Graph clustering and semi-supervised learning of non-binary, temporal and geometric networks

Maximilien Dreveton

UNIVERSITÉ CÔTE D'AZUR

ÉCOLE DOCTORALE
SCIENCES ET TECHNOLOGIES DE L'INFORMATION ET DE LA COMMUNICATION

THÈSE DE DOCTORAT

Détection de communautés et apprentissage semi-supervisé dans des réseaux non-binaires, temporels et géométriques

Maximilien DREVETON

Inria Sophia-Antipolis

Présentée en vue de l'obtention du grade de docteur en informatique
de l'Université Côte d'Azur.
Dirigée par : Konstantin AVRACHENKOV
Soutenue le : 6 avril 2022

Devant le jury, composé de :
Konstantin Avrachenkov, DR, Inria
Gérard Burnside, Research scientist, Nokia Bell Labs
Frédéric Havet, DR, Inria
Varun Jog, Assistant Professor, University of Cambridge
Marc Lelarge, DR, ENS / Inria
Lasse Leskelä, Associate Professor, Aalto University
Catherine Matias, DR, CNRS / Sorbonne Université
Dieter Mitsche, Professor, Institut Camille Jordan

Inria

Graph clustering and semi-supervised learning of non-binary, temporal and geometric networks

Détection de communautés et apprentissage semi-supervisé dans des réseaux non-binaires, temporels et géométriques

Jury :

Directeur de thèse :
Konstantin Avrachenkov, Directeur de Recherche, Institut National de Recherche en Informatique et en Automatique (INRIA)

Rapporteurs :
Catherine Matias, Directrice de Recherche, Centre National de Recherche Scientifique (CNRS); Sorbonne Université, Université de Paris
Dieter Mitsche, Professor, Institut Camille Jordan

Président du Jury :
Frédéric Havet, Directeur de Recherche, Institut National de Recherche en Informatique et en Automatique (INRIA)

Examinateurs :
Varun Jog, Assistant Professor, University of Cambridge
Marc Lelarge, Directeur de Recherche, Ecole Normale Supérieure (ENS) & Institut National de Recherche en Informatique et en Automatique (INRIA)
Lasse Leskelä, Associate Professor, Aalto University

Invité :
Gérard Bunrside, Research scientist, Nokia Bell Labs

En mémoire de mon père.

Résumé

La multiplication des données collectées a occasionné un intérêt multi-disciplinaire autour de l'étude statistique de systèmes complexes où les individus interagissent en pairs. Dans de tels réseaux, des individus similaires ont tendance à interagir similairement. Un important problème d'inférence statistique consiste donc à regrouper les individus similaires en *communautés* (aussi appelées *clusters*) en se basant uniquement sur l'observation des interactions entre individus. Ce problème d'apprentissage non-supervisé qui consiste à placer chaque nœud dans un groupe est appelé *détection des communautés*. Cette thèse a pour but d'étudier différents aspects de la détection de communautés dans des systèmes complexes.

En particulier, nous étudions des modèles de graphes aléatoires où chaque nœud appartient à une communauté (aussi appelé bloc) et où l'interaction entre deux nœuds dépend uniquement des blocs dans lesquels ces deux nœuds appartiennent. Pour ces modèles, nous établissons des résultats théoriques sur la possibilité et l'impossibilité de découverte des communautés. Notre étude autorise des interactions quelconques (non nécessairement binaires), ce qui rend le résultat applicable à de nombreuses situations (interaction pondérées, temporelles, multi-couches, etc.).

Dans le cas particulier où les interactions entre les nœuds varient au cours du temps, nous proposons plusieurs algorithmes. Plus précisément, nous présentons des méthodes spectrales utilisant les interactions persistantes ou des méthodes basées sur un calcul itératif de la vraisemblance.

