Research Statement for Kevin P. Costello

My research interests lie in the intersection of Combinatorics, Probability, and Linear Algebra. Typically I am interested in problems involving a combinatorial object (e.g. a matrix, a graph, or a polynomial) whose parameters involve a large number of independent random variables. The goal is then to obtain some sort of estimate on the probability distribution of various parameters involving the object (e.g. the rank of the matrix, or the performance of some algorithm on the graph), particularly an asymptotic estimate as some underlying parameter tends to infinity. In the following sections I describe some of this work, as well as some future questions in this area I hope to address.

1 Random Discrete Matrices

Here the objects under consideration are $n \times n$ matrices. The entries of the matrix are treated as random variables, and we aim to estimate the probability that the matrix is singular. If the distribution functions of the entries are independent and continuous, the singularity probability is automatically 0. However, matrices which do not meet this requirement are common, particularly in computer science, and it is of interest to obtain some sort of estimate on the probability of singularity in this case.

Perhaps the simplest example of such a matrix is that of a random Bernoulli matrix. Let $M_n$ be an $n \times n$ matrix whose entries are independently equal to either 0 or 1, with probability 1/2 each. As matrices with two equal rows are singular, it is easy to obtain a lower bound of $\Omega(n^2 2^{-n})$ on the probability $M_n$ is singular. Obtaining good upper bounds, however, seems to be a much more difficult problem.

The first upper bound for the singularity probability was provided by Komlós in 1967 [16], who originally only proved that the singularity probability tended to 0 as the size of the matrix tended to infinity. Although he later [1] improved his methods to obtain an upper bound of $O(n^{-1/2})$, it took several decades before Kahn, Komlós, and Szemerédi [14] obtained an exponential upper bound on the singularity probability using Fourier analysis. Somewhat surprisingly, the best current bound of $(\frac{1}{2} + o(1))^n$ due to Bourgain, Tao, and Wood ([2], extending results of [24]) comes from application of the tools of Arithmetic Combinatorics, and in particular on the development of a Freiman-type inverse theorem classifying subspaces containing many points of the hypercube.

In [8], Tao, Vu, and I analyzed the singularity question for random symmetric matrices. This is both one of the simplest examples of matrices with dependent entries and one of the ones most frequently encountered in practice. We obtained an analogue of Komlós’ original result in this case by showing the singularity probability tended to 0 (the current best bound is $O(n^{-1/2})$ due to the results of [6]). However, we are as yet unable to extend the more powerful methods of [14, 24] (which rely heavily on the independence of the rows of the matrix in question) to this model, and our bounds are correspondingly far weaker due to this.

An alternative interpretation of these symmetric matrices is as the adjacency matrices of Erdős-Rényi random graphs (graphs where each edge is treated independently included with probability $p$). A further avenue of my research has focused on considering the adjacency matrices of sparse graphs, where each entry is equal to 1 with a probability $p(n)$ which tends to 0 as $n$ grows large. In [10], Vu and I analyzed the singularity question for this model, proving that $p = \frac{\ln n}{n}$ is a sharp threshold for nonsingularity. Above this threshold, the matrix is almost surely nonsingular. Below it, the graph almost surely has isolated vertices and the matrix thus has rows which are entirely 0 and is trivially singular. In [11], we extended the analysis further by giving an exact characterization of all dependent sets of rows in the case where $p = \Omega(\frac{\ln n}{n})$.

There are two main ways in which I hope to strengthen the above results in the future. One goal would be to extend the results of [10, 11] to cover more general classes of matrices. A particular case of interest for me is the case of random regular graphs (graphs chosen uniformly at random from the collection of all
If $k = 2$, it is not hard to show that the adjacency matrix will almost surely be singular. However, the case $k = 3$ (and that of any larger $k$) remains open. Computer evidence suggests that for $k > 2$ the matrix will almost surely be non-singular, as does the following partial result: For any $k > 2$ there is a $c = c(k)$ such that almost surely the adjacency matrix of a random $k$-regular graph on $n$ vertices has no dependencies involving fewer than $cn$ rows.

