THE RANK OF RANDOM GRAPHS

KEVIN P. COSTELLO AND VAN H. VU

Abstract. We show that almost surely the rank of the adjacency matrix of the Erdős-Rényi random graph $G(n,p)$ equals the number of non-isolated vertices for any $c \ln n/n < p < 1/2$, where c is an arbitrary positive constant larger than $1/2$. In particular the giant component (a.s.) has full rank in this range.

1. Introduction

Let G be a (simple) graph on n points $\{1,\dots,n\}$. The adjacency matrix Q_G of G is the symmetric n by n matrix whose ij entry is one if the vertex i is connected to the vertex j and zero otherwise.

There are several models for random graphs. We will focus on the most popular one, the Erdős-Rényi $G(n,p)$ model (some other models will be discussed in the concluding remarks). In this model, one starts with the vertex set $\{1,\dots,n\}$ and puts an edge (randomly and independently) between any two distinct vertices i and j with probability p. We say that a property P holds almost surely for $G(n,p)$ if the probability that $G(n,p)$ possesses P goes to one as n tends to infinity.

We are interested in the rank of Q_G, where G is a graph chosen from $G(n,p)$. An interesting feature of this parameter is that, unlike many graph parameters (e.g. the connectivity or the chromatic number), the rank is not monotone under the addition of edges. For example, the path of length 4 has a full rank adjacency matrix, but adding an edge to create a 4-cycle decreases the rank by 2 (the matrix gains two pairs of equal rows). Nevertheless, the rank will turn out to exhibit many of the same threshold behaviors as these monotone properties.

A vertex v of a graph G is isolated if it has no neighbor. If v is isolated, then the row corresponding to v in Q_G is all-zero. Let $i(G)$ denote the number of isolated vertices of G. It is clear that

Fact 1.1. *For any graph G, rank$(Q_G) \le n - i(G)$.*

Kevin P. Costello is supported by an NSF Graduate Research Fellowship.
Van H. Vu is a Sloan Fellow and is supported by an NSF Career Grant.

The main result of this paper shows that for p sufficiently large, the above upper bound is tight for $G(n,p)$. In other words, any non-trivial linear dependence in $G(n,p)$ comes from the isolated vertices.

Theorem 1.2. *Let c be a constant larger than $\frac{1}{2}$. Then for any $c\ln n/n \leq p \leq \frac{1}{2}$, the following holds with probability $1-O((\ln\ln n)^{-1/4})$ for a random sample G from $G(n,p)$:*

$$\mathit{rank}(G) = n - i(G).$$

Remark 1.3. This result is sharp in two ways. Obviously, the estimate $\text{rank}(G) = n - i(G)$ cannot be improved. Furthermore, the bound $c > 1/2$ is also the best possible. For $c < 1/2$ and $p = c\ln n/n$, a random sample G from $G(n,p)$ satisfies the strict inequality

$$\text{rank}(G) < n - i(G)$$

almost surely. In order to see this, notice that in this range $G(n,p)$ almost surely contains two vertices u and v with degree one sharing a common neighbor. The rows corresponding to u and v are not zero, but they are equal and this reduces the rank further.

On the other hand, the upper bound $p < \frac{1}{2}$ can be replaced without much effort by $p < \alpha$ for any fixed $\alpha < 1$, though we do not claim any result for $p = 1 - o(1)$. Some sort of upper bound is needed, as the adjacency matrix of $G(n, 1 - \frac{\ln n}{3n})$ again almost surely contains pairs of equal nonzero rows.

Let us now deduce a few corollaries. It is well known that $p \geq c\ln n/n$ for some $c > 1/2$, then the random graph (a.s.) consists of a giant component and some isolated vertices.

Corollary 1.4. *Let c be a constant larger than $\frac{1}{2}$. Then for any $c\ln n/n < p < 1/2$, the adjacency matrix of the giant component of $G(n,p)$ is almost surely non-singular.*

Furthermore, if $c > 1$, then $G(n,p)$ almost surely is connected and contains no isolated vertices.

Corollary 1.5. *Let c be a constant larger than 1. Then for any $c\ln n/n < p < 1/2$, $G(n,p)$ is almost surely non-singular.*

It follows that the non-singularity of $G(n,p)$ has a sharp threshold at $p = \ln n/n$. For any positive constant ϵ, if $p < (1-\epsilon)\ln n/n$, then $G(n,p)$ is almost surely singular as it contains isolated vertices. On the other hand, the above corollary asserts that $G(n,p)$ is almost surely non-singular for $p > (1+\epsilon)\ln n/n$.

The special case $p = 1/2$ was a well known conjecture of B. Weiss, posed many years ago. This special case can be viewed as the symmetric version of a well-known theorem of Komlós on the non-singularity of (non-symmetric) random Bernoulli matrices [5] and was solved two years ago in [1]. The proof of Theorem 1.2 extends the ideas in that paper and combines them with arguments involving random graphs. In the next section, we outline this proof and present the key lemmas. In Section 3, we prove the theorem assuming the lemmas. Most of the rest of the paper is devoted to the proofs of the lemmas. The last section contains several remarks and open questions.

Notation. In the whole paper, we will always assume that n is sufficiently large. As usual, the asymptotic notation is used under the condition that $n \to \infty$. $\mathbf{P}$ denotes probability and $\mathbf{E}$ denotes expectation.

2. Outline of the proof and the main lemmas

We will assume that $p = o(1)$, as the case $p = \Omega(1)$ was presented as Theorem 6.5 of [1]. We denote by $Q(n, p)$ the adjacency matrix of $G(n, p)$.

Following the ideas from [5, 1], we are going to expose $Q(n, p)$ minor by minor. Letting Q_m denote the upper left $m \times m$ minor of $Q(n, p)$, we view Q_{m+1} as being formed by taking Q_m and augmenting by a column whose entries are chosen independently, along with the column's transpose. Denote by G_m the graph whose adjacency matrix is Q_m. In graph theoretic terms, we are considering the vertex exposure process of $G(n, p)$.

Our starting observation is that when a good portion of the vertices have been exposed, the matrix has rank close to its size.

Recall that $p \geq c \ln n/n$ for a constant $c > 1/2$. We can set a constant $0 < \delta < 1$ such that $1/2 < \delta c < 3/5$. Define $n' := \delta n$.

