Community detection on graphs

Leonardo Martins Bianco

Advised by Christine Keribin and Zacharie Naulet

Contents

1 Introduction **2**
- 1.1 Objectives of the internship . . . 2
- 1.2 Motivations for the internship . . . 2

2 Algorithms for community detection **6**
- 2.1 The stochastic block model . . . 6
- 2.2 Statistical approach . . . 9
- 2.3 Optimization approach . . . 12

3 Learning theory **17**
- 3.1 Agreement and degrees of recovery . . . 17
- 3.2 Asymptotic topologies . . . 18
- 3.3 Consistency of variational EM . . . 21
- 3.4 Consistency of spectral clustering . . . 21

4 Numerical experiments **25**
- 4.1 Numerical experiments . . . 25

5 Conclusion and outlook **26**
- 5.1 Conclusions . . . 26

A Proofs and calculations **27**
- A.1 Proof of Theorem 1 . . . 27
- A.2 Convergence properties . . . 28

B The case of two communities **30**
- B.1 The case of two communities* . . . 30

Chapter 1

Introduction

1.1 Objectives of the internship

This report succintly presents the work done during the internship concluding the Master 2 program MVA - *Mathématiques, Vision, Apprentissage*. The internship was academic in nature, and lasted six months. It will be followed by a PhD thesis on the same subject, by the same intern, under the same advisors. Its general goals were, sequentially,

1st To introduce the intern to the problem of community detection on graphs. This consists on getting a firm understanding on the different canonical approaches to the problem, gaining familiarity first with the field's classical and recent research literature, and gaining probabilistic as well as statistical intuitions that can be applied to related problems;

2nd To tackle a current research question, suggested by the advisors of the internship;

3rd To apply the knowledge obtained to think of new research directions autonomously.

1.2 Motivations for the internship

The field of community detection on graphs is rich both in theory and in applications. In this section, the motivations for the field and for the particular problem of the internship are explained.

1.2.1 Graphs with communities

A graph is a mathematical object expressing the interaction between entities. One of the most interesting and relevant structures a graph can have is that of *communities*. However, precisely defining this structure is subtle and not agreed upon. One common intu-

ition is to think of it as a set of vertices having more connections among themselves than with all other vertices. Such an intuition is useful in many situations, and it is called an *assortative* notion of community. In other cases, however, one's intuition of what such groups are does not fit the assortative case. Consider a party, where people dance in pairs. There are fifty men and fifty women, and assume that each man will pair up with some woman to dance. In this case, one can consider that there are two groups in the party, men and women. However, no two members of the same group connect. This is what is called an *dissortative* notion of community, as its members share a *pattern* of connection instead of denser connections. The conceptual difficulties do not stop here, as it may happen that the notion of what a community is depends on the *scale* (i.e., the amount of "zoom" into the graph) considered. It is clear then that the problem of community detection on graphs, also called graph clustering, is delicate.

1.2.2 Common approaches to community detection

Assume an arbitrary graph is observed, with the sole hypothesis that in it there are communities, i.e., there exists some particular partition of the graph corresponding to community assignments. This hypothesis is deliberately vague. Consider the following common, yet very distinct, approaches to building algorithms to find such partition.

Statistical approaches

In *statistical* approaches to community detection, one assumes that the communities assigned to nodes and the edges formed are random variables, and thus the graph observed arises as an observation of some model. One then analyzes its likelihood function to derive algorithms derived in order to estimate the model's parameters and infer the graph's communities. In this report, and in most recent research, the model assumed in this context is called the stochastic block model, or SBM for short. As it will be seen, directly computing the likelihood of an observation under such model is intractable, since doing so requires one to sum over all possible cluster assignments for the nodes. The number of terms in such a sum grows exponentially with the number of nodes, and there is no way to simplify it into a tractable form. A common way around this difficulty is to substitute the exact likelihood function by a variational approximation to it. This approximation can be built in various ways, the simplest one being the *mean-field* approximation. The optimization problem that arises can in theory be solved by an alternating optimization algorithm akin to the classical EM algorithm, and is called the Variational EM.

Optimization approaches

In *optimization* approaches to community detection, it is not assumed that the graph at hand is an observation of some (probabilistic) model. Instead, one derives algorithms

based on heuristics and approximations to optimization problems. One popular optimization problem for finding communities in graphs is the balanced min-cut problem, which searches for a partition such that the number of edges across classes is minimal. This agrees with the assortative intuition of what a community is. This problem is NP-hard and is popularly approximated by a relaxed version leading to the spectral clustering algorithm.

These spectral approaches can be seen as ways of embedding the graph in some vector space. In this vectorial representation of the graph, there are as many vectors as there are nodes, and as many dimensions as communities. In consequence, its dimensionality is typically low when compared to the complexity of the general graph representation. Moreover, under certain conditions, clustering the nodes in this representation (using classical vector space clustering algorithms such as k-means) could correspond to a clustering of nodes in the graph, thus revealing the communities.

1.2.3 Learning theory

These approaches are popular, but without a ground truth for the communities on the graph there is no way of measuring the accuracies of their results. In such an *unsupervised* setting, the answer to the question of what are the graph's communities must be the output of the algorithm itself.

One way of dealing with these difficulties is to consider a *generative* model. This provides a definite ground truth assignment of communities to the nodes of a graph arising as an observation from such model. As a consequence, one can develop, measure the accuracy, and compare algorithms designed to recover these communities. Arguably, the most popular such model is called the Stochastic Block Model (SBM). Essentially, it randomly assigns communities to nodes and then connects any pair of them with a probability depending on the communities of the pair, see Figure 2.1.

1.2.4 Applications

Data in the form of graphs and networks naturally appear in fields such as ecology, power grids, transportation systems, social networks, recommendation systems, neurology, telecommunications, and so on. Some interesting applications of community detection methods include the analysis of political blogospheres [4], analysis of criminal networks [5], cell profiling [10], analysis of ecological networks [9], and so on. There is a growing abundance of network data openly available online. Some useful resources are [11, 13, 12, 6].

This reports presents experiments performed in simulated data, as well as in real data coming from these sources.

1.2.5 Note on contributions

It is important to emphasize that this report aims to convey, beyond the effort put into answering any particular question, the view the intern ended up having of the field and the new challenges to be adressed as a continuation on his PhD. Several research problems were explored during these months. This was at times intentional, aiming to give the intern an understanding of the different open questions in the field, and at other times a result of the difficulty of the question at hand.

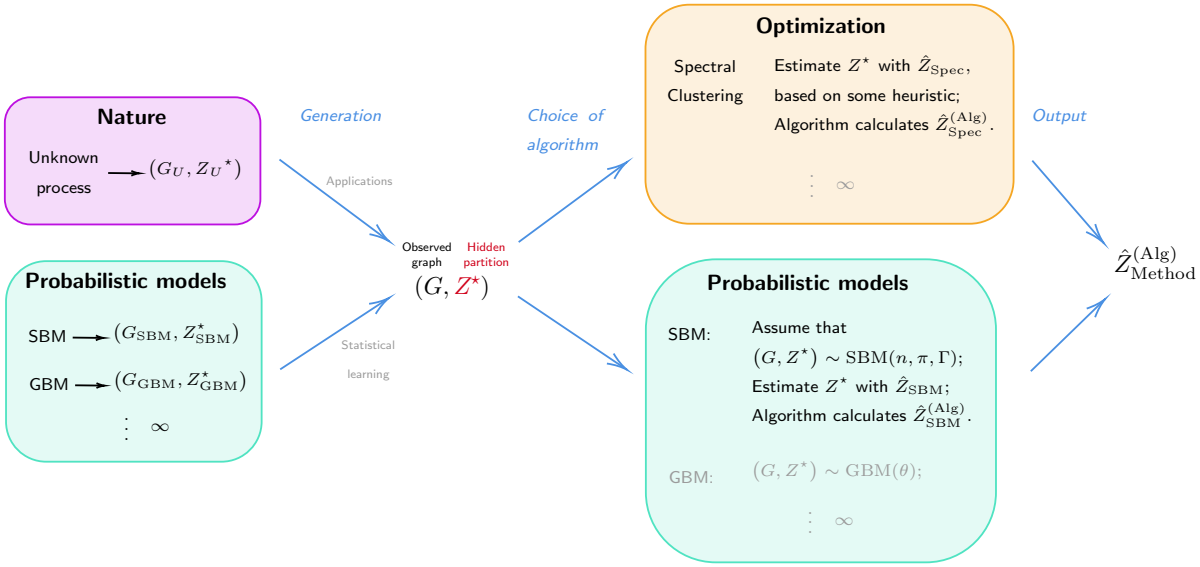


Figure 1.1: Diagram describing the steps taken from observing a graph to estimating the communities.

Chapter 2

Algorithms for community detection

2.1 The stochastic block model

2.1.1 The general SBM

The canonical probabilistic model for graphs with community structure is called the stochastic blockmodel, or SBM for short. For a lengthier discussion on the origins and variants of this model, refer to [1].