I would also like to obtain some sort of improvement to the upper bound on the singularity probability in the $p = 1/2$ case. Although an exponential bound in the lines of [14, 24] seems out of reach at the moment, it seems feasible to at least improve the power of $n$ in the polynomial bound, or even to prove a bound which is subpolynomial in the size of the matrix. I will go further into detail on this in the next section.

2 Polynomial Littlewood-Offord Problems

Littlewood and Offord [17], in their study of the roots of random polynomials, encountered the following question:

**Question 1.** Let $a_1, a_2, \ldots, a_n$ be real numbers satisfying $|a_i| \geq 1$. What is the maximum number of sums of the form

$$a_1 \pm a_2 \pm \cdots \pm a_n$$

that can lie in any single interval of length 1?

They obtained a bound of $O(\frac{\ln n}{\sqrt{n}}n)$, which was later improved to an exact bound of $\binom{n}{\lfloor n/2 \rfloor}$ by Erdős (with equality holding when the $a_i$ are all equal in magnitude). By using Stirling’s approximation and rescaling, we can restate Erdős’ result probabilistically as follows:

**Theorem 1.** Let $a_1, \ldots, a_n$ be nonzero constants, and let $x_1, \ldots, x_n$ be independently and uniformly drawn from $\{-1, 1\}$. Then

$$\sup_c P(\sum_{i=1}^{n} a_ix_i) = c = O(n^{-1/2}).$$

In this interpretation the $-1/2$ in the exponent is not surprising: If the $a_i$ are small integers, then the sum can be viewed as a random walk which will typically be about $\sqrt{n}$ away from 0 at time $n$.

Although Theorem 1 is in some sense sharp, there have been many subsequent improvements which all have a similar flavor: If we forbid some sort of structure from appearing in the $a_i$, then we can obtain an improved bound on the probability. For example, Sárközy and Szemerédi [23] showed that if we exclude the structure $a_i = a_j$ (i.e. require distinct $a_i$), then the maximum concentration drops to $n^{-3/2}$. Again this is unsurprising: If we assume the $a_i$ to be integers, their requirement ups the standard deviation of the sum to $n^{3/2}$. The same intuition holds for the following result of Halász: If we exclude the structure $a_i + a_j = a_k + a_l$, then the concentration drops to $n^{-5/2}$. More recently, Tao-Vu [26] and Rudelson-Vershynin [22] have generalized all of these results to obtain an inverse variant on Theorem 1 that almost completely classifies those coefficients leading to a linear form having polynomially large concentration on a single value: The only way this can happen is if the $a_i$ lie almost entirely in a structure known as a generalized arithmetic progression.

Results along this line are critical to the study of the behavior of $M_n$, as many of the properties of a matrix (e.g. the determinant, or the distance from one row to the span of the remaining rows) can be expressed as linear forms in the entry of any particular row or column of the matrix. Motivated originally by similar questions for a random symmetric matrix (where the determinant becomes a quadratic form in the entries of
any row or column as opposed to a linear one), I have been considering extensions of the Littlewood-Offord Lemma to higher degree polynomials.

For example, let $B$ be a matrix with many nonzero entries, and consider the bilinear form

$$x^T By = \sum_{i=1}^{n} \sum_{j=1}^{n} b_{ij} x_i y_j$$

If the entries of $B$ are small integers, than this form would typically be of magnitude about $n$ for random $x$ and $y$, so we would expect the concentration to be at most $n^{-1}$. However, the form $(x_1 \ldots x_n)(y_1 \ldots y_n)$ shows that this is not always the case. One recent result of mine in [6] shows that this is in some sense the only way in which the intuition breaks down:

**Theorem 2.** Let $\epsilon > 0$ be fixed, and let $x_1, \ldots, x_n, y_1, \ldots, y_n$ be uniformly and independently chosen from $\{-1, 1\}$. Let $B$ be an $n \times n$ matrix having at least $r$ nonzero entries in every row such that

$$\operatorname{sup}_{c} \operatorname{P}(x^T By = c) = O(n^{-1+\epsilon}).$$

Then $B$ contains a rank one square submatrix of size $n - O(\frac{r}{\log r})$.

