Finite Analogs of Szemerédi's Theorem

Doron ZEILBERGER[1] and *Paul RAFF*[1]

Szemerédi's Celebrated Theorem

One of the crowning achievements of combinatorics is

Szemerédi's Theorem ([S]): Given an integer $N \geq N$ and an integer $k \geq 3$, Let $r_k(N)$ denote the size of any largest subset S of $[N] := \{1, 2, \ldots, N\}$ for which there are **no** subsets of the form

$$\{i, i+d, i+2d, \ldots, i+(k-1)d\} \quad (i \geq 1, 1 \leq d < \infty) \quad ,$$

then $r_k(N) = o(N)$.

The depth and mainstreamness of this deep theorem can be gleaned by the fact that at least four Fields medalists (Klaus Roth, Jean Bourgain, Tim Gowers, and Terry Tao) and at least one Wolf prize winner (Hillel Furstenberg) made significant contributions.

This article is get another such contribution, and while it may not have the "depth" of the work of the above-mentioned human luminaries, it does have *one* advantage over them. We "cheat" and use a computer. It is true that, so far, we can only talk about *finite* analogs, but we do believe that the present approach could be eventually extended to sharpen the current rather weak bounds.

More Specifically, we prove:

Finite version of Szemerédi's Theorem ([S]): Given an integer $N \geq 2$ and integers $k \geq 3$, $D \geq 1$, let $R_{k,D}(N)$ denote the size of any largest subset S of $[N] := \{1, 2, \ldots, N\}$ for which there are **no** subsets of the form

$$\{i, i+d, i+2d, \ldots, i+(k-1)d\} \quad (i \geq 1, 1 \leq d \leq D) \quad ,$$

then there exists a rational number $\alpha_{k,D} = P_{k,D}/Q_{k,D}$ such that

$$\lim_{N \to \infty} \frac{R_{k,D}(N)}{N} = \alpha_{k,D} \quad .$$

We have computed $\alpha_{k,D}$ for small k and D in the table below. Paul please put table here

[1] Department of Mathematics, Rutgers University (New Brunswick), Hill Center-Busch Campus, 110 Frelinghuysen Rd., Piscataway, NJ 08854-8019, USA. `[praff,zeilberg] at math dot rutgers dot edu` , `http://www.math.rutgers.edu/~[praff,zeilberg]` . July 2, 2009. Accompanied by Maple package `ENDRE` and Mathematica package `X.m` downloadable from `http://www.math.rutgers.edu/~zeilberg/mamarim/mamarimhtml/szemeredi.html` .

The work of DZ was supported in part by the USA National Science Foundation.

It turns out that even more is true. $R_{k,D}(N)$ is a piecewise linear function and for $i = 1, \ldots, Q_{k,D}$ there exist integers $a_{k,D,i}$ between 0 and $P_{k,D} - 1$ such that

$$R_{k,D}(Q_{k,D}n + i) = P_{k,D}n + a_{k,D,i} \quad .$$

Our proof is algorithmic, and we show how to find these explicit expression using **rigorous experimental mathematics**.

Note that $a_{k,D}$ is a non-inreasing sequence in D, and Szemerédi's theorem is equivalent to the statement that

$$\lim_{D \to \infty} a_{k,D} = 0 \quad .$$

A Wordy Formulation

Every subset S of $[1, N] = \{1, 2, 3, \ldots, N\}$ corresponds to an N-letter word in the alphabet $\{0, 1\}$ defined by $w[i] = 1$ off $i \in S$. S as an arithmetical progression of size k if there is an *Equidistant Letter Sequence* in the sense of the **Bible Code** of the word 1^k. Denoting by 2 a place where the occupying letter may be either 0 or 1, we can say that the $R_k(N)$ of Szemerédi's theorem defined above asks to find the maximal number of words avoiding the infinitely many patterns

$$(12^d)^{k-1}1 \quad , \quad 0 \leq d < \infty.$$

Analogously, the $R_{k,D}(N)$ of the Finite-Version Szemerédi's theorem defined above asks we need to find the maximal number of words avoiding the finitely many patterns

$$(12^d)^{k-1}1 \quad , \quad 0 (\leq d \leq D).$$

Define the *weight* of a word w to be $t^{length} x^{\#of,1s}$. Let $F_{k,D}(x,t)$ be the weight-enumerator of all binary words avoiding the $D+1$ patterns $(12^d)^{k-1}1 \quad , \quad (0 \leq d \leq D)$. We will soon see that $F_{k,D}(x,t)$ is a rational function in (x,t).

