Community recovery in the stochastic block model using PECOK

Ismael Lemhadri, Youssouf Emin
Ecole Polytechnique *

November 5, 2018

Abstract

In many fields of science, we observe interactions between entities or individuals and we would like to recover the underlying community structure. The stochastic block model is a natural and canonical model for networks with communities. This paper proposes a new semi-definite program to recover communities in the stochastic block model, based on the *PECOK* algorithm introduced in (1). The *PECOK* algorithm is very general and flexible, and can work with a broad class of community models. It achieves exact recovery with high probability - as the number of observations tends to infinity - in the stochastick block model, when a simple variance-gap condition is satisfied.

1 Introduction

1.1 The problem of community recovery

The problem of community recovery is of utmost importance for many modern data-driven challenges. A huge amount of network data has been made available in the last few years from fields as diverse as social networks (2), protein to protein interaction (3), social science (4), webpage sorting (5), among many others. One of the main challenges in network data concerns the problem of *community detection*. The basic goal is that by observing the interactions between different entities (be they individuals or objects) we recover the underlying community structure.
With the increase in size and complexity of modern networks has arisen the need for more model-based approaches and better investigation of the statistical properties of networks. One common theme is to interpret a network as a random graph, the fundamental idea being that edges with similar connectivities should belong to the same group. This allows to interpret the problem of community detection as a special problem of unsupervised classification, leading the statistical community to postulate and fit various probabilistic models for the network structure.

1.2 The stochastic block model

Such models notably include the *stochastic block model*, for short SBM, a simple yet realistic and versatile framework for community structure.
Several approaches have been proposed to solve the problem of community recovery in the context of the stochastic block model. This paper is about this problem and focuses on an approach via semi-definite programming.
While the most direct approach may be to fit it by maximum likelihood, it turns out that fitting

*ismael.lemhadri@polytechnique.edu, youssouf.emin@polytechnique.edu

the SBM is a challenge: the problem of optimizing label assignment over all possibles classes is NP-hard. In practice, maximum likelihood estimation relies on the EM algorithm and suffers great sensitivity to starting points. Alternative methods using spectral clustering (6) are efficient for networks with large blocks but come with no guarantees for sparse networks, as the existing analyses (7, 8) postulate a denser network. We refer the reader to (9) for why spectral clustering is likely to be ineffective for sparse networks, and to section 2.2 for a more in-depth review of the literature.

1.3 Our approach

Whereas previous SDPs looked for similarity in the connectivity matrix, our program looks for similarity in terms of *common neighbors*. This is embodied in a semi-definite program whose solution is with high probability the exact configuration of groups, as soon as a condition on the 'within-between' covariance gap is satisfied. An important feature of this program is that it provides strong theoretical guarantees without requiring any knowledge of the model parameters other than the number of clusters.

2 The stochastic block model

2.1 Definitions and notations

Given a matrix $M \in \mathbb{R}^{p\times q}$, $M_{:j}$ and $M_{i:}$ stand for the j-th column and i-th line of M, respectively.
A random matrix $M \in \mathbb{R}^{n\times n}$ is said to be *symmetrically independent* if it is symmetric and the random variables $(M_{i,j})_{1\leq i\leq j\leq n}$ are independent.
The *stochastic block model* is a random graph model, first introduced by Holland (10), in reference to an older, non-stochastic model largely used in the social sciences.
The general stochastic block model is defined as follows:

- n is the number of vertices in the graph.
- Each vertex $i \in [n]$ is assigned one of k (hidden) labels $g_i \in [k]$ (i.e its community), leading to a partition $G = (G_k)_{1\leq k\leq K}$ of $\{1, .., n\}$.
- Each (unordered) pair of nodes $(u, v) \in V \times V$ is connected independently with probability Q_{g_i,g_j}, where the $K \times K$ *connectivity matrix* Q is symmetric with entries in $[0, 1]$.

As defined above, this is usually referred to as the *conditional stochastic block model*, since it assumes deterministic knowledge of the labels.
To every partition G, we associate the *membership matrix* $A \in \mathbb{R}^{n\times K}$ such that:

$$A_{i,j} = \begin{cases} 1, \text{ if } i \text{ and } j \text{ belong to the same community (i.e the same component of G)} \\ 0, \text{ otherwise} \end{cases}$$

(for all $1 \leq i \leq n$ and $1 \leq j \leq K$).
The corresponding *normalized membership matrix* $B \in \mathbb{R}^{n\times n}$ is such that

$$B_{i,j} = \begin{cases} \frac{1}{|G_k|}, \text{ if } i \text{ and } j \text{ belong to the same community} \\ 0, \text{ otherwise} \end{cases} .$$

Observe that the normalized membership matrix identifies perfectly the partition G.
From now on, G denotes the *exact* partition, A is its membership matrix, B its normalized membership matrix, Q is the connectivity matrix, and $m = \min_{1 \leq k \leq K} |G_k|$ is the size of the smallest community.
Let the random matrix $X \in \{0,1\}^{n \times n}$ represent one observation of the connections (or absence thereof) between the nodes. This means that $\mathbb{E}[X_{i,j}] = Q_{g_i,g_j}$ for every $i \neq j \in \{1, ..., n\}$. We also make the convention that a node does not connect with itself.In other terms, $\mathbb{E}[X] = AQA^t - \text{diag}(\text{AQA}^\text{t})$.
The problem of community detection becomes to recover G (up to a permutation of cluster indices; or, equivalently, recovering B) from one observation of X.

