UNIVERSITÉ PARIS-SACLAY

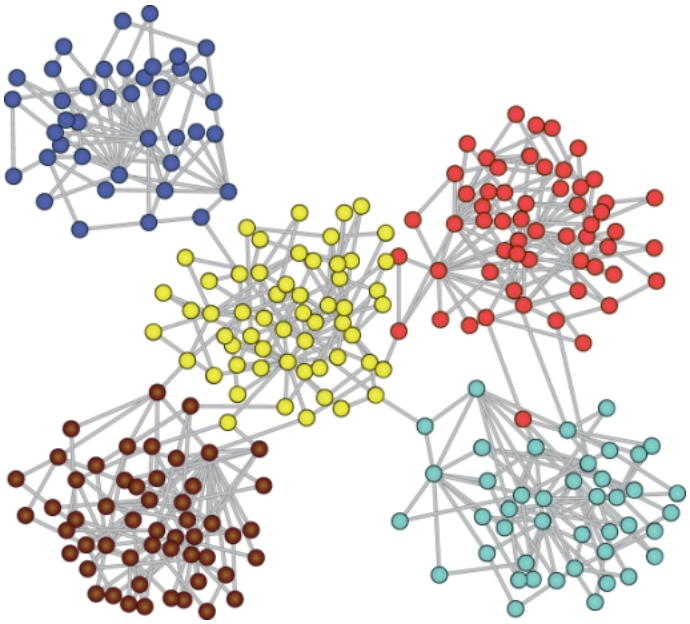
RANDOM MATRICES - COMMUNITY DETECTION IN NETWORKS - RESULTS FROM HIGH-DIMENSIONAL PROBABILITY (2018)

Community Detection in an Erdös-Rényi Graph

Authors:
Matthieu Dagès, Vincent Counathe

Supervisor:
Edouard Maurel-Segala

November 4, 2024



université PARIS-SACLAY

Mathématiques Orsay

Contents

1 Introduction and Motivations

In this brief note, and as a conclusion to the *Random Matrices* course taught by Edouard Maurel-Segala in 2024, we present a few results from Roman Vershynin, from his book *High Dimensional Probability* (2018). Notably, we focus on random vector analysis building on linear algebra, concentration results for random matrices and the concept of ϵ-net arguments, in relation to solving a community detection problem in a random Erdös-Rényi graph.

1.1 Context and Model

The context is as follows: a graph containing n vertices is split in two communities, each of size $n/2$ (we thus take n even). Vertices from the same community connect independently with probability p, $p \in (0,1)$, whilst vertices from different communities connect with probability q, $q \in (0,1)$. Let p be strictly greater than q. This setup is called the Stochastic Block Model. A visual representation of a graph G following the Stochastic Block Model is provided below.

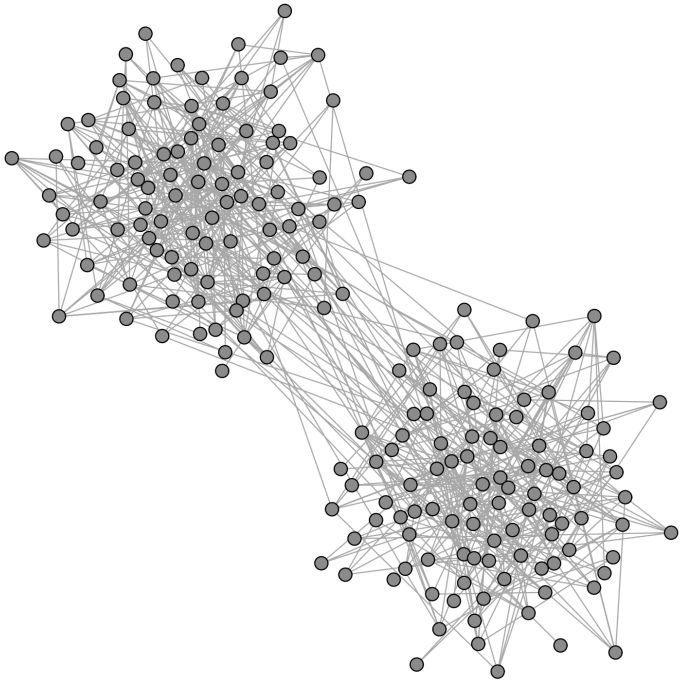


Figure 1: An illustration of a random graph G with parameters $n = 200, p = 1/20, q = 1/200$. From *High Dimensional Probability* (2018), page 94.

