The Erdős-Hanani Conjecture via Talagrand's Inequality

Joel Spencer

1 History.

For $2 \leq l < k < n$ let $m(n,k,l)$ denote the maximal size of a family F of k-element subsets of $\{1,\ldots,n\}$ with the property that no l points lie in more than one $A \in F$. Similarly, let $M(n,k,l)$ denote the minimal size of a family F of k-element subsets of $\{1,\ldots,n\}$ with the property that every l points lie in at least one $A \in F$. Elementary counting arguments give

$$m(n,k,l) \leq \frac{\binom{n}{l}}{\binom{k}{l}} \leq M(n,k,l)$$

Equality is achieved exactly when there is a family F so that every l points are in precisely one $A \in F$, what is called a tactical configuration. (For example, the case $k = 3, l = 2$ yields the well known Steiner Triple Systems.) In 1963 Paul Erdős and Haim Hanani [3]conjectured that these bounds were asymptotically achievable, more precisely that for any *fixed* k,l

$$\lim_{n\to\infty} m(n,k,l)\frac{\binom{k}{l}}{\binom{n}{l}} = 1 = \lim_{n\to\infty} M(n,k,l)\frac{\binom{k}{l}}{\binom{n}{l}} \tag{1}$$

This conjecture was proven in 1985 by Vojtech Rödl[4].

It will be convenient for us to formulate the Erdős-Hanani conjecture in hypergraph terms. Let G be the complete l-graph on n points. Then $m(n,k,l)$ is the maximal number of disjoint k-cliques that may be packed into G, and $M(n,k,l)$ is the minimal number of k-cliques that cover G.

The proof of Rödl was remarkable on several levels. It had long seemed clear that the conjecture called for a probabilistic argument. Rödl's argument was indeed probabilistic, but with a twist. Our description is given in more detail in [1]. Rather than randomly select all his k-cliques at once he randomly selected just so many k-cliques so that an l-edge was covered on average ϵ times, where ϵ was fixed and small. Some of those k-cliques would overlap but these contributed basically an ϵ^2 factor so that a proportion $(1-\epsilon)$ of the k-cliques chosen were mutually disjoint, indeed, isolated. Rödl then iterated this process, a notion now sometimes called the Rödl Nibble. The total number of iterations was a large but fixed integer. Let

G_i be the l-graph of edges not yet covered after i iterations. It turns out that the proportion of l-sets in G_i is roughly $p_i = e^{-i}$. One now wants to select randomly k-cliques from G_i. Here it is critical that in some sense G_i behaves like a random l-graph with edge probability p_i. After all, if one only knew the number of edges of G_i it might not have any k-cliques. The notions of "random-like" behavior that Rödl used were then studied by many authors and the notion of *quasirandomness* was developed. It turns out that there is quite a robust notion. The best development was given by Chung, Graham and Wilson[2] in 1989 in which several notions of quasirandomness were shown to be independent. Here is, roughly, one of the notions of quasirandomness when $l = 2$. We say G is quasirandom with probability p if for each fixed s almost all choices of distinct $x_1, \ldots, x_s$ have asymptotically $p^s n$ vertices z which are joined to all $x_1, \ldots, x_s$.

In this work we give a new proof of the Erdős-Hanani conjecture. The Rödl nibble will remain the same. However we will replace the notion of quasirandomness by a stronger notion that we will dub *superquasirandomess*. For $l = 2$, roughly, G would be superquasirandom with probability p if for each fixed s *all* choices of distinct $x_1, \ldots, x_s$ have asymptotically $p^s n$ vertices z which are joined to all $x_1, \ldots, x_s$. As with Rödl's proof the key will be an induction that if G_i is superquasirandom then G_{i+1} will be superquasirandom . To do this we use a new and far reaching probabilistic inequality of M. Talagrand[5], which is the subject of the next section.

From the logic of Rödl's result it is clear that the Erdős-Hanani conjecture must remain true even if we allow $k = k(n) \to \infty$ extremely slowly. It seemed that the nature of quasirandomness made it difficult to quantify that result. With these methods, however, is seems (though we don't do it here) possible to show that conclusion 1 still holds for, say, l constant and $k = \ln^c n$ for $c = c(l)$ appropriately small.