2.2 Parameter estimation in an SBM

Due to the high popularity of the SBM and its ease of applicability and interpretability, many procedures have been develop in order to estimate its parameters. Most algorithms broadly fall into one of three classes, and are assessed based on the detection thresholds for which they succeed, whether the recovery is *partial* or *exact*, and their computational tractability.

- The first class of methods relies on maximum likelihood estimation. However, since there are K^n possible assignments, solving the maximization problem is often computationally intractable, especially at the scale of today's modern datasets.

- The second class uses spectral clustering and has also been very popular in community detection (7, 11, 12). While spectral clustering can perform remarkably well on well-balanced datasets, its results strongly deteriorate for sparse networks.

- The last class re-writes the problem using semi-definite programming (SDP). Such an approach typically obtains the SDP formulation from a relaxation of the non-convex, NP-hard original problem. This class, while being the most recent, is nonetheless particularly promising. (13) shows that SDPs attain the information-theoretic bounds in several regimes, and (14) provides empirical evidence that SDPs outperform spectral methods for fitting SBM with a large number of blocks.

3 Main results

3.1 The PECOK algorithm

Our approach is based on the so-called PECOK algorithm (for PEnalized COnvex relaxation of Kmeans) in (1). The authors consider the framework of *G-latent models*, where they managed to achieve *exact* recovery of the community structure under a simple condition of identifiability on $\Delta_G(Q)$. This achievement encompasses the adaptive case where the number of clusters K is unknown and must also be estimated. In (15), the algorithm is successfully adapted to another probabilistic model that comprises mixture models.
This provides for a powerful, versatile and tractable approach to the problem of variable clustering. The present work develops a version of PECOK adapted to the stochastic block model.
TODO: adaptive estimation of the number of groups.

3.2 Semi-definite program

We consider the following semi-definite program:

$$\mathbf{P}: \max_{B \in \mathcal{C}} \langle X^2, B \rangle,$$

where

$$\mathcal{C} = \left\{ \begin{array}{c} B \succeq 0 \\ \Sigma_a B_{ab} = 1, \forall b \\ B_{ab} \geq 0, \forall a, b \\ \mathrm{tr}(B) = K \end{array} \right\} \subset \mathbb{R}^{p \times p}$$

and we let $\widehat{B}$ an optimal solution of $\mathbf{P}$.
Compared to previous SDPs, e.g (13) and (16), our formulation has the advantage that it does not require any knowledge about the size of the groups or the value of the parameters other than the number of clusters. In (13), an assumption is made that the within-community connection probability is greater than the across-community probability, but our program may recover the clusters even when this is not the case. This is because previous formulations attempt to find communities such that the number of within-community edges minus the across-community edges is largest. For instance, relaxing the maximum likelihood estimator leads to a semi-definite program roughly similar to $\mathbf{P}$ but with X instead of X^2, outlined in (17). In contrast, our formulation looks to cluster vertices based on their *common neighbors* (as $X^2_{i,j}$ represents the number of common neighbors of nodes i, j).

Definition: given a matrix $Q \in \mathbb{R}^{K \times K}$ and a partition G of $[n]$, define $m := \min_{1 \leq k \leq K} |G_k|$, the size of the smallest community, and:

$$\Delta_G(Q) = \min_{j \neq k} \sum_{l=1}^{K} |G_l| |Q_{lj} - Q_{lk}|^2.$$

Remark 1: The semi-definite program $\mathbf{P}$ may be seen as a relaxation of the $K-$Means problem. We refer the reader to (18) for more details about this. On the other hand, some spectral programs such as (6) interpret as a relaxation of the present PECOK algorithm. In terms of relaxation

hierarchies, PECOK therefore stands between $K-$Means and spectral clustering, with the added advantage that it does not require knowledge of the model parameters.

Remark 2: We note that $\Delta_{G^*}(Q) \geq 2 \min_{1\leq k\leq K} |G_k| \min_{\lambda\in \mathrm{Sp}(Q)} |\lambda|$. Compared to the spectral clustering algorithm of (6), this means that the PECOK algorithm does not require that Q be positive definite.

We need to make the following
Assumption: There exists a positive constant c for which:

$$\Delta_{G^*}(Q) \geq c(Q_\infty \ln n + \frac{\ln n + nQ_\infty}{m})$$

Our main result reads:

Theorem 1 *For any $r > 0$, there exist positive constants c and c_1 such that if the assumption holds for c, then $\widehat{B} = B^*$ with probability at least $1 - \frac{c_1}{n^r}$.*

This result is the consequence of several lemmas proved in 5.
We now compare our main assumption to some of the major previous SDP formulations, through the case study of the planted partition model.