The first step in progressing through the community detection problem is to identify the graph G with its adjacency matrix A, which fully characterises it. Below is provided the definition used by the author.

1.2 Introduction of Random Matrices

Definition 1. *(Adjacency matrix) The adjacency matrix of a graph G with n vertices is the $n \times n$ real symmetric matrix, whose entries A_{ij} are equal to 1 if vertex i and j are connected, and equal to 0 otherwise.*

A, as the adjacency matrix of a given graph G, is indeed the matrix of a given realization of the model. One might wonder what can be said about $\mathbb{E}(A)$. It is thus useful to split the adjancy matrix A between an "expected" part, which depends solely on the Bernouilli parameters p and q, and a residual part R, or noise.

Definition 2. *(Expected adjacency matrix) The expected adjacency matrix of a graph G with n vertices, following the Stochastic Block Model with two communities, is the $n \times n$ real symmetric matrix, whose entries D_{ij} are equal to p if vertex i and j connect with probability p, and equal to q if vertex i and j connect with probability q.*

Remark: we can assume here, to simplify computations that follow, and considering it does not alter the community detection problem in itself, that vertex i connects to itself with probability p. This will avoid the introduction of identity matrices in the computations.

Now that these definitions have been introduced, we pose $R := A - D$, which provides us with an expectation and noise decomposition, as mentioned above: $A = D + R$.

1.3 Expected Adjacency Matrix

In the case where $n = 4$, the expected adjacency matrix of a graph G following the Stochastic Block Model looks as follows:

$$D = \left[\begin{array}{cc|cc} p & p & q & q \\ p & p & q & q \\ \hline q & q & p & p \\ q & q & p & p \end{array}\right]$$

The eigenvalues and eigenvectors of D will allow us to progress in our problem. In the general case, we observe that:

$$\text{rank}(D) = 2, \quad \lambda_1 = \frac{p+q}{2}n \quad \lambda_2 = \frac{p-q}{2}n$$

where λ_1 and λ_2 denote eigenvalues. As regards eigenvectors, we have:

$$u_1 = \left[\begin{array}{c} 1 \\ \vdots \\ 1 \\ \hline 1 \\ \vdots \\ 1 \end{array}\right] \qquad u_2 = \left[\begin{array}{c} 1 \\ \vdots \\ 1 \\ \hline -1 \\ \vdots \\ -1 \end{array}\right]$$

Let us note that $\|u_1\|_2 = \|u_2\|_2 = \sqrt{n}$.

We observe the eigenvector u_2 associated with λ_2 fully characterises the communities of the graph, hence provides a solution to the community detection problem.

However, we know neither u_2 nor the expected adjacency matrix D. We would want to use $u_2(A)$, which can be computed explicitly as we have the information about matrix A, as a proxy for $u_2(D)$, however this requires the introduction of results on perturbation theory to ensure that the perturbation does not alter the problem too much. It would be useful to derive an upper bound for $\|u_2(D) - u_2(A)\|$.

2 Perturbation Theory and Random Eigenvector Bound

2.1 Perturbation Theory and Context

Perturbation theory serves as a useful tool in understanding how small changes in a system or mathematical object can affect its properties. In our case, we are interested in how perturbations to the expected adjacency matrix D affect its eigenvector structure. A key result is the Davis-Kahan theorem.

2.2 Davis-Kahan Theorem

The Davis-Kahan Theorem offers a crucial result which will, in our case, ensure the stability of random eigenvectors after perturbation. It postulates that for symmetric matrices S and T with identical dimensions, if a given eigenvalue of S is well-separated from the rest of the spectrum, the sine of the angle between the eigenvectors of S and T corresponding to the i-the largest eigenvalue is bounded. It is provided below, from Vershynin's *High Dimensional Probability*:

Theorem 1. *(Davis-Kahan Theorem)*
Let S and T be symmetric matrices with the same dimensions. Fix i and assume that the i-th largest eigenvalue of S is well separated from the rest of the spectrum:

$$\min_{j:j\neq i} |\lambda_i(S) - \lambda_j(S)| =: \delta > 0.$$

Then the angle between the eigenvectors of S and T corresponding to the i-th largest eigenvalues (as a number between 0 and $\frac{\pi}{2}$) satisfies

$$\sin\angle(v_i(S), v_i(T)) \leq \frac{2\|S-T\|}{\delta},$$

Remark: The eigenvalue separation hypothesis is critical in this theorem. Indeed, we want to make sure we are comparing eigenvectors corresponding to the same eigenvalue, before and after having added the perturbation R.

