is $\log k$.

In support of the generality of this conjecture, note the presence of peaks and valleys in geometric progression in the set of positive integers with no arithmetic progression of four terms, generated by the greedy algorithm (Figure 6; as usual the horizontal scale is in intervals of 10000).

All of the computations for this paper were done on a CDC Cyber 18/30. This machine was not ideal for the task, being relatively small and slow; its principal advantage was that I could use it for free on nights and weekends. I would like to thank H. Pritchett and E. R. Canfield for making this arrangement possible. I would also like to thank R. Chambers, who wrote two critical subroutines, one to replace

the standard FORTRAN WRITE routine (thus freeing enough memory to compute an additional 6000 elements of each set), and the other to store the array of elements on a disc when other people were using the machine (making some long computations possible; for example, Figure 3 required 72 hours of CPU time).

Department of Mathematics
University of Georgia
Athens, Georgia 30602

1. P. ERDÖS & P. TURAN, "On certain sequences of integers," *J. London Math. Soc.*, v. 11, 1936, pp. 261–264.

2. J. L. GERVER, "The sum of the reciprocals of a set of integers with no arithmetic progression of k terms," *Proc. Amer. Math. Soc.*, v. 62, 1977, pp. 211–214.

3. J. L. GERVER & L. T. RAMSEY, "Sets of integers with no long arithmetic progressions generated by the greedy algorithm," *Math. Comp.*, v. 33, 1979, pp. 1353–1359.

4. A. M. ODLYZKO & R. P. STANLEY, "Some curious sequences constructed with the greedy algorithm," unpublished Bell Laboratories report, January 1978.

5. A. M. ODLYZKO, private communication.

6. R. A. RANKIN, "Sets of integers containing not more than a given number of terms in arithmetical progression," *Proc. Roy. Soc. Edinburgh Sect.* A, v. 65, 1960/1961, pp. 332–344.