4 Comparison with previous results

The *planted partition model* assumes that all the groups have the same size $m = \frac{n}{K}$ and is governed by two parameters p, q, where p represents the probability of connection for two elements from the same group and q the probability of connection for two elements from (any) different groups.That is:

$$\begin{cases} Q_{kk} = p, \forall k \\ Q_{kl} = q, \forall k \neq l \end{cases}$$

Under these conditions and the "classical" assumption that $p \geq q$, the main result of this paper becomes:

$$(p-q)^2 \gtrsim \frac{p \ln n}{m} + \frac{Kp}{m} + \frac{\ln n}{m^2}.$$

The condition provided by (17) to achieve perfect recovery (with high probability) is summarized in their Theorem 4.1 and rewrites:

$$m^2(p-q)^2 \gtrsim mp \ln n + nq \ln n$$

A straightforward calculation shows that our assumption is weaker than the one above as soon as $np + \ln n \lesssim nq \ln n$, i.e $p \lesssim q \ln n$ and $q \gtrsim \frac{\ln n}{n}$. Note that to verify the "identifiability" condition we must have $p, q \gtrsim \frac{\ln n}{n}$ (see (17), Section 4.1). Hence it is sufficient to have $p \lesssim q \ln n$.

In (19) a weaker condition to obtain exact recover is provided when $p \gtrsim \frac{\ln n}{n}$, which is $m^2(p-q)^2 \gtrsim mp \ln n$ (see (19) *Proposition* 4). However, this condition turns out to be equivalent to our condition as soon as $m \gtrsim \frac{n}{\ln n}$, i.e when the number of clusters is less than $\ln n$).

Finally, (20) finds the exact recovery condition $m^2(p-q)^2 \gtrsim mp \ln n + qn$. This is a very weak condition, but it turns out to be equivalent to our condition as soon as $q \asymp p$ and $p \gtrsim \frac{\ln n}{n}$.

5 Proofs

5.1 Proof of Theorem 1

The proof of this theorem is adapted from (1), Section 8.4 with new results that use concentration inequalities for binary variables instead of Gaussian variables.

We begin by introducing the following Lemma 1, which is stated and proved in (1) as Lemma 3.

Lemma 1 *The collection $\mathcal{C}$ contains only one matrix whose support is included in $suppB^*$, that is:*

$$\mathcal{C} \cap \{B, suppB \subseteq suppB^*\} = \{B^*\}.$$

As a consequence of this result, we only need to prove that $\langle \Sigma, B^* - B \rangle > 0$ for all B such that $\text{supp}B \not\subseteq \text{supp}B^*$. Let $Z = AQ$ and let W, W_1 and W_2 be defined as follows. For all $a, b \in \{1, .., n\}$,

$$\begin{aligned}
[W]_{ab} &= \Sigma_{ab} - \frac{1}{2}(|[AZ^t]_{a:}|_2^2 + |[AZ^t]_{b:}|_2^2) + [AZ^tE]_{aa} + [AZ^tE]_{bb} \\
[W_1]_{ab} &= -\frac{1}{2}|[AZ^t]_{a:} - [AZ^t]_{b:}|_2^2 \\
[W_2]_{ab} &= [E_{b:} - E_{a:}]^t[[AZ^t]_{a:} - [AZ^t]_{b:}]
\end{aligned}$$

Then $W = W_1 + W_2 + E^2$ and for all B_1, B_2 in $\mathcal{C}$, $\langle W - \Sigma, B_1 - B_2 \rangle = 0$. We refer the reader to section 8.4 from (1) for more details about this fact.
As a consequence, our proof reduces to proving $\langle W_1 + W_2 + E^2, B^* - B \rangle > 0$ for all $B \in \mathcal{C}$ such that $\text{supp}(B) \not\subseteq \text{supp}(B^*)$.

First, note that (49) from [1] reads:

$$\langle W_1, B^* - B \rangle = \frac{1}{2} \sum_{j \neq k} |Z_{:j} - Z_{:k}|_2^2 |B_{G_j G_k}|_1. \tag{1}$$

We now want to upper-bound the two other terms. We will do that in Lemma 2 and Lemma 3

Lemma 2 *For all $r > 0$, there are positive constants c_2 and c_3 such that with probability larger than $1 - \frac{c_2}{n^r}$:*

$$|\langle W_2, B^* - B \rangle| \leq c_3 \sum_{j \neq k} \Big(\ln n |Z_{:j} - Z_{:k}|_\infty + \sqrt{\ln n . Q_\infty} |Z_{:j} - Z_{:k}|_2 \Big) |B_{G_j G_k}|_1$$

Lemma 3 *Let $d = \ln n + nQ_\infty$. For any $r > 0$, there are positive constants c_4 and c_5 such that with probability larger than $1 - \frac{c_4}{n^r}$:*

$$|\langle E^2, B^* - B \rangle| \leq c_5 \sum_{j \neq k} |B_{G_j G_k}|_1 \Big(\frac{d}{m} + \sqrt{\frac{Q_\infty d \ln n}{m}} \Big)$$