2 Talagrand.

Like many useful tools in the Probabilistic Method Talagrand's Inequality can be stated using only elementary probability. Let $\Omega = \prod_{i=1}^{w} \Omega_i$ be a product probability space, we write x_i for the i^{th} coordinate of an $x \in \Omega$. For $A \subset \Omega$ and t an arbitrary positive real we define Ω_t by saying $y \in \Omega_t$ if and only if for all $\alpha_1, \ldots, \alpha_w$ there exists an $x \in A$ with

$$\sum_{x_i \neq y_i} \alpha_i < t \sum_{1 \leq i \leq w} \alpha_i^2$$

Talagrand's Inequality:

$$\Pr[A](1 - \Pr[A_t]) \leq e^{-t^2/4} \tag{2}$$

The notation takes some getting used to. Suppose $\Omega_i = \{0, 1\}$ with the uniform distribution so that Ω is the Hamming w-cube with uniform distribution. If $y \in A_t$ then, taking all $\alpha_i = 1$, y must be within Hamming distance $t\sqrt{w}$ of A. In this sense the inequality is remniscent of isoperimetric inequalities. However there are certainly further conditions on membership in A_t that make this inequality far stronger.

We [1] now derive a corollary of this Inequality that will be more easy for us to apply. Let $h : \Omega \to R$. We call h *Lipschitz* if $|h(x) - h(y)| \leq 1$ for all $x, y \in \Omega$ which differ in only one coordinate. For $f : N \to N$ (e.g., $f(b) = b$) we call h f-certifiable if whenver $h(x) \geq s$ there exists an index set $I \subseteq \{1, \ldots w\}$ with $|I| \leq f(s)$ so that all $y \in \Omega$ that agree with x on the coordinates I have $h(y) \geq s$. (I.e., there will be a set of coordinate values of size $f(s)$ that will certify that $h \geq s$.) Let h satisfy the above and consider the random variable $X = h(x)$.

Corollary. Under the above assumptions and for all b, t

$$\Pr[X \leq b - t\sqrt{f(b)}] \Pr[X \geq b] \leq e^{-t^2/4} \tag{3}$$

Proof. Set $A = \{x : h(x) < b - t\sqrt{f(b)}\}$. Now suppose $h(y) \geq b$. We claim $y \notin A_t$. Let I be a set of indices of size at most $f(b)$ that certifies $h(y) \geq b$ as given above. Define $\alpha_i = 0$ when $i \notin I$, $\alpha_i = 1$ when $i \in I$. If $y \in A_t$ there exists a $z \in A$ that differs from y in at most $t\sqrt{f(b)}$ coordinates of I though at arbitrary coordinates outside of I. Let y' agree with y on I and agree with z outside of I. By the certification $h(y') \geq b$. Now y', z differ in at most $t\sqrt{f(b)}$ coordinates and so, by Lipschitz,

$$h(z) > h(y') - t\sqrt{f(b)} \geq b - t\sqrt{f(b)}$$

but then $z \notin A$, a contradiction. So $\Pr[X > b] \leq 1 - \Pr[A_t]$ so from 2.

$$\Pr[X < b - t\sqrt{f(b)}] \Pr[X \geq b] \leq e^{-t^2/4}$$

As the right hand side is continuous in t we may replace $<$ by $\leq$ giving the Corollary. □

[1] The development of Talagrand's Inequality into a form easily used by combinatorialists was a joint effort of several mathematicians visiting IMA in Fall 1993, including Svante Janson, Eli Shamir and Michael Steele

In applications one often takes b to be the median so that for t large the probability of being $t\sqrt{f(b)}$ under the median goes sharply to zero. But it works both ways, by parametrizing so that $d = b - t\sqrt{f(b)}$ is the median one usually gets $b \sim d + t\sqrt{f(d)}$ and that the probability of being $t\sqrt{f(b)}$ above the median goes sharply to zero. Martingales, via Azuma's Inequality, generally produce a concentration result around the mean of X while Talagrand's Inequality yields a concentration result about the median. Means tend to be easy to compute, medians notoriously difficult, but our tight concentration result will allow us to show that the mean and median are not far away.

3 The Critical Lemma.

We first quantify our notion of superrandomness adding a "tolerance" δ and limiting the scope of the conditions.

