Introduction and History

In 1936, Erdös and Turan [1] examined sequences of positive integers $0<a_1<a_2<\ldots\leq N$ with no arithmetic progression of three terms. They considered how dense such a sequence could be by bounding the maximum number of elements r(N).

One result they state, due to Szekeres, is that $r\left(\frac{3^k+1}{2}\right)\geq 2^k$. This is obtained by taking the elements a_i without a digit 2 in their ternary representations and forming the sequence $a_i + 1$ (to start with 1 instead of 0).

Szekeres' set can also be generated by a "greedy" algorithm. Greedy algorithms are called "greedy" because they generate elements greedily—the next element generated is the lowest number found that, in this case, does not form a three element arithmetic progression. So to generate Szekeres' set, begin with $a_1=1$. Then a_n, for $n\geq 2$, is the next positive integer (i.e. the least integer not yet considered) that doesn't form a progression of 3 elements with any of the previously found elements. So starting with the first element $a_1=1$, the second is $a_2=2$, and the third is $a_3=4$ because 3 would form the arithmetic progression 1, 2, 3.

Later, Odlyzko and Stanley [3] considered what happened if the greedy algorithm were applied starting with the *third* element of the set, with $a_1=1$ and a_2 an arbitrary positive integer (instead of 2). Odlyzko computed several thousand elements of the set with $a_2=5$, and found that the density of the set oscillated with peaks and valleys in geometric progression, rather than falling smoothly.

Szekeres' and Odlyzko's sets contained no three element arithmetic progressions, or in other words, they excluded subsets of the form $\{x, x+y, x+2y\}$. Gerver [2] examined Odlyzko's set and also considered other sets with different excluded subsets, such as $\{x, x+y, x+3y\}$ and $\{x, x+2y, x+3y\}$, but still generated by the greedy algorithm.

To talk more easily about sets such as Gerver and Odlyzko examined, I abbreviate phrases such as "the set generated by the greedy algorithm excluding subsets of the form $\{x, x+ay, x+by\}$" by "the set a,b" or "the a,b set."

The Sets Computed and how they were Analyzed

Gerver generated the members of Odlyzko's set 1,2 and the sets 1,3; 2,3; 3,4; and 3,5 up to 2.5 million and conjectured that the number of elements less than n was about $\sqrt{n \log n}$. In those subsets, the peaks were all in geometric ratios and the densities did approach the conjecture.

To further test that conjecture and to apply it to other sets, I consider the sets a,b with $a < b \leq 5$, i.e. Odlyzko's set 1,2 and the sets 1,3; 1,4; 1,5; 2,3; 2,5; 3,4; 3,5; and 4,5. I do not consider the set 2,4 because it should differ from 1,2 only when y is odd in the subset $\{x, x+y, x+2y\}$, otherwise yielding a very similar structure. I generated the subsets of these (infinite) sets with elements less than 120 million and analyzed the results*.

* Almost all of the set generation was done on Macintosh IIx computers publicly available to students at Princeton, taking from 3 to 14 hours depending on the set computed. The number of elements found varied from 65,000 to 110,000.

Figures 1 through 9 present the nine sets. In each case, I have graphed with a logarithmic horizontal scale, starting with n=1000 and going to 120 million. If the peaks are in geometric progression, they will be evenly spaced along the graph in arithmetic progression. In the bottom graph on each page, I graph the exponent $y(n) = \frac{\log \sigma_n}{\log n}$ (i.e. $n^{y(n)} = \sigma_n$), where σ_n is the number of elements of the set less than n. If Gerver's conjecture that $\sigma_n \to \sqrt{n \log n}$ is correct, then y(n) should approach $\frac{1}{2} + \frac{\log \log n}{2\log n}$, so in the top half of each figure, I graph $y(n) - \frac{\log \log n}{2\log n}$. If the conjectures are correct, the values in the graph should approach 0.5 and the peaks and valleys should be in arithmetic progression on the logarithmic horizontal scale.