A corollary for unit eigenvectors reveals that, subject to a sign difference, the difference between eigenvectors corresponding to the i-th largest eigenvalue of S and T can be bounded as below:

Corollary 1 (Davis-Kahan Corollary)**.** *For unit eigenvectors $v_i(S)$ and $v_i(T)$:*

$$\exists\theta \in \{-1,1\} : \|v_i(S) - \theta v_i(T)\|_2 \leq \frac{2\sqrt{2}\|S-T\|}{\delta}.$$

As we lack direct access to $D = \mathbb{E}(A)$ and to $u_2(D)$, but only have information about $A = D + R$, we will want to apply the Davis-Kahan theorem, with S and T being D and A in our case, respectively. This result implies that, after perturbation of our matrix D, accessing the random eigenvectors of the perturbed matrix A will make sense, since it will be close enough to the original corresponding eigenvector of D. We would then be able to bound the error by a constant times the operator norm of $A - D$.

To move forward, this requires to bound $\|A - D\| = \|R\|$.

3 Concentration Inequality and ϵ-net Argument

We are thus now left to derive an upper bound for $\|R\|$ with high probability.

While the author proves a more general result that holds for a matrix R composed of subgaussian variables, we rely on the fact that $R_{i,j} = A_{i,j} - \mathbb{E}[A_{i,j}]$ is a centred symmetric matrix with independent Bernoulli variables on the upper triangular half of the matrix, including the diagonal. We use the following characterisation of the spectral norm and one of Hoeffding's inequalities:

Proposition 1. *Let R an $n \times n$ real-valued matrix. Denote S^{n-1} the unit Euclidean sphere in $\mathbb{R}^n$. Then:*

$$\|R\| = \sup_{x,y\in S^{n-1}} \langle Rx, y\rangle$$

Theorem 2. *(Hoeffding) Let $X_1, \ldots, X_N$ be independent random variables. Assume that $X_i \in [m_i, M_i]$ for every i. Then, for any $t > 0$, we have*

$$P\left(\sum_{i=1}^{N}(X_i - \mathbb{E}[X_i]) \geq t\right) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{N}(M_i - m_i)^2}\right).$$

As stated above, the aim of this section is to find an upper bound for $P[\sup_{x,y\in S^{n-1}} \langle Rx, y\rangle \geq t], \forall\, t > 0$. We proceed in three steps.

- *Step 1*: Fix a direction $x, y \in \mathbb{R}^n$ and find an upper bound for $P\langle Rx, y\rangle \geq t], \forall\, t > 0$. We use Hoeffding's inequality to derive such a bound.

- *Step 2*: Find a subset $\mathcal{N} \subset S^{n-1}$ of finite cardinal such that $\sup_{x,y\in S^{n-1}} \langle Rx, y\rangle \leq C_\epsilon \max_{x,y\in\mathcal{N}} \langle Rx, y\rangle$, for an arbitrarily-small ϵ and $C_\epsilon \approx 1$.
- *Step 3*: Derive the following: $\forall t > 0, P[\sup_{x,y\in S^{n-1}} \langle Rx, y\rangle \geq t] \leq P[C_\epsilon \max_{x,y\in\mathcal{N}} \langle Rx, y\rangle \geq t]$ and use a union bound to derive:

$$\begin{aligned}
\forall t > 0, P\left[\sup_{x,y\in S^{n-1}} \langle Rx, y\rangle \geq t\right] &\leq P\left[C_\epsilon \max_{x,y\in\mathcal{N}} \langle Rx, y\rangle \geq t\right] \\
&= P\left[\bigcup_{x,y\in\mathcal{N}} \langle Rx, y\rangle \geq t\right] \\
&\leq \Sigma_{x,y\in\mathcal{N}} P\left[\langle Rx, y\rangle \geq \frac{t}{C_\epsilon}\right] \\
&\leq |\mathcal{N}| \max_{x,y\in\mathcal{N}} P\left[\langle Rx, y\rangle \geq \frac{t}{C_\epsilon}\right] \\
&\leq |\mathcal{N}| \sup_{x,y\in S^{n-1}} P\left[\langle Rx, y\rangle \geq \frac{t}{C_\epsilon}\right]
\end{aligned}$$

We can then bound this quantity using the upper bound found in step 1. We also hope to bound the cardinal of $\mathcal{N}$ by a known quantity.