Lemma 2.1. *For any constant $\epsilon > 0$ there exists a constant $\gamma > 0$ such that*

$$\mathbf{P}(rank(Q_{n'}) < (1 - \epsilon)n') = o(e^{-\gamma n \ln n})$$

Our plan is to show that the addition of the remaining $n - n'$ rows/columns is enough to remove all the linear dependencies from $Q_{n'}$, except those corresponding to the isolated vertices.

The next lemmas provide some properties of G_m for $n' \leq m \leq n$.

Definition 2.2. A graph G is **well separated** if it contains no pair of vertices of degree at most $\ln \ln n$ whose distance from each other is at most 2.

Lemma 2.3. *For any constant $\epsilon > 0$, G_m is well separated for every m between n' and n with probability $1 - O(n^{1-2c\delta+\epsilon})$.*

Here and later on, we always choose ϵ sufficiently small so that $1-2c\delta+\epsilon$ is negative.

Definition 2.4. A graph G is a **small set expander** if every subset S of the vertices of G with $|S| \leq \frac{n}{\ln^{3/2} n}$ containing no isolated vertices has at least $|S|$ edges connecting S to $\bar{S}$, its complement.

Lemma 2.5. *For any $m > n'$ the probability that G_m is well separated but is not a small set expander is $O(1/n^3)$.*

Remark 2.6. Lemmas 2.3 and 2.5 immediately imply that almost every graph encountered in our process after time n' will be a small set expander. We cannot expect this to occur for $p < \frac{(.5-\epsilon)\ln n}{n}$, as at this density the random graph will likely contain pairs of adjacent vertices of degree 1, leading to sets of size 2 without any edges at all leaving them.

Definition 2.7. A set S of the vertices of a graph G is **nice** if there are at least two vertices of G with exactly one neighbor in S.

Remark 2.8. The vertices of G may themselves be taken from S. For example, an isolated edge (a pair of adjacent degree one vertices) in G would form a nice set.

Set $k := \frac{\ln \ln n}{2p}$.

Definition 2.9. A graph G is **good** if the following two properties hold:

1. Every subset of the vertices of G of size at least 2 and at most k which contains no isolated vertices is nice.

2. At most $\frac{1}{p \ln n}$ vertices of G have degree less than 2.

A symmetric (0,1) matrix A is **good** if the graph for which it is an adjacency matrix is good.

The next lemma states that after we are far enough in the augmentation process we are likely to only run into good matrices.

Lemma 2.10. *Let ϵ be a positive constant. Then with probability $1 - O(n^{1-2c\delta+\epsilon})$, Q_m is good for every m between n' and n .*

Since each augmentation adds only one new row and one new column, we trivially have

$$\text{rank}_{Q_m} \leq \text{rank}(Q_{m+1}) \leq \min \text{rank}(Q_m) + 2, m + 1 \tag{1}$$

What our final two lemmas will say is that good matrices behave well under augmentation in the following sense: With high probability the second inequality in (1) in fact holds with equality.

Definition 2.11. A pair (A, A') of matrices is called **normal** if A' is an augmentation of A and every row of all 0's in A also contains only 0's in A' (in graph theoretic terms, the new vertex added by the augmentation is not adjacent to any vertices which were isolated before the augmentation).

Lemma 2.12. *Let A be any fixed, good $m \times m$ matrix with the property that $rank(A) + i(A) < m$. Then*

$$\mathbf{P}(\ rank(Q_{m+1}) -\ rank(Q_m) < 2 | (Q_m, Q_{m+1}) \text{ is normal} \wedge Q_m = A) = O((kp)^{-1/2}).$$

Lemma 2.13. *Let A be any fixed, good, $m \times m$ matrix with the property that $rank(A) + i(A) = m$. Then*

$$\mathbf{P}(\ rank(Q_{m+1}) - \ rank(Q_m) < 1 | (Q_m, Q_{m+1}) \textit{ is normal} \wedge Q_m = A) = O((kp)^{-1/4}).$$

Set $Y_m = m - (rank(Q_m) + i(Q_m))$. The above two lemmas force Y_m to stay near 0. Indeed, if Y_m is positive, then when the matrix is augmented, m increases by 1 but the rank of Q_m will likely increase by 2 (notice that $kp \to \infty$), reducing Y_m. On the other hand, if $Y_m = 0$, it is likely to stay the same after the augmentation.

In the next section, we will turn this heuristic into a rigorous calculation and prove Theorem 1.2, assuming the lemmas.

3. Proof of the Main Result from the Lemmas

In this section, we assume all lemmas are true. We are going to use a variant of an argument from [1].

Let B_1 be the event that the rank of $Q_{n'}$ is at least $n'(1 - \frac{1-\delta}{4\delta})$. Let B_2 be the event that Q_m is good for all $n' \leq m < n$. We therefore have

$$\mathbf{P}(\ \mathrm{rank}(Q_n) + i(Q_n) < n) \leq \mathbf{P}(\ \mathrm{rank}(Q_n) + i(Q_n) < n \wedge B_2 | B_1) + \mathbf{P}(\neg B_1) + \mathbf{P}(\neg B_2)$$

By Lemma 2.1 we have that $\mathbf{P}(\neg B_1) = o(e^{-\gamma n \ln n})$ and by Lemma 2.10 $\mathbf{P}(\neg B_2) = O(n^{1-2c\delta+\epsilon})$. Both probabilities are thus much smaller than the bound $O((\ln \ln n)^{-1/4})$ which we are trying to prove. So, it remains to bound the first term.

Let $Y_m = m - \ \mathrm{rank}(Q_m) - i(Q_m)$. Define a random variable X_m as follows:

- $X_m = 4^{Y_m}$ if $Y_m > 0$ and every Q_j with $n' \leq j \leq m$ is good;
- $X_m = 0$ otherwise.

The core of the proof is the following bound on the expectation of X_{m+1} given any fixed sequence $\mathcal{Q}_m$ of matrices $\{Q_{n'}, Q_{n'+1}, \ldots, Q_m\}$ encountered in the augmentation process.

Lemma 3.1. *For any sequence $\mathcal{Q}_m = \{Q_{n'}, Q_{n'+1}, \ldots, Q_m\}$ encountered in the augmentation process,*

$$\mathbf{E}(X_{m+1} | \mathcal{Q}_m) \leq \frac{3}{5} X_m + O((\ln \ln n)^{-1/4}).$$

Proof (of Lemma 3.1) If a matrix in the sequence $\mathcal{Q}_m = \{Q_{n'}, Q_{n'+1}, \ldots, Q_m\}$ is not good, then $X_{m+1} = 0$ by definition and there is nothing to prove. Thus, from now on we can assume that all matrices in the sequence are good.

