

# UC Irvine

## UC Irvine Previously Published Works

### Title

Concentration and regularization of random graphs

### Permalink

<https://escholarship.org/uc/item/1pv262ps>

### Journal

Random Structures and Algorithms, 51(3)

### ISSN

1042-9832

### Authors

Le, CM

Levina, E

Vershynin, R

### Publication Date

2017-10-01

### DOI

10.1002/rsa.20713

### Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at <https://creativecommons.org/licenses/by/4.0/>

Peer reviewed

# CONCENTRATION AND REGULARIZATION OF RANDOM GRAPHS

CAN M. LE, ELIZAVETA LEVINA AND ROMAN VERSHYNIN

**ABSTRACT.** This paper studies how close random graphs are typically to their expectations. We interpret this question through the concentration of the adjacency and Laplacian matrices in the spectral norm. We study inhomogeneous Erdős-Rényi random graphs on  $n$  vertices, where edges form independently and possibly with different probabilities  $p_{ij}$ . Sparse random graphs whose expected degrees are  $o(\log n)$  fail to concentrate; the obstruction is caused by vertices with abnormally high and low degrees. We show that concentration can be restored if we regularize the degrees of such vertices, and one can do this in various ways. As an example, let us reweight or remove enough edges to make all degrees bounded above by  $O(d)$  where  $d = \max np_{ij}$ . Then we show that the resulting adjacency matrix  $A'$  concentrates with the optimal rate:  $\|A' - \mathbb{E} A\| = O(\sqrt{d})$ . Similarly, if we make all degrees bounded below by  $d$  by adding weight  $d/n$  to all edges, then the resulting Laplacian concentrates with the optimal rate:  $\|\mathcal{L}(A') - \mathcal{L}(\mathbb{E} A')\| = O(1/\sqrt{d})$ . Our approach is based on Grothendieck-Pietsch factorization, using which we construct a new decomposition of random graphs. We illustrate the concentration results with an application to the community detection problem in the analysis of networks.

## 1. INTRODUCTION

Many classical and modern results in probability theory, starting from the Law of Large Numbers, can be expressed as concentration of random objects about their expectations. The objects studied most are sums of independent random variables, martingales, nice functions on product probability spaces and metric measure spaces. For a panoramic exposition of concentration phenomena in modern probability theory and related fields, the reader is referred to the books [25, 9].

This paper studies concentration properties of random graphs. The first step of such study should be to decide how to interpret the statement that a *random graph  $G$  concentrates near its expectation*. To do this, it will be useful to look at the graph  $G$  through the lens of the matrices classically associated with  $G$ , namely the adjacency and Laplacian matrices.

Let us first build the theory for the adjacency matrix  $A$ ; the Laplacian will be discussed in Section 1.5. We may say that  $G$  concentrates about its expectation if  $A$  is close to its expectation  $\mathbb{E} A$  in some natural matrix norm; we interpret the expectation of  $G$  as the weighted graph with adjacency matrix  $\mathbb{E} A$ . Various matrix norms could be of interest here. In this paper, we study concentration in the spectral norm. This automatically gives us a tight control of all eigenvalues and eigenvectors, according to Weyl's and Davis-Kahan perturbation inequalities (see [5, Sections III.2 and VII.3]).

Concentration of random graphs interpreted this way, and also of general random matrices, has been studied in several communities, in particular in random matrix theory, combinatorics and network science.

---

*Date:* August 10, 2016.

E. L. is partially supported by NSF grants DMS-1159005 and DMS-1521551. R. V. is partially supported by NSF grant 1265782 and U.S. Air Force grant FA9550-14-1-0009. This work was done while C. L. was a Ph.D. student at the University of Michigan.

We will study random graphs generated from an *inhomogeneous Erdős-Rényi model*  $G(n, (p_{ij}))$ , where edges are formed independently with given probabilities  $p_{ij}$ , see [7]. This is a generalization of the classical Erdős-Rényi model  $G(n, p)$  where all edge probabilities  $p_{ij}$  equal  $p$ . Many popular graph models arise as special cases of  $G(n, (p_{ij}))$ , such as the stochastic block model, a benchmark model in the analysis of networks [22] discussed in Section 1.7, and random subgraphs of given graphs.

Often, the question of interest is estimating some features of the probability matrix  $(p_{ij})$  from random graphs drawn from  $G(n, (p_{ij}))$ . Concentration of adjacency matrix and Laplacian matrix around their expectations, when it holds, guarantees that such features can be recovered. As an example of this use of our concentration results, we will show that if  $(p_{ij})$  has a block structure, the blocks can be accurately estimated from a single realization of  $G(n, (p_{ij}))$  even when the average vertex degree is bounded.

**1.1. Dense graphs concentrate.** The cleanest concentration results are available for the classical Erdős-Rényi model  $G(n, p)$  in the *dense* regime. In terms of the expected degree  $d = pn$ , we have with high probability that

$$\|A - \mathbb{E} A\| = 2\sqrt{d}(1 + o(1)) \quad \text{if } d \gg \log^4 n, \quad (1.1)$$

see [16, 44, 28]. Since  $\|\mathbb{E} A\| = d$ , we see that the typical deviation here behaves like the square root of the magnitude of expectation – just like in many other classical results of probability theory. In other words, *dense random graphs concentrate well*.

The lower bound on density in (1.1) can be essentially relaxed all the way down to  $d = \Omega(\log n)$ . Thus, with high probability we have

$$\|A - \mathbb{E} A\| = O(\sqrt{d}) \quad \text{if } d = \Omega(\log n). \quad (1.2)$$

This result was proved in [15] based on the method developed by J. Kahn and E. Szemerédi [17]. More generally, (1.2) holds for any inhomogeneous Erdős-Rényi model  $G(n, (p_{ij}))$  with maximal expected degree  $d = \max_i \sum_j p_{ij}$ . This generalization can be deduced from a recent result of S. Bandeira and R. van Handel [4, Corollary 3.6], while a weaker bound  $O(\sqrt{d \log n})$  follows from concentration inequalities for sums of independent random matrices [35]. Alternatively, an argument in [15] can be used to prove (1.2) for a somewhat larger but still useful value

$$d = \max_{ij} np_{ij}, \quad (1.3)$$

see [27, 12]. The same can be obtained by using Seginer’s bound on random matrices [20]. As we will see shortly, our paper provides an alternative and completely different approach to general concentration results like (1.2).

**1.2. Sparse graphs do not concentrate.** In the *sparse* regime, where the expected degree  $d$  is bounded, concentration breaks down. According to [24], a random graph from  $G(n, p)$  satisfies with high probability that

$$\|A\| = (1 + o(1))\sqrt{d(A)} = (1 + o(1))\sqrt{\frac{\log n}{\log \log n}} \quad \text{if } d = O(1), \quad (1.4)$$

where  $d(A)$  denotes the maximal degree of the graph (a random quantity). So in this regime we have  $\|A\| \gg \|\mathbb{E} A\| = d$ , which shows that *sparse random graphs do not concentrate*.

What exactly makes the norm  $A$  abnormally large in the sparse regime? The answer is, the vertices with too high degrees. In the dense case where  $d \gg \log n$ , all vertices typically have approximately the same degrees  $(1 + o(1))d$ . This no longer happens in the sparser

regime  $d \ll \log n$ ; the degrees do not cluster tightly about the same value anymore. There are vertices with too high degrees; they are captured by the second inequality in (1.4). Even a single high-degree vertex can blow up the norm of the adjacency matrix. Indeed, since the norm of  $A$  is bounded below by the Euclidean norm of each of its rows, we have  $\|A\| \geq \sqrt{d(A)}$ .

**1.3. Regularization enforces concentration.** If high-degree vertices destroy concentration, can we “tame” these vertices? One proposal would be to remove these vertices from the graph altogether. U. Feige and E. Ofek [15] showed that this works for  $G(n, p)$  – *the removal of the high degree vertices enforces concentration*. Indeed, if we drop all vertices with degrees, say, larger than  $2d$ , the remaining part of the graph satisfies

$$\|A' - \mathbb{E}A'\| = O(\sqrt{d}) \quad (1.5)$$

with high probability, where  $A'$  denotes the adjacency matrix of the new graph. The argument in [15] is based on the method developed by J. Kahn and E. Szemerédi [17]. It extends to the inhomogeneous Erdős-Rényi model  $G(n, (p_{ij}))$  with  $d$  defined in (1.3), see [27, 12]. As we will see, our paper provides an alternative and completely different approach to such results.

Although the removal of high degree vertices solves the concentration problem, such solution is not ideal, since those vertices are in some sense the most important ones. In real-world networks, the vertices with highest degrees are “hubs” that hold the network together. Their removal would cause the network to break down into disconnected components, which leads to a considerable loss of structural information.

Would it be possible to regularize the graph in a more gentle way – instead of removing the high-degree vertices, reduce the weights of their edges just enough to keep the degrees bounded by  $O(d)$ ? The main result of our paper states that this is true. Let us first state this result informally; Theorem 2.1 provides a more general and formal statement.

**Theorem 1.1** (Concentration of regularized adjacency matrices). *Consider a random graph from the inhomogeneous Erdős-Rényi model  $G(n, (p_{ij}))$ , and let  $d = \max_{ij} np_{ij}$ . For all high degree vertices of the graph (say, those with degrees larger than  $2d$ ), reduce the weights of the edges incident to them in an arbitrary way, but so that all degrees of the new (weighted) graph become bounded by  $2d$ . Then, with high probability, the adjacency matrix  $A'$  of the new graph concentrates:*

$$\|A' - \mathbb{E}A'\| = O(\sqrt{d}).$$

*Moreover, instead of requiring that the degrees become bounded by  $2d$ , we can require that the  $\ell_2$  norms of the rows of the new adjacency matrix become bounded by  $\sqrt{2d}$ .*

**1.4. Examples of graph regularization.** The regularization procedure in Theorem 1.1 is very flexible. Depending on how one chooses the weights, one can obtain as partial cases several results we summarized earlier, as well as some new ones.

1. *Do not do anything to the graph.* In the dense regime where  $d = \Omega(\log n)$ , all degrees are already bounded by  $2d$  with high probability. This means that the original graph satisfies  $\|A - \mathbb{E}A\| = O(\sqrt{d})$ . Thus we recover the result of U. Feige and E. Ofek (1.2), which states that *dense random graphs concentrate well*.
2. *Remove all high degree vertices.* If we remove all vertices with degrees larger than  $2d$ , we recover another result of U. Feige and E. Ofek (1.5), which states that *the removal of the high degree vertices enforces concentration*.
3. *Remove just enough edges from high-degree vertices.* Instead of removing the high-degree vertices with all of their edges, we can remove just enough edges to make all degrees bounded by  $2d$ . This milder regularization still produces the concentration bound (1.5).