3.1 Step 1: Getting a Bound for Each Direction x, y

Let $x, y \in \mathbb{R}^n$. We have: $\forall t > 0$,

$$P\left[\langle Rx, y\rangle \geq t\right] = P\left(\sum_{1\leq i\leq j\leq n} R_{i,j} \frac{(x_i y_j + x_j y_i)}{1 + \mathbb{1}_{i=j}} \geq t\right)$$

Since, $R_{i,j} \frac{(x_i y_j + x_j y_i)}{1+\mathbb{1}_{i=j}} \in [\pm \frac{|x_i y_j| + |x_j y_i|}{\mathbb{1}_{i=j}}]$ and are independent random variables for $1 \leq i \leq j \leq n$, we can now apply Hoeffding's inequality on a sum of independent bounded functions, that are in addition centred:

$$P\left[\langle Rx, y\rangle \geq t\right] \leq \exp\left(-\frac{2t^2}{4\sum_{1\leq i\leq j\leq n}\left(\frac{|x_i y_j| + |x_j y_i|}{1+\mathbb{1}_{i=j}}\right)^2}\right)$$

We have:

$$\begin{aligned}
\sum_{1\leq i\leq j\leq n}\left(\frac{|x_i y_j| + |x_j y_i|}{1 + \mathbb{1}_{i=j}}\right)^2 &= \sum_{i=1}^{N} |x_i y_i|^2 + \sum_{1\leq i<j\leq n} (|x_i y_j| + |x_j y_i|)^2 \\
&\leq \sum_{i=1}^{N} |x_i y_i|^2 + 2\sum_{1\leq i<j\leq n} (|x_i y_j|^2 + |x_j y_i|^2) \\
&= \sum_{1\leq i,j\leq n} |x_i y_j|^2 + \sum_{1\leq i<j\leq n} |x_i y_j|^2 \\
&\leq 2\sum_{1\leq i,j\leq n} |x_i y_j|^2 \\
&= 2, \text{ since } x, y \in S^{n-1}
\end{aligned}$$

Thus:

$$P\left[\langle Rx, y\rangle \geq t\right] \leq \exp\left(-\frac{t^2}{4}\right), \forall t > 0, \forall x, y \in S^{n-1}$$

We have now derived an upper bound in one arbitrary direction $x, y \in S^{n-1}$, and this upper bound does not depend on the direction x, y.

3.2 Step 2: Introducing ϵ-net Arguments

This step includes several preliminary definitions and properties, that will allow us to find a suitable subset $\mathcal{N}$ of S^{n-1}, of finite cardinal, such that a similar property to the above-mentioned approximation of $\sup_{x,y\in S^{n-1}} \langle Rx, y\rangle$.

We will position ourselves in a metric space (K,d).

Definition 3. *(ϵ-net and ϵ-separated subsets)*

- *A subset $N \subseteq K$ is called an ϵ-net of K if every point in K is within distance ϵ of some point of N, i.e.,*

$$\forall x \in K, \exists x_0 \in N : d(x, x_0) \leq \epsilon.$$

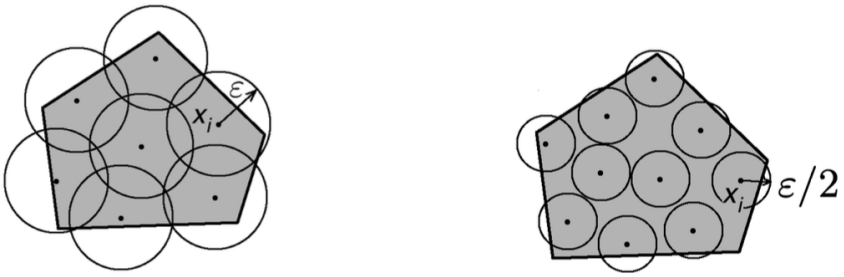
  Equivalently, N is an ϵ-net of K if and only if K can be covered by balls with centers in N and radii ϵ.