Let Z_m denote the number of vertices which were isolated in Q_m but not in Q_{m+1}. If Z_m is positive, then augmenting the matrix will increase Y_m by at most $Z_m + 1$ (m increases by 1, the number of isolated vertices decreases by at most Z_m, and the rank does not decrease). Furthermore, $Z_m = 0$ if and only if (Q_m, Q_{m+1}) is normal.

We thus have

$$\begin{aligned}\mathbf{E}(X_{m+1}|\mathcal{Q}_m) &= \mathbf{E}(X_{m+1}|Z_m > 0 \wedge \mathcal{Q}_m)\mathbf{P}(Z_m > 0|\mathcal{Q}_m) \\ &+ \mathbf{E}(X_{m+1}|\mathcal{Q}_m \wedge (Q_m, Q_{m+1}) \text{ is normal})\mathbf{P}((Q_m, Q_{m+1}) \text{ is normal}|\mathcal{Q}_m) \\ &\leq \mathbf{E}(X_{m+1}\,\chi(Z_m > 0)|\mathcal{Q}_m) + \mathbf{E}(X_{m+1}|\mathcal{Q}_m \wedge (Q_m, Q_{m+1}) \text{ is normal}) \\ &\leq \mathbf{E}(4^{Z_m+1+Y_m}\chi(Z_m > 0)|\mathcal{Q}_m) + \mathbf{E}(X_{m+1}|\mathcal{Q}_m \wedge (Q_m, Q_{m+1}) \text{ is normal}).\end{aligned}$$

Since Q_m is good, G_m has at most $\frac{1}{p\ln n}$ isolated vertices. Thus, we can bound Z_m by the sum of $\frac{1}{p\ln n}$ random Bernoulli variables, each of which is 1 with probability p. It follows that

$$\mathbf{P}(Z_m = i) \leq \mathbf{P}(Z_m \geq i) \leq \binom{(p\ln n)^{-1}}{i} p^i \leq (\ln n)^{-i}.$$

Adding up over all i, we have

$$\mathbf{E}(4^{Z_m+1}\chi(Z_m > 0)|\mathcal{Q}_m) \leq \sum_{i=1}^{\infty} 4^{i+1}(\ln n)^{-i} = O((\ln n)^{-1}).$$

If $Y_m = 0$ and (Q_m, Q_{m+1}) is normal, then by Lemma 2.13 (which applies since Q_m is good) X_{m+1} is either 0 or 4, with the probability of the latter being $O((\ln\ln n)^{-1/4})$. Therefore we have for any sequence $\mathcal{Q}_m = \{Q_{n'}, \dots Q_m\}$ of good matrices with $Y_m = 0$ that

$$\mathbf{E}(X_{m+1}|\mathcal{Q}_m) = O((\ln\ln n)^{-1/4} + (\ln n)^{-1}) = O((\ln\ln n)^{-1/4}). \qquad (2)$$

If $Y_m = j > 0$ and (Q_m, Q_{m+1}) is normal, then Y_{m+1} is $j-1$ with probability $1 - O((\ln\ln n)^{-1/2})$ by Lemma 2.12, and otherwise is at most $j+1$. Combining this with the bound on $\mathbf{E}(4^{Z_m+1}\chi(Z_m > 0)|\mathcal{Q}_m)$ we have

$$\mathbf{E}(X_{m+1}|\mathcal{Q}_m) = 4^{j-1} + 4^{j+1}O((\ln\ln n)^{-1/2}) + 4^j O((\ln n)^{-1}) \leq \frac{3}{5}4^j \qquad (3)$$

The lemma now follows immediately from (2) and (3).

■

Lemma 3.1 shows that for $n' < m$ we have

$$\mathbf{E}(X_{m+1}|Q_{n'}) < \frac{3}{5}\mathbf{E}(X_m|Q_{n'}) + O((\ln\ln n)^{-1/4}).$$

By induction on $m_2 - m_1$ we now have that for any $m_2 \geq m_1 \geq n'$

$$\mathbf{E}(X_{m_2}|Q_{n'}) < (\frac{3}{5})^{m_2-m_1}\mathbf{E}(X_{m_1}|Q_{n'}) + O((\ln\ln n)^{-1/4}).$$

In particular, by taking $m_2 = n$ and $m_1 = n'$ we get that

$$\mathbf{E}(X_n|Q_{n'}) < (\frac{3}{5})^{n-n'}X_{n'} + O((\ln\ln n)^{-1/4}).$$

If $Q_{n'}$ satisfies B_1, we automatically have $X_{n'} \leq 4^{\frac{(1-\delta)n'}{4\delta}} = (\sqrt{2})^{n-n'}$, so

$$\mathbf{E}(X_n|Q_{n'}) < (\frac{3\sqrt{2}}{5})^{n-n'} + O((\ln\ln n)^{-1/4}) = O((\ln\ln n)^{-1/4}).$$

By Markov's inequality, we therefore have

$$\mathbf{P}(X_n > 3|Q_{n'}) = O((\ln\ln n)^{-1/4})$$

for any $Q_{n'}$ satisfying B_1. It thus follows that

$$\mathbf{P}(X_n > 3|B_1) = O((\ln\ln n)^{-1/4}).$$

On the other hand, by definition $X_n \geq 4$ if $\text{rank}(Q_n) + i(Q_n) < n$ and B_2 holds. We therefore have that

$$\mathbf{P}(\ \text{rank}(Q_n) + i(Q_n) < n \wedge B_2|B_1) = O((\ln\ln n)^{-1/4}),$$

proving the theorem.

4. Proof of Lemma 2.1

By symmetry and the union bound

$$\mathbf{P}(\text{rank}(Q_{n'}) < (1-\epsilon)n') \leq \binom{n'}{\epsilon n'} \times \mathbf{P}(B_1^*),$$

where B_1^* denotes the event that the last $\epsilon n'$ columns of Q'_n are contained in the span of the remaining columns.