Definition. We call an l-graph G (p, k, δ)-superquasirandom if for every $l - 1 \leq t \leq k - 1$ and every set $x_1, \ldots, x_t$ of distinct vertices of G

$$1 - \delta < \frac{|N_G(x_1, \ldots, x_t)|}{(n-t)p^{\binom{l-1}{t}}} < 1 + \delta \tag{4}$$

where $N_G(x_1, \ldots, x_t)$ denotes those y so that all l-sets consisting of y and $l - 1$ of the the x's are hyperedges of G.

Let G be (p, k, δ)-superquasirandom as above. Let $\epsilon > 0$ be given. Let K be a random set of k-cliques of G where each k-clique of G is placed in K with probability

$$q = \frac{\epsilon p \binom{n}{l} / \binom{k}{l}}{\binom{n}{k} p^{\binom{k}{l}}}$$

Let H be the l-graph of hyperedges of G that are not in any clique $C \in K$. That is, from G delete each k-clique with independent probability q, producing the random l-graph H. Lets first motivate the choice of q. As G behaves like a random l-graph with probability p it has roughly $\binom{n}{k} p^{\binom{k}{l}}$ k-cliques and so roughly $p\epsilon \binom{n}{l} / \binom{k}{l}$ cliques are chosen for K, these cover *with repetition* $p\epsilon \binom{n}{l}$ edges. But G has roughly $p\binom{n}{l}$ edges so each edge is covered on average ϵ times. The number of times an edge is covered is given roughly by a Poisson distribution with mean roughly ϵ and an edge "survives" into H with probability roughly $e^{-\epsilon}$. Thus one would hope that H would be superquasirandom with $p' = pe^{-\epsilon}$ and this, basically, is the case.

Lemma. Fix $2 \leq l < k$ and fix $p, \epsilon \in (0,1)$. Let Gbe (p,k,δ)-superquasirandom and let K be the random family of k-cliques and H the random graph of surviving hyperedges as defined above. *Then, with high probability, H will be $(p', k, O^*(\delta))$-superquasirandom* where $p' = pe^{-\epsilon}$ and $O^*(\delta)$ is defined below.

Special Notation: $O^*(\delta)$ represents a function which for $\delta > 0$ sufficiently small is less than $c\delta$, c dependent on (at most) p, ϵ, k, l. Note that

$$(1 + O^*(\delta))^c = 1 + O^*(\delta) \text{ and } \exp[O^*(\delta)] = 1 + O^*(\delta) \tag{5}$$

We shall use these both in the argument.

Proof. Fix t with $l - 1 \leq t \leq k$ and fix distinct $x_1, \ldots, x_t$. Consider the random variable

$$X = |N_H(x_1, \ldots, x_t)|$$

Our object will be to show

$$\Pr\left[X \neq (n-l)(p')^{\binom{t}{l-1}}(1 + O^*(\delta))\right] = o(n^{-t}) \tag{6}$$

From this the Lemma will immediately follow since there are only $\binom{n}{t} = O(n^t)$ choices of $x_1 \ldots, x_t$ and the probability of H not being $(p', k, O^*(\delta))$-superquasirandom would be $\sum_t O(n^t)o(n^{-t}) = o(1)$. In fact we will get that this failure probability is at most e^{-cn}.

We begin by estimating $E[X]$. For $y \in N_G$ (for convenience we omit the arguments $x_1, \ldots, x_t$ henceforth) let I_y be the indicator random variable for $y \in N_H$ so that

$$X = \sum_{y \in N_G} I_y$$

Equation 4 bounds $|N_G|$. By successive applications of 4 every l-edge of G is in between $(1-\delta)^{k-l}A$ and $(1+\delta)^{k-l}A$ k-cliques where $A = \binom{n-l}{k-l}p^{\binom{k}{l}-1}$. (I.e., A is the expected number of such extensions in a random graph.) Similarly every $l+1$ vertices are in at most

$$B = (1+\delta)^{k-l-1}\binom{n-l-1}{k-l-1}p^{\binom{k}{l}-l-1}$$

k-cliques. For $y \in N_G$ let C_y denote the number of k-cliques of G containing at least $l-1$ of the $x_1, \ldots, x_t$. Inclusion-Exclusion gives

$$\binom{t}{l-1}(1-\delta)^{k-l}A - \binom{t}{l-1}^2 B \leq C_y \leq \binom{t}{l-1}(1+\delta)^{k-l}A$$