- *A subset N of K is ϵ-separated if $d(x,y) > \epsilon$ for all distinct points $x, y \in \mathcal{N}$*

Definition 4. *(Covering and Packing Numbers)*

- *The smallest possible cardinality of an ϵ-net of K is called the covering number of K and is denoted $N(K,d,\epsilon)$. Equivalently, $N(K,d,\epsilon)$ is the smallest number of closed balls with centers in K and radii ϵ whose union covers K.*

- *The largest possible cardinality of an ϵ-separated subset of a given set K is called the packing number of K and is denoted $P(K,d,\epsilon)$. In a normed space, we have the equivalent definition: it is the largest number of closed disjoint balls with centers in K and radii $\frac{\epsilon}{2}$.*



(a) This covering of a pentagon K by seven ε-balls shows that $\mathcal{N}(K,\varepsilon) \leq 7$.

(b) This packing of a pentagon K by ten $\varepsilon/2$-balls shows that $\mathcal{P}(K,\varepsilon) \geq 10$.

Figure 2: Example of ϵ-net and ϵ-separated subset of a compact K. From *High Dimensional Probability* (2018), page 82.

On the left-side, one can see how a compact set can be approximated by a finite number of points forming an ϵ-net of the set. In fact, one important properties of ϵ-nets is that the covering number, i.e. the smallest cardinality of such a subset, is finite if and only if the set K is precompact.

The use of the covering and packing numbers is important in the application of our problem.

Indeed, in a Euclidean space, they have properties that relate them to the volume of a set and of the Euclidean closed balls. We establish in this context a partial equivalence between the volume of an ϵ-net and the volume of the whole set.

Lemma 1 (Equivalence of covering and packing numbers)**.** *For any normed set K and any $\epsilon > 0$, we have*

$$P(K,d,2\epsilon) \leq N(K,d,\epsilon) \leq P(K,d,\epsilon).$$

Proof: The upper bound follows from the fact a maximal ϵ-separated subset $\mathcal{N}$ is necessarily an ϵ-net, since adding any point to $\mathcal{N}$ would make it lose its separation property, thus implying that points in K are at a distance of at least ϵ to a point in $\mathcal{N}$.

For the lower bound, we use the characterisation of the packing number of P, an arbitrary 2ϵ-separated subset of K in terms of disjoint balls of radii ϵ. We choose N an ϵ-net of minimal cardinality (i.e equal to the covering number of K for ϵ). By definition of a net, each point of P is contained in a closed ball of center in N, and each ball cannot contain more than one point of P, since P is 2ϵ-separated, and the triangle inequality applies in a metric space. Therefore $|P| \leq |N|$, therefore

we have the result.

This lemma, relating the covering and packing numbers, allows to establish a powerful result in the context of our problem, that is the partial equivalence between covering numbers and volume in an Euclidean space.

Proposition 2 (Covering numbers and volume)**.** *Let K be a subset of $\mathbb{R}^n$ and $\epsilon > 0$. Then*

$$\frac{|K|}{|\epsilon B_2^n|} \leq N(K,\epsilon) \leq P(K,\epsilon) \leq \frac{|(K + (\epsilon/2)B_2^n)|}{|(\epsilon/2)B_2^n|}.$$

Here $|\cdot|$ denotes the volume in $\mathbb{R}^n$, B_2^n denotes the unit Euclidean ball in $\mathbb{R}^n$, so ϵB_2^n is a Euclidean ball with radius ϵ.

Proof: We will not prove the result extensively as it is already detailed in Vershynin's book.

Let us simply remark that the lower bound can be explained as: the volume of set K, with covering number $N(K,\epsilon)$ is upper-bounded by the volume of $N(K,\epsilon)$ closed balls of radius ϵ.

The upper bound crucially relies on the previous lemma. Indeed, in order to prove that the volume of a slightly extended set of K is larger than the volume of $N(K,\epsilon)$ closed balls of radius $\frac{\epsilon}{2}$, one sufficiently upper-bounds the packing number of K, which we know from the lemma to be larger than the covering number. And, since the space is normed, one uses the characterisation of the packing number in terms of the maximum number of disjoint balls of radii $\frac{\epsilon}{2}$ with centers in K. This naturally allows us to upper bound the covering number. Both the lower and upper-bounds of $N(K,\epsilon)$ are relatively easy to compute, which will help us derive an easily-usable upper-bound. As an immediate Corollary, we have:

Corollary 2 (Covering numbers of the Euclidean ball)**.** *The covering numbers of the unit Euclidean ball B_2^n satisfy the following for any $\epsilon > 0$:*

$$\left(\frac{1}{\epsilon}\right)^n \leq N(B_2^n,\epsilon) \leq \left(\frac{2}{\epsilon}+1\right)^n.$$

The same upper bound is true for the unit Euclidean sphere S^{n-1}.