We view $Q_{n'}$ as a block matrix,

$$Q_{n'} = \left[\begin{array}{c|c} A & B \\ \hline B^T & C \end{array}\right],$$

where A is the upper left $(1-\epsilon)n' \times (1-\epsilon)n'$ sub-matrix and C has dimension $\epsilon n' \times \epsilon n'$. We obtain an upper bound on $\mathbf{P}(B_1^*)$ by bounding the probability of B_1^* conditioned on any fixed A and B (treating C as random).

B_1^* cannot hold unless the columns of B are contained in the span of those of A, meaning the equation $B = AF$ holds for some (not necessarily unique) matrix F. If this is the case, then B_1^* will hold only when we also have $C = B^T F$. This means that each entry of C is forced by our choice of A, B and our assumption that B_1^* holds.

However, C is still random, and the probability that any given entry takes on its forced value is at most $1 - p$. The entries are not all independent (due to the symmetry of C), but those on or above the main diagonal are. Therefore the probability that B_1^* holds for any fixed A and B is at most $(1-p)^{\frac{(\epsilon n')^2}{2}}$.

We therefore have

$$\begin{aligned}
\mathbf{P}(\mathrm{rank}(Q_{n'}) < (1-\epsilon)n') &\leq \binom{n'}{\epsilon n'}(1-p)^{\frac{(\epsilon n')^2}{2}} \\
&\leq (\frac{n'e}{\epsilon n'})^{\epsilon n'} e^{\frac{-p(\epsilon n')^2}{2}} \\
&\leq {c_2}^n e^{-c_1 n \ln n},
\end{aligned}$$

where c_1 and c_2 are positive constants depending on ϵ, δ, and c (but independent of n). The result follows.

Remark 4.1. The same argument gives an upper bound of ${c_2}^n e^{-c_1 n^2 p}$ on the probability for any n and p. Holding ϵ fixed, we see that the probability becomes $o(1)$ for $p = y/n$ with sufficiently large fixed y. In particular, if $p \to 0$ and $np \to \infty$, then $\mathrm{rank}(Q_{n,p})/n \to 1$.

5. Proof of Lemma 2.3

If p is at least $(\ln n)^2/n$ then $G(m, p)$ will with probability at least $1 - o(1/n^3)$ have no vertices with degree at most $\ln \ln n$, in which case the lemma is trivially true. Therefore we can assume $p \leq (\ln n)^2/n$

If G_m fails to be well separated for some m between n' and n there must be a first m_0 with this property. We are going to bound the probability that a fixed m is m_0.

Case 1: $m_0 = n'$. The probability that G'_n fails to be well separated is at most n^2 times the probability that any particular pair of vertices v and w are both of small degree and at distance at most 2 from each other.

The probability that v has sufficiently small degree is at most

$$\begin{aligned}\sum_{i=0}^{\ln\ln n}\binom{n'-1}{i}p^i(1-p)^{n'-i} &\leq (1+o(1))\sum_{i=0}^{\ln\ln n}(n'p)^i(1-p)^{n'-i}\\ &\leq (1+o(1))(n'p)^{\ln\ln n}e^{-n'p},\\ &\leq \frac{(\ln n)^{\ln\ln n}}{n^{c\delta}},\end{aligned}$$

where the last inequality comes from our lower bound on p since the second to last line is decreasing in p. The same holds for w even if we assume v has small degree (although the degree of v and that of w aren't quite independent, we can bound the probability w has small degree by the probability it has at most $\ln\ln n$ neighbors not including v).

Since a given pair of vertices both being of small degree is a monotone decreasing graph property, the FKG inequality gives that the probability of an edge being present between v and w is at most p even after we condition on them both being small degree. Similarly, the probability of the existence of an x adjacent to both v and w is at most np^2. Combining these facts, the probability of two small degree vertices being close is at most

$$n^2\Big(\frac{(\ln n)^{\ln\ln n}}{n^{c\delta}}\Big)^2(p+np^2) \leq \frac{(\ln n)^{2\ln\ln n}\ln^4 n}{n^{2c\delta-1}} = o(n^{1-2c\delta+\epsilon}).$$

Case 2: $m_0 = m$ for some $n' < m \leq n$. We bound the probability that m satisfies the somewhat weaker condition that G_{m-1} is well separated but G_m is not. Let w be the vertex newly added to the graph. There are only two ways that the addition of w can cause G_m to lose well separatedness: either w serves as the link between two low degree vertices v_1 and v_2 that were previously unconnected, or w is itself a low degree vertex of distance at most 2 from a previous low degree vertex v_0.

Applying (4) twice, the probability that any particular v_1 and v_2 both have low degree and w is connected to both of them is at most $p^2(\frac{(\ln n)^{\ln\ln n}}{n^{c\delta}})^2$.

Again by (4) the probability that w is of low degree is at most $\frac{(\ln n)^{\ln\ln n}}{n^{c\delta}}$, as is the probability that any particular choice of candidate for v_0 has low degree. By the FKG inequality, the probability that w and our candidate v_0 share a common neighbor given they both have small degree is at most np^2, while the probability they are themselves connected is p.

Since there are at most n^2 choices for v_1 and v_2 and at most n choices for v_0, applying the union bound over all these choices, we obtain that the probability G_{m-1} is well connected but G_m is not is at most

$$\Big(\frac{(\ln n)^{\ln\ln n}}{n^{c\delta}}\Big)^2(2n^2p^2+p) = o(n^{-2c\delta+\epsilon})$$

Applying the union bound over all possible m (there are at most n values for m), we obtain that the probability of the existence of such an m_0 is $o(n^{1-2c\delta+\epsilon})$. The proof is complete.

6. Proof of Lemma 2.5

In order to prove the edge expansion property we first show that almost surely all small subgraphs of $G(n, \frac{c \ln n}{n})$ will be very sparse.

Lemma 6.1. *For fixed c the probability that $G(n, \frac{c \ln n}{n})$ has a subgraph containing at most $\frac{n}{\ln^{3/2} n}$ vertices with average degree at least 8 is $O(n^{-4})$*

Proof Let q_j be the probability that there exists a subset of size j inducing at least $4j$ edges. By the union bound, this is at most $\binom{n}{j}$ times the probability of a particular subset having at least $4j$ edges, so

$$\begin{aligned} q_j &\leq \binom{n}{j}\binom{j^2/2}{4j}p^{4j} \\ &\leq (\frac{ne}{j})^j(\frac{ejc\ln n}{8n})^{4j} \\ &\leq (\frac{ce^5 j^3 \ln^4 n}{n^3})^j. \end{aligned}$$

For $j < n^{1/4}$ this gives $q_j \leq n^{-2j}$, while for $j > n^{1/4}$ we have (using our upper bound on j) $q_j \leq (\ln n)^{-j/2} = o(n^{-5})$. By adding up over all j at least 2, we can conclude that the failure probability is $O(n^{-4})$, completing the proof. ■

Armed with this lemma we can now prove Lemma 2.5; we do so in two cases depending on the value of p.