The reason I look only at the exponent graphs, instead of looking directly at the density, is because they are much smoother but still allow correct determination of the positions of the peaks. In intervals where the density is below average, the exponent will decrease; if the density is above average, the exponent will increase. So by looking at the points where the exponent begins to decrease, I can get near the end of a peak in density, and similarly for a valley. And if all I measure is the ratio of peaks, then it doesn't matter where in the cycle I begin as long as I am consistent from cycle to cycle.

In addition to the graphing the sets, I also did Fourier transforms to determine the frequencies (and hence the ratios) of the peaks (graphs of the Fourier transforms follow the figures). I took the whole set, (in a logarithmic scale) scaled to get 500 horizontal points, x from 1 to 500, and took f(x) to be the value of the exponent minus the conjecture (the values in the top graph in the figures). Then I computed the Fourier coefficients a_n and b_n, and graphed $\sqrt{a_n^2 + b_n^2}$. I also did the same thing with the last 1/5 of the graphs: scaled to get 500 points, then did the Fourier transform on them; this second transform should avoid the noise from the beginnings of the sets and should more clearly display the behavior of the interesting areas at the end (especially in 2,3; 3,4; and 4,5, where there are subpeaks in the graphs).

If there is a large coefficient in the first Fourier graph at a frequency of k, then that means there are k wavelengths from 1000 to 120,000,000 on the logarithmic scale. There are thus k wavelengths in a factor of 120,000, and so each wave should have peaks at ratios of s, where s^k=120,000, or s= $\sqrt[k]{120{,}000}$. In the second Fourier graphs (those of the last 1/5 of the subsets I generated), a frequency of k corresponds to a frequency of 5k in the first (whole set) graphs.

Results of the Analysis

The following table presents a summary of my findings; they are discussed in more detail below.

Set	The ratio of peaks measured (see Appendix)	The largest frequency components overall	The ratios giving those frequencies
1,2	2.52±0.04	13 25	2.46 1.60

1,3	4.1±0.3	8 19	4.31 1.85
1,4	2.2±0.2	7 15	5.32 2.18
1,5	6.3±1.2	13 27	2.46 1.54
2,3	1.72±0.02	22 44	1.70 1.30
2,5	3.33±0.08	9 19	3.67 1.85
3,4	1.47±0.01	31 61	1.46 1.21
3,5	1.40±0.01	36	1.38
4,5	1.347±0.004	39	1.35

The first set I consider is Odlyzko's set 1,2 with the greedy algorithm delayed by making the second element $a_2 = 5$. When Gerver generated this set to 2.5 million, he found that four peaks were in a ratio of about 2.53, close to Odlyzko's conjectured value of $\phi^2 = 2.62$. However, in my further computation of the set I measured a ratio of 2.52 over the last 11 peaks. From the Fourier transform, the fundamental frequency is around 13, or a ratio of 2.46; a ratio of 2.62 would be a frequency of 12, which does not have as high a component in the transform.

Further evidence against Odlyzko's conjecture comes from the fact that when his argument is applied to other sets it gives ratios nowhere near the ratios computed from the sets themselves. Odlyzko reasoned as follows: since the peaks and valleys are empirically found to be in geometric progressions, assume that the ratio from a peak to the next valley and the ratio from that valley to the next peak are both r. Then two peaks and the following valley are in the ratios $1:r^2:r^3$. If the peaks and valley are in the pattern of the excluded subset, then the peaks would reinforce the valley. In general, if the excluded subset {x, x + ay, x + by} is in the pattern $1:r^2:r^3$ then r must satisfy $a(r^3 - r^2) = (b - a)(r^2 - 1)$, or dividing out the solution r=1, $ar^2 - (b - a)r - (b - a) = 0$.