4. *Reduce the weight of edges proportionally to the excess of degrees.* Instead of removing edges, we can reduce the weight of the existing edges, a procedure which better preserves the structure of the graph. For instance, we can assign weight  $\sqrt{\lambda_i \lambda_j}$  to the edge between vertices  $i$  and  $j$ , choosing  $\lambda_i := \min(2d/d_i, 1)$  where  $d_i$  is the degree of vertex  $i$ . One can check that this makes the  $\ell_2$  norms of all rows of the adjacency matrix bounded by  $2d$ . By Theorem 1.1, such regularization procedure leads to the same concentration bound (1.5).

**1.5. Concentration of Laplacian.** So far, we have looked at random graphs through the lens of their adjacency matrices. A different matrix that captures the geometry of a graph is the (symmetric, normalized) Laplacian matrix, defined as

$$\mathcal{L}(A) = D^{-1/2}(D - A)D^{-1/2} = I - D^{-1/2}AD^{-1/2}. \quad (1.6)$$

Here  $I$  is the identity matrix and  $D = \text{diag}(d_i)$  is the diagonal matrix with degrees  $d_i = \sum_{j=1}^n A_{ij}$  on the diagonal. The reader is referred to [13] for an introduction to graph Laplacians and their role in spectral graph theory. Here we mention just two basic facts: the spectrum of  $\mathcal{L}(A)$  is a subset of  $[0, 2]$ , and the smallest eigenvalue is always zero.

Concentration of Laplacians of random graphs has been studied in [35, 11, 39, 23, 18]. Just like the adjacency matrix, the Laplacian is known to concentrate in the dense regime where  $d = \Omega(\log n)$ , and it fails to concentrate in the sparse regime. However, the obstructions to concentration are opposite. For the adjacency matrices, as we mentioned, the trouble is caused by high-degree vertices. For the Laplacian, the problem lies with *low-degree vertices*. In particular, for  $d = o(\log n)$  the graph is likely to have isolated vertices; they produce multiple zero eigenvalues of  $\mathcal{L}(A)$ , which are easily seen to destroy the concentration.

In analogy to our discussion of adjacency matrices, we can try to regularize the graph to “tame” the low-degree vertices in various ways, for example remove the low-degree vertices, connect them to some other vertices, artificially increase the degrees  $d_i$  in the definition (1.6) of Laplacian, and so on. Here we will focus on the following simple way of regularization proposed in [3] and analyzed in [23, 18]. Choose  $\tau > 0$  and add the same number  $\tau/n$  to all entries of the adjacency matrix  $A$ , thereby replacing it with

$$A_\tau := A + (\tau/n)\mathbf{1}\mathbf{1}^\top$$

in the definition (1.6) of the Laplacian. This regularization raises all degrees  $d_i$  to  $d_i + \tau$ . If we choose  $\tau \sim d$ , the regularized graph does not have low-degree vertices anymore.

The following consequence of Theorem 1.1 shows that such regularization indeed forces Laplacian to concentrate. Here we state this result informally; Theorem 4.1 provides a more formal statement.

**Theorem 1.2** (Concentration of the regularized Laplacian). *Consider a random graph from the inhomogeneous Erdős-Rényi model  $G(n, (p_{ij}))$ , and let  $d = \max_{i,j} np_{ij}$ . Choose a number  $\tau \sim d$ . Then, with high probability, the regularized Laplacian  $\mathcal{L}(A_\tau)$  concentrates:*

$$\|\mathcal{L}(A_\tau) - \mathcal{L}(\mathbb{E} A_\tau)\| = O\left(\frac{1}{\sqrt{d}}\right).$$

We will deduce this result from Theorem 1.1 in Section 4. Theorem 1.2 is an improvement upon a bound in [18] that had an extra  $\log d$  factor, and it was conjectured there that the logarithmic factor is not needed. Theorem 1.2 confirms this conjecture.

**1.6. A numerical experiment.** To conclude our discussion of various ways to regularize sparse graphs, let us illustrate the effect of regularization by a numerical experiment. Consider an inhomogeneous Erdős-Rényi graph with  $n = 1000$  vertices, 90% of which have expected degrees 7 and 10% percent have expected degrees 35. We then regularize the graph by reducing the weights of edges proportionally to the excess of degrees – just as we described in Section 1.4 item 4, except that we use the overall average degree (approximately 10) instead of  $d$  (which results in a more severe weight reduction suitable for our illustration purpose).

Figure 1 shows the histogram of the spectrum of  $A$  (left) and  $A'$  (right). As we can see, the high degree vertices lead to the long tails in the histogram of the eigenvalues, and regularization shrinks these tails toward the bulk.

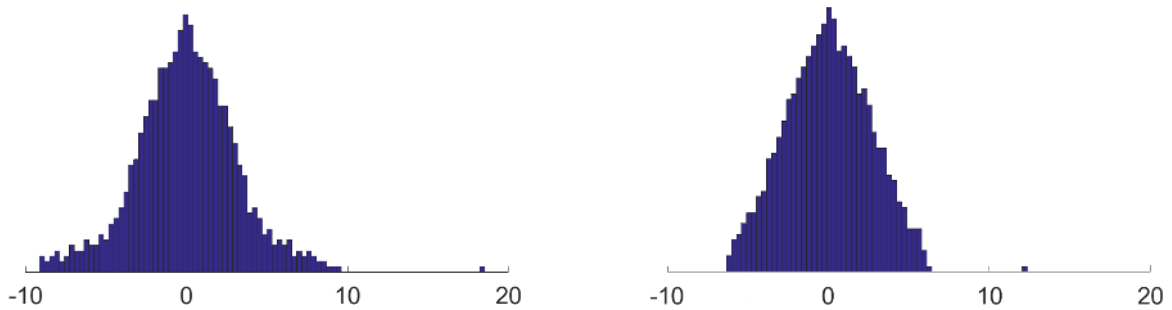


Figure 1. Histogram of the spectrum of adjacency matrix  $A$  (left) and regularized adjacency matrix  $A'$  (right) for a sparse random graph generated from the inhomogeneous Erdős-Rényi model with  $n = 1000$  vertices, 90% of which have expected degrees 7 and 10% percent have expected degrees 35.

**1.7. Application: community detection in networks.** Concentration of random graphs has an important application to statistical analysis of networks, in particular to the problem of community detection. A common way of modeling communities in networks is the *stochastic block model* [22], which is a special case of the inhomogeneous Erdős-Rényi model considered in this paper. For the purpose of this example, we focus on the simplest version of the stochastic block model  $G(n, \frac{a}{n}, \frac{b}{n})$ , also known as the balanced planted partition model, defined as follows. The set of vertices is divided into two subsets (communities) of size  $n/2$  each. Edges between vertices are drawn independently with probability  $a/n$  if they are in the same community and with probability  $b/n$  otherwise.

The community detection problem is to recover the community labels of vertices from a single realization of the random graph model. A large literature exists on both the recovery algorithms and the theory establishing when a recovery is possible [14, 33, 34, 32, 1, 29, 8]. There are methods that perform *better than a random guess* (i.e. the fraction of misclassified vertices is bounded away from 0.5 as  $n \rightarrow \infty$  with high probability) under the condition

$$(a - b)^2 > 2(a + b),$$

and no method can perform better than a random guess if this condition is violated.

Moreover, *strong consistency*, or exact recovery (labeling *all* vertices correctly with high probability) is possible when the expected degree  $(a + b)/2$  is of order  $\log n$  or larger and  $a$  and  $b$  are sufficiently separated, see [32, 30, 6, 20, 10]. *Weak consistency* (the fraction of mislabeled vertices going to 0 with high probability) is achievable if and only if

$$(a - b)^2 > C_n(a + b) \quad \text{with} \quad C_n \rightarrow \infty,$$

see [32]. Many of these results hold in the non-asymptotic regime, for graphs of fixed size  $n$ . Thus, for any  $\varepsilon > 0$  there exists  $C_\varepsilon$  (which only depends on  $\varepsilon$ ) such that one can recover communities up to  $\varepsilon n$  mislabeled vertices as long as

$$(a - b)^2 > C_\varepsilon(a + b).$$

In particular, recovery of communities is possible even for very sparse graphs – those with bounded expected degrees. Several types of algorithms are known to succeed in this regime, including non-backtracking walks [33, 29, 8], spectral methods [12] and methods based on semidefinite programming [19, 31].

As an application of the new concentration results, we show that the *regularized spectral clustering* [3, 23], one of the simplest most popular algorithms for community detection, can recover communities in the sparse regime. In general, spectral clustering works by computing the leading eigenvectors of either the adjacency matrix or the Laplacian or their regularized versions, and running the  $k$ -means clustering algorithm on these eigenvectors to recover the node labels. In the simple case of the model  $G(n, \frac{a}{n}, \frac{b}{n})$ , one can simply assign nodes to communities based on the sign (positive or negative) of the corresponding entries of the eigenvector  $v_2(\mathcal{L}(A_\tau))$  corresponding to the second smallest eigenvalue of regularized Laplacian matrix  $\mathcal{L}(A_\tau)$  (or the regularized adjacency matrix  $A'$ ).

Let us briefly explain how our concentration results validate regularized spectral clustering. If the concentration of random graphs holds and  $\mathcal{L}(A_\tau)$  is close to  $\mathcal{L}(\mathbb{E} A_\tau)$ , then the standard perturbation theory (Davis-Kahan theorem below) shows that  $v_2(\mathcal{L}(A_\tau))$  is close to  $v_2(\mathcal{L}(\mathbb{E} A_\tau))$ , and in particular, the signs of these two eigenvectors must agree on most vertices. An easy calculation shows that the signs of  $v_2(\mathcal{L}(\mathbb{E} A_\tau))$  recover the communities exactly: this vector is a positive constant on one community and a negative constant on the other. Therefore, the signs of  $v_2(\mathcal{L}(A_\tau))$  must recover the communities up to a small fraction of misclassified vertices.

Before stating our result, let us quote a simple version of the Davis-Kahan theorem perturbation theorem (see e.g. [5, Theorem VII.3.2]).