Case 1: $p < \frac{12 \ln n}{n}$:

Suppose that G failed to expand properly. If this is the case, there must be a minimal subset S_0 with fewer than $|S_0|$ edges leaving it. If any vertex in S_0 were adjacent to no other vertex in S_0, it would have a neighbor outside S_0 (since S_0 contains no isolated vertices), and dropping it would lead to a smaller non-expanding set, a contradiction. Therefore every vertex in S_0 has a neighbor in S_0. By the well separatedness assumption the vertices of degree at most $\ln \ln n$ are non-adjacent and share no common neighbors. Thus it follows that at most half the vertices in S_0 are of degree at most $\ln \ln n$. Since at most $|S_0|$ edges leave S_0, it follows that there are $\Omega(|S_0| \ln \ln n)$ edges between vertices of S_0. But by Lemma 6.1 the probability an S_0 with this many edges exists is $O(\frac{1}{n^4})$, competing the proof for this case.

Case 2: $p \geq \frac{12 \ln n}{n}$: We estimate the probability that there is a non-expanding small set directly by using the union bound over all sets of size $i < n \ln^{-3/2} n$. The probability in question can be bounded from above by

$$\sum_{i=1}^{n \ln^{-3/2} n} \binom{n}{i} \binom{i(n-i)}{i-1} (1-p)^{i(n-i)-(i-1)} \leq \sum_{i=1}^{n \ln^{-3/2} n} n^i (en)^{i-1} e^{-inp(1+o(1))}$$
$$= \frac{1}{en} \sum_{i=1}^{n \ln^{-3/2} n} (n^2 e^{-np(1+o(1))})^i.$$

The lower bound on p guarantees that the summand is $O(n^{-(4+o(1))i})$, so the probability for any p in this range is $o(\frac{1}{n^4})$. Notice that in this case we do not need the well separatedness assumption.

7. Proof of Lemma 2.10

Let C_0 be the event that G_m is good for every m between n' and n. Let C_1 be the event that G_m has at most $\frac{1}{p \ln n}$ vertices of degree less than 2 for every m between n' and n, C_2 be the event that G_m has maximum degree at most $10np$ for each m, and C_3 be the event that G_m is well separated and a small set expander for every m between n' and n. We have

$$\mathbf{P}(\neg C_0) \leq \mathbf{P}(\neg C_0 \wedge C_1 \wedge C_2 \wedge C_3) + \mathbf{P}(\neg C_1) + \mathbf{P}(\neg C_2) + \mathbf{P}(\neg C_3).$$

It suffices to bound each term on the right hand side separately, and we will do so in reverse order.

Lemmas 2.5 and 2.3 together show that $\mathbf{P}(\neg C_3) = O(n^{1-2c\delta+\epsilon})$.

$\mathbf{P}(\neg C_2)$ is at most the expected number of vertices of degree at least $10np$ in G_n, which is at most

$$n \binom{n}{10np} p^{10np} \leq n(e/10p)^{10np} p^{10np} \leq ne^{-10np} = o(n^{-4}).$$

To bound $\mathbf{P}(\neg C_1)$, we note that the probability that some G_m contains a set of vertices of degree less than 2 of size $s = p^{-1} \ln^{-1} n$ is bounded from above by the probability that at least s vertices in G_n each have fewer than 2 neighbors amongst the vertices of $G_{n'}$, which by Markov's inequality is at most

$$\frac{1}{s} n (n'p(1-p)^{n'-2} + (1-p)^{n'-1}) = O(s^{-1} n^2 p e^{-\delta np}) = o(n^{-1/3}).$$

It follows that $\mathbf{P}(\neg C_1) = o(n^{-1/3})$.

It remains to estimate the first term, which we will do by the union bound over all m. Since property C_1 implies G_m has few vertices of small degree, it suffices to estimate the probability G_m contains a non-nice set while still satisfying properties C_1, C_2, and C_3.

Let p_j be the probability that conditions C_1, C_2, and C_3 hold but some subset of exactly j vertices without isolated vertices is not nice. Symmetry and the union bound give that p_j is at most $\binom{m}{j}$ times the probability that the three conditions hold and some fixed set S of j vertices is not nice. We will do this in three cases depending on the size of j.

Case 1: $\frac{1}{p\sqrt{\ln n}} \leq j \leq k = \frac{\ln \ln n}{2p}$.

Direct computation of the probability that a fixed set of j vertices has either 0 or 1 vertices adjacent to exactly one vertex in the set gives:

$$\begin{aligned} p_j &\leq \binom{m}{j}((1-jp(1-p)^{j-1})^m + mjp(1-p)^{j-1}(1-jp(1-p)^{j-1})^{m-1}) \\ &\leq (mep\sqrt{\ln n})^j((1-jp(1-p)^{j-1})^m + mjp(1-p)^{j-1}(1-jp(1-p)^{j-1})^{m-1}) \\ &\leq (mep\sqrt{\ln n})^j((1-jpe^{-jp(1+o(1))})^m + mjpe^{-jp(1+o(1))}(1-jpe^{-jp(1+o(1))})^{m-1}) \\ &\leq (mep\sqrt{\ln n})^j(e^{-mjp(1+o(1))e^{-jp(1+o(1))}})(1+mjpe^{-jp(1+o(1))}). \end{aligned}$$

It follows from our bounds on j and p that $mjpe^{-jp}$ tends to infinity, so the second half dominates the last term of the above sum and we have:

$$p_j \leq (mep\sqrt{\ln n})^j(e^{-mjp(1+o(1))e^{-jp(1+o(1))}})(2mjpe^{-jp(1+o(1))}).$$

Taking logs and using $\delta n \leq m \leq n$ gives:

$$\begin{aligned} \ln(p_j) &\leq (1+o(1))j(\ln(enp\sqrt{\ln n}) - \delta npe^{-jp(1+o(1))} - p + \frac{\ln(2njp)}{j}) \\ &\leq (1+o(1))j(4\ln(np) - \delta npe^{-kp(1+o(1))}) \\ &= (1+o(1))j(4\ln(np) - \frac{\delta np}{(\ln n)^{\frac{1}{2}+o(1)}}). \end{aligned}$$

Since $np > 0.5 \ln n$, taking n large gives that the probability of failure for any particular j in this range is $o(1/n^4)$, and adding up over all j and m gives that the probability of a failure in this range is $o(1/n^2)$.