Definition 1 (Stochastic blockmodel)**.** Let $n \in \mathbb{N}$, $k \in \mathbb{N}$, $\pi = (\pi_1, \dots, \pi_k)$ be a probability vector on $[k] := \{1, \dots, k\}$ and Γ be a $k \times k$ symmetric matrix with entries $\gamma_{ij} \in [0, 1]$. A pair (Z, G) is said to be *drawn under a* SBM(n, π, Γ) if

- $Z = (Z_1, \dots, Z_n)$ is an n-tuple of $\mathbb{N}^k$-valued random variables $Z_i \sim \mathcal{M}(1, \pi)$,
- G is a simple graph with n vertices whose symmetric adjacency matrix has zero diagonal and for $j > i$, $A_{ij}|Z \sim$ Ber(γ_{Z_i, Z_j}), the lower triangular part being completed by symmetry.

Remark. The quantity n should be thought of as being the number of nodes in G, k should be thought of as being the number of communities in G, π should be thought of as being a prior on the community assignments Z, and Γ should be thought of as a matrix of intra-cluster and inter-cluster connectivities. The random variables of the model are the n community assignments Z and the $\binom{n}{2}$ entries A_{ij} of the adjacency matrix.

Remark. Although each community assignment Z_i is a vector, the same Z_i can be used to denote the *number* of the community that node i is assigned to, to make notation lighter.

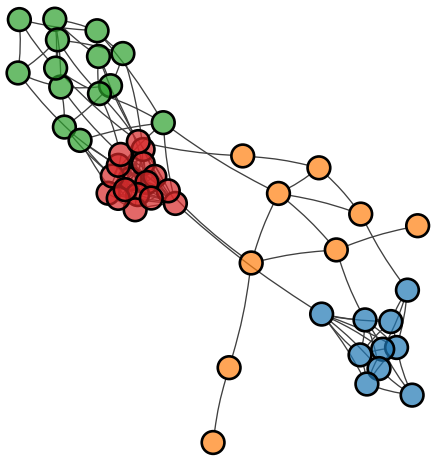


Figure 2.1: An example of an SBM graph with assortative communities

It is important to emphasize that although intuition frequently refers to the assortative case, such as in Figure 2.1, the SBM is versatile and can reproduce many other characteristics of graphs with communities. For instance, the SBM can generate bipartite graphs as a model for the example given in the introduction, where couples dance in a party; it can also generate graphs with "stars", and reproduce the "core-periphery" phenomenon. See Figure 2.2.

2.1.2 The symmetric SBM

Even though the SBM is a simple and intuitive model for graphs with communities, the calculations associated with it can already yield long expressions and present subtleties. For this reason, it is desirable to have a yet simpler version of the SBM where one can test intuitions and perform preliminary calculations. The symmetric SBM is precisely such a model.

Definition 2 (Symmetric SBM)**.** The pair (X, G) is drawn from SSBM (n, k, p, q) if it is drawn from an SBM model with $\pi = \frac{1}{k}\mathbf{1}_k$ and Γ taking values p on the diagonal and q outside the diagonal.

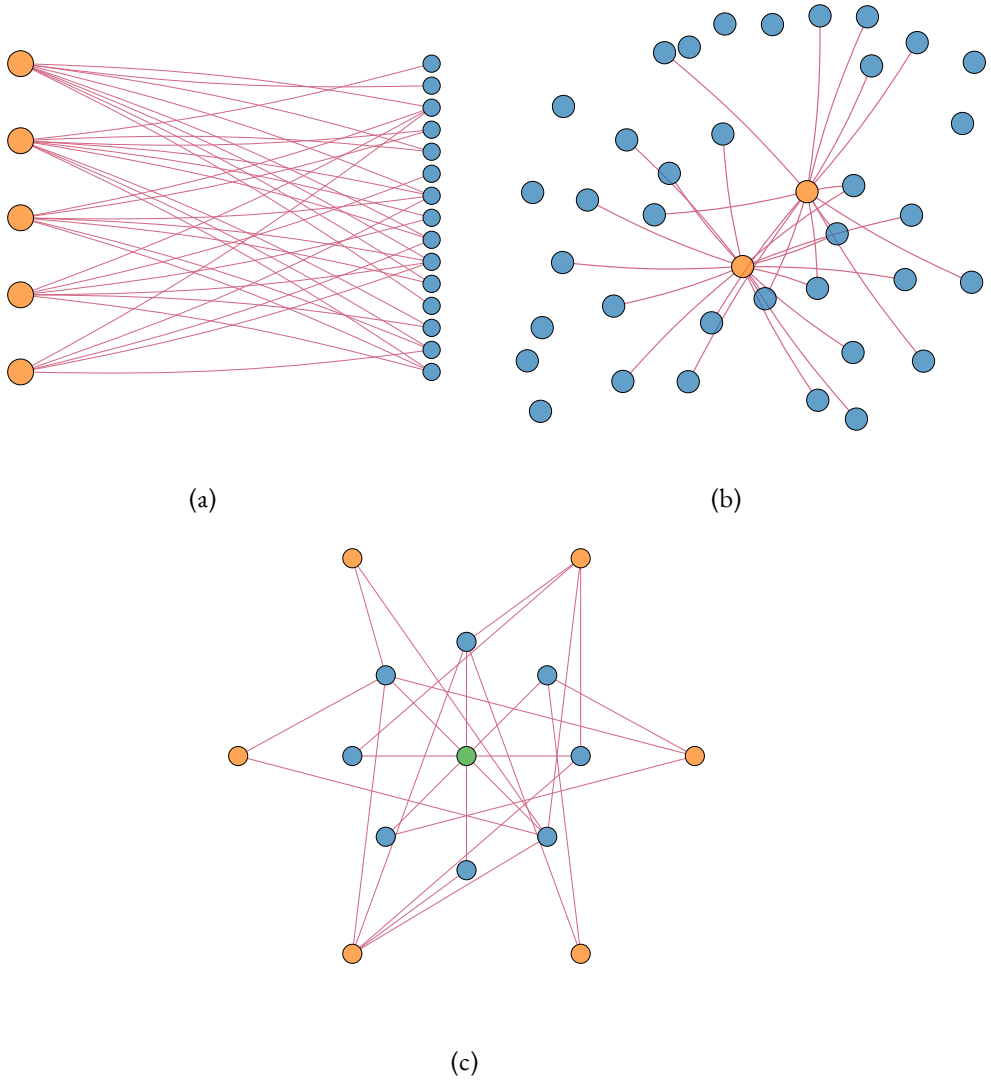


Figure 2.2: The SBM is versatile and can give rise to different features, such as (a) bipartite structures, (b) star structures, (c) core-periphery structures.

2.2 Statistical approach

In this section, the general framework for *statistical* approaches to community detection is laid down. It consists in assuming a model for the observed graph, then estimating the parameters of the model by approximately maximizing the likelihood using an algorithm similar to the classical EM, called the *variational EM algorithm*. The communities can then immediately be inferred from the variational parameters optimized during the inference procedure. Here, the model chosen is the SBM due to its popularity.

2.2.1 Model likelihood

Complete model likelihood

The community assignment variable Z is latent, in the sense that in practice it is not observed. However, it *is* a random variable of the model that is determined at the moment of sampling of the graph observed, and thus different possible assignments have different probabilities associated to them. What is called the *complete* model likelihood is the likelihood taking Z into account as a variable. It writes

$$p(A, Z) = \prod_{i=1}^{n} \pi_{Z_i} \prod_{\substack{i=1 \\ j>i}}^{n} \gamma_{Z_i Z_j}^{A_{ij}} (1 - \gamma_{Z_i Z_j})^{1 - A_{ij}}. \tag{2.1}$$

Of course it cannot be calculated from an observation, since the Z are unknown.

Remark. Equation (2.1) is an example of the abuse of notation described in the remarks below Definition 1.

Observed model likelihood

In order to have a likelihood associated to an observation, one needs to marginalize the latent variables present in Equation (2.1). That is,

$$p(A) := \sum_{Z \in K^n} p(A, Z). \tag{2.2}$$

This sum over the whole latent space is intractable, since it has an exponential number of terms and it cannot be analytically evaluated to an useful simpler form. Therefore, it will be strictly necessary to approximate this observed likelihood in order to perform estimation and inference on SBMs from a probabilistic point of view.

2.2.2 Variational decomposition and mean field approximation

The likelihood of an observation under the SBM, in Equation (2.2), is complex to deal with for mainly two reasons. First, it is a sum over all latent configurations, and thus it has an exponential number of terms, making it intractable. Second, it can have multiple local optima. Therefore, approximations are needed in order to work with this model. A common one is the variational approximation to the likelihood, which will deal with the problem of summing an exponential number of terms. This is still complicated in all its generality, so a second "mean field" approximation is used on top of the first variational one. This consists in searching the solution to the variational approximation amidst factorizable distributions.

Deriving the variational decomposition.

Let Z denote the vector of latent variables, A an observation, and θ the parameters of an SBM. The following *variational decomposition* leads to an useful approximation to the likelihood.

Definition 3. For any two distributions $p(z), q(z)$,

$$\mathrm{KL}(q\|p) := -\int_Z \log\left(\frac{p(z)}{q(z)}\right) q(z)\, dz \tag{2.3}$$

is called the *Kullback-Leibler divergence* from $p(z)$ to $q(z)$.