**Theorem 1.3** (Davis-Kahan theorem). *Let  $X, Y$  be symmetric matrices such that the second smallest eigenvalues of  $X$  and  $Y$  have multiplicity one and they are of distance at least  $\delta > 0$  from the remaining eigenvalues of  $X$  and  $Y$ . Denote by  $x$  and  $y$  the eigenvectors of  $X$  and  $Y$  corresponding to the second largest eigenvalues of  $X$  and  $Y$ , respectively. Then*

$$\min_{\beta=\pm 1} \|x + \beta y\| \leq \frac{2\|X - Y\|}{\delta}.$$

**Corollary 1.4** (Community detection in sparse graphs). *Let  $\varepsilon > 0$  and  $r \geq 1$ . Let  $A$  be the adjacency matrix drawn from the stochastic block model  $G(n, \frac{a}{n}, \frac{b}{n})$ . Assume that*

$$(a - b)^2 > C_\varepsilon(a + b) \tag{1.7}$$

where  $C_\varepsilon = Cr^4\varepsilon^{-2}$  and  $C$  is an appropriately large absolute constant. Choose  $\tau$  to be the average degree of the graph, i.e.  $\tau = (d_1 + \dots + d_n)/n$  where  $d_i$  are vertex degrees. Then with probability at least  $1 - e^{-r}$ , we have

$$\min_{\beta=\pm 1} \|v_2(\mathcal{L}(A_\tau)) + \beta v_2(\mathcal{L}(\mathbb{E} A_\tau))\| \leq \varepsilon.$$

*In particular, the signs of the entries of  $v_2(\mathcal{L}(A_\tau))$  correctly estimate the partition into the two communities, up to at most  $\varepsilon n$  misclassified vertices.*

*Proof.* We apply Theorem 1.3 with  $X = \mathcal{L}(A_\tau)$  and  $Y = \mathcal{L}(\mathbb{E} A_\tau)$ . A simple calculation shows that the spectral gap  $\delta$  defined in Theorem 1.3 is of the order  $(a - b)/(a + b)$ . The claim of Corollary 1.4 then follows from Davis-Kahan Theorem 1.3, Concentration Theorem 4.1 (which is a formal version of Theorem 1.2) and condition (1.7).  $\square$

**1.8. Organization of the paper.** In Section 2, we state a formal version of Theorem 1.1. We show there how to deduce this result from a new decomposition of random graphs, which we state as Theorem 2.6. We prove this decomposition theorem in Section 3. In Section 4, we state and prove a formal version of Theorem 1.2 about the concentration of the Laplacian. We conclude the paper with Section 5 where we propose some questions for further investigation.

**Acknowledgement.** The authors are grateful to Ramon van Handel for several insightful comments on the preliminary version of this paper.

## 2. FULL VERSION OF THEOREM 1.1, AND REDUCTION TO A GRAPH DECOMPOSITION

In this section we state a more general and quantitative version of Theorem 1.1, and we reduce it to a new form of graph decomposition, which can be of interest on its own.

**Theorem 2.1** (Concentration of regularized adjacency matrices). *Consider a random graph from the inhomogeneous Erdős-Rényi model  $G(n, (p_{ij}))$ , and let  $d = \max_{ij} np_{ij}$ . For any  $r \geq 1$ , the following holds with probability at least  $1 - n^{-r}$ . Consider any subset consisting of at most  $10n/d$  vertices, and reduce the weights of the edges incident to those vertices in an arbitrary way. Let  $d'$  be the maximal degree of the resulting graph. Then the adjacency matrix  $A'$  of the new (weighted) graph satisfies*

$$\|A' - \mathbb{E} A\| \leq Cr^{3/2}(\sqrt{d} + \sqrt{d'}).$$

Moreover, the same bound holds for  $d'$  being the maximal  $\ell_2$  norm of the rows of  $A'$ .

In this result and in the rest of the paper,  $C, C_1, C_2, \dots$  denote absolute constants whose values may be different from line to line.

*Remark 2.2* (Theorem 2.1 implies Theorem 1.1). The subset of  $10n/d$  vertices in Theorem 2.1 can be completely arbitrary. So let us choose the high-degree vertices, say those with degrees larger than  $2d$ . There are at most  $10n/d$  such vertices with high probability; this follows by an easy calculation, and also from Lemma 3.5. Thus we immediately deduce Theorem 1.1.

*Remark 2.3* (Tight upper bound). If we do not reduce weights of any edges and  $d$  is bounded, then the upper bound in Theorem 2.1 is tight (up to a constant depending on  $r$ ). This is because of (1.4), which states that the adjacency matrix does not concentrate in the sparse regime without regularization.

*Remark 2.4* (Method to prove Theorem 2.1). One may wonder if Theorem 2.1 can be proved by developing an  $\epsilon$ -net argument similar to the method of J. Kahn and E. Szemerédi [17] and its versions [2, 15, 27, 12]. Although we can not rule out such possibility, we do not see how this method could handle a general regularization. The reader familiar with the method can easily notice an obstacle. The contribution of the so-called light couples becomes hard to control when one changes, and even reduces, the individual entries of  $A$  (the weights of edges).

We will develop an alternative and somewhat simpler approach, which will be able to handle a general regularization of random graphs. It sheds light on the specific structure of graphs that enables concentration. We are going to identify this structure through a *graph*



*decomposition* in the next section. But let us pause briefly to mention the following useful reduction.

*Remark 2.5* (Reduction to directed graphs). Our arguments will be more convenient to carry out if the adjacency matrix  $A$  has all independent entries. To be able to make this assumption, we can decompose  $A$  into the upper-triangular and a lower-triangular parts, both of which have independent entries. If we can show that each of these parts concentrate about its expectation, it would follow that  $A$  concentrate about  $\mathbb{E}A$  by triangle inequality.

In other words, we may prove Theorem 2.1 for *directed* inhomogeneous Erdős-Rényi graphs, where edges between any vertices and in any direction appear independently with probabilities  $p_{ij}$ . In the rest of the argument, we will only work with such random directed graphs.

**2.1. Graph decomposition.** In this section, we reduce Theorem 2.1 to the following decomposition of inhomogeneous Erdős-Rényi directed random graphs. This decomposition may have an independent interest. Throughout the paper, we denote by  $B_{\mathcal{N}}$  the matrix which coincides with a matrix  $B$  on a subset of edges  $\mathcal{N} \subset [n] \times [n]$  and has zero entries elsewhere.

**Theorem 2.6** (Graph decomposition). *Consider a random directed graph from the inhomogeneous Erdős-Rényi model, and let  $d$  be as in (1.3). For any  $r \geq 1$ , the following holds with probability at least  $1 - 3n^{-r}$ . One can decompose the set of edges  $[n] \times [n]$  into three classes  $\mathcal{N}$ ,  $\mathcal{R}$  and  $\mathcal{C}$  so that the following properties are satisfied for the adjacency matrix  $A$ .*

- The graph concentrates on  $\mathcal{N}$ , namely  $\|(A - \mathbb{E}A)_{\mathcal{N}}\| \leq Cr^{3/2}\sqrt{d}$ .
- Each row of  $A_{\mathcal{R}}$  and each column of  $A_{\mathcal{C}}$  has at most  $32r$  ones.

Moreover,  $\mathcal{R}$  intersects at most  $n/d$  columns, and  $\mathcal{C}$  intersects at most  $n/d$  rows of  $[n] \times [n]$ .

Figure 2 illustrates a possible decomposition Theorem 2.6 can provide. The edges in  $\mathcal{N}$  form a big “core” where the graph concentrates well even without regularization. The edges in  $\mathcal{R}$  and  $\mathcal{C}$  can be thought of (at least heuristically) as those attached to high-degree vertices.

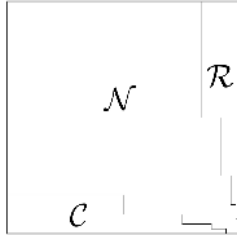


Figure 2. An example of graph decomposition in Theorem 2.6.

We will prove Theorem 2.6 in Section 3; let us pause to deduce Theorem 2.1 from it.

**2.2. Deduction of Theorem 2.1.** First, let us explain informally how the graph decomposition could lead to Theorem 2.1. The regularization of the graph does not destroy the properties of  $\mathcal{N}$ ,  $\mathcal{R}$  and  $\mathcal{C}$  in Theorem 2.6. Moreover, regularization creates a new property for us, allowing for a good control of the *columns* of  $\mathcal{R}$  and *rows* of  $\mathcal{C}$ . Let us focus on  $A_{\mathcal{R}}$  to be specific. The  $\ell_1$  norms of all columns of this matrix are at most  $d'$ , and the  $\ell_1$  norms of all rows are  $O(r)$  by Theorem 2.6. By a simple calculation which we will do in Lemma 2.7, this implies that  $\|A_{\mathcal{R}}\| = O(\sqrt{rd'})$ . A similar bound can be proved for  $\mathcal{C}$ . Combining  $\mathcal{N}$ ,  $\mathcal{R}$  and  $\mathcal{C}$  together will lead to the error bound  $O(r^{3/2}(\sqrt{d} + \sqrt{d'}))$  in Theorem 2.1.

To make this argument rigorous, let us start with the simple calculation we just mentioned.

**Lemma 2.7.** *Consider a matrix  $B$  in which each row has  $\ell_1$  norm at most  $a$ , and each column has  $\ell_1$  norm at most  $b$ . Then  $\|B\| \leq \sqrt{ab}$ .*

*Proof.* Let  $x$  be a vector with  $\|x\|_2 = 1$ . Using Cauchy-Schwarz inequality and the assumptions, we have

$$\begin{aligned} \|Bx\|_2^2 &= \sum_i \left( \sum_j B_{ij} x_j \right)^2 \leq \sum_i \left( \sum_j |B_{ij}| \sum_j |B_{ij}| x_j^2 \right) \\ &\leq \sum_i \left( a \sum_j |B_{ij}| x_j^2 \right) = a \sum_j x_j^2 \sum_i |B_{ij}| \leq a \sum_j x_j^2 b = ab. \end{aligned}$$

Since  $x$  is arbitrary, this completes the proof.  $\square$

*Remark 2.8* (Riesz-Thorin interpolation theorem implies Lemma 2.7). Lemma 2.7 can also be deduced from Riesz-Thorin interpolation theorem (see e.g. [40, Theorem 2.1]), since the maximal  $\ell_1$  norm of columns is the  $\ell_1 \rightarrow \ell_1$  operator norm, and the maximal  $\ell_1$  norm of rows is the  $\ell_\infty \rightarrow \ell_\infty$  operator norm.

We are ready to formally deduce the main part of Theorem 2.1 from Theorem 2.6; we defer the “moreover” part to Section 3.6.

*Proof of Theorem 2.1 (main part).* Fix a realization of the random graph that satisfies the conclusion of Theorem 2.6, and decompose the deviation  $A' - \mathbb{E}A$  as follows:

$$A' - \mathbb{E}A = (A' - \mathbb{E}A)_{\mathcal{N}} + (A' - \mathbb{E}A)_{\mathcal{R}} + (A' - \mathbb{E}A)_{\mathcal{C}}. \quad (2.1)$$

We will bound the spectral norm of each of the three terms separately.