Case 2: $\frac{10}{2c\delta - 1} \leq j \leq \frac{1}{p\sqrt{\ln n}}$.

Let b be the number of vertices outside S adjacent to at least one vertex in S, and let a be the number of edges between S and the vertices of G outside S. If G_m is to satisfy the properties C_2 and C_3 it must be true that $j \le a \le 10jnp$.

Next, we note that if S is not nice, then at least $b-1$ of the neighbors of S must be adjacent to at least two vertices in S. This implies that $b \le \frac{a+1}{2}$. It follows that

$$p_j \le \binom{m}{j}\Big(\max_{10j<w\le 10jnp} \mathbf{P}(b \le \frac{w+1}{2}|a=w) + \mathbf{P}(a \le 10j) \max_{j\le w\le 10j} \mathbf{P}(b \le \frac{w+1}{2}|a=w)\Big) \tag{4}$$

To bound $\mathbf{P}(b \le \frac{w+1}{2}|a=w)$, we fix a set of $\frac{w+1}{2}$ vertices and bound the probability that w consequentially randomly selected vertices were in that set. We will view the w vertices as being chosen uniformly with replacement. However, allowing edges to be repeated can only increase the probability that few destination vertices are chosen. Using the union bound over all possible sets of $\frac{w+1}{2}$ vertices, we have

$$\begin{aligned}
\mathbf{P}(b \le \frac{w+1}{2}|a=w) &\le \binom{m-j}{\frac{w+1}{2}}(\frac{w+1}{2(m-j)})^w \\
&\le (\frac{2e(m-j)}{w-1})^{\frac{w+1}{2}}(\frac{w+1}{2(m-j)})^w \\
&\le (\frac{4w}{m})^{\frac{w-1}{2}}.
\end{aligned}$$

This last bound is decreasing in w for the entire range under consideration (our bounds on j guarantee w is at most $\frac{10n}{\sqrt{\ln n}}$). Therefore we can plug in the smallest values of w in (4) to get

$$\begin{aligned}
p_j &\le \binom{m}{j}(\mathbf{P}(a<10j)(\frac{4j}{m})^{\frac{j-1}{2}} + (\frac{40j}{m})^{\frac{10j-1}{2}}) \\
&\le 3\sqrt{n}\binom{m}{j}(\mathbf{P}(a<10j)(\frac{4j}{m})^{\frac{j}{2}} + (\frac{40j}{m})^{5j}) \\
&\le 3\sqrt{n}(\frac{me}{j})^j(\mathbf{P}(a<10j)(\frac{4j}{m})^{\frac{j}{2}} + (\frac{40j}{m})^{5j}) \\
&\le 3\sqrt{n}(\mathbf{P}(a<10j)(\frac{4e^2n}{j})^{\frac{j}{2}} + (\frac{130j}{n'})^{4j}).
\end{aligned}$$

a here is the sum of at least $n'j(1+o(1))$ independent Bernoulli variables each with probability of success at least $\frac{c\ln n}{n}$. We therefore have

$$\begin{aligned} \mathbf{P}(a \le 10j) &\le \sum_{i=0}^{10j} \binom{n'j}{i} (\frac{c\ln n}{n})^i (1 - \frac{c\ln n}{n})^{n'j(1+o(1))} \\ &\le \sum_{i=0}^{10j} (\frac{cej\ln n}{i})^i n^{-jc\delta(1+o(1))} \\ &= n^{-jc\delta(1+o(1))}, \end{aligned}$$

yielding

$$p_j \le 4\sqrt{n}((\frac{2e}{\sqrt{j}n^{(1+o(1))c\delta-1/2}})^j + (\frac{130j}{n'})^{4j}).$$

Both terms are decreasing in j in the range under consideration, and plugging in the lower endpoint of our range gives that $p_j = o(1/n^4)$ for each j and m in the range. By the union bound the probability of failure in this range is $o(1/n^2)$.

Case 3: $2 \le j \le \frac{10}{2c\delta-1}$

Let a and b be as in case 2. We again bound the probability of failure for any fixed set of vertices by the probability that $b \le \frac{a+1}{2}$.

We first note that if condition C_3 is to be satisfied then this inequality cannot be satisfied any time when a is at most $10j$. This is because if a is in this range it follows that every vertex in our set is of degree below $\ln\ln n$, and the well separatedness condition then guarantees that each edge leaving our set must go to a different vertex.

Because of this, we can rewrite equation (4) as

$$\begin{aligned} p_j &\le \binom{m}{j} (\max_{10j<w\le 10jnp} \mathbf{P}(b \le \frac{w+1}{2} | a = w) \\ &\le 4\sqrt{n}(\frac{130j}{n'})^{4j}, \end{aligned}$$

where the second inequality comes from our computations in case 2. Adding up over all j in this range gives that the probability of failure in this range is $o(n^{-3})$.

8. Some Littlewood-Offord-Type Results

The proof of the remaining two lemmas rely on modifications to the following lemma of Littlewood and Offord [3]:

Lemma 8.1. *Let a_i be fixed constants, at least q of which are nonzero. Let $z_1, z_2, \dots z_n$ be random, independent Bernoulli variables which take on 0 and 1 each with probability 1/2. Then for any fixed c,*

$$\mathbf{P}(\sum_{i=1}^{n} a_i z_i = c) = O(q^{-1/2}),$$

where the implied constant is independent of n, the a_i, and c.

The variables we are now considering, however, are not equally likely to be 0 and 1. Thus we need the following special case of a more general result in [4].

Lemma 8.2. *Let a_i be fixed constants, at least q of which are nonzero. Let $z_1, z_2, \dots z_n$ be random, independent Bernoulli variables which take on 1 with probability $p < \frac{1}{2}$, 0 with probability $1 - p$. Then for any fixed c,*

$$\mathbf{P}(\sum_{i=1}^{n} a_i z_i = c) = O((qp)^{-1/2}),$$

where the implied constant is absolute.