Theorem 1. *The observed likelihood can be decomposed as*

$$\log p(A;\theta) = F(q,\theta) + \mathrm{KL}(q\|p(Z|A;\theta)), \tag{2.4}$$

where

$$F(q,\theta) := \int_Z \log\left(\frac{p\,(Z,A;\theta)}{q(Z)}\right) q(Z)\, dZ \tag{2.5}$$

is called the evidence lower bound, or ELBO for short.

For the proof, see Section A.1. Equation (2.4) forms the basis of the classical EM algorithm for estimating θ, where one performs alternate minimization of the KL term and subsequent maximization of the term $F\,(q,\theta)$. A classical result proves that the KL is always positive. Therefore, the ELBO is indeed a lower bound for the log likelihood being decomposed. Given the independence of the left hand side with respect to q, observe that maximizing the ELBO amounts to minimizing the KL term, and this can be done by setting $q = p\,(Z|A;\theta)$. However, in the case of the SBM, this conditional probability is itself intractable, therefore this step of EM must be performed differently.

The ELBO is also called the "free energy" in the literature.

Mean field approximation.

A common strategy used to deal with the problem of having an untractable solution q to the variational approximation is called the "mean field approximation". It consists in trying to find a q distribution maximizing the ELBO constrained to a family of tractable distributions. Here, tractability means factorizability: consider distributions of the form

$$q(Z) = \prod_{i=1}^{n} q_i(Z_i).$$

There are other "correlated" mean field approximations, see [8].

2.2.3 Variational estimation of the SBM

The ELBO in the SBM case

In the case of the SBM, each factor in the mean field approximation must be multinomial distribution, and they differ only by their parameters, that is,

$$q(Z) = \prod_{i=1}^{n} m(Z_i; \tau_i), \tag{2.6}$$

where $m(\cdot, \tau_i)$ is the probability mass function of a multinomial distribution with parameter τ_i. It is then possible to find an explicit form for the ELBO of an observation assuming the SBM model.

Proposition 1. *Given an adjacency matrix A and assuming as* $\mathrm{SBM}(n, \pi, \Gamma)$*, the mean-field ELBO is given by*

$$F_A(\tau, \pi, \Gamma) = \sum_{i=1}^{n}\sum_{k=1}^{K}\left[\tau_{ik}\log\frac{\pi_k}{\tau_{ik}} + \frac{1}{2}\sum_{j=1}^{n}\sum_{l=1}^{K}\tau_{ik}\tau_{jl}\left(A_{ij}\log\gamma_{kl} + (1-\delta_{ij}-A_{ij})\log(1-\gamma_{kl})\right)\right]. \tag{2.7}$$

For the proof, see Section **??**.

The variational EM algorithm

Traditionally, numerically maximizing the ELBO in models with latent variables is done via the EM algorithm. In this case, since there is the extra step of approximating the q distribution within the mean-field variational family, the algorithm is called the *variational*

EM. Its steps are

$$\tau_{t+1} \coloneqq \operatorname*{argmax}_{\tau} F(q_\tau, \theta_t) \qquad \text{(E step)}$$

$$\theta_{t+1} \coloneqq \operatorname*{argmax}_{\theta} \mathbb{E}_{q_{\tau_{t+1}}}[\log p(x, z; \theta_t)] \qquad \text{(M step)}.$$

One performs this iteratively until convergence or some stopping criterion.

It is possible to explicitly describe these steps. The E step can be calculated by solving the fixed point relation

$$\hat{\tau}_{ik} \propto \pi_k \prod_{\substack{j>i \\ l=1,\dots,K}} \left(\gamma_{kl}^{A_{ij}} \left(1-\gamma_{kl}\right)^{(1-A_{ij})}\right)^{\tau_{jl}}. \tag{2.8}$$

The M step can be calculated directly by

$$\hat{\gamma}_{kl} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} \tau_{ik}\tau_{jl}A_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{n} \tau_{ik}\tau_{jl}\left(1-\delta_{ij}\right)}. \tag{2.9}$$

For a proof of these statements, see Section **??**.

2.3 Optimization approach

2.3.1 The spectral point of view

The use of transformations has proven to be, time and again, very useful in mathematics. Classical examples are the Fourier transform and its uses in signal processing, and the Laplace transform with its uses in differential equations and probability. The idea is always to bring a dual point of view to the problem. For example, the Fourier transform allows one to analyze a signal processing problem in the time domain or in the frequency domain, according to which one is more convenient.

The same idea applies for graphs. On the one hand, a graph can be analyzed in the *topological* domain, and quantities based on the nodes' connectivity can be calculated. Examples of topological quantities are the diameter, the degree, the clustering coefficient, the girth, but there are many others. On the other hand, the graph can be represented by a symmetric adjacency matrix A that is completely characterized by its eigensystem. This follows from the spectral decomposition $A = X\Lambda X^t$.

There are reasons for which the spectral domain might present advantages for analyzing a graph. The topological quantities described are usually correlated and dependent, while the eigenvectors of A are orthogonal and its eigenvalues are independent quantities. The spectral domain representation of a graph lies in an Euclidean space, so if structures appear in it it is possible to use classical algorithms on such spaces. These tend to be simpler, faster, and more interpretable. This is precisely the strategy of spectral clustering.

2.3.2 Graph Laplacians

Defining graph Laplacians

Although the adjacency matrix completely describes the graph, other matrices (or "*graph operators*"), can be used for constructing spectral representations of G. One popular class of matrices used are the different graph Laplacians. The unnormalized and normalized Laplacians are of particular interest.

Definition 4. Let G be a simple unweighted graph. Denote by A its adjacency matrix. Define the *degree matrix* as the diagonal matrix D such that $D_{ii} \coloneqq \sum_{ij} A_{ij}$ for each $i \in \{1, \dots, n\}$. Define the unnormalized and normalized Laplacians respectively by

$$\begin{aligned} L_{\text{unn}} &\coloneqq D - A \\ L_{\text{sym}} &\coloneqq I - D^{-1/2} A D^{-1/2}. \end{aligned} \tag{2.10}$$

There are several different ways of motivating this popularity. In what follows, this is explained from a community detection perspective, in the intuitive case of assortative communities.

The Laplacian and connectivity

Assume a disconnected graph G having k connected components. A particular instance of this is when there are k assortative communities that are completely separated. The kernel of the Laplacian contains precisely the connectivity information of the graph, and in this degenerate case this coincides with the community information.

Proposition 2. *Let G be a simple unweighted graph with k connected components $\Omega_1, \dots, \Omega_k$. Then the algebraic multiplicity of the eigenvalue 0 of L_{unn} equals k and the indicator vectors $\mathbf{1}_{\Omega_1}, \dots, \mathbf{1}_{\Omega_k}$ span its null space.*

Remark. An analogous proposition holds for L_{sym}, with the sole difference being that it is the vectors $\{D^{1/2}\mathbf{1}_{\Omega_i}\}_{i=1,\dots,n}$ that span the null space of the Laplacian.

The Laplacian and graph cuts

Moving on from the degenerate case, consider now that the graph is perturbed and the k communities now communicate weakly, that is, that there are some edges across them. Then it makes sense to propose a method of finding the communities by seeking a partition which minimizes the number of edges crossing classes, while still keeping the partitions with a reasonable size to avoid degenerate solutions. Formalizing this yields the following definitions.

Definition 5. Let $G = (V, E)$ be a simple graph. The *cut* is a function associating any partition P of G to the number of edges connecting nodes belonging to different classes, i.e.,

$$\text{cut}(P) \coloneqq \frac{1}{2} \sum_{i=1}^{n} \sum_{j : P(i) \neq P(j)} A_{ij}. \tag{2.11}$$

The first step is to normalize this metric with respect to class sizes, to avoid taking unbalanced solutions. The *ratio cut* is a possible normalization of the cut. The results that follow will be associate it to the unnormalized Laplacian. Taking the alternative NCut normalization yields their analogous version for the symmetric normalized Laplacian. For the sake of simplicity, only the results using the ratio cut will be presented.

Definition 6. Let $\text{cut}(P)_{ij}$ denote the cut between classes i and j of partition P, assumed to have k classes. The ratio cut is defined as

$$\text{RatioCut}_k(P) \coloneqq \frac{1}{2} \sum_{i=1}^{k} |P_i|^{-1} \sum_{\substack{j=1 \\ j \neq i}}^{k} \text{cut}(P)_{ij}. \tag{2.12}$$

Definition 7. The *balanced min-cut problem* is the following optimization problem:

$$\min_{P \in \mathcal{P}_k(G)} \text{RatioCut}_k(P), \tag{2.13}$$

where $\mathcal{P}_k(G)$ denotes the set of all partitions of G into k classes.

This problem can be rewritten in terms of the Laplacian.

Proposition 3. *The balanced min-cut problem can be rewritten in terms of the Laplacian as*

$$\begin{aligned} &\min_{A_1, \ldots, A_k} \text{Tr}(H^t L H) \\ &\textit{s.t. } H^t H = I, \end{aligned} \tag{2.14}$$

where $H \in \mathbb{R}^{n \times k}$ is the matrix

$$h_{ij} \coloneqq \begin{cases} 1/\sqrt{|A_{ij}|} & \textit{if node } i \in A_j \\ 0 & \textit{otherwise.} \end{cases} \tag{2.15}$$

2.3.3 The spectral clustering algorithm

The problem in Proposition 3 is NP-hard due to its hard constraint (2.15) on the form of the matrix H. It is natural to drop this constraint to get a solvable approximation to the problem.