**Step 1. Deviation on  $\mathcal{N}$ .** Let us further decompose

$$(A' - \mathbb{E}A)_{\mathcal{N}} = (A - \mathbb{E}A)_{\mathcal{N}} - (A - A')_{\mathcal{N}}. \quad (2.2)$$

By Theorem 2.6,  $\|(A - \mathbb{E}A)_{\mathcal{N}}\| \leq Cr^{3/2}\sqrt{d}$ . To control the second term in (2.2), denote by  $\mathcal{E} \subset [n] \times [n]$  the subset of edges that are reweighted in the regularization process. Since  $A$  and  $A'$  are equal on  $\mathcal{E}^c$ , we have

$$\begin{aligned} \|(A - A')_{\mathcal{N}}\| &= \|(A - A')_{\mathcal{N} \cap \mathcal{E}}\| \leq \|A_{\mathcal{N} \cap \mathcal{E}}\| \quad (\text{since } 0 \leq A - A' \leq A \text{ entrywise}) \\ &\leq \|(A - \mathbb{E}A)_{\mathcal{N} \cap \mathcal{E}}\| + \|\mathbb{E}A_{\mathcal{N} \cap \mathcal{E}}\| \quad (\text{by triangle inequality}). \end{aligned} \quad (2.3)$$

Further, a simple restriction property implies that

$$\|(A - \mathbb{E}A)_{\mathcal{N} \cap \mathcal{E}}\| \leq 2\|(A - \mathbb{E}A)_{\mathcal{N}}\|. \quad (2.4)$$

Indeed, restricting a matrix onto a product subset of  $[n] \times [n]$  can only reduce its norm. Although the set of reweighted edges  $\mathcal{E}$  is not a product subset, it can be decomposed into two product subsets:

$$\mathcal{E} = (I \times [n]) \cup (I^c \times I) \quad (2.5)$$

where  $I$  is the subset of vertices incident to the edges in  $\mathcal{E}$ . Then (2.4) holds; right hand side of that inequality is bounded by  $Cr^{3/2}\sqrt{d}$  by Theorem 2.6. Thus we handled the first term in (2.3).

To bound the second term in (2.3), we can use another restriction property that states that the norm of the matrix with non-negative entries can only reduce by restricting onto any subset of  $[n] \times [n]$  (whether a product subset or not). This yields

$$\|\mathbb{E}A_{\mathcal{N} \cap \mathcal{E}}\| \leq \|\mathbb{E}A_{\mathcal{E}}\| \leq \|\mathbb{E}A_{I \times [n]}\| + \|\mathbb{E}A_{I^c \times I}\| \quad (2.6)$$

where the second inequality follows by (2.5). By assumption, the matrix  $\mathbb{E} A_{I \times [n]}$  has  $|I| \leq 10n/d$  rows and each of its entries is bounded by  $d/n$ . Hence the  $\ell_1$  norm of all rows is bounded by  $d$ , and the  $\ell_1$  norm of all columns is bounded by 10. Lemma 2.7 implies that  $\|\mathbb{E} A_{I \times [n]}\| \leq \sqrt{10d}$ . A similar bound holds for the second term of (2.6). This yields

$$\|\mathbb{E} A_{\mathcal{N} \cap \mathcal{E}}\| \leq 5\sqrt{d},$$

so we handled the second term in (2.3). Recalling that the first term there is bounded by  $Cr^{3/2}\sqrt{d}$ , we conclude that  $\|(A - A')_{\mathcal{N}}\| \leq 2Cr^{3/2}\sqrt{d}$ .

Returning to (2.2), we recall that the first term in the right hand is bounded by  $Cr^{3/2}\sqrt{d}$ , and we just bounded the second term by  $2Cr^{3/2}\sqrt{d}$ . Hence

$$\|(A' - \mathbb{E} A)_{\mathcal{N}}\| \leq 4Cr^{3/2}\sqrt{d}.$$

**Step 2. Deviation on  $\mathcal{R}$  and  $\mathcal{C}$ .** By triangle inequality, we have

$$\|(A' - \mathbb{E} A)_{\mathcal{R}}\| \leq \|A'_{\mathcal{R}}\| + \|\mathbb{E} A_{\mathcal{R}}\|.$$

Recall that  $0 \leq A'_{\mathcal{R}} \leq A_{\mathcal{R}}$  entrywise. By Theorem 2.6, each of the rows of  $A_{\mathcal{R}}$ , and thus also of  $A'_{\mathcal{R}}$ , has  $\ell_1$  norm at most  $32r$ . Moreover, by definition of  $d'$ , each of the columns of  $A'$ , and thus also of  $A'_{\mathcal{R}}$ , has  $\ell_1$  norm at most  $d'$ . Lemma 2.7 implies that  $\|A'_{\mathcal{R}}\| \leq \sqrt{32rd'}$ .

The matrix  $\mathbb{E} A_{\mathcal{R}}$  can be handled similarly. By Theorem 2.6, it has at most  $n/d$  entries in each row, and all entries are bounded by  $d/n$ . Thus each column of  $\mathbb{E} A_{\mathcal{R}}$  has  $\ell_1$  norm at most 1, and each row has  $\ell_1$  norm at most  $d$ . Lemma 2.7 implies that  $\|\mathbb{E} A_{\mathcal{R}}\| \leq \sqrt{d}$ .

We showed that

$$\|(A' - \mathbb{E} A)_{\mathcal{R}}\| \leq \sqrt{32rd'} + \sqrt{d}.$$

A similar bound holds for  $\|(A' - \mathbb{E} A)_{\mathcal{C}}\|$ . Combining the bounds on the deviation of  $A' - \mathbb{E} A$  on  $\mathcal{N}$ ,  $\mathcal{R}$  and  $\mathcal{C}$  and putting them into (2.1), we conclude that

$$\|A' - \mathbb{E} A\| \leq 4Cr^{3/2}\sqrt{d} + 2(\sqrt{32rd'} + \sqrt{d}).$$

Simplifying this inequality, we complete the proof of the main part of Theorem 2.1.  $\square$

### 3. PROOF OF DECOMPOSITION THEOREM 2.6

**3.1. Outline of the argument.** We will construct the decomposition in Theorem 2.6 by an iterative procedure. The first and crucial step is to find a big block<sup>1</sup>  $\mathcal{N}' \subset [n] \times [n]$  of size at least  $(n - n/d) \times n/2$  on which  $A$  concentrates, i.e.

$$\|(A - \mathbb{E} A)_{\mathcal{N}'}\| = O(\sqrt{d}).$$

To find such block, we first establish concentration in  $\ell_\infty \rightarrow \ell_2$  norm; this can be done by standard probabilistic techniques. Next, we can automatically upgrade this to concentration in the spectral norm ( $\ell_2 \rightarrow \ell_2$ ) once we pass to an appropriate block  $\mathcal{N}'$ . This can be done using a general result from functional analysis, which we call Grothendieck-Pietsch factorization.

Repeating this argument for the transpose, we obtain another block  $\mathcal{N}''$  of size at least  $n/2 \times (n - n/d)$  where the graph concentrates as well. So the graph concentrates on  $\mathcal{N}_0 := \mathcal{N}' \cup \mathcal{N}''$ . The ‘‘core’’  $\mathcal{N}_0$  will form the first part of the class  $\mathcal{N}$  we are constructing.

It remains to control the graph on the complement of  $\mathcal{N}_0$ . That set of edges is quite small; it can be described as a union of a block  $\mathcal{C}_0$  with  $n/d$  rows, a block  $\mathcal{R}_0$  with  $n/d$  columns and

<sup>1</sup>In this paper, by block we mean a product set  $I \times J$  with arbitrary index subsets  $I, J \subset [n]$ . These subsets are not required to be intervals of successive integers.

an exceptional  $n/2 \times n/2$  block; see Figure 3b for illustration. We may consider  $\mathcal{C}_0$  and  $\mathcal{R}_0$  as the first parts of the future classes  $\mathcal{C}$  and  $\mathcal{R}$  we are constructing.

Indeed, since  $\mathcal{C}_0$  has so few rows, the expected number of ones in each column of  $\mathcal{C}_0$  is bounded by 1. For simplicity, let us think that all columns of  $\mathcal{C}_0$  have  $O(1)$  ones as desired. (In the formal argument, we will add the bad columns to the exceptional block.) Of course, the block  $\mathcal{R}_0$  can be handled similarly.

At this point, we decomposed  $[n] \times [n]$  into  $\mathcal{N}_0, \mathcal{R}_0, \mathcal{C}_0$  and an exceptional  $n/2 \times n/2$  block. Now we repeat the process for the exceptional block, constructing  $\mathcal{N}_1, \mathcal{R}_1$ , and  $\mathcal{C}_1$  there, and so on. Figure 3c illustrates this process. At the end, we choose  $\mathcal{N}, \mathcal{R}$  and  $\mathcal{C}$  to be the unions of the blocks  $\mathcal{N}_i, \mathcal{R}_i$  and  $\mathcal{C}_i$  respectively.

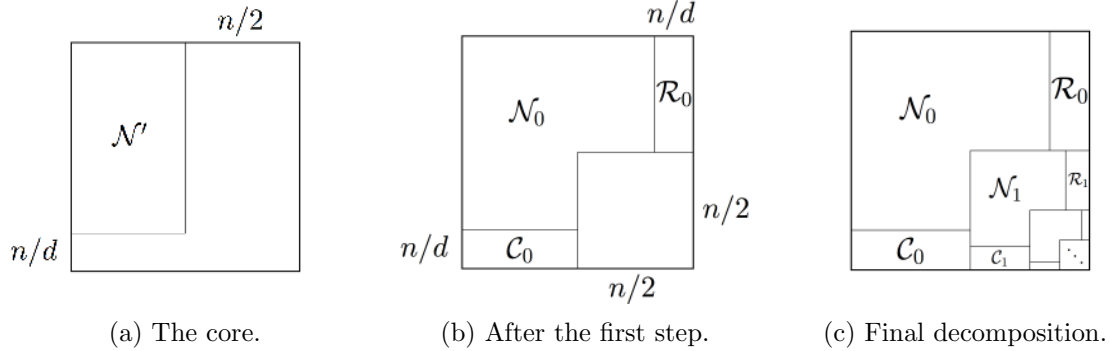


Figure 3. Constructing decomposition iteratively in the proof of Theorem 2.6.

Two precautions have to be taken in this argument. First, we need to make concentration on the core blocks  $\mathcal{N}_i$  *better at each step*, so that the sum of those error bounds would not depend of the total number of steps. This can be done with little effort, with the help of the exponential decrease of the size of the blocks  $\mathcal{N}_i$ . Second, we have a control of the sizes but not locations of the exceptional blocks. Thus to be able to carry out the decomposition argument inside an exceptional block, we need to make the argument valid *uniformly* over all blocks of that size. This will require us to be delicate with probabilistic arguments, so we can take a union bound over such blocks.

**3.2. Grothendieck-Pietsch factorization.** As we mentioned in the previous section, our proof of Theorem 2.6 is based on Grothendieck-Pietsch factorization. This general and well known result in functional analysis [36, 37] has already been used in a similar probabilistic context, see [26, Proposition 15.11].

Grothendieck-Pietsch factorization compares two matrix norms, the  $\ell_2 \rightarrow \ell_2$  norm (which we call the spectral norm) and the  $\ell_\infty \rightarrow \ell_2$  norm. For a  $k \times m$  matrix  $B$ , these norms are defined as

$$\|B\| = \max_{\|x\|_2=1} \|Bx\|_2, \quad \|B\|_{\infty \rightarrow 2} = \max_{\|x\|_\infty=1} \|Bx\|_2 = \max_{x \in \{-1,1\}^m} \|Bx\|_2.$$

The  $\ell_\infty \rightarrow \ell_2$  norm is usually easier to control, since the supremum is taken with respect to the discrete set  $\{-1,1\}^m$ , and any vector there has all coordinates of the same magnitude.