Remark 8.3. The theorem is also true (with near identical proof) if one replaces the distribution of the z_i by one with $\mathbf{P}(z_i = 1) = \mathbf{P}(z_i = -1) = p, \mathbf{P}(z_i = 0) = 1 - 2p$

Proof

Let r_i be Bernoulli random variables taking on 1 with probability $2p$, and 0 with probability $1 - 2p$. Let s_i be random variables taking on 1 and 0 with equal probability, and replace z_i by $r_i s_i$ (which has the same distribution). We thus have

$$\mathbf{P}(\sum_{i=1}^{n} (a_i r_i) s_i = c) \leq \mathbf{P}(\sum_{i=1}^{n} a_i r_i s_i = c | \sum_{i=1}^{n} r_i \geq qp) + \mathbf{P}(\sum_{i=1}^{n} r_i < qp)$$

Since $\mathbf{E}(\sum_{i=1}^{n} r_i) = 2qp$ and $Var(\sum_{i=1}^{n} r_i) \leq 2qp$, by Chebyshev's inequality the second term on the right is $O((qp)^{-1}) = O((qp)^{-1/2})$. In the first term there are at least qp nonzero $a_i r_i$, so the bound follows immediately from the original Littlewood-Offord lemma. ∎

The other modified Littlewood-Offord result we need is a similar modification of the Quadratic Littlewood-Offord lemma in [1]:

Lemma 8.4. *Let a_{ij} be fixed constants such that there are at least q indices j each with at least q indices i for which $a_{ij} \neq 0$. Let $z_1, z_2, \dots z_n$ be as in Lemma 8.2. Then for any fixed c*

$$\mathbf{P}(\sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} z_i z_j = c) = O((qp)^{-1/4}), \tag{5}$$

where the implied constant is absolute.

Proof

The proof of Lemma 8.4 relies on the use of the following application of the Cauchy-Schwartz inequality:

Lemma 8.5. *Let X and Y be random variables, and let $E(X,Y)$ be an event depending on X and Y. Let X' be an independent copy of X. Then*

$$\mathbf{P}(E(X,Y)) \leq (\mathbf{P}(E(X,Y) \wedge E(X',Y)))^{1/2}$$

Proof

We will assume here that Y takes a finite number of values $y_1, \dots, y_n$ (the general case is almost identical in proof). Note that

$$\mathbf{P}(E(X,Y)) = \sum_{i=1}^{n} \mathbf{P}(E(X,y_i))\mathbf{P}(Y = y_i)$$

and

$$\mathbf{P}(E(X,Y) \wedge E(X',Y)) = \sum_{i=1}^{n} \mathbf{P}(E(X,y_i))^2 \mathbf{P}(Y = Y_i),$$

and the result follows immediately from Cauchy-Schwartz. ∎

Without loss of generality we can assume that the q indices j given in the assumptions of our lemma are $1 \leq j \leq q$.

Define $X := (z_i)_{i>q/2}$, $Y := (z_i)_{i \leq q/2}$. Let $Q(X,Y)$ be the quadratic form in (5), and let $E(X,Y)$ be the event that that form is c. By Lemma 8.5 we have

$$\mathbf{P}(Q(X,Y) = c)^2 \leq \mathbf{P}(Q(X,Y) = Q(X',Y) = c) \leq \mathbf{P}(Q(X,Y) - Q(X',Y) = 0)$$

Thus it is enough to show the right hand side of this is $O((qp)^{-1/2})$. To estimate the right hand side, we note that

$$Q(X,Y) - Q(X',Y) = \sum_{j \leq q/2} W_j z_j + f(X,X'),$$

where

$$W_j = \sum_{i>q/2} a_{ij}(z_i - z_i')$$

and f is a quadratic form independent of Y. As in Lemma 8.2, we next condition on the number of nonzero W_j. Let I_j be the indicator variable of the event $W_j = 0$.

We have that

$$\begin{aligned}&\mathbf{P}(\sum_{j\le d/2} W_j z_j) = -f(X,X')\\ &\le \mathbf{P}(\sum_{j\le d/2} W_j z_j = -f(X,X')|\sum I_j < q/4) + \mathbf{P}(\sum I_j \ge q/4)\end{aligned}$$

By Lemma 8.2 the first term on the right hand side is $O((qp)^{-1/2})$ for any fixed value of X, so it immediately follows that the same holds true for X random.

For the second term, we note that since Y only involves $q/2$ indices, each index in Y must have at least $q/2$ indices in X with $a_{ij} \neq 0$. If follows from the remark following Lemma 8.2 that $\mathbf{E}(I_j) = O((qp)^{-1/2})$, so $\mathbf{E}(\sum I_j) = O(q(qp)^{-1/2})$. By Markov's inequality, the second term is also $O((qp)^{-1/2})$, and we are done. ■

9. Proofs of lemmas 2.12 and 2.13

The assumption that the pair (Q_m, Q_{m+1}) is normal means that the rows in Q_m which are entirely 0 have no bearing on the rank of Q_{m+1}. Thus without loss of generality we can drop those rows/columns and assume that A has no rows which are all 0, at which point A will still be singular in Lemma 2.12, but will have become nonsingular in Lemma 2.13.

Proof of Lemma 2.12.

If the new column is independent from the columns of A, then the rank increases by two after the augmentation (since the matrices are symmetric). Thus if the rank fails to increase by two then adding a new column does not increase the rank.

Assume, without loss of generality, that the rank of A is D and the first D rows $x_1, \dots, x_D$ of A are linearly independent. Then the last row x_m can be written as a linear combination of those rows in a unique way

$$x_m = \sum_{i=1}^{D} a_i x_i.$$

By throwing away those a_i which are zero, we can assume that there is some $D' \le D$ such that

$$x_m = \sum_{i=1}^{D'} a_i x_i,$$

where all $a_i \neq 0$. Recall that we defined $k = \frac{\ln \ln n}{2p}$. If $D' + 1 < k$, then there is a vertex j which is adjacent to exactly one vertex from $S = \{1, \dots, D', m\}$, thanks to the goodness of Q_m. But this is a contradiction as the jth coordinates of x_m and $\sum_{i=1}^{D'} a_i x_i$ do not match (exactly one of them is zero). Thus we can assume that $D' \geq k - 1$.