Nous examinons aussi le problème de la détection semi-supervisées des communautés. Dans ce cas, un oracle nous renseigne sur la communauté d'un petit nombre de nœuds. Ces renseignements s'ajoutent aux interactions observées entre les nœuds, et facilitent le problème initial (l'apprentissage des communautés des nœuds).

Enfin, nous étudions la situation où les nœuds sont positionnés dans un espace métrique. Dans ce cas, nous montrons que les méthodes spectrales classiques (telles que Spectral Clustering) peuvent totalement échouer, et nous analysons une parade basée sur la sélection de vecteurs propres d'ordre supérieur.

Mots clés : détection de communautés, modèle des blocs stochasique, réseaux temporels, réseaux géometriques, apprentissage semi-supervisé, spectral clustering.

Abstract

The massive explosion of data collection led to a multi-disciplinary interest in the statistical inference of complex systems. In these systems, agents interact by pairs. Since similar agents tend to interact similarly, an important unsupervised learning problem consists of grouping the agents into *communities* or *clusters* based on the pairwise interactions. This thesis explore various aspects of this learning task.

In particular, we study random graph models in which each node belongs to a community (also called block) and the interactions between node pairs depend on the community structure. For those *stochastic block models*, we establish consistency thresholds for community recovery. These results allow for non-binary interactions, such as weighted, temporal or multiplex networks.

We propose several algorithms for clustering temporal networks, such as spectral methods based on the persisting edges, or methods based on an online computation of the likelihood.

We also study graph clustering in a semi-supervised setting. In this setting, an oracle provides the community memberships of a few nodes. This extra information helps to recover the community labels of the rest of the nodes.

Finally, we investigate networks in which the nodes have a position in a metric space. In such *geometric networks*, we show that standard spectral methods (such as Spectral Clustering) fail at recovering the communities. We propose and analyse a spectral algorithm based on a higher-order eigenvector.

Keywords: graph clustering, stochastic block model, temporal networks, geometric networks, semi-supervised learning, spectral clustering.

Remerciements

Tout d'abord, je tiens à remercier chaleureusement Konstantin Avrachenkov pour avoir été mon encadrant de thèse, et pour tous les savoirs, conseils et opportunités qu'il a su me transmettre. Par ailleurs, cette thèse n'aurait pas été possible sans le soutien de l'Inria et du financement de Nokia Bell Labs.

Je remercie aussi Lasse Leskelä pour m'avoir chaleureusement accueilli à plusieurs reprises à Helsinki, et pour toutes les heures que l'on a pu passer à travailler ensemble. Ce projet sans fin s'est étalé sur plus de 2 ans et a débouché sur une large portion du présent manuscript ainsi que plusieurs articles.

Je remercie tout naturellement les membres du jury, et plus spécifiquement Catherine Matias et Dieter Mitsche pour avoir accepté d'être rapporteur.

Ces années de thèse ont été enrichies de tous les bons moments passés avec les membres de l'équipe NEO. Je pense tout particulièrement aux retraites, aux repas à la cantine et aux goûters partagés. Ainsi je tiens à remercier, sans ordre particulier, les permanents de l'équipe (Alain, Sara, Giovanni, Samir, Laurie, Jane) ainsi que tous les non-permanents qui se sont succédés durant mon passage chez NEO, et dont la liste est trop longue et le risque d'oubli trop important. Je remercie néanmoins particulièrement Abhishek, Dimitra, Guilherme, Mikhaïl, Chuan, Andreï, Tareq, Othmane, Youness et Angelo.

Enfin, je tiens à remercier Tracy Trepy pour son soutien moral sans faille et ses conseils éclairés, ainsi que ma famille et tout particulièrement ma maman.