Definition 8. The *relaxed balanced min-cut problem* is defined by replacing the constraint 2.15 on the form of H by a more general orthogonality constraint:

$$\begin{aligned} &\min_{H \in \mathbb{R}^{n \times k}} \operatorname{Tr}(H^t L H) \\ &\text{s.t. } H^t H = I \end{aligned} \tag{2.16}$$

This kind of problem has a known solution, given by the Rayleigh-Ritz theorem.

Proposition 4 (Rayleigh-Ritz)**.** *Problem* (2.16) *is solved by the matrix H having the first k eigenvectors of L as columns.*

Originally, H's columns indicated the communities, by constraint (2.15). One might expect that this solution to the relaxed problem might still contain this information. Therefore the final step is to go from this approximate solution to community assignments. Originally, the fact that H's columns were indicatrices for the communities means that there were only k distinct rows in it. That is, by seeing the rows of H as vectors, there were only k distinct vectors. One might expect that, seeing the rows of the approximate solution H_{approx} as vectors, these vectors still "fluctuate" around the original k distinct vectors. If this is the case, clustering these vectors might yield the community information. This clustering can be done by k-means, for example. This procedure is what is commonly called the spectral clustering method. TO DO: Enhance this paragraph.

Remark. Spectral clustering *tries* to find a balanced minimum cut partition. It may succeed in applications, but theoretically there are simple counterexamples showing that the quality of this approximation can be arbitrarily bad, see [17]. Searching other algorithms can only partially help, as there is no general efficient algorithm for solving graph cut problems TO DO: cite.

Long story short, here are the steps of the spectral clustering algorithm. TO DO: Write the algorithm.

```
Step 1: Calculate the Laplacian L from the adjacency matrix A;
Step 2: Calculate the k first eigenvectors of L, make them columns
            of a matrix H;
Step 3: Cluster the rows of H with an algorithm such as k-means;
Step 4: Return cluster assignments as community assignments for
            the nodes;
```

The diagram below illustrates the intuition and logic steps behind the deduction of spectral clustering algorithms.

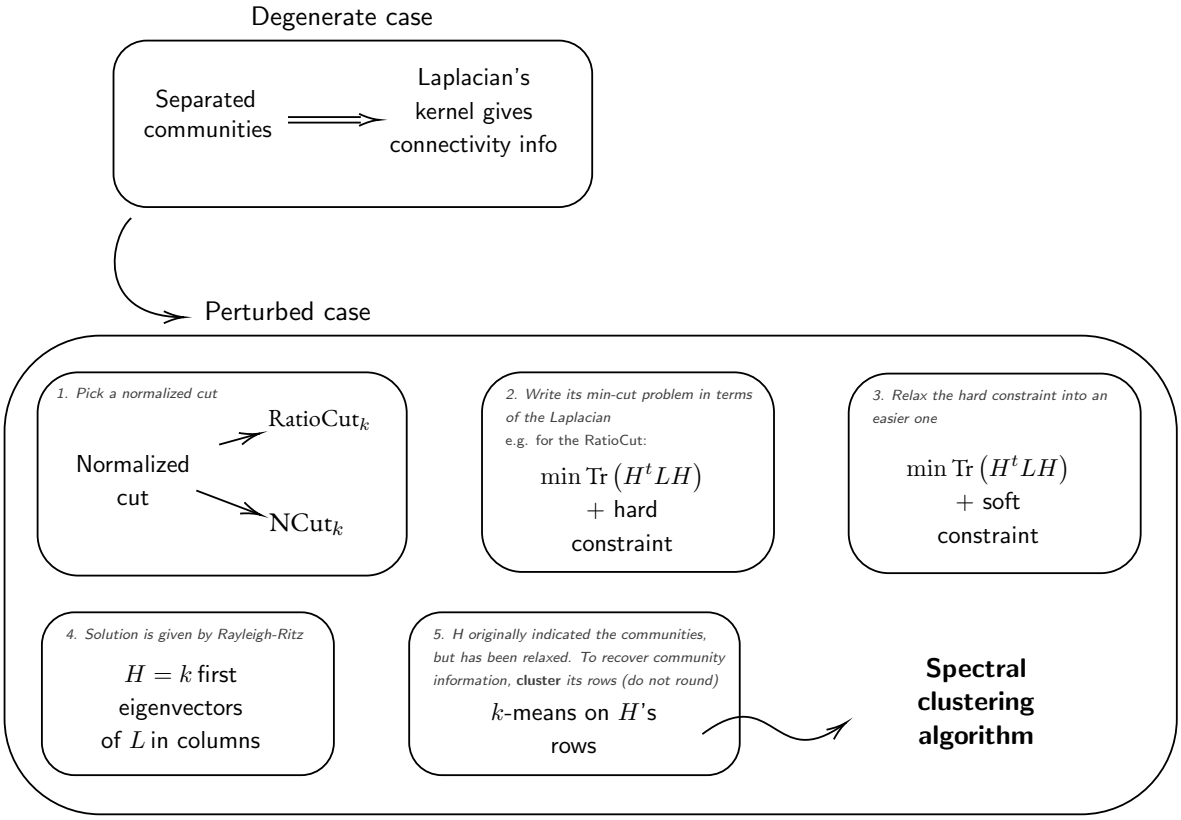


Figure 2.3: Diagram describing the steps taken to derive spectral algorithms for community detection.

Chapter 3

Learning theory

3.1 Agreement and degrees of recovery

3.1.1 Agreement

Consider the task of evaluating the answer returned by any algorithm for community detection. Assume knowledge of the true vector of community assignments $Z^\star$. Given an estimate $\hat{Z}$ for $Z^\star$ how can one measure the quality of such an estimation? The most intuitive metric for this is the *agreement*, which is simply the average number of nodes whose communities were estimated correctly, up to an arbitrary relabeling of the communities. This relabeling must be taken into account since the choice of the integer associated to a community is arbitrary.

Definition 9 (Agreement)**.** Let $Z^\star$ and $\hat{Z}$ be, respectively, the true and an arbitrary vector of community assignments. Let also S_k denote the group of permutations of $[k]$. Define the *agreement* between $Z^\star$ and $\hat{Z}$ to be

$$A(Z^\star, \hat{Z}) \coloneqq \frac{1}{n} \max_{\sigma \in S_k} \sum_{i=1}^{n} \mathbf{1}(Z_i^\star = \sigma \hat{Z}_i). \tag{3.1}$$

However, when studying weaker forms of recovery, under general (asymmetric) SBMs, a *normalized* version of this metric is actually needed.

Definition 10 (Normalized agreement)**.** Let $Z^\star$ and $\hat{Z}$ be, respectively, the true and an arbitrary vector of community assignments. Let also S_k denote the group of permutations of $[k]$. Define the *normalized agreement* between $Z^\star$ and $\hat{Z}$ to be

$$\tilde{A}(Z^\star, \hat{Z}) \coloneqq \frac{1}{k} \max_{\sigma \in S_k} \sum_{k=1}^{K} \frac{\sum_{i=1}^{n} \mathbf{1}(Z_i^\star = \sigma \hat{Z}_i) \mathbf{1}(Z_i^\star = k)}{\sum_{i=1}^{n} \mathbf{1}(Z_i^\star = k)}. \tag{3.2}$$

3.1.2 Degrees of recovery

Definition 9 can be used to measure different degrees of performance of algorithms whose task is to estimate the communities $Z^\star$. The degree to which an algorithm is capable of recovering the communities is captured in what are called the different (asymptotic) *degrees of recovery*. Notice these are all defined asymptotically.

Definition 11 (Degrees of recovery)**.** Let $(Z^\star, G) \sim \text{SBM}(n, \pi^\star, \Gamma^\star)$, and $\hat{Z}$ be the output of an algorithm taking $(G, \pi^\star, \Gamma^\star)$ as input. Then, the following *degrees of recovery* are said to be *solved* if, asymptotically on n, one has

- Exact recovery $\leftrightarrow \mathbb{P}(A(Z^\star, \hat{Z}) = 1) = 1 - o(1)$,
- Almost exact recovery $\leftrightarrow \mathbb{P}(A(Z^\star, \hat{Z}) = 1 - o(1)) = 1 - o(1)$,
- Partial recovery $\leftrightarrow \mathbb{P}(\tilde{A}(Z^\star, \hat{Z}) \geq \alpha) = 1 - o(1), \alpha \in (1/k, 1)$,
- Weak recovery (also called detection) $\leftrightarrow \mathbb{P}(\tilde{A}(Z^\star, \hat{Z}) \geq 1/k + \Omega(1)) = 1 - o(1)$.

Remark. There is an intuition for why $\alpha > 1/k$ in the definitions of partial and weak recovery above. If one assumes knowledge of the true parameters $(\pi^\star, \Gamma^\star)$ of the model when designing an estimation algorithm, then the trivial algorithm of simply assigning each node a random community according to $\pi^\star$ will achieve an agreement of $\|\pi\|_2^2$, by the law of large numbers. In particular, in the case where the communities are uniform, $\pi = \mathbf{1}_k/k$ and the trivial agreement reached will be $A = 1/k$.