Now look at the new column $(y_1, \dots, y_m)$. Since the rank does not increase, we should have

$$x'_m = \sum_{i=1}^{D'} a_i x_i,$$

where x'_i is the extension of x_i. This implies

$$y_m = \sum_{i=1}^{D'} a_i y_i.$$

Since all a_i are non zero, by Lemma 8.2 the probability that this happens is $O((Dp)^{-1/2}) = O((kp)^{-1/2})$, concluding the proof.

Proof of Lemma 2.13. Let A be a good non-singular symmetric matrix of order m. Let A' be the $m+1$ be $m+1$ symmetric matrix obtained from A by adding a new random $(0, 1)$ column u of length $m+1$ as the $m+1$st column and its transpose as the $m+1$st row.

Let $x_1, \dots, x_{m+1}$ be the coordinates of u; x_{m+1} is the lower-right diagonal entry of A' and is zero. The determinant for A' can be expressed as

$$\sum_{i=1}^{m} \sum_{j=1}^{m} c_{ij} x_i x_j = \sum_{i=1}^{m} \sum_{j=1}^{m} c_{ij} x_i x_j = Q$$

where c_{ij} is the ij cofactor of A. It suffices to bound the probability that $Q = 0$. We can do this using Lemma 8.4 if we can show that many of the c_{ij} are nonzero.

Since A is now nonsingular, dropping any of the columns of A will lead to a $m \times (m-1)$ matrix whose rows admit (up to scaling) precisely one nontrivial linear combination equal to 0. If any of the rows in that combination are dropped, we will be left with an $(m-1) \times (m-1)$ nonsingular matrix, i.e. a nonzero cofactor.

As above, that combination cannot involve between 2 and k rows (since A is good, any set of between 2 and k rows has at least two columns with exactly one nonzero entry, and even after a column is removed there will still be one left). The combination will involve exactly 1 row only when the column removed corresponds to the only neighbor of a degree 1 vertex (which becomes isolated upon the removal of its neighbor). But by assumption there are only $\frac{1}{p \ln n} = O(\frac{n}{\ln^2 n})$ possibilities for such a neighbor.

It follows that for each index j except at most $O(\frac{n}{\ln^2 n})$ indices there are at least k indices i for which $c_{ij} \neq 0$, and we can therefore apply Lemma 8.4 with $q = k$ to get that $\mathbf{P}(Q = 0)$ is $O((pk)^{-1/4})$, proving Lemma 2.13.

10. Open Problems and Avenues for Further Research

Theorem 1.5 gives that $\frac{\ln n}{n}$ is a threshold for the singularity of $Q(n, p)$ but it would still be of interest to describe the sources of singularity once p drops below the threshold. While we have obtained such a description for $p = \Omega(\frac{\ln n}{n})$ [2], it is still unclear what happens when p drops below this range.

As noted in Remark 4.1, it is still the case at this point that $\text{rank}(Q_{n,p})/n \to 1$, and this will continue to occur until $pn = O(1)$. For y fixed and $p = y/n$, we have from consideration of isolated vertices and the bounds in Lemma 2.1 that

$$1 - O(\ln y/y) \leq (1 + o(1)) \text{ rank}(Q_{n,p})/n \leq 1 - e^{-y} = 1 - i(G)/n$$

It seems likely that $\mathbf{E}(\text{ rank}(Q_{n,y/n}))/n$ tends to some function $g(y)$ as $n \to \infty$, and it would be of interest to compute g. Azuma's inequality applied to the vertex exposure process guarantees the ratio is highly concentrated around this g, whatever it may be.

Let us now consider the case when p is above the threshold $\ln n/n$. What is the probability that $Q(n, p)$ is singular ? The current proof gives bounds which tends to zero rather slowly. For $p > n^{-\alpha}$ we can prove the singularity probability is $O(n^{-1/4(1-2\alpha)})$. However, for $p < n^{-1/2}$, we can only prove $O((\ln \ln n)^{-1/4})$. While it is certain that these bounds can be improved by tightening the arguments, it is not clear how to obtain a significant improvement. For instance, we conjecture that in the case $p = \Theta(1)$, the singularity probability is exponentially small. Such bounds are known for non-symmetric random matrices [6, 7, 8], but the proofs do not extend for the symmetric case.

The assumption of independence between the edges of G seems crucial to 1.5. In particular, the results in this paper do not yet apply to the model of random regular graphs.

Question 10.1. *For what d will the adjacency matrix of the random d-regular graph on n vertices almost surely be nonsingular?*

For $d = 1$, the matrix is trivially non-singular. For $d = 2$, the graph is union of cycles and the matrix will almost surely be singular (any cycle of length a multiple of 4 leads to a singular adjacency matrix). We conjecture that for $3 \leq d \leq n-4$, the matrix is again almost surely nonsingular.

References

[1] K. Costello, T. Tao and V. Vu, *Random symmetric matrices are almost surely non-singular*, Duke Math J. **135** (2006), 395-413.

[2] K. Costello and V. Vu, *On the rank of random sparse matrices*, Preprint.

[3] P. Erdös, *On a lemma of Littlewood and Offord*, Bull. Amer. Math. Soc. **51** (1945), 898–902.

[4] G. Halász, *Estimates for the concentration function of combinatorial number theory and probability*, Period. Math. Hungar. **8** (1977), no. 3-4, 197-211.

[5] J. Komlós, *On the determinant of* $(0,1)$ *matrices*, Studia. Sci. Math. Hungar. **2** (1967), 7-22.

[6] J. Kahn, J. Komlós, E. Szemerédi, *On the probability a random* ± 1 *matrix is singular*, J. Amer. Math Soc. **8** (1995), 223-240.

[7] T. Tao and V. Vu, *On random* ± 1 *matrices: Singularity and Determinant*, Random Structures and Algorithms **28** (2006), 1-23.

[8] T. Tao and V. Vu, *On the singularity probability of random Bernoulli matrices*, J. Amer. Math. Soc. **20** (2007), 603-628.

Institute for Advanced Study, Princeton, NJ 08540

E-mail address: kcostell@math.ias.edu

Department of Mathematics, Rutgers, Piscataway, NJ 08854

E-mail address: vanvu@math.rutgers.edu