Contents

List of Notations

Symbol	Meaning
δ_x	Dirac measure at x
δ_{ab}, $\delta(a,b)$	Kronecker delta
$1(\cdot)$	Indicator function
η	Scale parameter
n or N	Number of nodes
K	Number of communities (blocks)
$\mathcal{S}$	Space of interaction types (*e.g.*, $\mathcal{S} = \{0,1\}^T$ for temporal networks)
T	Number of snapshots (for temporal networks)
$A = (A_{ij})$	data array ($A \in \mathcal{S}^{N\times N}$)
L, $\mathcal{L}$	Standard and Normalized graph Laplacians
σ	Node labelling ($\sigma \in [K]^N$)
i, j	Node indices
k, ℓ	Community indices
f, g	intra-block and inter-block interaction distribution
μ_a, ν_a	Initial intra-block and inter-block interaction distributions, $a \in \{0,1\}$
P_{ab}, Q_{ab}	Probability of transition $a \to b$ ($a, b \in \{0,1\}$) for intra- and inter-block interactions
$\mathrm{D}_\alpha(f\|g)$	Rényi-divergence of order α
$Z_\alpha(f\|g)$	Hellinger integrals of order α
$\mathrm{Hel}(f,g)$	Hellinger distance
$\mathrm{d}_{\mathrm{Ham}}(\sigma_1,\sigma_2)$	Hamming distance
$\mathrm{d}^*_{\mathrm{Ham}}(\sigma_1,\sigma_2)$	Absolute classification error (permutation-corrected Hamming distance)

Chapter 1

Introduction

1.1 Networks

A *network* is a collection of interconnected objects. This definition covers a lot of real-world phenomena, as data sets in many application domains consist of pairwise interactions between objects. Examples include human interactions in sociology and epidemiology, brain activity measurements in neuroscience, protein interactions in molecular biology and financial interactions in economics.

When the pairwise interactions are binary, the network can mathematically be represented by a *graph*. A graph G is a pair (V, E) where V is the set of objects (called *nodes* or *vertices*) and E is the set of pairs of nodes (called *edges* or *links*). This graph representation can easily be extended to non-binary networks. The following examples show that many situations involve non-binary interactions as well.

Examples of networks

Binary networks The *political blog* data set [AG05] is a simple example of a network with binary interactions. This data set represents the linking patterns of political bloggers during the U.S. Presidential Election of 2004. The data set is composed of 1494 blogs, 759 liberals and 735 conservatives, and the interactions identify whether one blog references another blog. As shown in Figure 1.1, the difference between liberal and conservative blogospheres is clear. Indeed, 90% of the interactions occur between blogs belonging to the same political community.

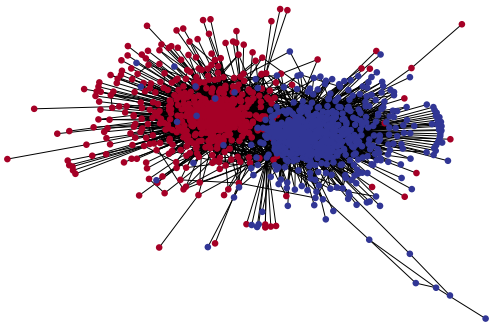


Figure 1.1: Network of political blogs.

Weighted networks A binary interaction can be enriched by a weight, representing the interaction strength. A *weighted network* is a network whose interactions are positive real numbers. Examples encompass transportation network between cities, where the weights correspond to the number of passengers going from one city to another.

Weighted networks can also be built directly from data. Indeed, in machine learning problems, data often comes as a collection $x_1, \cdots, x_n$ of n data points in an Euclidean space (*e.g.*, $x_i \in \mathbb{R}^m$). A common way to define a weight w_{ij} between two data points x_i and x_j is via a threshold Gaussian kernel

$$w_{ij} = \begin{cases} \exp\left(-\frac{\|x_i - x_j\|^2}{\tau^2}\right), & \text{if } \|x_i - x_j\|^2 \leq \kappa, \\ 0, & \text{otherwise,} \end{cases}$$

where τ and κ are some parameters and $\|\cdot\|$ is a distance between the data points. The cutoff parameter κ prevents having a dense network with many small weights. Another familiar method is to connect each vertex to its k-nearest neighbours. We refer to [GP10, Chapter 4] and [Sta+20] for the description of other methods.