Gathering (1), Lemma 2 and Lemma 3 we obtain that with probability larger than $1 - \frac{c_1}{n^r}$:

$$\begin{aligned}
\langle W_1 + W_2 + E^2, B^* - B \rangle \geq \sum_{j \neq k} |B_{G_j G_k}|_1 \Bigg(\frac{1}{2} |Z_{:j} - Z_{:k}|_2^2 - c_0 \Big(\ln(n) |Z_{:j} - Z_{:k}|_\infty + \sqrt{\ln(n) . Q_\infty} |Z_{:j} - Z_{:k}|_2 \\
+ \frac{d}{m} + \sqrt{\frac{Q_\infty d \ln n}{m}} \Big) \Bigg)
\end{aligned}$$

for some positive constants c_0 and c_1.
Therefore it is enough to prove that

$$\frac{1}{2}|Z_{:j} - Z_{:k}|_2^2 > c_0\Big(\ln n|Z_{:j} - Z_{:k}|_\infty + \sqrt{\ln n.Q_\infty}|Z_{:j} - Z_{:k}|_2 + \frac{d}{m} + \sqrt{\frac{Q_\infty d \ln n}{m}}\Big)$$

for some $j \neq k$. This is in turn equivalent to proving that the left-hand-side is greater than each of the four terms of the righ-hand-side.
Let the assumption hold for some constant $c > 0$. Note that $\Delta_G(Q) = \min_{j\neq k}|Z_{:j} - Z_{:k}|_2^2$.
First, it is straightforward to check that $|Z_{:j} - Z_{:k}|_\infty \leq Q_\infty$, which is enough to deal with the first term of the right-hand-side. The second and third term are bounded immediately using the assumption. As to the last term, note that applying the inequality $2xy \leq x^2 + y^2$ to Assumption 1 gives :

$$|Z_{:j} - Z_{:k}|_2^2 \geq c\Big(Q_\infty \ln n + \frac{d}{m}\Big) \geq 2c\sqrt{\frac{Q_\infty d \ln n}{m}}$$

which is enough to conclude for an appropriate choice of c.

5.1.1 Proof of Lemma 2

The first steps of this proof are analogous to the proof of Theorem 1 in (1).
We define $E = X - AQA^t$ and $D = \mathrm{diag}(\mathrm{AQA^t})$, and we note that since X is symmetrically independent, $\mathcal{E} := E + D$ is also symmetrically independent. In addition, $\mathcal{E}$ is a centered random matrix (i.e $\mathbb{E}[\mathcal{E}] = 0$) with null diagonal.

Consider any a and b in $\{1, .., n\}$ and let j and k be such that $a \in G_k$ and $b \in G_j$.
If $j = k$, $(W_2)_{ab} = 0$. Otherwise:

$$\begin{aligned}(W_2)_{ab} &= [E_{b:} - E_{a:}].[Z_{:j} - Z_{:k}] \\ &= \sum_{c\neq a, c\neq b} (\mathcal{E}_{bc} - \mathcal{E}_{ac}).(Z_{cj} - Z_{ck}) + \mathcal{E}_{ab}(Z_{aj} + Z_{bk} - Z_{bj} - Z_{ak}) \\ &- D_{bb}.(Z_{bj} - Z_{bk}) - D_{aa}.(Z_{aj} - Z_{ak}).\end{aligned}$$

Note that since a and b are not in the same group, the variables $(\mathcal{E}_{bc} - \mathcal{E}_{ac})_{c\neq a, c\neq b}$ are independent. In addition $Var(\mathcal{E}_{ec}) \leq Q_\infty$ for all $e, c \in \{1, .., n\}$ and thus:

$$\begin{aligned}\mathrm{Var}\Big[\sum_{c\neq a, c\neq b} (\mathcal{E}_{bc} - \mathcal{E}_{ac}).(Z_{cj} - Z_{ck})\Big] &= \sum_{c\neq a, c\neq b} (Var(\mathcal{E}_{bc}) + Var(\mathcal{E}_{ac})).(Z_{cj} - Z_{ck})^2 \\ &\leq 2Q_\infty|Z_{:j} - Z_{:k}|_2^2\end{aligned}$$

Therefore, using Bernstein's inequality, there are two positive constants $c_2^{(1)} > 0$ and $c_3^{(1)}$ such that with probability larger than $1 - \frac{c_2^{(1)}}{n^{r+2}}$:

$$\sum_{c\neq a, c\neq b} (\mathcal{E}_{bc} - \mathcal{E}_{ac}).(Z_{cj} - Z_{ck}) \leq c_3^{(1)}\Big(|Z_{:j} - Z_{:k}|_\infty \ln n + |Z_{:j} - Z_{:k}|_2\sqrt{Q_\infty \ln n}\Big).$$

Going back to $(W_2)_{ab}$, we obtain a constant $c_2 > 0$ such that with probability larger than $\frac{1}{n^{r+2}}$:

$$|(W_2)_{ab}| \leq c_3\Big(|Z_{:j} - Z_{:k}|_\infty \ln n + |Z_{:j} - Z_{:k}|_2\sqrt{Q_\infty \ln n}\Big)$$

Now apply union bound over all coefficients of W_2. It must then be that with probability larger than $1 - \frac{c_2}{n^r}$,

$$|\langle W_2, B^* - B\rangle| \leq c_3 \sum_{j\neq k} \Big(\ln n |Z_{:j} - Z_{:k}|_\infty + \sqrt{\ln n . Q_\infty} |Z_{:j} - Z_{:k}|_2 \Big) |B_{G_j G_k}|_1,$$

which concludes the proof of Lemma 2.