Proof: The proof is immediate from the previous proposition.
In particular, if we take $\epsilon = \frac{1}{4}$, we have:

$$N(S^{n-1}, \frac{1}{4}) \leq 9^n$$

The final justification for introducing these objects in the context of our problem is the following proposition:

Proposition 3. *(Partial equivalence of the spectral norm and sup over an ϵ-net)*
$\forall \epsilon > 0$, for any ϵ-net $\mathcal{N}$ of S^{n-1}, for any nxn matrix R,

$$\sup_{x,y\in\mathcal{N}} \langle Rx, y\rangle \leq \sup_{x,y\in S^{n-1}} \langle Rx, y\rangle \leq \frac{1}{1-2\epsilon} \sup_{x,y\in\mathcal{N}} \langle Rx, y\rangle$$

Proof: The lower bound is trivial since $\mathcal{N}$ is a subset of S^{n-1} by definition.

For the upper bound: take $x, y \in S^{n-1}$ and $\epsilon > 0$ and $\mathcal{N}$ an ϵ-net of S^{n-1}. There exists $x_0, y_0 \in \mathcal{N}$, such that $\|x - x_0\| \leq \epsilon$ and $\|y - y_0\| \leq \epsilon$.

$$\begin{aligned}
\langle Rx, y\rangle &= \langle R(x - x_0), y\rangle + \langle Rx_0, y\rangle \\
&\leq \|R(x - x_0)\|\|y\| + \langle Rx_0, y\rangle, \text{ by Cauchy-Schwarz} \\
&\leq \epsilon\|R\| + \langle Rx_0, y_0\rangle + \langle Rx_0, y - y_0\rangle \\
&\leq \epsilon\|R\| + \langle Rx_0, y_0\rangle + \|Rx_0\|\|y - y_0\| \\
&\leq \langle Rx_0, y_0\rangle + 2\epsilon\|R\|
\end{aligned}$$

So $\forall x, y \in S^{n-1}, \exists x_0, y_0 \in \mathcal{N}$, such that $\langle Rx, y\rangle - 2\epsilon\|R\| \leq \langle Rx_0, y_0\rangle$

Thus, taking the sup, we have: $\sup_{x,y\in S^{n-1}} \langle Rx, y\rangle \leq \frac{1}{1-2\epsilon} \sup_{x,y\in\mathcal{N}} \langle Rx, y\rangle$. This completes the proof.

3.3 Step 3: Using an ϵ-net Argument and the Union Bound

We have shown in step 2, that if we take an ϵ-net $\mathcal{N}$ of S^{n-1} with minimum cardinality (in particular, finite cardinality), we can bound its cardinal, using $\epsilon = \frac{1}{4}$, by 9^n. Thus, since we have $\|R\| \leq \frac{1}{1-2\epsilon} \max_{x,y\in\mathcal{N}} \langle Rx, y\rangle$, we can derive, using the same logic as previously detailed in the sketching of step 3: $\forall\, t > 0, \forall \epsilon > 0$,

$$\begin{aligned} P\left[\|R\| \geq t\right] &\leq P\left[\frac{1}{1-2\epsilon} \max_{x,y\in\mathcal{N}} \langle Rx, y\rangle \geq t\right] \\ &\leq |\mathcal{N}| \sup_{x,y\in S^{n-1}} P\left[\langle Rx, y\rangle \geq (1-2\epsilon)\, t\right] \\ &\leq |\mathcal{N}| \exp\left(-\frac{t^2(1-2\epsilon)^2}{4}\right) \end{aligned}$$

For $\epsilon = \frac{1}{4}$, we get $\forall\, t > 0$:

$$P\left[\|R\| \geq t\right] \leq 9^n \exp\left(\frac{-t^2}{16}\right)$$

Take $u = C(\sqrt{n} + t)$, with C an sufficiently large absolute constant, we get that $\|R\| \leq C(\sqrt{n} + t)$, with probability $1 - e^{-t^2}$, $\forall\, t > 0$.