For the bilinear form, $B$ having rank one is equivalent to the form factoring. What this result says is that if the concentration is significantly larger than $n^{-1}$, then we can turn the form into one which factors by setting only a small portion of its variables equal to 0. Its proof relies both on the prior inverse linear results of [26] and [22] and on a careful analysis of the following question: If $B$ is a matrix such that for random $y$ the vector $By$ is likely to be structured (likely to have entries drawn from a short arithmetic progression), what does this imply about the structure of $B$?

A second result from the same paper gives a bound for the concentration of quadratic forms:

**Theorem 3.** Let $x_i$ be as above. For any fixed $\epsilon > 0$ and any $n \times n$ matrix $A$ with at least $\frac{n}{10}$ nonzero entries per row,

$$\operatorname{sup}_{c} \operatorname{P}(x^T Ax = c) = O(n^{-1/2+\epsilon}).$$

In one sense this theorem is almost tight: The form $(x_1 + \cdots + x_n)^2$ has concentration $\Omega(n^{-1/2})$. However, the high concentration of this form again seems to be an artifact of its degeneracy, and my conjecture is that there should be a result for quadratic forms similar to that for bilinear ones: The concentration is at most $n^{-1}$ unless the form is almost degenerate.

Very recently results along the lines of Theorem 2 (giving a weaker structure on forms where the concentration is at least $n^{-C}$ instead of at least $n^{-1}$) were used by Nguyen [20] and Vershynin [27] to provide further improved bounds on the singularity probability and of random symmetric matrices. However, it remains an open (and tantalizing) question to find an analogue of the exponential bounds on the singularity probability obtained in the asymmetric case by the results of [14] and its successors.

A second avenue which seems worth investigating is the question of what happens for higher degree polynomials. The methods of [8] and [6] are in some sense based on lifting from linear polynomials to quadratic ones. An inductive analogue of this argument can be used to give a bound on the concentration of higher degree polynomials on a single value, but the loss in each step of the lift means that the bounds get much worse as the degree of the polynomial increases (roughly $n^{-f(k)}$, where $f(k)$ is exponential in $k^2$). Alternative methods (for example via exploiting the polynomial version of the Berry-Esseen inequality [18] that has recently found use in theoretical computer science) can improve the dependence on $k$. But in reality we would expect the increase of the degree to decrease the concentration on one value, not increase it (modulo non-degeneracy assumptions as in Theorem 2), as a higher degree means more non-zero coefficients, which in turn leads to the polynomial spreading the input over a larger range.
3 Randomized Greedy Algorithms

Another focus of my recent research has involved the study of the performance of a certain type of “randomized derandomized” algorithm, which can perhaps best be explained by example.

For the Maximum Satisfiability problem (or MAX-SAT) problem, we are given a set of disjunctive clauses (such as “\((x_1 \lor x_2 \lor \neg x_4)\)”) along with a weight for each clause. The goal is then to set each variable to true or false so as to maximize the total weight of satisfied clauses. Since this problem is NP-HARD to solve in general, it is common to instead look for approximation algorithms.

The first and simplest approximation comes by assigning each variable to either “true” or “false” with equal probability. Since every clause is satisfied with probability at least \(1/2\) by this algorithm (in fact, a clause of length \(i\) is satisfied with probability \(1 - 2^{-i}\)), this algorithm satisfies at least half the total weight of clauses on average, and therefore also at least half the weight of the optimum assignment. Johnson [13] proposed the following deterministic variant of this algorithm that can be thought of as the greedy derandomization of the above algorithm.

**Johnson’s Algorithm:** Place the variables in an arbitrary order \(\{x_1, \ldots, x_n\}\). At each step, assign \(x_i\) to either true or false so as to maximize the expected weight of satisfied clauses, treating the assignments of \(\{x_1, \ldots, x_{i-1}\}\) as being fixed and \(\{x_{i+1}, \ldots, x_n\}\) as being assigned uniformly at random. For example, given the clause set
\[
(x_1 \lor \neg x_2), (x_1 \lor \neg x_3), (\neg x_1)
\]
the greedy algorithm can begin by setting \(x_1\) to true, not realizing that in doing so it will only satisfy two clauses that could have been satisfied by \(x_2\) and \(x_3\). In other words, the greedy algorithm can sometimes satisfy only two clauses when an optimal assignment satisfies three.