5.1.2 Proof of Lemma 3

Again, the first steps are identical, and we refer the reader to (1) for completeness. The arguments that lead to (58) in (1) remain valid by replacing $\widetilde{\Gamma} - \Gamma$ by E^2, which gives:

$$|\langle E^2, (I - B^*)B(I - B^*)\rangle| \leq 2\Big[\sum_{j\neq k} |B_{G_j G_k}|\Big] \Big(\frac{||E^2||_{op}}{2m} + 3|B^* E^2|_\infty \Big)$$

Applying *Theorem 5.2* of (6), there is a constant $C > 0$ such that with probability larger than $1 - n^{-r}$, $||\mathcal{E}|| < C\sqrt{d}$. Additionally, note that $||E^2||_{op} \leq 2(||\mathcal{E}||^2_{op} + ||D||^2_{op})$, so there is a constant $c_5^{(1)} > 0$ such that $||E^2||_{op} \leq c_5^{(1)} d$.
As to the second term, write

$$B^*.E^2 = B^*\mathcal{E}^2 + B^*.\mathcal{E}.D + B^*.D.\mathcal{E} + B^*.D^2.$$

First, a straightfoward calculation gives that $|B^* D^2|_\infty \leq \frac{1}{m}$. Secondly, for all a in G_k and b in G_j:

$$[B^*.\mathcal{E}.D]_{ab} = \frac{1}{|G_k|} \sum_{l\in G_k} \mathcal{E}_{lb} D_{bb} \text{ and } [B^*.D.\mathcal{E}]_{ab} = \frac{1}{|G_k|} \sum_{l\in G_k} \mathcal{E}_{lb} D_{ll}.$$

Consequently, using Bernstein's inequality, there is a constant $c_5^{(2)} > 0$ such that with probability larger than $1 - \frac{2}{n^{r+2}}$:

$$[B^*.\mathcal{E}.D]_{ab} \leq \frac{c_5^{(2)}}{|G_k|} \Big(|D|_\infty \ln n + \sqrt{\mathrm{Var}\Big(\sum_{l\in G_k} \mathcal{E}_{lb} D_{bb} \Big) \ln n} \Big)$$
$$\text{and } [B^*.D.\mathcal{E}]_{ab} \leq \frac{c_5^{(2)}}{|G_k|} \Big(|D|_\infty \ln n + \sqrt{\mathrm{Var}\Big(\sum_{l\in G_k} \mathcal{E}_{lb} D_{ll} \Big) \ln n} \Big)$$

The variance can be bounded as follows:

$$\mathrm{Var}\Big(\sum_{l\in G_k} \mathcal{E}_{lb} D_{bb} \Big) = \sum_{l\in G_k} D_{bb}^2 \mathrm{Var}(\mathcal{E}_{lb}) \leq |D|^2_\infty Q_\infty |G_k| \leq n Q_\infty \leq d$$
$$\text{and } \mathrm{Var}\Big(\sum_{l\in G_k} \mathcal{E}_{lb} D_{ll} \Big) = \sum_{l\in G_k} D_{ll}^2 \mathrm{Var}(\mathcal{E}_{lb}) \leq |D|^2_\infty Q_\infty |G_k| \leq n Q_\infty \leq d$$

Since $\ln n \leq d$, there is therefore a constant $c_5^{(3)} > 0$ such that with probability larger than $1 - \frac{2}{n^{r+2}}$:

$$[B^*.\mathcal{E}.D]_{ab} \leq c_5^{(3)} \frac{d}{m} \text{ and } [B^*.D.\mathcal{E}]_{ab} \leq c_5^{(3)} \frac{d}{m}.$$

Finally, union bound gives again that with probability larger than $1 - \frac{2}{n^r}$:

$$[B^*.\mathcal{E}.D]_\infty + [B^*.D.\mathcal{E}]_\infty \leq 2c_5^{(3)} \frac{d}{m}$$

.