To compare the two norms, one can start with the obvious inequality

$$\frac{\|B\|_{\infty \rightarrow 2}}{\sqrt{m}} \leq \|B\| \leq \|B\|_{\infty \rightarrow 2}.$$

Both parts of this inequality are optimal, so there is an unavoidable slack between the upper and lower bounds. However, Grothendieck-Pietsch factorization allows us to tighten the inequality by changing  $B$  slightly. The next two results offer two ways to change  $B$  – introduce weights and pass to a sub-matrix.

**Theorem 3.1** (Grothendieck-Pietsch’s factorization, weighted version). *Let  $B$  be a  $k \times m$  real matrix. Then there exist positive weights  $\mu_j$  with  $\sum_{j=1}^m \mu_j = 1$  such that*

$$\|B\|_{\infty \rightarrow 2} \leq \|BD_{\mu}^{-1/2}\| \leq \sqrt{\pi/2} \|B\|_{\infty \rightarrow 2}, \quad (3.1)$$

where  $D_{\mu} = \text{diag}(\mu_j)$  denotes the  $m \times m$  diagonal matrix with weights  $\mu_j$  on the diagonal.

This result is a known combination of the Little Grothendieck Theorem (see [41, Corollary 10.10] and [38]) and Pietsch Factorization (see [41, Theorem 9.2]). In an explicit form, a version of this result can be found e.g. in [26, Proposition 15.11]. The weights  $\mu_j$  can be computed algorithmically, see [42].

The following related version of Grothendieck-Pietsch’s factorization can be especially useful in probabilistic contexts, see [26, Proposition 15.11]. Here and in the rest of the paper, we denote by  $B_{I \times J}$  the sub-matrix of a matrix  $B$  with rows indexed by a subset  $I$  and columns indexed by a subset  $J$ .

**Theorem 3.2** (Grothendieck-Pietsch factorization, sub-matrix version). *Let  $B$  be a  $k \times m$  real matrix and  $\delta > 0$ . Then there exists  $J \subseteq [m]$  with  $|J| \geq (1 - \delta)m$  such that*

$$\|B_{[k] \times J}\| \leq \frac{2\|B\|_{\infty \rightarrow 2}}{\sqrt{\delta m}}.$$

*Proof.* Consider the weights  $\mu_i$  given by Theorem 3.1, and choose  $J$  to consist of the indices  $j$  satisfying  $\mu_j \leq 1/(\delta m)$ . Since  $\sum_j \mu_j = 1$ , the set  $J$  must contain at least  $(1 - \delta)m$  indices as claimed. Furthermore, the diagonal entries of  $(D_{\mu}^{-1/2})_{J \times J}$  are all bounded from below by  $\sqrt{\delta m}$ , which yields

$$\|(BD_{\mu}^{-1/2})_{[k] \times J}\| \geq \sqrt{\delta m} \|B_{[k] \times J}\|.$$

On the other hand, by (3.1) the left-hand side of this inequality is bounded by  $2\|B\|_{\infty \rightarrow 2}$ . Rearranging the terms, we complete the proof.  $\square$

**3.3. Concentration on a big block.** We are starting to work toward constructing the core part  $\mathcal{N}$  in Theorem 2.6. In this section we will show how to find a big block on which the adjacency matrix  $A$  concentrates. First we will establish concentration in  $\ell_{\infty} \rightarrow \ell_2$  norm, and then, using Grothendieck-Pietsch factorization, in the spectral norm.

The lemmas of this and next section should be best understood for  $m = n$ ,  $I = J = [n]$  and  $\alpha = 1$ . In this case, we are working with the entire adjacency matrix, and trying to make the first step in the iterative procedure. The further steps will require us to handle smaller blocks  $I \times J$ ; the parameter  $\alpha$  will then become smaller in order to achieve better concentration for smaller blocks.

**Lemma 3.3** (Concentration in  $\ell_{\infty} \rightarrow \ell_2$  norm). *Let  $1 \leq m \leq n$  and  $\alpha \geq m/n$ . Then for  $r \geq 1$  the following holds with probability at least  $1 - n^{-r}$ . Consider a block  $I \times J$  of size  $m \times m$ . Let  $I'$  be the set of indices of the rows of  $A_{I \times J}$  that contain at most  $\alpha d$  ones. Then*

$$\|(A - \mathbb{E}A)_{I' \times J}\|_{\infty \rightarrow 2} \leq C\sqrt{\alpha d m r \log(en/m)}. \quad (3.2)$$

*Proof.* By definition,

$$\|(A - \mathbb{E} A)_{I' \times J}\|_{\infty \rightarrow 2}^2 = \max_{x \in \{-1, 1\}^m} \sum_{i \in I'} \left( \sum_{j \in J} (A_{ij} - \mathbb{E} A_{ij}) x_j \right)^2 = \max_{x \in \{-1, 1\}^m} \sum_{i \in I} (X_i \xi_i)^2 \quad (3.3)$$

where we denoted

$$X_i := \sum_{j \in J} (A_{ij} - \mathbb{E} A_{ij}) x_j, \quad \xi_i := \mathbf{1}_{\{\sum_{i \in J} A_{ij} \leq \alpha d\}}.$$

Let us first fix a block  $I \times J$  and a vector  $x \in \{-1, 1\}^m$ . Let us analyze the independent random variables  $X_i \xi_i$ . Since  $|X_i| \leq \sum_{j \in J} |A_{ij} - \mathbb{E} A_{ij}| \leq \sum_{j \in J} A_{ij}$ , it follows by definition of  $\xi_i$  that

$$|X_i \xi_i| \leq \alpha d. \quad (3.4)$$

A more useful bound will follow from Bernstein's inequality. Indeed,  $X_i$  is a sum of  $m$  independent random variables with zero means and variances at most  $d/n$ . By Bernstein's inequality, for any  $t > 0$  we have

$$\mathbb{P}\{|X_i \xi_i| > tm\} \leq \mathbb{P}\{|X_i| > tm\} \leq 2 \exp\left(\frac{-mt^2/2}{d/n + t/3}\right), \quad t \geq 0. \quad (3.5)$$

For  $tm \leq \alpha d$ , this can be further bounded by  $2 \exp(-m^2 t^2 / 4\alpha d)$ , once we use the assumption  $\alpha \geq m/n$ . For  $tm > \alpha d$ , the left-hand side of (3.5) is automatically zero by (3.4). Therefore

$$\mathbb{P}\{|X_i \xi_i| > tm\} \leq 2 \exp\left(\frac{-m^2 t^2}{4\alpha d}\right), \quad t \geq 0. \quad (3.6)$$

We are now ready to bound the right-hand side of (3.3). By (3.6), the random variable  $X_i \xi_i$  is sub-gaussian<sup>2</sup> with sub-gaussian norm at most  $\sqrt{\alpha d}$ . It follows that  $(X_i \xi_i)^2$  is sub-exponential with sub-exponential norm at most  $2\alpha d$ . Using Bernstein's inequality for sub-exponential random variables (see Corollary 5.17 in [43]), we have

$$\mathbb{P}\left\{\sum_{i \in I} (X_i \xi_i)^2 > \varepsilon m \alpha d\right\} \leq 2 \exp[-c \min(\varepsilon^2, \varepsilon) m], \quad \varepsilon \geq 0. \quad (3.7)$$

Choosing  $\varepsilon := (10/c)r \log(en/m)$ , we bound this probability by  $(en/m)^{-5rm}$ .

Summarizing, we have proved that for fixed  $I, J \subseteq [n]$  and  $x \in \{-1, 1\}^m$ , with probability at least  $1 - (en/m)^{-5rm}$ , the following holds:

$$\sum_{i \in I} (X_i \xi_i)^2 \leq (10/c)r \log(en/m) \cdot m \alpha d. \quad (3.8)$$

Taking a union bound over all possibilities of  $m, I, J, x$  and using (3.3), (3.8), we see that the conclusion of the lemma holds with probability at least

$$1 - \sum_{m=1}^n 2^m \binom{n}{m}^2 \left(\frac{en}{m}\right)^{-5rm} \geq 1 - n^{-r}.$$

The proof is complete.  $\square$

Applying Lemma 3.3 followed by Grothendieck-Piesch factorization (Theorem 3.2), we obtain the following.

<sup>2</sup>For definitions and basic facts about sub-gaussian random variables, see e.g. [43].

**Lemma 3.4** (Concentration in spectral norm). *Let  $1 \leq m \leq n$  and  $\alpha \geq m/n$ . Then for  $r \geq 1$  the following holds with probability at least  $1 - n^{-r}$ . Consider a block  $I \times J$  of size  $m \times m$ . Let  $I'$  be the set of indices of the rows of  $A_{I \times J}$  that contain at most  $\alpha d$  ones. Then one can find a subset  $J' \subseteq J$  of at least  $3m/4$  columns such that*

$$\|(A - \mathbb{E}A)_{I' \times J'}\| \leq C \sqrt{\alpha d r \log(en/m)}. \quad (3.9)$$

**3.4. Restricted degrees.** The two simple lemmas of this section will help us to handle the part of the adjacency matrix outside the core block constructed in Lemma 3.4. First, we show that almost all rows have at most  $O(\alpha d)$  ones, and thus are included in the core block.

**Lemma 3.5** (Degrees of subgraphs). *Let  $1 \leq m \leq n$  and  $\alpha \geq \sqrt{m/n}$ . Then for  $r \geq 1$  the following holds with probability at least  $1 - n^{-r}$ . Consider a block  $I \times J$  of size  $m \times m$ . Then all but  $m/\alpha d$  rows of  $A_{I \times J}$  have at most  $8r\alpha d$  ones.*

*Proof.* Fix a block  $I \times J$ , and denote by  $d_i$  the number of ones in the  $i$ -th row of  $A_{I \times J}$ . Then  $\mathbb{E}d_i \leq md/n$  by the assumption. Using Chernoff's inequality, we obtain

$$\mathbb{P}\{d_i > 8r\alpha d\} \leq \left(\frac{8r\alpha d}{emd/n}\right)^{-8r\alpha d} \leq \left(\frac{2\alpha n}{m}\right)^{-8r\alpha d} =: p.$$

Let  $S$  be the number of rows  $i$  with  $d_i > 8r\alpha d$ . Then  $S$  is a sum of  $m$  independent Bernoulli random variables with expectations at most  $p$ . Again, Chernoff's inequality implies

$$\mathbb{P}\{S > m/\alpha d\} \leq (ep\alpha d)^{m/\alpha d} \leq p^{m/2\alpha d} = \left(\frac{2\alpha n}{m}\right)^{-4rm}.$$

The second inequality here holds since  $e\alpha d \leq p^{-1/2}$ . (To see this, notice that the definition of  $p$  and assumption on  $\alpha$  imply that  $p^{-1/2} = (2\alpha n/m)^{4r\alpha d} \geq 2^{4r\alpha d}$ .)