Chen, Friesen, and Zheng [4] showed that this example was in some sense the worst case possible: Johnson’s algorithm always satisfies at least \(2/3\) as many clauses as the optimal assignment. They also observed that the above example was in some sense due to a very unfortunate ordering of the variables: The algorithm performs perfectly on this “worst case”, as long as it starts by trying to assign any variable other than \(x_1\). Motivated by this, they asked:

**Question 2.** Suppose we added an additional randomization step at the beginning of Johnson’s algorithm so that the variables are considered in a random order instead of an arbitrary one. Will this (on average) improve the performance ratio above \(2/3\) on any set of clauses?

In [7], I showed with Shapira and Tetali that this is indeed the case

**Theorem 4.** There is a \(c > 0\) such that this randomized greedy algorithm always satisfies (in expectation) at least \((2/3 + c)\) of the weight satisfied by the optimal assignment for any set of clauses.

The value of \(c\) we obtain in the above result is very tiny, and we still do not know the optimum performance ratio of this algorithm, though we do know it does not perform as well as more complicated semi-definite programming techniques.

As a second example, consider the following “MAX-CUT” problem: Given a graph \(G\), find a partition of the vertices of \(G\) into two classes \((L, R)\) in such a way as to maximize the number of edges between \(L\) and \(R\). Again there is a simple randomized algorithm: If each vertex is randomly assigned to \(L\) or \(R\) with equal probability, on average half the edges will be cut. This algorithm can also be greedily derandomized by considering the vertices in some arbitrary order \(\{x_1, \ldots, x_n\}\) and always assigning \(x_i\) on the side opposite the majority of its neighbors among \(\{x_1, \ldots, x_{i-1}\}\).
In the worst case, this algorithm still only cuts half the edges of $G$. For example, in the bipartite graph $K_{2n, 2n}$, the algorithm could start by considering all the vertices on the left side of the bipartition and splitting them evenly between $L$ and $R$ (as the greedy algorithm doesn’t look ahead, it doesn’t see any edges to motivate it to place the vertices any other way). Once this happens, there’s no way for the algorithm to recover. Any time it encounters a vertex from the original right side of the bipartition that vertex already has half its neighbors in $L$ and half in $R$, so we cannot cut more than half the edges no matter which side it is placed on.

However, in this particular case the failure stems in large part from an unfortunate choice of the order in which we had considered the vertices. We could just as easily chosen $x_1$ and $x_2$ to come from opposite sides of the original bipartition. In that case the greedy algorithm would place $x_1$ in $L$ and $x_2$ in $R$ (or vice versa), and everything else would cascade in place; any vertex from the left side would have neighbors in $R$ and get placed in $L$, and vice versa, and all the edges would end up cut. Motivated by this example, Mathieu and Schudy [19] considered the following randomized variant of the greedy algorithm.

Algorithm 2: At each step choose an unassigned vertex uniformly at random, and assign it so as to cut the most edges between that vertex and the vertices already assigned.

Mathieu and Schudy showed that this simple-looking algorithm actually cuts within a $(1 - \epsilon)$ fraction of the maximum number of edges for any $\epsilon$ as the size of the graph tends to infinity, assuming the original graph is dense (has $\delta n^2$ edges for some $\delta > 0$). Motivated by this, they asked their algorithm could at least provide a $(\frac{1}{2} + \epsilon)$ approximation for arbitrary graphs. Given the performance of the algorithm on dense graphs and the behavior of the similar looking randomized variant of Johnson’s Algorithm, this seemed quite possible. Unfortunately, the answer turns out to be negative.

**Theorem 5.** [7] There are bipartite graphs on $n$ vertices for which Algorithm 2 fails to cut a $(\frac{1}{2} + \epsilon)$ edges with probability $1 - \exp(-O(\epsilon(\log^3 n)))$.

The construction for the above theorem is a random bipartite graph having very close to constant degree (average degree $\sqrt{\log n}$). The proof involves a martingale stopping-time argument similar in flavor to the differential-equations method for random graphs.

Given how well this algorithm performs on dense graphs, it still seems a very interesting question to analyze its performance on graphs of intermediate edge density (e.g. graphs on $n$ vertices and $n^{1+a}$ edges, where $0 < a < 1$), or if there is some alternative variant of this algorithm that provides a better-than-$1/2$ approximation on general graphs.