As $B = O(n^{k-l-1}) = o(A)$ and $(1 \pm \delta)^c = 1 + O^*(\delta)$

$$C_y = \binom{t}{l-1} A(1 + O^*(\delta))$$

But $y \in N_H$ exactly when none of these C_y cliques are placed in K and this occurs with probability $(1-q)^{C_y}$. We have chosen q so that $e^{-qA} = e^{-\epsilon}$. As $q = O(n^{-1})$ approximation of $1-q$ by e^{-q} yields a negligible error. Then

$$(1-q)^{C_y} \sim e^{-qC_y} = e^{-\epsilon\binom{t}{l-1}(1+O^*(\delta))} = e^{-\epsilon\binom{t}{l-1}}(1 + O^*(\delta))$$

Factoring in the uncertainity of $|N_G|$ still

$$E[X] = (n-l)(p')^{\binom{t}{l-1}}(1 + O^*(\delta)) \tag{7}$$

Now we employ the new technique, Talagrand's Inequality. Let $C_1 \ldots, C_w$ denote the k-cliques of G containing at least one $y \in N_G$ and at least $l-1$ of the $x_1 \ldots, x_t$. For $1 \leq j \leq w$ set $\epsilon_j =$ yes if $C_j \in K$ and $\epsilon_j =$ no if $C_j \notin K$. Then $\Omega = \{(\epsilon_1, \ldots, \epsilon_w)\}$ is a product probability space. Define $h : \Omega \to R$ by $h(x) = |N_G| - |N_H|$, noting that the choices on these cliques determines N_H. Changing ϵ_j from no to yes can delete at most $k-l+1$ vertices from N_H, the extreme being if C_j contains $k-l+1$ vertices from N_G and $l-1$ from $x_1 \ldots, x_t$. Set $h^*(x) = h(x)/k$ so that h^* is Lipschitz with room to spare. If $h^*(x) \geq s$ then $h(x) \geq ks$. Then there are sk vertices $y \in N_G$ (maybe more) that are not in N_H. To each there is a clique $C \in K$ which contains y and at least $l-1$ of the $x_1 \ldots, x_t$. Let I be the set of indices of these cliques. (Note: if there are more than sk vertices y take just sk of them. To each there may be many such cliques C but take just one of them.) Then $|I| \leq sk$ and I provides a certificate that $h^*(x) \geq s$ as once those cliques are placed in K their corresponding vertices cannot be in N_H. That is, h^* is f-certifiable where $f(s) = sk$. By 3

$$\Pr[X < b - \lambda\sqrt{kb}] \Pr[X > b] < e^{-\lambda^2/4} \tag{8}$$

for all b and all positive λ.

Take $\lambda = n^{.01}$. Then for any $0 \leq b \leq n$ (the range of X)

$$\Pr[X \leq b - n^{.52}] \Pr[X \geq b] < e^{-n^{.02}/4} \tag{9}$$

Let m denote the *median* of X. Applying 9 with $b = m$ and $b = m + n^{.52}$ gives

$$\Pr[|X - m| \geq n^{.52}] < 4e^{-n^{.02}/4} < e^{-n^{.015}}$$

The median is therefore quite close to the mean:

$$|E[X] - m| \leq E[|X - m|] \leq n^{.52} + ne^{-n^{.015}} \leq 2n^{.52}$$

From 7 $E[X] = \Theta(n)$ and $2n^{.52} = o(n)$ so

$$m = (n - l)(p')^{\binom{t}{l-1}}(1 + O^*(\delta))$$

and therefore

$$\Pr[X \neq (n - l)(p')^{\binom{t}{l-1}}(1 + O^*(\delta)] < \Pr[|X - b| > nO^*(\delta)] < e^{-cn\delta}$$

and so 6 is shown with plenty of room.