It remains to take a union bound over all blocks  $I \times J$ . We obtain that the conclusion of the lemma holds with probability at least

$$1 - \sum_{m=1}^n \binom{n}{m}^2 \left(\frac{2\alpha n}{m}\right)^{-4rm} \geq 1 - n^{-r}.$$

In the last inequality we used the assumption that  $\alpha \geq \sqrt{m/n}$ . The proof is complete.  $\square$

Next, we handle the block of rows that do have too many ones. We show that most *columns* of this block have  $O(1)$  ones.

**Lemma 3.6** (More on degrees of subgraphs). *Let  $1 \leq m \leq n$  and  $\alpha \geq \sqrt{m/n}$ . Then for  $r \geq 1$  the following holds with probability at least  $1 - n^{-r}$ . Consider a block  $I \times J$  of size  $k \times m$  with some  $k \leq m/\alpha d$ . Then all but  $m/4$  columns of  $A_{I \times J}$  have at most  $32r$  ones.*

*Proof.* Fix  $I$  and  $J$ , and denote by  $d_j$  the number of ones in the  $j$ -th column of  $A_{I \times J}$ . Then  $\mathbb{E}d_j \leq kd/n \leq m/\alpha n$  by assumption. Using Chernoff's inequality, we have

$$\mathbb{P}\{d_j > 32r\} \leq \left(\frac{32r}{em/\alpha n}\right)^{-32r} \leq \left(\frac{10\alpha n}{m}\right)^{-32r} =: p.$$

Let  $S$  be the number of columns  $j$  with  $d_j > 32r$ . Then  $S$  is a sum of  $m$  independent Bernoulli random variables with expectations at most  $p$ . Again, Chernoff's inequality implies

$$\mathbb{P}\{S > m/4\} \leq (4ep)^{m/4} \leq p^{m/6} \leq \left(\frac{10\alpha n}{m}\right)^{-5rm}.$$

The second inequality here holds since  $4e < p^{1/2}$ , which in turn follows by assumption on  $\alpha$ .

It remains to take a union bound over all blocks  $I \times J$ . It is enough to consider the blocks with largest possible number of columns, thus with  $k = \lceil m/\alpha d \rceil$ . We obtain that the conclusion of the lemma holds with probability at least

$$1 - \sum_{m=1}^n \binom{n}{m} \binom{n}{\lceil m/\alpha d \rceil} \left(\frac{10\alpha n}{m}\right)^{-5rm} \leq 1 - n^{-r}.$$

In the last inequality we used the assumption that  $\alpha \geq \sqrt{m/n}$ . The proof is complete.  $\square$

**3.5. Iterative decomposition: proof of Theorem 2.1.** Finally, we combine the tools we developed so far, and we construct an iterative decomposition of the adjacency matrix the way we outline in Section 3.1. Let us set up one step of this procedure, where we consider an  $m \times m$  block and decompose almost all of it (everything except an  $m/2 \times m/2$  block) into classes  $\mathcal{N}$ ,  $\mathcal{R}$  and  $\mathcal{C}$  satisfying the conclusion of Theorem 2.6. Once we can do this, we repeat the procedure for the  $m/2 \times m/2$  block, etc.

**Lemma 3.7** (Decomposition of a block). *Let  $1 \leq m \leq n$  and  $\alpha \geq \sqrt{m/n}$ . Then for  $r \geq 1$  the following holds with probability at least  $1 - 3n^{-r}$ . Consider a block  $I \times J$  of size  $m \times m$ . Then there exists an exceptional sub-block  $I_1 \times J_1$  with dimensions at most  $m/2 \times m/2$  such that the remaining part of the block, that is  $(I \times J) \setminus (I_1 \times J_1)$ , can be decomposed into three classes  $\mathcal{N}$ ,  $\mathcal{R} \subset (I \setminus I_1) \times J$  and  $\mathcal{C} \subset I \times (J \setminus J_1)$  so that the following holds.*

- The graph concentrates on  $\mathcal{N}$ , namely  $\|(A - \mathbb{E}A)_{\mathcal{N}}\| \leq Cr^{3/2} \sqrt{\alpha d \log(en/m)}$ .
- Each row of  $A_{\mathcal{R}}$  and each column of  $A_{\mathcal{C}}$  has at most  $32r$  ones.

Moreover,  $\mathcal{R}$  intersects at most  $n/\alpha d$  columns and  $\mathcal{C}$  intersects at most  $n/\alpha d$  rows of  $I \times J$ .

After a permutation of rows and columns, a decomposition of the block stated in Lemma 3.7 can be visualized in Figure 4c.

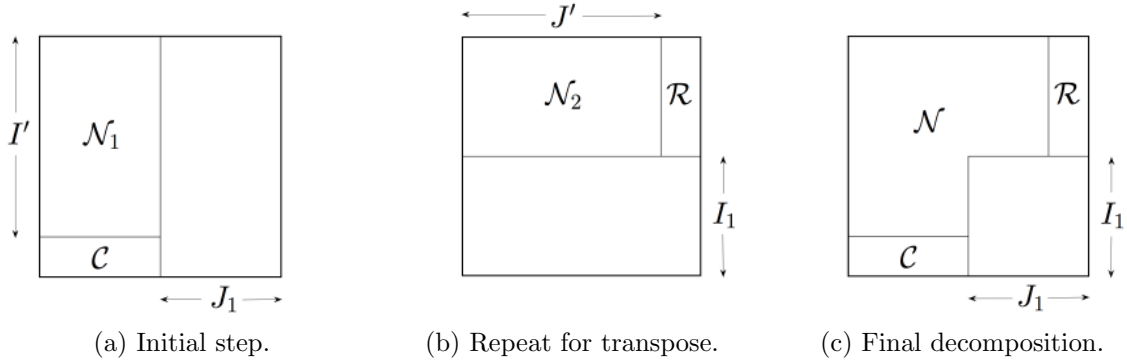


Figure 4. Construction of a block decomposition in Lemma 3.7.

*Proof.* Since we are going to use Lemmas 3.4, 3.5 and 3.6, let us fix realization of the random graph that satisfies the conclusion of those three lemmas.

By Lemma 3.5, all but  $m/\alpha d$  rows of  $A_{I \times J}$  have at most  $8r\alpha d$  ones; let us denote by  $I'$  the set of indices of those rows with at most  $8r\alpha d$  ones. Then we can use Lemma 3.4 for the block  $I' \times J$  and with  $\alpha$  replaced by  $8r\alpha$ ; the choice of  $I'$  ensures that all rows have small numbers of ones, as required in that lemma. To control the rows outside  $I'$ , we may use Lemma 3.6 for  $(I \setminus I') \times J$ ; as we already noted, this block has at most  $m/\alpha d$  rows as required in that



lemma. Intersecting the good sets of columns produced by those two lemmas, we obtain a set of at most  $m/2$  exceptional columns  $J_1 \subset J$  such that the following holds.

- On the block  $\mathcal{N}_1 := I' \times (J \setminus J_1)$ , we have  $\|(A - \mathbb{E}A)_{\mathcal{N}_1}\| \leq Cr^{3/2} \sqrt{\alpha d \log(en/m)}$ .
- For the block  $\mathcal{C} := (I \setminus I') \times (J \setminus J_1)$ , all columns of  $A_{\mathcal{C}}$  have at most  $32r$  ones.

Figure 4a illustrates the decomposition of the block  $I \times J$  into the set of exceptional columns indexed by  $J_1$  and good sets  $\mathcal{N}_1$  and  $\mathcal{C}$ .

To finish the proof, we apply the above argument to the transpose  $A^T$  on the block  $J \times I$ . To be precise, we start with the set  $J' \subset J$  of all but  $m/\alpha d$  small columns of  $A_{I \times J}$  (those with at most  $8r\alpha d$  ones); then we obtain an exceptional set  $I_1 \subset I$  of at most  $m/2$  rows; and finally we conclude that  $A$  concentrates on the block  $\mathcal{N}_2 := (I \setminus I_1) \times J'$  and has small rows on the block  $\mathcal{R} := (I \setminus I_1) \times (J \setminus J')$ . Figure 4b illustrates this decomposition.

It only remains to combine the decompositions for  $A$  and  $A^T$ ; Figure 4c illustrates a result of the combination. Once we define  $\mathcal{N} := \mathcal{N}_1 \cup \mathcal{N}_2$ , it becomes clear that  $\mathcal{N}$ ,  $\mathcal{R}$  and  $\mathcal{C}$  have the required properties.<sup>3</sup>  $\square$

*Proof of Theorem 2.6.* Let us fix a realization of the random graph that satisfies the conclusion of Lemma 3.7. Applying that lemma for  $m = n$  and with  $\alpha = 1$ , we decompose the set of edges  $[n] \times [n]$  into three classes  $\mathcal{N}_0$ ,  $\mathcal{C}_0$  and  $\mathcal{R}_0$  plus an  $n/2 \times n/2$  exceptional block  $I_1 \times J_1$ . Apply Lemma 3.7 again, this time for the block  $I_1 \times J_1$ , for  $m = n/2$  and with  $\alpha = \sqrt{1/2}$ . We decompose  $I_1 \times J_1$  into  $\mathcal{N}_1$ ,  $\mathcal{C}_1$  and  $\mathcal{R}_1$  plus an  $n/4 \times n/4$  exceptional block  $I_2 \times J_2$ .

Repeat this process for  $\alpha = \sqrt{m/n}$  where  $m$  is the running size of the block; we halve this size at each step, and so we have  $\alpha_i \leq 2^{-i/2}$ . Figure 3c illustrates a decomposition that we may obtain this way. In a finite number of steps (actually, in  $O(\log n)$  steps) the exceptional block becomes empty, and the process terminates. At that point we have decomposed the set of edges  $[n] \times [n]$  into  $\mathcal{N}$ ,  $\mathcal{R}$  and  $\mathcal{C}$ , defined as the union of  $\mathcal{N}_i$ ,  $\mathcal{C}_i$  and  $\mathcal{R}_i$  respectively, which we obtained at each step. It is clear that  $\mathcal{R}$  and  $\mathcal{C}$  satisfy the required properties.

It remains to bound the deviation of  $A$  on  $\mathcal{N}$ . By construction,  $\mathcal{N}_i$  satisfies

$$\|(A - \mathbb{E}A)_{\mathcal{N}_i}\| \leq Cr^{3/2} \sqrt{\alpha_i d \log(e\alpha_i)}.$$

Thus, by triangle inequality we have

$$\|(A - \mathbb{E}A)_{\mathcal{N}}\| \leq \sum_{i \geq 0} Cr^{3/2} \sqrt{\alpha_i d \log(e\alpha_i)} \leq C'r^{3/2} \sqrt{d}.$$

In the second inequality we used that  $\alpha_i \leq 2^{-i/2}$ , which forces the series to converge. The proof of Theorem 2.6 is complete.  $\square$

**3.6. Replacing the degrees by the  $\ell_2$  norms in Theorem 2.1.** Let us now prove the “moreover” part of Theorem 2.1, where  $d'$  is the maximal  $\ell_2$  norm of the rows and columns of the regularized adjacency matrix  $A'$ . This is clearly a stronger statement than in the main part of the theorem. Indeed, since all entries of  $A'$  are bounded in absolute value by 1, each degree, being the  $\ell_1$  norm of a row, is bounded below by the  $\ell_2$  norm squared.