Now the quadratic term. For any t, u $\in \mathbb{R}_+$, it is clear that

$$\mathbb{P}(|[B^*\mathcal{E}^2]_{ab}| > t) \leq \mathbb{P}(\sum_c \mathcal{E}_{cb}^2 > u) + \mathbb{P}(|[B^*\mathcal{E}^2]_{ab}| > t, \sum_c \mathcal{E}_{cb}^2 \leq u).$$

The first term can be bounded immediately using Bernstein's inequality:

$$\begin{aligned}\mathbb{P}(\sum_c \mathcal{E}_{cb}^2 > u) &= \mathbb{P}(\sum_c \mathcal{E}_{cb}^2 - \mathbb{E}\Big[\sum_c \mathcal{E}_{cb}^2\Big] > u - \mathbb{E}\Big[\sum_c \mathcal{E}_{cb}^2\Big]) \\ &\leq 2\exp\Big(-\frac{\frac{1}{2}(u - \mathbb{E}\Big[\sum_c \mathcal{E}_{cb}^2\Big])^2}{nQ_\infty + \frac{1}{3}(u - \mathbb{E}\Big[\sum_c \mathcal{E}_{cb}^2\Big])}\Big)\end{aligned}$$

On the other hand, we write:

$$\begin{aligned}|G_k|[B^*\mathcal{E}^2]_{ab} &= \sum_{l \in G_k} [\mathcal{E}^2]_{lb} = \sum_{1 \leq l \neq c \leq n} \mathcal{E}_{cb} 1_{l \in G_k} \mathcal{E}_{lc} \\ &= \sum_{1 \leq l < c \leq n} (\mathcal{E}_{cb} 1_{l \in G_k} + \mathcal{E}_{lb} 1_{c \in G_k}) \mathcal{E}_{lc}\end{aligned}$$

$$\begin{aligned}\mathbb{E}\Big[[B^*\mathcal{E}^2]_{ab}.|G_k|\Big|(\mathcal{E}_{cb})_c\Big] &= \mathbb{E}\Big[\sum_{1 \leq l < c \leq n} (\mathcal{E}_{cb} 1_{l \in G_k} + \mathcal{E}_{lb} 1_{c \in G_k}) \mathcal{E}_{lc}\Big|(\mathcal{E}_{cb})_c\Big] \\ &= 1_{b \in G_k} . \sum_c \mathcal{E}_{cb}^2\end{aligned}$$

and thus

$$\begin{aligned}
\mathrm{Var}\Big[\sum_{1\le l<c\le n}(\mathcal{E}_{cb}.1_{l\in G_k}+\mathcal{E}_{lb}1_{c\in G_k})\mathcal{E}_{lc}\Big|(\mathcal{E}_{cb})_c\Big] &= \sum_{1\le l<c\le n}\mathrm{Var}[\mathcal{E}_{lc}|(\mathcal{E}_{cb})_c].(\mathcal{E}_{cb}1_{l\in G_k}+\mathcal{E}_{lb}1_{c\in G_k})^2\\
&\le \sum_{1\le l<c\le n}\mathrm{Var}(\mathcal{E}_{lc}).(\mathcal{E}_{cb}.1_{l\in G_k}+\mathcal{E}_{lb}1_{c\in G_k})^2\\
&\le Q_\infty\sum_{1\le l<c\le n}(\mathcal{E}_{cb}.1_{l\in G_k}+\mathcal{E}_{lb}1_{c\in G_k})^2\\
&= Q_\infty\sum_{1\le l<c\le n}\Big(\mathcal{E}_{cb}^2 1_{l\in G_k}+\mathcal{E}_{lb}^2 1_{c\in G_k}+2\mathcal{E}_{cb}.\mathcal{E}_{lb}1_{l\in G_k}1_{c\in G_k}\Big)\\
&= Q_\infty\sum_{1\le l\ne c\le n}(\mathcal{E}_{cb}^2 1_{l\in G_k}+\mathcal{E}_{cb}.\mathcal{E}_{lb}1_{l\in G_k}1_{c\in G_k})\\
&= Q_\infty\Big(|G_k|\sum_{1\le c\le n}\mathcal{E}_{cb}^2+(\sum_{l\in G_k}\mathcal{E}_{lb})^2\Big)\le 2Q_\infty|G_k|\sum_c\mathcal{E}_{cb}^2,
\end{aligned}$$

where we used the Cauchy-Schwarz inequality for the last step.

Consequently, conditioning by $(\mathcal{E}_{cb})_{c\in\{1,..,n\}}$ and applying Bernstein's inequality, the third term can be bounded as follows:

$$\begin{aligned}
&\mathbb{P}(|[B^*\mathcal{E}^2]_{ab}|>t,\sum_c\mathcal{E}_{cb}^2\le u)=\mathbb{P}(|[B^*\mathcal{E}^2]_{ab}|.|G_k|>t|G_k|,\sum_c\mathcal{E}_{cb}^2\le u)\\
&\quad=\mathbb{E}\left[\mathbb{P}(|[B^*\mathcal{E}^2]_{ab}|.|G_k|>t.|G_k|,\sum_c\mathcal{E}_{cb}^2\le u\Big|(\mathcal{E}_{cb})_{c\in\{1,..,n\}})\right]\\
&\quad=\mathbb{E}\left[\mathbb{P}\Big((|[B^*\mathcal{E}^2]_{ab}|.|G_k|>t|G_k|\Big|(\mathcal{E}_{cb})_{c\in\{1,..,n\}}\Big)1_{\sum_c\mathcal{E}_{cb}^2\le u}\right]\\
&\quad\le\mathbb{E}\left[\mathbb{P}\Big(|[B^*\mathcal{E}^2]_{ab}.|G_k|-\mathbb{E}\Big[[B^*\mathcal{E}^2]_{ab}.|G_k|\Big|(\mathcal{E}_{cb})_c\Big]|>t|G_k|-|\mathbb{E}\Big[[B^*\mathcal{E}^2]_{ab}.|G_k|\Big|(\mathcal{E}_{cb})_c\Big]|\Big|(\mathcal{E}_{cb})_c\Big)1_{\sum_c\mathcal{E}_{cb}^2\le u}\right]\\
&\quad\le\mathbb{E}\left[\mathbb{P}\Big(|[B^*\mathcal{E}^2]_{ab}.|G_k|-\mathbb{E}\Big[[B^*\mathcal{E}^2]_{ab}.|G_k|\Big|(\mathcal{E}_{cb})_c\Big]|>t|G_k|-u\Big|(\mathcal{E}_{cb})_c\Big)1_{\sum_c\mathcal{E}_{cb}^2\le u}\right]\\
&\quad\le\mathbb{E}\left[2\exp\left(-\frac{\frac{1}{2}(t.|G_k|-u)^2}{Var\left[\sum_{1\le l<c\le n}(\mathcal{E}_{cb}.1_{l\in G_k}+\mathcal{E}_{lb}1_{c\in G_k})\mathcal{E}_{lc}\Big|(\mathcal{E}_{cb})_c\right]+\frac{2}{3}(t|G_k|-u)}\right)1_{\sum_c\mathcal{E}_{cb}^2\le u}\right]\\
&\quad\le\mathbb{E}\left[2\exp(-\frac{\frac{1}{2}(t|G_k|-u)^2}{2|G_k|Q_\infty\sum_{1\le l\le n}\mathcal{E}_{lb}^2+\frac{2}{3}(t|G_k|-u)})1_{\sum_c\mathcal{E}_{cb}^2\le u}\right]\\
&\quad\le 2\exp(-\frac{\frac{1}{2}|G_k|(t-\frac{u}{|G_k|})^2}{2Q_\infty u+\frac{2}{3}(t-\frac{u}{|G_k|})})
\end{aligned}$$

Gathering the above altogether yields

$$\mathbb{P}(|[B^*\mathcal{E}^2]_{ab}| > t) \leq 2\Bigg(\exp\Big(- \frac{\frac{1}{2}|G_k|(t - \frac{u}{|G_k|})^2}{2Q_\infty u + \frac{2}{3}(t - \frac{u}{|G_k|})}\Big) + \exp\Big(- \frac{\frac{1}{2}(u - \mathbb{E}\Big[\sum_c \mathcal{E}^2_{cb}\Big])^2}{nQ_\infty + \frac{1}{3}(u - \mathbb{E}\Big[\sum_c \mathcal{E}^2_{cb}\Big])}\Big)\Bigg)$$

If we choose:

$$u = (c_4^{(1)} + 1).(\ln n + n.Q_\infty) = (c_4^{(1)} + 1)d$$
$$t = \frac{u}{m} + c_4^{(2)}.(\frac{\ln n}{m} + \sqrt{\frac{dQ_\infty.\ln n}{m}})$$

and since $\mathbb{E}\Big[\sum_c \mathcal{E}^2_{cb}\Big] \leq nQ_\infty$, we have:

$$\begin{aligned}
\exp\Big(- \frac{\frac{1}{2}(u - \mathbb{E}\Big[\sum_c \mathcal{E}^2_{cb}\Big])^2}{nQ_\infty + \frac{1}{3}(u - \mathbb{E}\Big[\sum_c \mathcal{E}^2_{cb}\Big])}\Big) &\leq \exp\Big(- \frac{\frac{1}{2}(u - nQ_\infty)^2}{nQ + \frac{1}{3}(u - nQ_\infty)}\Big) \\
&\leq \exp\Big(- \frac{\frac{1}{2}(c_4^{(1)}(nQ_\infty + \ln n))^2}{(c_4^{(1)} + 1)(nQ_\infty + \ln n)}\Big) \\
&\leq \exp\Big(- \frac{(c_4^{(1)})^2}{2(c_4^{(1)} + 1)}(\ln n + nQ_\infty)\Big) \leq n^{-\frac{(c_4^{(1)})^2}{2(c_4^{(1)}+1)}}
\end{aligned}$$