3.2 Asymptotic topologies

3.2.1 The case of the Erdős-Rényi model

The first random graph model was the Erdős-Rényi model, denoted $G(n, p)$, over graphs with n vertices [3]. Under this model, the presence of an edge between each pair of nodes is determined by a Bernoulli random variable of parameter p. See Figure 3.1.

Although this model does not present clusters of nodes, it is nevertheless of great interest, since it reveals a key phenomenon: there exist tightly defined and distinct "asymptotic topologies" for random graphs arising from this model as $n \to \infty$. Which one arises is a function of the growth of p with respect to n.

Theorem 2. *content...*

In the SBM, essentially the same phenomenon happens, and this affects directly the performance of algorithms, and has led to ideas of how to increase their robustness.

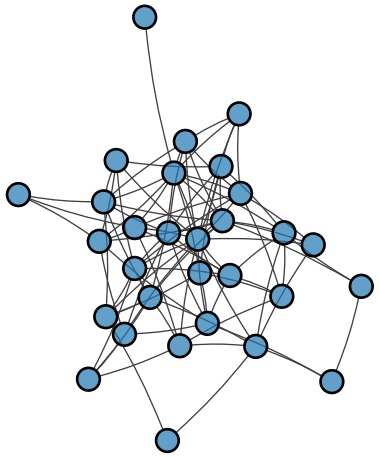
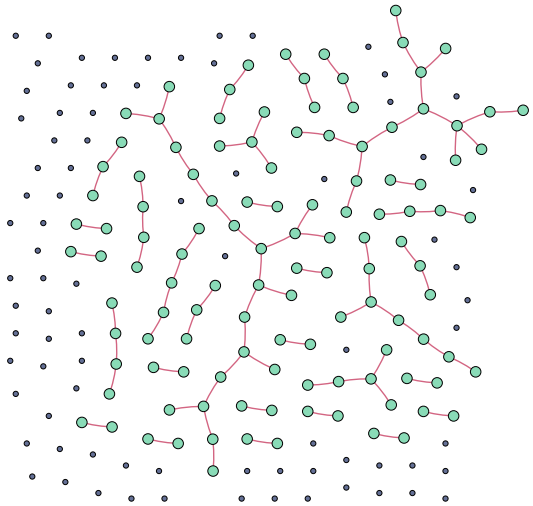


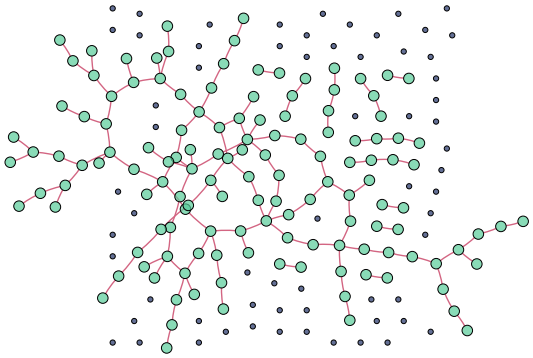
Figure 3.1: An observation from an Erdős-Rényi model $G(30, 0.2)$

3.2.2 The case of the SBM

3.2.3 Why does this matter?



(a)



(b)

Figure 3.2: (a) At $np = 0.8 < 1$, there are some small trees of size at most $O(\log(n))$. (b) At $np = 1.33 > 1$ a giant component appears, of size $O(n^{2/3})$.

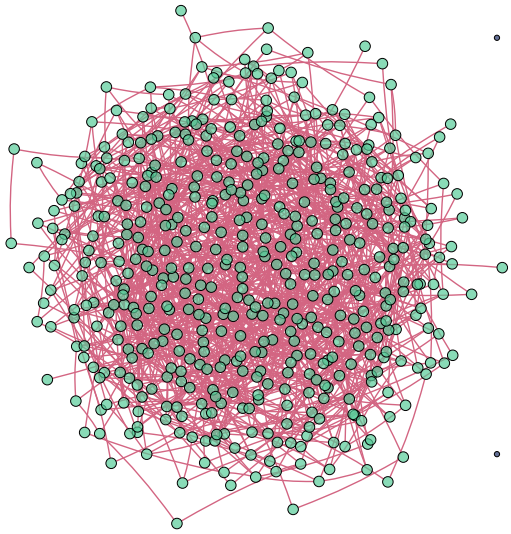


Figure 3.3: At $p = 0.011 < 0.012$ there exists almost surely an isolated vertex, and the graph is disconnected. When $p = 0.013 > 0.012$, isolated vertices disappear almost surely, and the graph finally becomes connected.

3.3 Consistency of variational EM

Your text.

3.4 Consistency of spectral clustering

3.4.1 Notations, conventions, definitions.

Motivation

When studying the relationship between SBMs and spectral clustering algorithms, it is natural to ask when is the spectral clustering algorithm capable of recovering the community partitions of graphs generated under some SBM. The seminal work of [14] was among the first to study this question. Figure 3.4 is a schematic diagram linking the results presented in it. Lemma 3.1. will be of particular interest for connections with probabilistic approaches.

General latent space models

It is important to notice that all the results concerning the first objective in Figure 3.4 are valid for latent space models, which is a class of models more general than the SBM.

Definition 12 (Latent space model)**.** For i.i.d. random vectors $z_1, \dots, z_n \in \mathbb{R}^k$ and random adjacency matrix $A \in \{0,1\}^{n \times n}$, let $\mathbb{P}(A_{ij}|z_i, z_j)$ be the probability mass

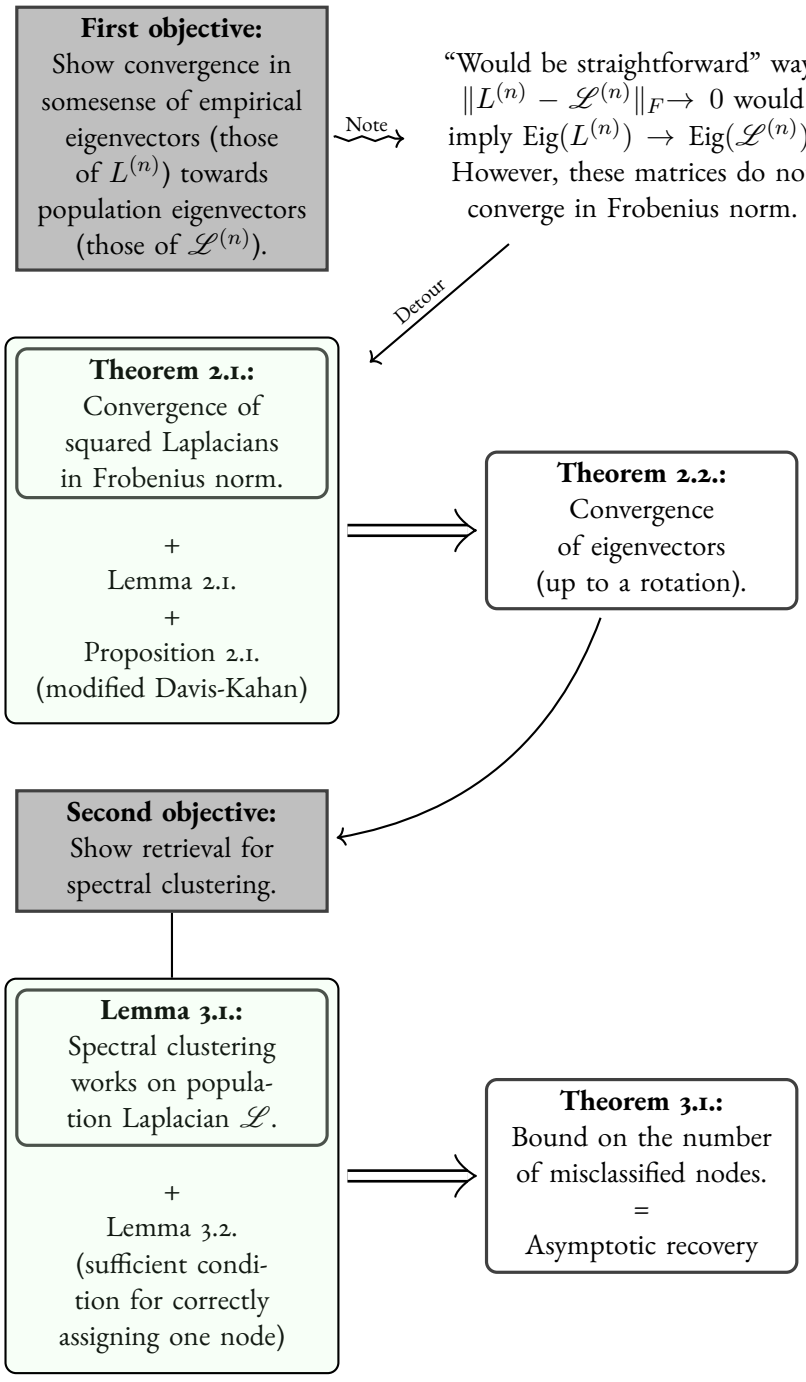


Figure 3.4: Diagram of results in [14].

function of A_{ij} conditioned on z_i, z_j. If a probability distribution on A has the conditional dependence relationships

$$\mathbb{P}(A|z_1, \dots, z_n) = \prod_{i<j} \mathbb{P}(A_{ij}|z_i, z_j),$$

and $\mathbb{P}(A_{ii} = 0) = 1$ for all i, then it is called an *undirected latent space model*.