This strengthening is in fact easy to check. To do so, note that the definition of  $d'$  was used only once in the proof of Theorem 2.1, namely in Step 2 where we bounded the norms of  $A'_{\mathcal{R}}$  and  $A'_{\mathcal{C}}$ . Thus, to obtain the strengthening, it is enough to replace the application of Lemma 2.7 there by the following lemma.

<sup>3</sup>It may happen that an entry ends up in more than one class  $\mathcal{N}$ ,  $\mathcal{R}$  and  $\mathcal{C}$ . In such cases, we split the tie arbitrarily.

**Lemma 3.8.** *Consider a matrix  $B$  with entries in  $[0, 1]$ . Suppose each row of  $B$  has at most a non-zero entries, and each column has  $\ell_2$  norm at most  $\sqrt{b}$ . Then  $\|B\| \leq \sqrt{ab}$ .*

*Proof.* To prove the claim, let  $x$  be a vector with  $\|x\|_2 = 1$ . Using Cauchy-Schwarz inequality and the assumptions, we have

$$\begin{aligned} \|Bx\|_2^2 &= \sum_j \left( \sum_i B_{ij} x_i \right)^2 \leq \sum_j \left( \sum_{i: B_{ij} \neq 0} B_{ij}^2 \sum_{i: B_{ij} \neq 0} x_i^2 \right) \\ &\leq \sum_j \left( b \sum_{i: B_{ij} \neq 0} x_i^2 \right) = b \sum_i x_i^2 \sum_{j: B_{ij} \neq 0} 1 \leq b \sum_i x_i^2 a = ab. \end{aligned}$$

Since  $x$  is arbitrary, this completes the proof.  $\square$

#### 4. CONCENTRATION OF THE REGULARIZED LAPLACIAN

In this section, we state the following formal version of Theorem 1.2, and we deduce it from concentration of adjacency matrices (Theorem 2.1).

**Theorem 4.1** (Concentration of regularized Laplacians). *Consider a random graph from the inhomogeneous Erdős-Rényi model, and let  $d$  be as in (1.3). Choose a number  $\tau > 0$ . Then, for any  $r \geq 1$ , with probability at least  $1 - e^{-r}$  one has*

$$\|\mathcal{L}(A_\tau) - \mathcal{L}(\mathbb{E} A_\tau)\| \leq \frac{Cr^2}{\sqrt{\tau}} \left(1 + \frac{d}{\tau}\right)^{5/2}.$$

*Proof.* Two sources contribute to the deviation of Laplacian – the deviation of the adjacency matrix and the deviation of the degrees. Let us separate and bound them individually.

**Step 1. Decomposing the deviation.** Let us denote  $\bar{A} := \mathbb{E} A$  for simplicity; then

$$E := \mathcal{L}(A_\tau) - \mathcal{L}(\bar{A}_\tau) = D_\tau^{-1/2} A_\tau D_\tau^{-1/2} - \bar{D}_\tau^{-1/2} \bar{A}_\tau \bar{D}_\tau^{-1/2}.$$

Here  $D_\tau = \text{diag}(d_i + \tau)$  and  $\bar{D}_\tau = \text{diag}(\bar{d}_i + \tau)$  are the diagonal matrices with degrees of  $A_\tau$  and  $\bar{A}_\tau$  on the diagonal, respectively. Using the fact that  $A_\tau - \bar{A}_\tau = A - \bar{A}$ , we can represent the deviation as

$$E = S + T \quad \text{where} \quad S = D_\tau^{-1/2} (A - \bar{A}) D_\tau^{-1/2}, \quad T = D_\tau^{-1/2} \bar{A}_\tau D_\tau^{-1/2} - \bar{D}_\tau^{-1/2} \bar{A}_\tau \bar{D}_\tau^{-1/2}.$$

Let us bound  $S$  and  $T$  separately.

**Step 2. Bounding  $S$ .** Let us introduce a diagonal matrix  $\Delta$  that should be easier to work with than  $D_\tau$ . Set  $\Delta_{ii} = 1$  if  $d_i \leq 8rd$  and  $\Delta_{ii} = d_i/\tau + 1$  otherwise. Then entries of  $\tau\Delta$  are upper bounded by the corresponding entries of  $D_\tau$ , and so

$$\tau\|S\| \leq \|\Delta^{-1/2} (A - \bar{A}) \Delta^{-1/2}\|.$$

Next, by triangle inequality,

$$\tau\|S\| \leq \|\Delta^{-1/2} A \Delta^{-1/2} - \bar{A}\| + \|\bar{A} - \Delta^{-1/2} \bar{A} \Delta^{-1/2}\| =: R_1 + R_2. \quad (4.1)$$

In order to bound  $R_1$ , we use Theorem 2.1 to show that  $A' := \Delta^{-1/2} A \Delta^{-1/2}$  concentrates around  $\bar{A}$ . This should be possible because  $A'$  is obtained from  $A$  by reducing the degrees that are bigger than  $8rd$ . To apply the ‘‘moreover’’ part of Theorem 2.1, let us check the magnitude of the  $\ell_2$  norms of the rows  $A'_i$  of  $A'$ :

$$\|A'_i\|_2^2 = \sum_{j=1}^n \frac{A_{ij}}{\Delta_{ii} \Delta_{jj}} \leq \frac{d_i}{\Delta_{ii}} \leq \max(8rd, \tau).$$

Here in the first inequality we used that  $\Delta_{jj} \geq 1$  and  $\sum_j A_{ij} = d_i$ ; the second inequality follows by definition of  $\Delta_{ii}$ . Applying Theorem 2.1, we obtain with probability  $1 - n^{-r}$  that

$$R_1 = \|A' - \bar{A}\| \leq C_1 r^2 (\sqrt{d} + \sqrt{\tau}).$$

To bound  $R_2$ , we note that by construction of  $\Delta$ , the matrices  $\bar{A}$  and  $\Delta^{-1/2} \bar{A} \Delta^{-1/2}$  coincide on the block  $I \times I$ , where  $I$  is the set of vertices satisfying  $d_i \leq 8rd$ . This block is very large – indeed, Lemma 3.5 implies that  $|I^c| \leq n/d$  with probability  $1 - n^{-r}$ . Outside this block, i.e. on the small blocks  $I^c \times [n]$  and  $[n] \times I^c$ , the entries of  $\bar{A} - \Delta^{-1/2} \bar{A} \Delta^{-1/2}$  are bounded by the corresponding entries of  $\bar{A}$ , which are all bounded by  $d/n$ . Thus, using Lemma 2.7, we have

$$R_2 \leq \|\bar{A}_{I^c \times [n]}\| + \|\bar{A}_{[n] \times I^c}\| \leq 2\sqrt{d}.$$

Substituting the bounds for  $R_1$  and  $R_2$  into (4.1), we conclude that

$$\|S\| \leq \frac{C_2 r^2}{\tau} (\sqrt{d} + \sqrt{\tau})$$

with probability at least  $1 - 2n^{-r}$ .

**Step 3. Bounding  $T$ .** Bounding the spectral norm by the Hilbert-Schmidt norm, we get

$$\|T\| \leq \|T\|_{\text{HS}} = \sum_{i,j=1}^n T_{ij}^2, \quad \text{where } T_{ij} = (\bar{A}_{ij} + \tau/n) \left[ 1/\sqrt{\delta_{ij}} - 1/\sqrt{\bar{\delta}_{ij}} \right]$$

and  $\delta_{ij} = (d_i + \tau)(d_j + \tau)$  and  $\bar{\delta}_{ij} = (\bar{d}_i + \tau)(\bar{d}_j + \tau)$ . To bound  $T_{ij}$ , we note that

$$0 \leq \bar{A}_{ij} + \tau/n \leq \frac{d + \tau}{n} \quad \text{and} \quad \left| 1/\sqrt{\delta_{ij}} - 1/\sqrt{\bar{\delta}_{ij}} \right| = \left| \frac{\delta_{ij} - \bar{\delta}_{ij}}{\delta_{ij} \sqrt{\delta_{ij}} + \bar{\delta}_{ij} \sqrt{\bar{\delta}_{ij}}} \right| \geq \frac{|\delta_{ij} - \bar{\delta}_{ij}|}{2\tau^3}.$$

Recalling the definition of  $\delta_{ij}$  and  $\bar{\delta}_{ij}$  and adding and subtracting  $(d_i + \tau)(\bar{d}_j + \tau)$ , we have

$$\delta_{ij} - \bar{\delta}_{ij} = (d_i + \tau)(d_j - \bar{d}_j) + (\bar{d}_j + \tau)(d_i - \bar{d}_i).$$

So, using the inequality  $(a + b)^2 \leq 2(a^2 + b^2)$  and bounding  $\bar{d}_j + \tau$  by  $d + \tau$ , we obtain

$$\|T\|^2 \leq \frac{(d + \tau)^2}{n^2 \tau^6} \left[ \sum_{i=1}^n (d_i + \tau)^2 \sum_{j=1}^n (d_j - \bar{d}_j)^2 + n(d + \tau)^2 \sum_{i=1}^n (d_i - \bar{d}_i)^2 \right]. \quad (4.2)$$

We claim that

$$\sum_{j=1}^n (d_j - \bar{d}_j)^2 \leq C_3 r^2 n d \quad \text{with probability } 1 - e^{-2r}. \quad (4.3)$$

Indeed, since the variance of each  $d_i$  is bounded by  $d$ , the expectation of the sum in (4.3) is bounded by  $nd$ . To upgrade the variance bound to an exponential deviation bound, one can use one of the several standard methods. For example, Bernstein's inequality implies that  $X_i = d_j - \bar{d}_j$  satisfies  $\mathbb{P} \left\{ X_i > C_4 t \sqrt{d} \right\} \leq e^{-t}$  for all  $t \geq 1$ . This means that the random variable  $X_i^2$  belongs to the Orlicz space  $L_{\psi_{1/2}}$  and has norm  $\|X_i^2\|_{\psi_{1/2}} \leq C_3 d$ , see [26]. By triangle inequality, we conclude that  $\|\sum_{i=1}^n X_i^2\|_{\psi_{1/2}} \leq C_3 n d$ , which in turn implies (4.3).