Note that the function $v \to (r+2)\ln n(2Q_\infty.u + \frac{2}{3}v) - \frac{1}{2}mv^2$ is a decreasing function on $[c(r)\ln n/m, +\infty[$, with $c(r)$ a constant depending only on r. Therefore, if $c_4^{(2)} \geq c(r)$:

$$\begin{aligned}
(r+2)\ln n(2Q_\infty.u + \frac{2}{3}(t - \frac{u}{|G_k|})) - \frac{1}{2}(t - \frac{u}{|G_k|})^2.m &\leq (r+2)\ln n(2Q_\infty.u + \frac{2}{3}(t - \frac{u}{m})) - \frac{1}{2}(t - \frac{u}{m})^2.m \\
&= (r+2)\ln n.(2.Q_\infty(c_4^{(1)} + 1).(\ln n + n.Q_\infty) + \frac{2}{3}c_4^{(2)}.(\frac{\ln n}{m} + \sqrt{\frac{dQ_\infty.\ln n}{m}})) - \frac{1}{2}(c_4^{(2)})^2.(\frac{\ln n}{m} + \sqrt{\frac{dQ_\infty.\ln n}{m}})^2 \\
&= 2.dQ_\infty((r+2)(c_4^{(1)} + 1) - \frac{1}{4}(c_4^{(2)})^2)\ln n + (\frac{2}{3}c_4^{(2)}(r+2) - \frac{1}{2}(c_4^{(2)})^2)(\frac{\ln n}{m} + \sqrt{\frac{dQ_\infty.\ln n}{m}})\ln n
\end{aligned}$$

After choosing $c_4^{(1)}$ and $c_4^{(2)}$ so that

$$\frac{(c_4^{(1)})^2}{2(c_4^{(1)} + 1)} \geq r + 2$$
$$(r+2)(c_4^{(1)} + 1) - \frac{1}{4}(c_4^{(2)})^2 \leq 0$$
$$\frac{2}{3}c_4^{(2)}(r+2) - \frac{1}{2}(c_4^{(2)})^2 \leq 0$$

we get:

$$\exp\Big(- \frac{\frac{1}{2}(u - \mathbb{E}\Big[\sum_c \mathcal{E}^2_{cb}\Big])^2}{nQ_\infty + \frac{1}{3}(u - \mathbb{E}\Big[\sum_c \mathcal{E}^2_{cb}\Big])}\Big) \leq n^{-(r+2)}$$
$$\text{i.e}\exp\Big(- \frac{\frac{1}{2}((t - \frac{u}{|G_k|})^2.|G_k|)}{2Q_\infty.u + \frac{2}{3}(t - \frac{u}{|G_k|})}\Big) \leq n^{-(r+2)}$$

Finally, there are two constants $c_5^{(1)}$, and $c_4^{(3)}$ such that:

$$\mathbb{P}\Big(|[B^*\mathcal{E}^2]_{ab}| > c_4^{(3)}\Big(\frac{d}{m} + \sqrt{\frac{dQ\ln n}{m}}\Big)\Big) \leq \frac{c_5^{(1)}}{n^{r+2}}$$

and Lemma 3 follows after an union bound.

Acknowledgements

We would like to thank Christophe Giraud for fruitful mentoring, and Martin Royer for generously providing numerical experiment support.

References

[1] F. Bunea, C. Giraud, M. Royer, and N. Verzelen. Pecok: a convex optimization approach to variable clustering. 2016.

[2] X. Zhang, C. Moore, and M.E.J Newman. Random graph models for dynamic networks. 2016.

[3] M.S. Cline and al. Integration of biological networks and gene expression data using cytoscape. 2007.

[4] A. Goldenberg, A.X. Zheng, S.E. Fienberg, and E.M. Airoldi. A survey of statistical network models. 2010.

[5] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Trawling the web for emerging cyber-communities. 1999.

[6] J Lei and A. Rinaldo. Consistency of spectral clustering in stochastic block models. *Carnegie Mallon University*, 2014.

[7] K. Rohe, S. Chatterjee, and B. Yu. Spectral clustering and the high-dimensional stochastic block model. 2011.

[8] D.E. Fishkind, D. L. Sussman, M. Tang, J.T. Vogelstein, and C.E. Priebe. Consistent adjacency-spectral partitioning for the stochastic block model when the model parameters are unknown. 2013.

[9] P. Zhang. Robust spectral detection of global structures in the data by learning a regularization. 2016.

[10] P.W. Holland, K.B. Laskey, and S. Leinhardt. Stochastic blockmodels: first steps. 1983.

[11] T. Qin and K. Rohe. Regularized spectral clustering under the degree-corrected stochastic blockmodel. 2013.

[12] K. Chaudhuri, F. Chung, and A. Tsiatas. Spectral clustering of graphs with general degrees in the extended planted partition model. 2012.

[13] B. Hajek, Y. Wu, and J. Xu. Semidefinite programs for exact recovery of a hidden community. 2016.

[14] A.A. Amini and E. Levina. On semidefinite relaxations for the block model. 2016.

[15] M. Royer. Adaptive clustering through semidefinite programming. 2017.

[16] E. Abbe, A.S. Bandeira, and G. Hall. Exact recovery in the stochastic block model. 2016.

[17] A.A. Amini and E. Levina. On semidefinite relaxations for the block model. 2016.

[18] J. Peng and Y. Wei. Approximating k-means-type clustering via semidefinite programming. 2005.

[19] A Perry and A.S. Wein. A semidefinite program for unbalanced multisection in the stochastic block model. 2016.

[20] B. Hajek, Y. Wu, and J. Xu. Achieving exact cluster recovery threshold via semidefinite programming: Extensions. 2016.