A closely related alternative is to assume that the excluded subset is formed by a peak with the *second* following peak excluding the next valley, so the ratios would be $1:r^4:r^5$. The corresponding equation is $a(r^5 - r^4) = (b - a)(r^4 - 1)$, or (again dividing out the solution r=1) $ar^4 - (b - a)(r^3 + r^2 + r + 1) = 0$. The following table shows these different equations and the predicted peak ratio r^2, computed from the positive root r. From the table we see that this kind of argument does not work to predict the peak ratios (furthermore, in a brief inspection of other ratios such as $1:r^2:r^5$ I found none that were predictive of the ratios found by generating the sets themselves.)

set	Equation from $1:r^2:r^3$	peak ratio (r^2)	Equation from $1:r^4:r^5$	peak ratio (r^2)	Actual ratio
1,2	$r^2 - r - 1 = 0$	2.62	$r^4 - (r^3 + r^2 + r + 1) = 0$	3.72	2.52
1.3	$r^2 - 2r - 2 = 0$	7.46	$r^4 - 2(r^3 + r^2 + r + 1) = 0$	8.85	4.1
1,4	$r^2 - 3r - 3 = 0$	14.37	$r^4 - 3(r^3 + r^2 + r + 1) = 0$	15.91	2.2
1,5	$r^2 - 4r - 4 = 0$	23.31	$r^4 - 4(r^3 + r^2 + r + 1) = 0$	24.94	6.3
2,3	$2r^2 - r - 1 = 0$	1	$2r^4 - (r^3 + r^2 + r + 1) = 0$	1.82	1.72

2,5	$2r^2 - 3r - 3=0$	4.78	$2r^4 - 3(r^3 + r^2 + r + 1)=0$	6.05	3.33
3,4	$3r^2 - r - 1 =0$	0.59	$3r^4 - (r^3 + r^2 + r + 1)=0$	1.27	1.47
3,5	$r^2 - 2r - 2 =0$	1.48	$3r^4 - 2(r^3 + r^2 + r + 1)=0$	2.41	1.40
4,5	$4r^2 - r - 1 =0$	0.41	$4r^4 - (r^3 + r^2 + r + 1)=0$	1	1.347

The strongest peak in the Fourier transform is at a frequency of 13, or a ratio of $\sqrt[13]{120000}$ =2.46. Other strong frequencies seen in the graph are 25 and 47, which are most likely the first and third harmonics of the fundamental frequency; because of the wideness of the peaks, it is hard to tell exactly where the harmonics should be.

On the second Fourier transform graph of the set 1,2, derived from the last 1/5 of the graph at 5 times the resolution, the fundamental frequency is visible at 3 (which would be 15 on the first transform graph). The first and third harmonics are at approximate frequencies of 5 and 8, but no further frequency components (that would have been hidden in high frequency noise on the other graph) are apparent.

In the Fourier transform of the set 1,3 the fundamental frequency is obvious at 8. Small frequency peaks are visible at 19 and 30, although they could be due to random noise at the beginning of the set. In fact, the Fourier transform of the last section (of the computed part) of the 1,3 set shows that the only periodicity present in the set is the fundamental frequency around 2. This is, however, not a well-defined frequency, as can be seen from the wideness of the peak.

The set 1,4 itself doesn't show strong peaks—the exponent graph is practically a straight line, indicating a rather uniformly decreasing density. The Fourier transform bears this out in its uniformity; if there are any real frequency peaks, they are extremely wide. There is, though, one interesting thing about the graph: at a frequency of 13 the Fourier coefficient is very small compared to nearby ones, looking like an absorption line. The last fifth Fourier graph also shows a smaller coefficient at 3, which might be the representation of the absorption line in that graph. But because the coefficients in that graph are so small compared to other graphs (about 0.1 vs. 0.5), I suggest that all this absorption line shows is a random factor akin to the error involved in sampling a continuous distribution. A higher resolution treatment would probably show it to be an artifact.