### 4 Monochromatic Solutions to Linear Equations

Steve Butler, Ron Graham, and myself, have recently been considering the following question in arithmetic Ramsey theory:

**Question 3.** Under what conditions on the coefficients $\{a_1, a_2, \ldots, a_k\}$ will there be a 2-coloring of $[1, n]$ such that the number of monochromatic solutions to

$$a_1x_1 + \cdots + a_kx_n = 0 \quad (1)$$

is significantly less than that of a random coloring?

---

1 Finding any purely combinatorial algorithm that provides such an approximation was a long-standing problem that has only recently been solved [15].
Here “significantly less than” should be thought of as differing by at least \( cn^{k-1} \) for some \( c \) as \( k \) tends to infinity, that is to say by some constant fraction of the total number of solutions. When the \( a_i \) are such that there is such a coloring, that immediately raises a second question

**Question 4.** Given \( \{a_1, a_2, \ldots, a_k\} \), what 2-coloring of \([1, n]\) minimizes the number of monochromatic solutions to the above equation?

As of right now, we only know the answer to the first question in only a few special cases.

- If \( k = 3 \), there is always such a coloring. Butler, Graham, and I have proven this for \( a_1 + a_2 + a_3 = 0 \) in [3], and can be proven for the general case without much additional effort.
- For the equation \( a_1 + a_2 = a_3 + a_4 \) (the so-called “additive energy”), there is no such coloring. In this case the additive energy is counted by the sum of the fourth powers of Fourier coefficients, and the fact that a random coloring is optimal becomes equivalent to the inequality \( x^4 + y^4 \geq \frac{1}{8}(x + y)^4 \).

However, the method for these cases seems to be unfortunately ad-hoc, and we are still looking for a method that can handle the general case. One tantalizing conjecture for the general answer (suggested by Asaf Shapira) consistent with all our known examples is the following:

**Conjecture 1.** There is a coloring with significantly fewer monochromatic solutions to (1) then the random coloring if and only if the coefficient set can be partitioned into disjoint, nonempty sets \( S \) and \( T \) such that \( S \) and \( T \) both sum to 0.

The second question has thus far remained intractable for almost all cases. Even for the simple looking equation \( x + y = 2z \) (corresponding to monochromatic 3-term arithmetic progressions), it remains open, with the best bounds due to Parillo, Robertson, and Saracino [21]. Recently we have developed a method to experimentally determine a likely optimal coloring, given an arbitrary coefficient set [3]. However, our method only provides a rigorous guarantee of local optimality, so it only proves an upper bound on the number of solutions and does not fully answer the second question.

## 5 Estimating Optimums Over Large Amounts of Input

I have also been interested in questions which involve estimating the minimum or maximum of a large number of functions dependent on (comparatively) few variables. Often the extremum is being taken over so many arguments that it is difficult to compute in practice for most particular sets of inputs. However, it may still be possible to determine what the distribution of the extreme value is for random input is, and that may be the best guide available. A prototypical example of this phenomenon is the clique number of a graph. The clique number of a particular graph may be quite hard to find, but the clique number of a random graph is for large \( n \) and fixed edge probability concentrated around only one or two known values [1].

Another problem that fits this paradigm is the following “balancing vectors” problem. Let \( x_1, \ldots, x_n \) be independent Gaussian vectors in \( R^d \) (\( d \) fixed). The parameter in question is

\[
\min \| \pm x_1 \pm x_2 \pm \cdots \pm x_n \|
\]

where the minimum is taken over all \( 2^n \) possible sign sequences and \( \| \cdot \| \) is any (fixed) norm on \( R^d \). Any particular sum is unlikely to be small, and in fact will typically be have norm of order \( \sqrt{n} \). However, the minimum is being taken over a very large number of sums. In [5] the probability distribution of the smallest norm over all sign sequences is determined in the limit as \( n \) tends to infinity, and in particular it is shown that almost surely there are sign sequences obtaining a norm of \( (\frac{1}{2} + o(1))^n/d \). However, the proof is entirely nonconstructive.
References