Choice of Laplacian

They use the matrix $L = D^{-1/2}AD^{-1/2}$ as Laplacian. This is justified, since this matrix has the same eigenvectors as the more common normalized Laplacian $\tilde{L} = I - L$, and the eigenvectors are the only thing that matters in the spectral clustering algorithm. However, due care should be taken when translating their results to those obtained when using the unnormalized Laplacian. The population adjacency matrix is defined as $\mathscr{A} \coloneqq \mathbb{E}[A|Z^\star]$, and the population degree matrix is the diagonal matrix with diagonal entries $\mathscr{D}_{ii} = \sum_k \mathscr{A}_{ik}$. This allows the definition of the population Laplacian $\mathscr{L} = \mathscr{D}^{-1/2}\mathscr{A}\mathscr{D}^{-1/2}$. Their work consists of two parts.

3.4.2 First part: convergence of eigenvectors

The first part consists in showing that the eigenvectors of the empirical Laplacian converge in some sense to the eigenvectors of the population Laplacian. This would be immediate if these matrices converged in Frobenius norm, i.e., if $\|L - \mathscr{L}\|_F \to 0$. However, they do not converge in such a norm, and so a "detour" needs to be made in order to achieve this result. They show instead that the *squared* version of these matrices converge in Frobenius norm, and then show directly that this implies that *up to a rotation* the eigenvectors of L converge to those of $\mathscr{L}$.

3.4.3 Second part: retrieval for spectral clustering

The results of this part focus on the special case of an SBM, and shows the asymptotic consistency when using the spectral clustering algorithm to estimate the community assignment clatent variables $Z^\star$. It starts with the following lemma, which shows that applying the algorithm to the expected Laplacian given the assignments recovers precisely the partitions of the SBM. Notice that this fact is non-asymptotic.

Lemma 1. *Consider the Stochastic Blockmodel with k blocks,*

$$\mathscr{A} = Z\Gamma Z^t \in \mathbb{R}^{n\times n} \text{ for } \Gamma \in \mathbb{R}^{k\times k} \text{ and } Z \in \{0,1\}^{n\times k},$$

and let $\mathscr{L}$ be the expected Laplacian given the true assignments $Z^\star$. Then, there exists a matrix $\mu \in \mathbb{R}^{k\times k}$ such that the eigenvectors of $\mathscr{L}$ corresponding to the nonzero eigenvalues are the columns of the $Z\mu$. Furthermore,

$$z_i\mu = z_j\mu \iff z_i = z_j, \tag{3.3}$$

where z_i is the i-th row of Z.

Remark. The equivalence in Equation 3.3 means that rows i and j of $Z\mu$ are equal if, and only if, the corresponding rows of Z are equal, that is, if nodes i and j belong to the same community. Since there are k communities, this implies that there can be at most k unique rows in the matrix $Z\mu$ of eigenvectors of $\mathscr{L}$. Spectral clustering applies k-means to these vectors, and thus these become precisely the centroids of k-means (since one is applying k-means to at most k different vectors). The rows of $Z\mu$ will then obviously be attributed to the centroid they are equal to, and by the equivalence in Equation 3.3, this implies that spectral clustering perfectly identifies the clusters in the expected Laplacian $\mathscr{L}$.

Proof. Here is a sketch of the proof, which is quite algebraic and not much motivated.

1. Factor $\mathscr{L}$ as $\mathscr{L} = Z\Gamma_L Z^t$ for some matrix $\Gamma_L \in \mathbb{R}^{k\times k}$.
2. Consider now the (different) matrix $(Z^tZ)^{1/2}\Gamma_L(Z^tZ)^{1/2}$: this is the decomposition given for $\mathscr{L}$ under the change $Z \to (Z^tZ)^{1/2}$, which can be thought of as a "square matrix version" of Z.
3. Show that $(Z^tZ)^{1/2}\Gamma_L(Z^tZ)^{1/2}$ is symmetric and positive-definite, implying the spectral decomposition $(Z^tZ)^{1/2}\Gamma_L(Z^tZ)^{1/2} = V\Lambda V^t$.
4. Multiply the spectral decomposition on both sides by $(Z^tZ)^{-1/2}Z^t$, revealing that
$$Z\Gamma_L Z^t = \mathscr{L} = (Z\mu)\Lambda(Z\mu)^t \tag{3.4}$$
for $\mu \coloneqq (Z^tZ)^{-1/2}V$.
5. Together with the fact that $(Z\mu)^t(Z\mu) = I_k$, where I_k is the $k \times k$ identity, Equation 3.4 is precisely the eigenvector equation for $\mathscr{L}$. This shows that the columns of $Z\mu$ are the eigenvectors of $\mathscr{L}$ associated to the non-zero eigenvalues.
6. Finally, the equivalence is a direct consequence of the fact that μ is invertible:
$$\det(\mu) = \det((Z^tZ)^{-1/2})\det(V) > 0.$$

□

Chapter 4

Numerical experiments

4.1 Numerical experiments

Chapter 5

Conclusion and outlook

5.1 Conclusions

Your text.

Appendix A

Proofs and calculations

A.1 Proof of Theorem 1

The posterior is defined as

$$p\left(Z|A;\theta\right)=\frac{p\left(Z,A;\theta\right)}{p\left(A;\theta\right)},$$

which implies

$$p\left(A;\theta\right)=\frac{p\left(Z,A;\theta\right)}{p\left(Z|A;\theta\right)}.$$

Thus,

$$\log p\left(A;\theta\right)=\log p\left(Z,A;\theta\right)-\log p\left(Z|A;\theta\right).$$

Taking the expectation of this expression with respect to some proposal distribution $q(Z)$ depending only on Z, one has

$$\begin{aligned}\log p\left(A;\theta\right) &= \int_{Z}\log p\left(Z,A;\theta\right)q\left(Z\right)\,dZ-\int_{Z}\log p\left(Z|A;\theta\right)q\left(Z\right)\,dZ\\ &= \int_{Z}\left(\log\left(\frac{p\left(Z,A;\theta\right)}{q\left(Z\right)}\right)-\log q\left(Z\right)\right)q\left(Z\right)\,dZ\\ &\quad -\int_{Z}\left(\log\left(\frac{p\left(Z|A;\theta\right)}{q\left(Z\right)}\right)-\log q\left(Z\right)\right)q\left(Z\right)\,dZ.\end{aligned}$$

Therefore,

$$\log p\left(A;\theta\right)=\int_{Z}\log\left(\frac{p\left(Z,A;\theta\right)}{q(Z)}\right)q(Z)\,dZ-\int_{Z}\log\left(\frac{p\left(Z|A;\theta\right)}{q(Z)}\right)q(z)\,dZ,$$

that is,

$$\log p\left(A;\theta\right)=F\left(q,\theta\right)+\mathrm{KL}\left(q\|p\left(Z|A;\theta\right)\right).$$

A.2 Convergence properties

Section 2.2.3 allows one to estimate the communities in any graph assuming that its underlying generative process is an SBM. Of course, the true generative process and the true communities in the graph are unknown, so there is no benchmark for the results obtained. Now assume the observation graphs whose generative process *does* come from an SBM model. What is the behavior of the algorithm proposed?

A.2.1 Convergence to a local maximum

Showing that the algorithm increases the likelihood at each step (and therefore achieves some local maximum) is very similar to the case of the classical EM.

Proposition 5. *The VEM algorithm increases the likelihood at each step.*

Proof. For any fixed θ_0, the variational decomposition is

$$\log p(x;\theta_0) = F(q_\tau,\theta_0) + \mathrm{KL}(q_\tau \| p(\cdot|x;\theta_0)). \tag{A.1}$$

Observe that in general the KL term can no longer be zero, but it can be minimized, leading to the best approximation to the posterior within the mean-field family. Minimizing the KL still is equivalent to maximizing the ELBO, thus the new *variational* E step consists in finding τ, fixed θ_0:

$$\tau_0 = \operatorname*{argmax}_\tau F(q_\tau,\theta_0). \tag{A.2}$$

Now fix τ_0 and consider a general θ in the variational decomposition:

$$\log p(x;\theta) = F(q_{\tau_0},\theta) + \mathrm{KL}(q_{\tau_0} \| p(\cdot|x;\theta)).$$

If you take $\theta_1 \coloneqq \operatorname{argmax}_\theta F(q_{\tau_0},\theta)$, then

$$\begin{aligned}\log p(x;\theta_1) &= F(q_{\tau_0},\theta_1) + \mathrm{KL}(q_{\tau_0} \| p(\cdot|x;\theta_1)) \\ &\geq F(q_{\tau_0},\theta_0) + \mathrm{KL}(q_{\tau_0} \| p(\cdot|x;\theta_0)) \\ &= \log p(x;\theta_0).\end{aligned}$$

That is, this choice of a next θ makes the observed log-likelihood grow. This maximization is the variational analogue of the M step. Notice that the observed log-likelihood does not depend on the q_τ chosen, thus the M-step keeps the observed log-likelihood constant, while the E-step makes it grow, and so overall it must grow after each EM alternation. Notice that the ELBO itself grows in both steps. □

However, there can be multiple uninformative local maxima, and it is known [15] that in closely related algorithms these bad optima can attract almost all initializations for the parameters. Consider then the question whether this algorithm converges to the *true* global maximum as $n \to \infty$, that is, the question concerning its asymptotic consistency.