The MNIST database [LCB98] is a database of 70,000 images of handwritten digits commonly used as a benchmark in machine learning. Figure 1.2 represents a network built from 300 images of digits $0, 1, 2$ using a Gaussian kernel as weight function. More precisely, we first compute a k-nearest neighbour graph ($k = 8$) with weights

$$w_{ij} = \begin{cases} \exp\left(-\frac{4\|x_i - x_j\|^2}{\tau_i}\right) & \text{if } x_j \text{ is in the } k\text{-th nearest neighbour of } x_i, \\ 0 & \text{otherwise,} \end{cases}$$

where τ_i represents the distance between x_i and its k-th nearest neighbour. The weight matrix is finally symmetrised by replacing W with $\frac{1}{2}(W + W^T)$.

Figure 1.2: Network constructed from 300 images of digits 0, 1 and 2 taken from the MNIST database.

Edge-labelled and multiplex networks Complex relational databases and networks often include interactions of different types. For example, chemical reactions may be exothermic or endothermic, movie-user associations in collaborative filtering typically come with user ratings; communication between individuals may be cold, formal, or familiar. Such networks are called *edge-labelled* networks, where the interaction is an element of the set $\{0, 1, \cdots, L\}$, the type 0 denoting no interaction.

Alternatively, an edge-labelled network can be represented by a multiplex network. Multiplex networks refer to networks with several layers in which all the layers share the same node-set, and each layer corresponds to an interaction type. This view is sometimes more convenient. For example, two social networks users might be connected through Facebook, Twitter, or LinkedIn, where each social network creates one layer of a user-user interaction network.

Temporal networks Interactions between node pairs might vary with time. For example, the *high school data sets* represent close proximity encounters between students in a French high school. Students-to-students interactions are recorded every 20 seconds through wearable sensors, and the experiment span several school days. The same experiment was performed three consecutive years [FB14; MFB15], and the dimension of each data set is given in Table 1.1. We also plot in Figure 1.3 the weighted network for the year 2013, where the weights correspond to the number of interactions recorded between two students.

We notice that this time-aggregation can lose important temporal information. For example, Figure 1.4 shows snapshot per snapshot the average number of interactions per student on a given day. The peaks observed coincide with the starting and ending times of the breaks between courses since students leave the classrooms.

Year	n	K	T
2011	118	3	5609
2012	180	5	11273
2013	327	9	7375

Table 1.1: Dimensions of three data sets of interacting high school students: number of students n, number of classes K and number of snapshots T.

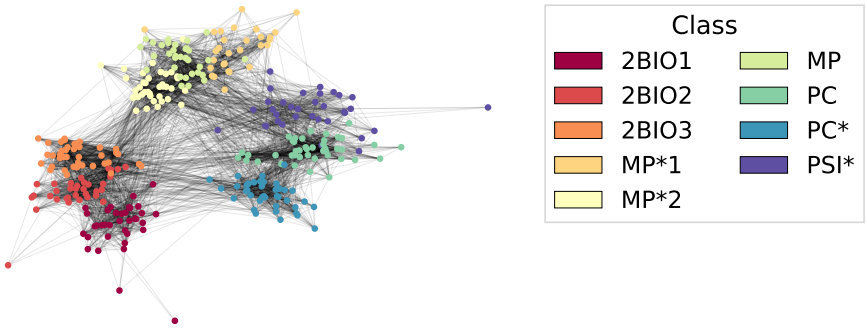


Figure 1.3: Time-aggregated network obtained from the high-school interaction network (year 2013).