Let's treat the more general case of an *arbitrary* set of *generalized patterns*. But let's first define it.

Definition: A *generalized pattern* is a word in the alphabet $\{0, 1, 2\}$, where 2 stands for "space".

Now let's say what it means to *contain* a pattern.

Definition: A word $w = w_1 w_2 \ldots w_n$ in the alphabet $\{0, 1\}$ contains the pattern $p = p_1 p_2 \ldots p_m$ if there exists a place i $(1 \leq i \leq n - m + 1$ such that

$$w_{i+j-1} = p_j \quad , if \quad p_j \neq 2 \quad , \quad j = 1, \ldots, m \quad .$$

For example, the word 011101101 contains the pattern 12221 (with $i = 3$).

A word w avoids a set of generalized patterns P if w avoids all the members of P.

Of course the same definitions, and the arguments below, equally hold for an arbitrary finite alphabet, where we can use SPACE instead of 2, but for the sake of definiteness and simplicity, we will stick to the 2-letter alphabet $\{0,1\}$ and denote SPACE by 2.

The General Problem

Consider a finite alphabet A together with a symbol SPACE not in A. We are interested in weight-enumerating the set of words that avoid a set of patterns P, according to the weight

$$weight(w_1 w_2 \dots w_n) = x[w_1]x[w_2]\cdots x[w_n] \quad ,$$

where $x[a]$ ($a \in A$) are *commuting indeterminates*. For example, $weight(PAUL) = x[P]x[A]x[U]x[L] = x[A]x[L]x[P]x[U]$, $weight(DORON) = x[D]x[N]x[O]^2x[N]$.

Let F be the weight-enumerator (sum of weights of its members, a formal power series in the variables $\{x[a], a \in A\}$) of the the set of such words (that avoid P), let's call it, for reasons to shortly become clear, $S[P,\emptyset]$. A word belonging to it is either empty, or else starts with one of the letters of our alphabet. If you chop that letter, what remains is a shorter word in $S[P,\emptyset]$, but with *more* conditions, it can not *start* with a "chopped pattern" obtained by chopping-off the first letter for all those patterns of P that happen to start with that letter or with SPACE.

This motivates the following definition.

Given a word or pattern $w = w_1 w_2 \dots w_n$, let $BEHEAD(w) := w_2 \dots w_n$. For example, $BEHEAD(DORON) = ORON$, $BEHEAD(PAUL) = AUL$, $BEHEAD(__L_OVE) = _L_OVE$ (here _ denots SPACE).

Let P be a set of patterns, and let a be any letter of our alphabet A, then

$$P/a := \{BEHEAD(p) \,|\, p \in P, p_1 = a \quad or p_1 = _\} \quad .$$

For example, if the alphabet is $\{0,1\}$, and

$$P = \{000, 0_0_0, 0__0__0, 111, 1_1_1, 1__1__1, __101\} \quad ,$$

then

$$P/0 = \{00, _0_0, _0__0, __101\} \quad ,$$

$$P/1 = \{11, _1_1, _1__1, _101\} \quad .$$

So if w belongs to our set $S[P,\emptyset]$ and it starts with the letter a then the chopped word obviously also avoids P but in addition avoids P/a at the very beginning. This motivates us to make yet another

Definition: Let P and P' be sets of patterns. The set $S[P,P']$ consists of all words avoiding the patterns in P and in addition avoiding the patterns P' at the very beginning.

Since every word in $S[P, P']$ must be either empty or begin with one of the letters of our alphabet a, we have the linear equation, for the weight-enumerators $F[P, P'](\{x_a\})$,

$$F[P, P'] = 1 + \sum_{a \in A} x_a F[P, P/a \cup P'/a] \quad .$$

If P' (or P) contains an empty pattern, then of course we have the **initial condition** $F[P, P'] = 0$, since not even the empty word avoids the empty word as a factor.

Of course, we only care about $F[P, \emptyset]$, but in order to compute it, we need to set up a system of linear equations featuring lots of $F[P, P']$ with various P', but nevertheless finitely many of them. Since the P' of the right side always contain shorter patterns, and eventually we can use the initial conditions, we get as many equations as unknowns. Also, since we know from the outset that a solution exists (from the combinatorics), it follows that the system of equations is non-singular, and by Cramer's rule that we have a *rational function* in the variables $\{x[a]\}$ $(a \in A)$.