Furthermore, (4.3) implies

$$\sum_{i=1}^n (d_i + \tau)^2 \leq 2 \sum_{i=1}^n (d_i - \bar{d}_i)^2 + 2 \sum_{i=1}^n (\bar{d}_i + \tau)^2 \leq 2C_3 r^2 n d + 2n(d + \tau)^2 \leq C_5 r^2 n (d + \tau)^2.$$

Substituting this bound and (4.3) into (4.2) we conclude that

$$\|T\|^2 \leq \frac{(d+\tau)^2}{n^2\tau^6} \cdot C_3 r^2 n d \left[ C_5 r^2 n (d+\tau)^2 + n (d+\tau)^2 \right] \leq \frac{C_6 r^4}{\tau} \left(1 + \frac{d}{\tau}\right)^5.$$

It remains to substitute the bounds for  $S$  and  $T$  into the inequality  $\|E\| \leq \|S\| + \|T\|$ , and simplify the expression. The resulting bound holds with probability at least  $1 - n^{-r} - n^{-r} - e^{-2r} \geq 1 - e^{-r}$ , as claimed.  $\square$

## 5. FURTHER QUESTIONS

**5.1. Optimal regularization?** The main point of our paper was that regularization helps sparse graphs to concentrate. We have discussed several kinds of regularization in Section 1.4 and mentioned some more in Section 1.4. We found that any meaningful regularization works, as long as it reduces the too high degrees and increases the too low degrees. Is there an optimal way to regularize a graph? Designing the best “preprocessing” of sparse graphs for spectral algorithms is especially interesting from the applied perspective, i.e. for real world networks.

On the theoretical level, can regularization of sparse graphs produce the same optimal bound  $2\sqrt{d}(1 + o(1))$  that we saw for dense graphs in (1.1)? Thus, an ideal regularization should bring all parasitic outliers of the spectrum into the bulk. If so, this would lead to a potentially simple spectral clustering algorithm for community detection in networks which matches the theoretical lower bounds. Algorithms with optimal rates exist for this problem [33, 29], but their analysis is very technical.

**5.2. How exactly concentration depends on regularization?** It would be interesting to determine how exactly the concentration of Laplacian depends on the regularization parameter  $\tau$ . The dependence in Theorem 4.1 is not optimal, and we have not made efforts to improve it. Although it is natural to choose  $\tau \sim d$  as in Theorem 1.2, choosing  $\tau \gg d$  could also be useful [23]. Choosing  $\tau \ll d$  may be interesting as well, for then  $\mathcal{L}(\mathbb{E} A_\tau) \approx \mathcal{L}(\mathbb{E} A)$  and we obtain the concentration of  $\mathcal{L}(A_\tau)$  around the Laplacian of the expectation of the original (rather than regularized) matrix  $\mathbb{E} A$ .

**5.3. Average expected degree?** Both concentration results of this paper, Theorems 1.1 and 1.2, depend on  $d = \max_{ij} np_{ij}$ . Would it be possible to reduce  $d$  to the maximal expected degree  $d_{ave} = \max_i \sum_j p_{ij}$ ?

**5.4. From random graphs to random matrices?** Adjacency matrices of random graphs are particular examples of random matrices. Does the phenomenon we described, namely that regularization leads to concentration, apply for general random matrices? Guided by Theorem 1.1, we might expect the following for a broader general class of random matrices  $B$  with mean zero independent entries. First, the only reason the spectral norm of  $B$  is too large (and that it is determined by outliers of spectrum) could be the existence of a large row or column. Furthermore, it might be possible to reduce the norm of  $B$  (and thus bring the outliers into the bulk of spectrum) by regularizing in some way the rows and columns that are too large. For related questions in random matrix theory, see the recent work [4, 21].

## REFERENCES

- [1] E. Abbe, A. S. Bandeira, and G. Hall. Exact recovery in the stochastic block model. *IEEE Transactions on Information Theory*, 62(1):471–487, 2016.
- [2] N. Alon and N. Kahale. A spectral technique for coloring random 3-colorable graphs. *SIAM J. Comput.*, (26):1733–1748, 1997.

- [3] A. A. Amini, A. Chen, P. J. Bickel, and E. Levina. Pseudo-likelihood methods for community detection in large sparse networks. *The Annals of Statistics*, 41(4):2097–2122, 2013.
- [4] A. Bandeira and R. V. Handel. Sharp nonasymptotic bounds on the norm of random matrices with independent entries. *Annals of Probability, to appear*, 2014.
- [5] R. Bhatia. *Matrix Analysis*. Springer-Verlag New York, 1996.
- [6] P. J. Bickel and A. Chen. A nonparametric view of network models and Newman-Girvan and other modularities. *Proc. Natl. Acad. Sci. USA*, 106:21068–21073, 2009.
- [7] B. Bollobas, S. Janson, and O. Riordan. The phase transition in inhomogeneous random graphs. *Random Structures and Algorithms*, 31:3–122, 2007.
- [8] C. Bordenave, M. Lelarge, and L. Massoulié. Non-backtracking spectrum of random graphs: community detection and non-regular Ramanujan graphs. *arxiv:1501.06087*, 2015.
- [9] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: a nonasymptotic theory of independence*. Oxford University Press, 2013.
- [10] T. Cai and X. Li. Robust and computationally feasible community detection in the presence of arbitrary outlier nodes. *Ann. Statist.*, 43(3):1027–1059, 2015.
- [11] K. Chaudhuri, F. Chung, and A. Tsiatas. Spectral clustering of graphs with general degrees in the extended planted partition model. *Journal of Machine Learning Research Workshop and Conference Proceedings*, 23:35.1 – 35.23, 2012.
- [12] P. Chin, A. Rao, and V. Vu. Stochastic block model and community detection in the sparse graphs : A spectral algorithm with optimal rate of recovery. *arXiv:1501.05021*, 2015.
- [13] F. R. K. Chung. *Spectral Graph Theory*. CBMS Regional Conference Series in Mathematics, 1997.
- [14] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Physical Review E*, 84:066106, 2011.
- [15] U. Feige and . Ofek. Spectral techniques applied to sparse random graphs. *Wiley InterScience*, 2005.
- [16] Z. Fredi and J. Komlós. The eigenvalues of random symmetric matrices. *Combinatorica*, 1:3:233–241, 1980.
- [17] J. Friedman, J. Kahn, and E. Szemerédi. On the second eigenvalue in random regular graphs. *Proc Twenty First Annu ACMSymp Theory of Computing*, pages 587–598, 1989.
- [18] C. Gao, Z. Ma, A. Y. Zhang, and H. H. Zhou. Achieving optimal misclassification proportion in stochastic block model. *arXiv:1505.03772*, 2015.
- [19] O. Guédon and R. Vershynin. Community detection in sparse networks via grothendieck’s inequality. *Probability Theory and Related Fields, to appear*, 2014.
- [20] B. Hajek, Y. Wu, and J. Xu. Achieving exact cluster recovery threshold via semidefinite programming. *arXiv:1412.6156*, 2014.
- [21] R. V. Handel. On the spectral norm of inhomogeneous random matrices. *arXiv:1502.05003*, 2015.
- [22] P. W. Holland, K. B. Laskey, and S. Leinhardt. Stochastic blockmodels: first steps. *Social Networks*, 5(2):109–137, 1983.
- [23] A. Joseph and B. Yu. Impact of regularization on spectral clustering. *Ann. Statist.*, 44(4):1765–1791, 2016.
- [24] M. Krivelevich and B. Sudakov. The largest eigenvalue of sparse random graphs. *Combin Probab Comput*, 12:61–72, 2003.
- [25] M. Ledoux. *The Concentration of Measure Phenomenon*, volume 89 of *Mathematical Surveys and Monographs*. Amer. Math. Society, 2001.
- [26] M. Ledoux and M. Talagrand. *Probability in Banach spaces: Isoperimetry and processes*. Springer-Verlag, Berlin, 1991.
- [27] J. Lei and A. Rinaldo. Consistency of spectral clustering in stochastic block models. *Ann. Statist.*, 43(1):215–237, 2015.
- [28] L. Lu and X. Peng. Spectra of edge-independent random graphs. *The electronic journal of combinatorics*, 20(4), 2013.
- [29] L. Massoulié. Community detection thresholds and the weak Ramanujan property. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, STOC ’14, pages 694–703, 2014.
- [30] McSherry. Spectral partitioning of random graphs. *Proc. 42nd FOCS*, pages 529–537, 2001.
- [31] A. Montanari and S. Sen. Semidefinite programs on sparse random graphs and their application to community detection. *arXiv:1504.05910*, 2015.
- [32] E. Mossel, J. Neeman, and A. Sly. Consistency thresholds for binary symmetric block models. *arXiv:1407.1591*, 2014.

- [33] E. Mossel, J. Neeman, and A. Sly. A proof of the block model threshold conjecture. *arXiv:1311.4115*, 2014.
- [34] E. Mossel, J. Neeman, and A. Sly. Reconstruction and estimation in the planted partition model. *Probability Theory and Related Fields*, 2014.
- [35] R. Oliveira. Concentration of the adjacency matrix and of the laplacian in random graphs with independent edges. *arXiv:0911.0600*, 2010.
- [36] A. Pietsch. *Operator Ideals*. North-Holland Amsterdam, 1978.
- [37] G. Pisier. *Factorization of linear operators and geometry of Banach spaces*. Number 60 in CBMS Regional Conference Series in Mathematics. AMS, Providence, 1986.
- [38] G. Pisier. Grothendiecks theorem, past and present. *Bulletin (New Series) of the American Mathematical Society*, 49(2):237–323, 2012.
- [39] T. Qin and K. Rohe. Regularized spectral clustering under the degree-corrected stochastic blockmodel. In *Advances in Neural Information Processing Systems*, pages 3120–3128, 2013.
- [40] E. M. Stein and R. Shakarchi. *Functional Analysis: Introduction to Further Topics in Analysis*. Princeton University Press, 2011.
- [41] N. Tomczak-Jaegermann. *Banach-Mazur distances and finite-dimensional operator ideals*. John Wiley & Sons, Inc., New York, 1989.
- [42] J. A. Tropp. Column subset selection, matrix factorization, and eigenvalue optimization. *Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 978–986, 2009.
- [43] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Y. Eldar and G. Kutyniok, editors, *Compressed sensing: theory and applications*. Cambridge University Press. Submitted.
- [44] V. Vu. Spectral norm of random matrices. *Combinatorica*, 27(6):721–736, 2007.

DEPARTMENT OF STATISTICS, UNIVERSITY OF CALIFORNIA, DAVIS, ONE SHIELDS AVE, DAVIS, CA 95616, U.S.A.

*E-mail address:* `canle@ucdavis.edu`

DEPARTMENT OF STATISTICS, UNIVERSITY OF MICHIGAN, 1085 S. UNIVERSITY AVE, ANN ARBOR, MI 48109, U.S.A.

*E-mail address:* `elevina@umich.edu`

DEPARTMENT OF MATHEMATICS, UNIVERSITY OF MICHIGAN, 530 CHURCH ST, ANN ARBOR, MI 48109, U.S.A.

*E-mail address:* `romanv@umich.edu`