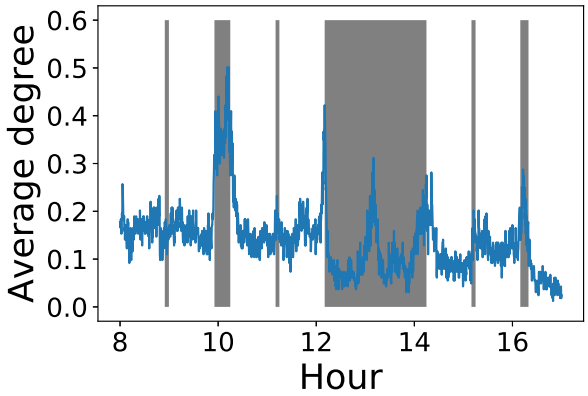


Figure 1.4: Average degree on Thursday. The hashed regions show the breaks between classes.

1.2 Clustered random graph models

Numerous random graph models have been proposed to describe real-world networks. In this thesis, we will focus on *clustered random graphs*. A *cluster* (also called community or block) is a group of nodes that share common properties, and thus play a similar role in the network. The previous examples of networks showed that clusters are naturally present in real work networks: political affiliation of blogs, digits in handwritten images, classes of high school students, etc.

Stochastic Block Model

The Stochastic Block Model is one of the oldest and simplest random graph models with community structure. Although this model was originally defined for multiplex graphs [HLL83], it has been later on restricted to binary interactions. We define here the general setting where the interactions belong to a measurable space $\mathcal{S}$.

In this model, the node-set $V = [N]$ is partitioned into K disjoints sets, called blocks. This partition is conveniently represented by a node labelling $\sigma\colon [N] \to [K]$ so that $\sigma(i)$ indicates the block which contains node i. The model is further parametrized by a collection $f = (f_{k\ell})_{k,\ell\in[K]}$ of probability distributions over $\mathcal{S}$, such that $f_{k\ell} = f_{\ell k}$. These parameters specify the probability distribution

$$\mathbb{P}(A \mid \sigma) = \prod_{1\le i<j\le N} f_{\sigma(i)\sigma(j)}(A_{ij})$$

of the interaction tensor $A \in \mathcal{S}^{N\times N}$, verifying $A_{ij} = A_{ji}$ for all i, j. Our main focus is on *homogeneous* models in which the probability distribution $f_{k\ell}$ equals f if $k = \ell$, and g otherwise and the node labels are chosen uniformly at random.

In this framework, a mutliplex network can be represented by choosing $\mathcal{S} = \{0,1\}^M$ where M is the number of layers, while a temporal network with T snapshots correspond to $\mathcal{S} = \{0,1\}^T$. Other important possible choices for the interaction space are $\mathcal{S} = \mathbb{R}_+$ (weighted SBMs), $\mathcal{S} = \{0, 1\ldots, L\}$ (edge-labelled SBM) or $\mathcal{S} = \mathbb{Z}_+$.

The *inference problem* consists in the recovery of the node labelling σ given the interaction tensor A, and possibly the number of blocks and the interaction probabilities f, g. The recovery of the blocks by an estimator is *consistent* if the fraction of mis-clustered nodes goes to zero when the number of nodes N goes to infinity, and *strongly consistent* if the number of mis-clustered nodes goes to zero. Intuitively, this task becomes harder when f and g become too similar and is even trivially impossible when $f = g$. This implies the existence of *fundamental limits*, that is conditions under which the blocks can or cannot be recovered. For example, consistent recovery can be impossible (in the sense that no algorithm can attain it) if some condition is not verified, and possible (there exists an algorithm achieving it) if another condition is verified.

When $\mathcal{S} = \{0,1\}$ existing work on community recovery in the binary homogeneous SBM provides a strong information-theoretic foundation [ZZ16; Gao+17; ABH16; MNS16]. In particular, many sharp fundamental limits (or *phase transitions*) have been derived. Denoting by $\mathrm{D}_{1/2}(f,g)$ the Rényi divergence between f and g, consistent recovery is possible if $N\,\mathrm{D}_{1/2}(f,g) \gg 1$ and is impossible otherwise. Similarly, strong consistency is possible if $\mathrm{D}_{1/2}(f,g) \ge (1+\Omega(1))\,\frac{K\log N}{N}$, and impossible if $\mathrm{D}_{1/2}(f,g) \le (1-\Omega(1))\,\frac{K\log N}{N}$.