Specializing

Going back to the Szemerédi scenario, we have a two-letter alphabet $\{0, 1\}$ with weight $x[0] = t, x[1] = zt$. For *any* set of forbidden patterns, in particular, those that avoid arithmetical progression of size k with spacings $\leq d$, the generating function is of the form

$$R(x, t) = \frac{P(x, t)}{Q(x, t)} \quad ,$$

where t keeps track of the length and x keeps track of the number of 1s.

Expanding $R(x, t)$ as a power-series of t, we get

$$R(x, t) = \sum_{n=0}^{\infty} r_n(x) t^n \quad ,$$

and $r_i(x)$ is a polynomial whose *degree* (in x) is the largest number 1's in an n-letter word. By looking at the monomials of the denominator, $Q(x, t)$ and searching for the monomial $x^i t^j$ with *largest* ratio $r := i/j$, we get that asymptotically the largest number of 1's in an n-letter word in $\{0, 1\}$ is nr, and more precisely we have the behavior described above for $R_{k,D}$.

An Experimental-Yet-Rigorous Shortcut

Solving a huge system of linear equations with *symbolic* coefficients is very time- and memory-consuming. Restricting attention to the alphabet $\{0, 1\}$, and letting $f(P, P')(n)$ be the maximum number of 1's in an n-letter word that avoids the patterns in P and in addition, at the beginning, the patterns in P', we get:

$$f(P, P')(n) = max(f(P, P/0 \cup P'/0)(n-1), f(P, P/1 \cup P'/1)(n-1) + 1) \quad .$$

We ask the computer to *first* find the **scheme**, in terms of a binary tree where the left-child of P' is $P/0 \cup P'/0$ and its right-child is $P/1 \cup P'/1$. Then we ask the computer to *crank-out* lots of data, say, the first 300 terms (or whatever is needed), and then the computer automatically *guesses* explicit expressions in the form

$$R_{k,D}(Q_{k,D}n + i) = P_{k,D}n + a_{k,D,i} \quad .$$

Once guessed, the computer *automatically* gives a fully rigorous proof, *a posteriori* by checking all the above equations, this time *symbolically*.

Supporting Software

All this is implemented in the Maple package `ENDRE`, and the Mathematica package `X.m`. See the webpage `http://www.math.rutgers.edu/~zeilberg/mamarim/mamarimhtml/szemeredi.html` for these packages, as well as sample input and output.

Exact Enumeration

From Sloane's point of view, it is interesting to crank-out as many terms as possible of $R_{k,D}(n)$, both for their own sake, and also because they offer upper bounds for $R_k(n)$. The interesting and efficient methods of the recent paper [GGK], that treats $R_3(n)$, may be useful.

Pipe dreams

For a fixed k, $a_{k,D}$ gets harder and harder to compute as D gets larger and larger, **but** we believe that a clever analysis of the max equations, might lead, one day, to a *quantitative* understanding of how $a_{k,D}$ decreases with D, that may, who knows? lead to an easier proof of Szemerédi's theorem, and more importantly, improved lower bounds on $R_k(n)$.

What we are essentially doing is solving a system of recurrences of the form

$$f_i(n) = max(f_{a(i)}(n-1) + 1, f_{b(i)}(n-1)) \quad ,$$

for N sequences $\{f_i(n)\}$, $i = 1..N$. Here $a(i)$ $b(i)$ are some functions from $[1, N]$ to $[1, N]$. It may be worthwhile to study such recurrences *for their own sake*, abstractly, and come up with a study of the asymptotic density as they depend on $a(i), b(i)$. Hopefully, we can get some general theorems, and since $a(i)$ and $b(i)$ are *arbitrary*, you have lots of elbow-room for induction.

Then we check that the *particular* $a(i)$, $b(i)$ that shows up satisfy some general conditions that would enable us to get upper bounds on $a_{k,D}$ as a function of D.

References

[GGK] W. Gasarch, J. Glenn, C.P. Kruskal, *Finding large 3-free sets I: The small n case*, Journal of Computer and System Sciences **74** (2008), 628-655. `http://www.cs.loyola.edu/ jglenn/Papers/3apI.pdf`

[S] E. Szemerédi, *On sets of integers containing no k elements in arithmetic progression*, Acta Arith. **27**(1975), 199-245.