4 The Rest of the Proof.

We only sketch the remainder of the proof, which follows the ideas of Rödl almost directly. Fix $0 < \epsilon < 1$ and K a positive integer. (Think of ϵ small and K large.) For $0 \leq i \leq K$ define $p_i = e^{-i\epsilon}$. Set $\delta_0 = \delta$ and let $\delta_{i+1} = O^*(\delta_i)$, with the constant so as to make the Critical Lemma work with $p = p_i$. Note, as K is fixed, all $\delta_i = O^*(\delta)$

Lemma. Let $0 \leq i < K$ and let G_i be (p_i, k, δ_i)-superquasirandom . Then there exists a set K_i of k-cliques of G_i so that, letting G_{i+1} denote the graph of edges of G_i not in any clique of K_i

- G_{i+1} is $(p_{i+1}, k, \delta_{i+1})$-superquasirandom .
- $|K_i|$ is asymptotically $\epsilon_i p_i \binom{n}{l} / \binom{k}{l}$.
- A proportion at least $1 - 2\binom{k}{l}\epsilon$ of the cliques of K_i are isolated in that there is no other clique of K_i that has an edge in common with it.

Proof. We consider the random K generated as in the lemma of §3. That lemma gives that almost surely the first condition would hold. The size of K_i has a Binomial Distribution so that the second condition holds almost surely by basic tail estimates. For the third condition, fix a k-clique C of G_i and condition on it being in K_i. For C not to be isolated some other clique of G_i that overlaps with it must be put in K_i. There are $\binom{k}{l}$ edges e of C. Each edge is in an expected number $\sim \epsilon$ of other cliques of K_i. Thus the total number of overlapping cliques has expectation $\binom{k}{l}\epsilon$ so C has probability at most $\binom{k}{l}\epsilon$ of not being isolated. The expected number of non-isolated k-cliques is $\binom{k}{l}\epsilon$ times the expected number of k-clique and so with probability at least $.5 + o(1)$ the third property will hold.

As the first two properties holds almost surely, the three properties hold simultaneously with probability at least $.5 - o(1)$ which is positive and so, by the Probabilistic Method, there exists a set K_i with these properties. □

Now we prove the Erdős-Hanani Conjecture for the packing function $m(n,k,l)$. (In their original paper Erdős and Hanani showed that this would imply the conjecture for covering M and conversely.) For fixed ϵ, K we let G_0 be the complete l-graph on n vertices and set $p_0 = 1$ so that G_0 is certainly (p_0, k, δ_0)-superquasirandom , indeed we could have even taken $\delta_0 = 0$. We set $p_i = e^{-i\epsilon}$. Let $\delta_{i+1} = O^*(\delta_i)$ with the constant so that the above lemma holds. We pick $\delta = \delta_0$ sufficiently small so that all of the δ_i are in turn sufficiently small that the δ_{i+1} can be taken to be constant multiples of δ_i and so that, say, $\delta_K \leq 1$. We apply the Lemma for $0 \leq i < K$ giving graphs G_i and collections K_i of k-cliques. When $i < j$ the cliques in K_i, K_j are edge-disjoint since all the edges of K_i were removed from G_i and are not in G_j. Let K_i^I denote the isolated k-cliques of K_i. Then let K^I denote the union of the K_i^I, this is a family of edge disjoint k-cliques. The proportion of edges in G_K would be $e^{-K\epsilon}$ if G_K were random, as $\delta_K \leq 1$ we can say it has at most a proportion $2e^{-K\epsilon}$ of the $\binom{n}{l}$ original edges. At each stage K_i^I covers all but at most a proportion $2\binom{k}{l}\epsilon$ of the edges covered by K_i. So in total the proportion of edges missed by K^I is at most $2\binom{k}{l}\epsilon + 2e^{-K\epsilon}$. As

$$\lim_{\epsilon \to 0^+} \lim_{K \to \infty} 2\binom{k}{l}\epsilon + 2e^{-K\epsilon} = 0$$

we can select ϵ, K so that this proportion is arbitrarily small, completing the proof.

References

[1] N. Alon and J. H. Spencer, **The Probabilistic Method**, Wiley, 1991.

[2] F.R.K. Chung, R.L. Graham, and R.M. Wilson, Quasi-random graphs, *Combinatorica* **9** (1989), 345-362

[3] P. Erős and H. Hanani, On a limit theorem in combinatorial analysis, *Publ. Math. Decebren* **10** (1963), 10-13.

[4] V. Rödl, On a packing and covering problem, *European Journal of Combinatorics* **6** (1985), 69-78.

[5] Michel Talagrand, Concentration of Measures and Isoperimetric Inequalities in Product Space, preprint. (esp. Thm. 4.1.1 and Lemma 4.1.2)