Taking t at the right scale, i.e. $\sqrt{n}$, we have $\|R\| \leq C\sqrt{n}$, with probability $1 - e^{-n}$, with C an absolute constant:

Theorem 3. *(Upper-bound on the spectral norm of R)*

$$\|R\| \leq C\sqrt{n}$$

with probability $1 - e^{-n}$*, with C an absolute constant.*

Remark: We wanted to bound the probability of a sup over an infinite non-countable set, we have introduced simple approximations of the set, in this case ϵ-nets and ϵ-separated subsets, to actually consider the max over a finite cardinal set, whose cardinal we are able to bound. This has enabled us to bound the probability of a sup, by the finite sum of the probabilities in all directions, multiplied by the upper bound of the cardinal of the simple approximated set. We then obtain a sup of the probabilities in all directions of the set, which we rather easily bounded in step 1 using Hoeffding's lemma. This is the key message that we would like to emphasise from this proof and reasoning.

4 Deriving an Upper Bound on the Number of Errors in the Community Detection Problem

In tackling the issue of detecting communities within a network, tools to conclude and derive a result are now available. We use the Davis-Kahan corollary, setting $S = D$ and $T = A = D + R$, specifically focusing on the second-largest eigenvalue, as u_2 is the eigenvector of focus (considering it contains information about the community structure).

It is key to ensure that λ_2 is separated from D's spectrum, i.e. from 0 and λ_1. This separation, denoted as δ, is given by:

$$\delta = \min(\lambda_2 - 0, \lambda_1 - \lambda_2) = \min\left(\frac{p-q}{2}, q\right) n =: \mu n.$$

Thus, with the established limit on $R = A - D$ and invoking the Davis-Kahan corollary, we're able to bound $\|u_2(D) - \theta u_2(A)\|$. We first write the following bound for unit eigenvectors:

There exists a scalar $\theta \in \{-1, 1\}$ such that

$$\|v_2(D) - \theta v_2(A)\|_2 \leq \frac{C\sqrt{n}}{\mu n} = \frac{C}{\mu\sqrt{n}}$$

with probability of at least $1 - e^{-n}$.

As $\|u_2(D)\| = \sqrt{n}$, multiplying the previous bound by $\sqrt{n}$ results in:

$$\| u_2(D) - \theta u_2(A) \|_2 \leq \frac{C}{\mu}.$$

And finally by squaring the previous bound:

$$\sum_{j=1}^{n} |u_2(D)_j - \theta u_2(A)_j|^2 \leq \frac{C^2}{\mu^2},$$

with $u_2(D)_j$ strictly being $\pm 1 \forall j$.

This allows us to conclude: a discord in signs between $\theta v_2(A)_j$ and $v_2(D)_j$ adds at least 1 to the sum. Therefore, the count of disagreeing signs is upper-bounded by

$$\frac{C^2}{\mu^2}.$$

Which, in the community detection problem, means that the eigenvector $u_2(A)$ can serve as a good proxy for $u_2(D)$. The sign patterns of $u_2(A)$ approximately delineates the two communities. This approach, known as spectral clustering, can be summarized in the following theorem:

Theorem 4. *(Spectral Clustering under the Stochastic Block Model)*
Given a graph $G \sim G(n, p, q)$ *with* $p > q$ *and* $\min(q, \frac{p-q}{2} = \mu > 0$, *the spectral clustering method, with a probability exceeding* $1 - e^{-n}$, *accurately discerns G's communities save for a margin of* C^2/μ^2 *misclassified vertices.*

5 Conclusion

In conclusion, exploring the community detection problem, specifically under the Stochastic Block Model has highlighted the efficacy of the spectral clustering method. Introducing a perturbation matrix R and bounding its spectral norm allows to derive probabilistic bounds helpful in recovering community structures within a random graph, with low error.

One interesting insight from this chapter is the Vershynin's use of ϵ-nets, central to derive a probabilistic bound on the operator norm of a matrix, usefully enabling to bound a quantity in one direction in order to bound the supremum of all quantities over an uncountable set.

Finally, Vershynin's result highlights the interplay between linear algebra, random matrices theory, and general probability theory, and how these domains can be mixed to address an apparently simple network analysis problem.

References

[1] R. Vershynin. *High-Dimensional Probability*. Cambridge University Press, 2018.