A.2.2 Asymptotic consistency

TO DO: Write here about Bickel, Christine, cite the results without proof, comment on how they depend on the growing regime of the connectivities.

Appendix B

The case of two communities

B.1 The case of two communities*

B.1.1 Rewriting the ELBO

Suppose now that $k = 2$. Then, for any node of index i, $\tau_{i2} = 1 - \tau_{i1}$. Thus one can work only with the first component τ_{i1}, which will be denoted from now on by τ_i. A similar logic applies to π since $\pi_2 = 1 - \pi_1$, thus subsequently work only with π_1 which will be denoted simply π. The ELBO from 2.7 writes

$$\begin{aligned}\mathcal{L}_{(k=2)} = \sum_{i=1}^{n} \tau_i \log \frac{\pi}{\tau_i} + (1-\tau_i) \log\left(\frac{1-\pi}{1-\tau_i}\right) + \frac{1}{2}\sum_{j\neq i} A_{ij} \left(\tau_i\tau_j \log \gamma_{11}\right. \\ + \tau_i(1-\tau_j)\log\gamma_{12} + (1-\tau_i)\tau_j \log\gamma_{21} + (1-\tau_i)(1-\tau_j)\log\gamma_{22}) \\ + \frac{1}{2}\sum_{j\neq i}(1-A_{ij})\left(\tau_i\tau_j\log(1-\gamma_{11}) + \tau_i(1-\tau_j)\log(1-\gamma_{12})\right. \\ \left. + (1-\tau_i)\tau_j\log(1-\gamma_{21}) + (1-\tau_i)(1-\tau_j)\log(1-\gamma_{22})\right).\end{aligned}$$

This expression can be grouped differently:

$$\begin{aligned}\mathcal{L}_{(k=2)} = \sum_{i=1}^{n} \tau_i \log \frac{\pi}{\tau_i} + (1-\tau_i)\log\left(\frac{1-\pi}{1-\tau_i}\right) \\ + \frac{1}{2}\sum_{j\neq i}\tau_i\tau_j\left(A_{ij}\log\gamma_{11} + (1-A_{ij})\log(1-\gamma_{11})\right) \\ + \sum_{j\neq i}\tau_i(1-\tau_j)\left(A_{ij}\log\gamma_{12} + (1-A_{ij})\log(1-\gamma_{12})\right) \\ + \frac{1}{2}\sum_{j\neq i}(1-\tau_i)(1-\tau_j)\left(A_{ij}\log\gamma_{22} + (1-A_{ij})\log(1-\gamma_{22})\right).\end{aligned}$$

To make things simpler, write this in matrix notation. Let $\mathbf{1}_n \coloneqq (1, \dots, 1)$ be a vector of dimension n, I_n be the identity matrix of size $n \times n$, and $J \coloneqq \mathbf{1}_n \mathbf{1}_n^t - I_n$ be the matrix with zeros on the diagonal and ones everywhere else. The previous expression for the ELBO becomes

$$\begin{aligned}\mathcal{L}_{(k=2)} = &\sum_{i=1}^{n} \tau_i \log \frac{\pi}{\tau_i} + (1-\tau_i) \log\left(\frac{1-\pi}{1-\tau_i}\right) \\ &+ \frac{1}{2}\left(\tau^t A \tau \log \gamma_{11} + \tau^t (J-A)\tau \log(1-\gamma_{11})\right) \\ &+ \left(\tau^t A (\mathbf{1}_n - \tau) \log \gamma_{12} + \tau^t (J-A)(\mathbf{1}_n - \tau) \log(1-\gamma_{12})\right) \\ &+ \frac{1}{2}\left((\mathbf{1}_n - \tau)^t A (\mathbf{1}_n - \tau) \log \gamma_{22}\right. \\ &\qquad \left. + (\mathbf{1}_n - \tau)^t (J-A)(\mathbf{1}_n - \tau) \log(1-\gamma_{22})\right).\end{aligned} \tag{B.1}$$

B.1.2 The Φ function

When one observes a graph from an SBM, it is typically not the case that any parameter of the model is known, thus it is natural to consider the function $\Phi(\tau) \coloneqq \sup_{\pi,\gamma} \mathcal{L}(\tau; \pi, \gamma)$. It is then natural to consider its empirical version by substituting the parameters by their estimators found in equations **??** and **??**. Notice that in the binary case, the expression in **??** becomes

$$\hat{\pi} = \frac{\mathbf{1}_n^t \tau}{n}, \tag{B.2}$$

while the expression in **??** becomes

$$\hat{\gamma}_{11} = \frac{\tau^t A \tau}{\tau^t J \tau}, \tag{B.3}$$

$$\hat{\gamma}_{12} = \hat{\gamma}_{21} = \frac{(\mathbf{1}_n - \tau)^t A \tau}{(\mathbf{1}_n - \tau)^t J \tau}, \tag{B.4}$$

$$\hat{\gamma}_{22} = \frac{(\mathbf{1}_n - \tau)^t A (\mathbf{1}_n - \tau)}{(\mathbf{1}_n - \tau)^t J (\mathbf{1}_n - \tau)}. \tag{B.5}$$

Notice also that the equation $\hat{\gamma}_{21} = \hat{\gamma}_{12}$ comes from the symmetry of A and J. Substituting these estimators in the expression for the ELBO in order to find an expression for

$\hat{\Phi}(\tau)$, one obtains the following equation after some straightforward calculations:

$$\begin{aligned}\hat{\Phi}(\tau) = &\sum_{i=1}^{n} H(\tau_i) - nH\left(\frac{\mathbf{1}_n^t \tau}{n}\right) \\ &- \frac{\tau^t J \tau}{2} H\left(\frac{\tau^t A \tau}{\tau^t J \tau}\right) \\ &- \tau^t J(\mathbf{1}_n - \tau) H\left(\frac{\tau^t A(\mathbf{1}_n - \tau)}{\tau^t J(\mathbf{1}_n - \tau)}\right) \\ &- \frac{(\mathbf{1}_n - \tau)^t J(\mathbf{1}_n - \tau)}{2} H\left(\frac{(\mathbf{1}_n - \tau)^t A(\mathbf{1}_n - \tau)}{(\mathbf{1}_n - \tau)^t J(\mathbf{1}_n - \tau)}\right),\end{aligned} \tag{B.6}$$

where $H(x) \coloneqq -x \log x - (1-x)\log(1-x)$ is the entropy of a Bernoulli. Notice that for τ lying on the corners of the hypercube this corresponds to a "sure" assignment of the nodes of the graph to one of the two communities. Of course, trying to optimize such a function on the corners of the cube is NP-hard. To simplify this expression, introduce the notations

$$E_\tau \coloneqq \frac{\tau^t A \tau}{2} \qquad E_{\bar{\tau}} \coloneqq \frac{\bar{\tau}^t A \bar{\tau}}{2} \qquad E_M \coloneqq \tau^t A \bar{\tau} \tag{B.7}$$

$$C_\tau \coloneqq \frac{\tau^t J \tau}{2} \qquad C_{\bar{\tau}} \coloneqq \frac{\bar{\tau}^t J \bar{\tau}}{2} \qquad C_M \coloneqq \tau^t J \bar{\tau}. \tag{B.8}$$

This is motivated by the fact that in the particular case of τ lying on the vertices of the cube these quantities equal the number of edges in the community determined by τ (the nodes i such that $\tau_i = 1$) and the number of edges that there would be in the complete graph determined by these same nodes (likewise for $\bar{\tau} \coloneqq \mathbf{1}_n - \tau$). The edges and would-be edges between different communities are included in this notation by dropping the $1/2$ factor. A last (simple) piece of notation is $n_\tau \coloneqq \mathbf{1}_n^t \tau$, the number of nodes in the community of nodes with $\tau_i = 1$. Using these notations, the objective in B.6 can be more elegantly expressed as

$$\begin{aligned}-\frac{\hat{\Phi}(\tau)}{C} = &-\frac{1}{C}\sum_{i=1}^{n} H(\tau_i) + \frac{n}{C} H\left(\frac{n_\tau}{n}\right) \\ &+ \frac{C_\tau}{C} H\left(\frac{E_\tau}{C_\tau}\right) + \frac{C_M}{C} H\left(\frac{E_M}{C_M}\right) + \frac{C_{\bar{\tau}}}{C} H\left(\frac{E_{\bar{\tau}}}{C_{\bar{\tau}}}\right).\end{aligned} \tag{B.9}$$

The minus sign is introduced so that this becomes an objective function to minimize, as is standard in optimization. The division by $C \coloneqq \binom{n}{2}$ is for normalization purposes. This is a non-convex function on τ, which complicates its optimization.

B.1.3 Expected ELBO as objective function

Direct maximization of the $\hat{\Phi}$ function can be challenging, as it is not convex, and its solution might be in the interior of the hypercube. However, it is reasonable to expect that the maximum of the ELBO $\mathcal{L}$ should converge to the maximum of the expected ELBO $\mathbb{E}[\mathcal{L}|Z]$, where the expectation is taken with respect to the randomness of A and assuming knowledge of the model parameters and Z. The expected ELBO should be simpler to treat, and intuitively its maximum should be the assignments Z. Numerical simulation supports this intuition.