1,5 is a more interesting set. From the graph of the set, the fundamental frequency appears to be about 6, and the Fourier transform has a peak at 6, albeit a wide one. In the graph of the set, it looks like there are weaker frequencies than 6. The Fourier transform has a narrow peak at the first harmonic at 13 and a wider (and lower) peak around 27, near the third harmonic of the fundamental frequency.

Moving on to the 2,3 set, from the graph of the set we see that it has much better defined peaks and a higher frequency than the 1,n sets above. In about the last half of the graph there are even prominent subpeaks, clearly demonstrating even higher frequency components. From the Fourier coefficient graphs, the fundamental frequency of 22 is rather well-defined (the peak is narrow). This first harmonic at 44 is quite strong, and there are perhaps even slight peaks at the harmonics at 66 and 87.

These harmonics are also visible in the Fourier transform graph of the last 1/5 of the set. It has coefficient peaks around 4, 9, and 22, which would be 20, 45, and 110 when scaled to the larger first transform graph. Because these are all harmonics of the fundamental frequency, the structure of the peaks is probably again due to the fundamental frequency and its harmonics.

The 2,5 set has the strangest looking peaks of any of the sets—wide peaks with two subpeaks each. The Fourier transform has peaks at 9, 11, 19, and 29, with 29 being very spread out. These peaks also appear as the peaks at 2, 4, and 6 in the transform of the last 1/5 of the generated subset. This set, with frequencies 9 and 11, is one of the two I found with strong frequencies that aren't all harmonics (or very nearly so) of one fundamental frequency (the other is the 3,4 set below).

3,4 has a high fundamental frequency of 31, but unlike the 2,3 set, the higher harmonics appear weak in the whole subset Fourier transform; only the first harmonic shows up at all (at 61). But because the peaks break into subpeaks towards the end, the Fourier transform of the end of the graph shows more strong frequencies. It has the highest coefficient at a frequency of 6 (which is the 31 peak on the other transform), and harmonics at frequencies of 12 and 18.

The frequencies of 33, 39, and 45 are also especially strong on the end transform graph. Yet they are nowhere near being integer multiples of the fundamental frequency of 6; instead they are closest to $5\frac{1}{2}$, $6\frac{1}{2}$, and $7\frac{1}{2}$ times the frequency, half integer multiples of it.

Frequencies for the 3,5 set Fourier transform are quite spread out; there is only a relatively small peak at the frequency of 36; this is the frequency of the peaks that only begin to be visible around 500,000. This peak is much more evident in the Fourier transform of the end of my subset, where it appears at 7 with the first harmonic very weak at 14.

Finally, the 4,5 set has, like 2,3 and 3,4, a finer structure visible at the end of the subset I generated. Also like those sets, it has a high fundamental frequency, 39 (8 on the end 1/5 graph). The finer structure shows up as higher frequency harmonics around 15, 23, 53, and 60 on the second graph.

Analysis of the $\sqrt{n \log n}$ density conjecture

If Gerver's conjecture about the density is correct, then the top graphs in figures 1 through 9 should all have their bases around .5, or should at least be heading that way. Odlyzko's 1,2 set fits the conjecture very well, with the valleys already having exponents around .503.

The other sets' density exponents are all above .5 and decreasing, but it is hard to tell if they are approaching .5 or another limit, or if they will continue to fall towards zero. The exponents themselves (bottom graphs) all appear to be levelling off. Perhaps further computation of the sets or a different theoretical approach can help decide Gerver's conjecture.

References

1. Paul Erdös and Paul Turan, “On Some Sequences of Integers,” *Journal of the London Mathematical Society*, v. 11 (1936), pp. 261–264.

2. Joseph Gerver, “Irregular Sets of Integers Generated by the Greedy Algorithm,” *Mathematics of Computation*, v. 40 (1983), pp. 667–676.

3. A. M. Odlyzko and R. P. Stanley, unpublished Bell Laboratories report, cited in Gerver [2].