Thus, proceed in two steps. First, show that the maximum of the expected ELBO is indeed the vector of assignments Z. This will be done first on the simple case of the SBM with two communities. Then, properly show the convergence of the maxima of the ELBO to Z.

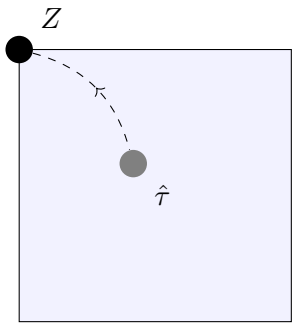


Figure B.1: The maximum $\hat{\tau}$ of the ELBO should converge to the maximum of the expected ELBO, which intuitively should be Z.

Starting from equation B.1, by linearity it suffices to substitute A by $\mathscr{A} \coloneqq \mathbb{E}[A|Z]$. This matrix has a simple structure. If $J_k \coloneqq \mathbf{1}_k^t \mathbf{1}_k - I_k$, then

$$\mathscr{A} = \begin{pmatrix} \gamma_{11} J_{n_1} & \gamma_{12} \mathbf{1}_{n_1 \times n_2} \\ \gamma_{12} \mathbf{1}_{n_2 \times n_1} & \gamma_{22} J_{n_2} \end{pmatrix}.$$

Notice also that $\bar{A} \coloneqq J - A$ becomes

$$J - \mathscr{A} = \begin{pmatrix} (1-\gamma_{11}) J_{n_1} & (1-\gamma_{12}) \mathbf{1}_{n_1 \times n_2} \\ (1-\gamma_{12}) \mathbf{1}_{n_2 \times n_1} & (1-\gamma_{22}) J_{n_2} \end{pmatrix},$$

which is the expected adjacency matrix of complementary graphs to the graphs originating from the model. This matrix will be denoted $\bar{\mathscr{A}} \coloneqq J - \mathscr{A}$. In order to organize the calculations, break the ELBO B.1 in two parts:

$$\mathcal{L}_{(k=2)} = \mathcal{L}^{\log}_{(k=2)} + \mathcal{L}^{\text{sym}}_{(k=2)}, \tag{B.10}$$

where

$$\begin{aligned}\mathcal{L}^{\text{sym}}_{(k=2)} \coloneqq & \frac{1}{2}\left(\tau^t A\tau \log\gamma_{11} + \tau^t \bar{A}\tau \log\left(1-\gamma_{11}\right)\right) \\ & + \left(\tau^t A\left(\mathbf{1}_n - \tau\right)\log\gamma_{12} + \tau^t \bar{A}\left(\mathbf{1}_n-\tau\right)\log\left(1-\gamma_{12}\right)\right) \\ & + \frac{1}{2}\left(\left(\mathbf{1}_n-\tau\right)^t A\left(\mathbf{1}_n-\tau\right)\log\gamma_{22} + \left(\mathbf{1}_n-\tau\right)^t \bar{A}\left(\mathbf{1}_n-\tau\right)\log\left(1-\gamma_{22}\right)\right),\end{aligned}$$

and $L^{\log}_{(k=2)}$ are the remaining non-random (in A) terms. Denote $\tau = (\tau_1, \tau_2)$, where $\tau_1 \in [0,1]^{n_1}$, $\tau_2 \in [0,1]^{n_2}$. Finally, denote $H(a,b) \coloneqq a\log b + (1-a)\log 1 - b$. The symmetric term above expands to

Symmetric term.

$$\begin{aligned}\mathcal{L}^{\text{sym}}_{(K=2)} = & \frac{1}{2}\left[H(\gamma_{11},\gamma_{11})\tau_1 J_{n_1}\tau_1 + H(\gamma_{22},\gamma_{11})\tau_2 J_{n_2}\tau_2\right. \\ & \qquad \left. + 2H(\gamma_{12},\gamma_{11})\tau_1 \mathbf{1}_{n_1\times n_2}\tau_2\right] \\ & + \left[H(\gamma_{11},\gamma_{12})\tau_1 J_{n_1}\bar{\tau}_1 + H(\gamma_{12},\gamma_{12})\tau_1\mathbf{1}_{n_1\times n_2}\bar{\tau}_2\right. \\ & \qquad \left. + H(\gamma_{12},\gamma_{12})\tau_2\mathbf{1}_{n_2\times n_1}\bar{\tau}_1 + H(\gamma_{22},\gamma_{12})\right] \\ & + \frac{1}{2}\left[H(\gamma_{11},\gamma_{22})\bar{\tau}_1 J_{n_1}\bar{\tau}_1 + H(\gamma_{22},\gamma_{22})\bar{\tau}_2 J_{n_2}\bar{\tau}_2\right. \\ & \qquad \left. + 2H(\gamma_{12},\gamma_{22})\bar{\tau}_1\mathbf{1}_{n_1\times n_2}\bar{\tau}_2\right].\end{aligned}$$

Now, it holds that $H(a,b) \leq H(a,a)$, implying (after straightforward simplification) that

TO DO: Put this inequality as a proposition !

$$\mathcal{L}^{\text{sym}}_{(k=2)} \leq C_{n_1} H(\gamma_{11}) + C_{n_2} H(\gamma_{22}) + n_1 n_2 H(\gamma_{12}),$$

using the notation from B.7. Notice the right-hand side is constant in τ. Finally, one can check that substituting $\tau^\star = (\mathbf{1}_{n_1}, 0_{n_2})$ achieves this upper bound. Thus it maximizes the symmetric part of the ELBO and corresponds to Z, the true community labels. As for the other term, notice that at $\tau^\star$ it becomes

Logarithmic term.

$$\mathcal{L}^{\log}_{(k=2)}(\tau^\star) = n_1 \log\pi + n_2\log\left(1-\pi\right).$$

This is not the maximum that such term can reach, since if one takes $\tau = \pi\mathbf{1}_n$ then this nonpositive term vanishes. However, notice that in order to analyze the ELBO asymptotically, proper normalization is required. Dividing the ELBO by $C = \binom{n}{2}$, this logarithmic term vanishes (since it grows linearly on the size of the communities).

Therefore, assigning the true labels $\tau = Z$ *asymptotically* maximizes the ELBO *in expectation*.

Important conclusion.

Bibliography

[1] Emmanuel Abbe. Community detection and stochastic block models: recent developments. *arXiv:1703.10146 [cs, math, stat]*, March 2017. arXiv: 1703.10146.

[2] Shaofeng Deng, Shuyang Ling, and Thomas Strohmer. Strong Consistency, Graph Laplacians, and the Stochastic Block Model. Technical Report arXiv:2004.09780, arXiv, April 2020. arXiv:2004.09780 [cs, stat] type: article.

[3] P. Erdös and A. Rényi. On random graphs i. *Publicationes Mathematicae Debrecen*, 6:290, 1959.

[4] Pierre Latouche, Etienne Birmelé, and Christophe Ambroise. Overlapping stochastic block models with application to the French political blogosphere, April 2011. arXiv:0910.2098 [stat].

[5] Sirio Legramanti, Tommaso Rigon, Daniele Durante, and David B. Dunson. Extended Stochastic Block Models with Application to Criminal Networks, April 2022. arXiv:2007.08569 [stat].

[6] Jure Leskovec. Stanford large network dataset collection. https://snap.stanford.edu/data/#communities. Accessed: 2022-08-25.

[7] Mahendra Mariadassou, Stéphane Robin, and Corinne Vacher. Uncovering latent structure in valued graphs: A variational approach. *The Annals of Applied Statistics*, 4(2):715–742, June 2010. Publisher: Institute of Mathematical Statistics.

[8] Marc Mezard and Andrea Montanari. *Information, physics, and computation*. OUP Oxford, 2009.

[9] Vincent Miele and Catherine Matias. Revealing the hidden structure of dynamic ecological networks. *Royal Society Open Science*, 4(6):170251, June 2017.

[10] Leonardo Morelli, Valentina Giansanti, and Davide Cittaro. Nested Stochastic Block Models applied to the analysis of single cell data. *BMC Bioinformatics*, 22(1):576, November 2021.

[11] Mark Newman. Network data. http://www-personal.umich.edu/~mejn/netdata/. Accessed: 2022-08-25.

[12] Tiago Peixoto. Netzschleuder. https://networks.skewed.de/. Accessed: 2022-08-25.

[13] Roldan Pozo. Complex network resources. https://math.nist.gov/~RPozo/complex_datasets.html. Accessed: 2022-08-25.

[14] Karl Rohe, Sourav Chatterjee, and Bin Yu. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, 39(4), August 2011. arXiv: 1007.1684.

[15] Purnamrita Sarkar, Y. X. Rachel Wang, and Soumendu S. Mukherjee. When random initializations help: a study of variational inference for community detection. *Journal of Machine Learning Research*, 22(22):1–46, 2021.

[16] A.W. van der Vaart. *Asymptotic Statistics*. Asymptotic Statistics. Cambridge University Press, 2000.

[17] Ulrike von Luxburg. A Tutorial on Spectral Clustering. *arXiv:0711.0189 [cs]*, November 2007. arXiv: 0711.0189.