Analogous results have been obtained, sometimes with additional technical conditions on the connectivity functions f, g, for multiplex SBMs with independent layers [PC16], edge-labelled SBMs [YP16; JL15], and weighted SBMs [XJL20]. One contribution of this thesis is to extend these information-theoretic results to $\mathcal{S}$-valued SBM, where $\mathcal{S}$ is any measurable space and might depend on the scale parameter. One possible application is temporal networks with

T dependent layers, where the number of nodes N and the number of layers T both grow unbounded.

Clustered geometric graphs

In many situations, nodes have geometric attributes in addition to community labelling. These geometric attributes are represented by a position in a metric space. The interaction between a pair of nodes, therefore, depends not only on the community labelling but also on the distance between the two nodes. We can model this by assigning to each node a position, chosen in a metric space. Then, the probability of an edge appearance between two nodes will depend both on the community labelling and on the positions of these nodes. Recent proposals of random geometric graphs with community structure include the Geometric Block Model (GBM) [Gal+18] and Euclidean random geometric graphs [ABS17].

In the GBM, two nodes i, j in the same community are interacting if they are at a distance d_{ij} less than r_{in}, while two nodes in different communities are interacting if their distance is less than r_{out}, where $0 \leq r_{\text{out}} < r_{\text{in}}$ are two parameters. In Euclidean random geometric graphs, the probability of connection between two nodes i, j in the same community equals $f(d_{ij})$ where d_{ij} denotes the distance between the two nodes. Furthermore, the probability of interaction between two nodes i, j in different communities is $g(d_{ij})$ (here $f, g \colon \mathbb{R}_+ \to [0, 1]$ are two functions). We obtain the GBM as a particular case (by letting $f(x) = 1(|x| \leq r_{\text{in}})$ and $g(x) = 1(|x| \leq r_{\text{out}})$), as well as the binary SBM (by simply letting $f(x) = p_{\text{in}}$ and $g(x) = p_{\text{out}}$ where $p_{\text{in}}, p_{\text{out}} \in [0, 1]$).

The nodes' interactions in geometric models are no longer independent: two interacting nodes are likely to have many common neighbours. While this is more realistic ("friends of my friends are my friends"), this also renders the theoretical study more challenging.

1.3 Graph clustering

Many powerful clustering methods exist for binary or weighted networks, such as spectral methods [Von07], modularity maximisation [BC09; Blo+08; GN02], belief propagation [MM09; Moo17], Bayesian methods [HW08; Pei19], likelihood-based methods [WB17]). We refer to [For10] for an overview of such methods.

Spectral clustering

Motivation

Spectral clustering is one of the simplest clustering algorithms. We will give it here a quick overview, and we refer to [Von07] for a more detailed presentation. Let A be the graph adjacency matrix and $d_i = \sum_{j=1}^{N} A_{ij}$ be the degree of node i. We denote by $D = \text{diag}(d_1, \cdots, d_n)$ the *degree matrix*, and define the *standard Laplacian* $L = D - A$ and the *normalized Lapla-*

cian $\mathcal{L} = I - D^{-1/2}AD^{-1/2}$. In the following, $V_1, \dots, V_K$ denote a partition of the node-set $V = [N]$ into K non-overlapping clusters, that is $V_1 \cup \cdots \cup V_K = V$ and $V_k \cap V_\ell = \emptyset$ for $k \neq \ell$. The *graph-Cut* associated to such partition is defined as

$$\mathrm{Cut}(A, V_1, \dots, V_K) \;=\; \sum_{k=1}^{K} \mathrm{Cut}(A, V_k),$$

where $\mathrm{Cut}(A, V_k) \;=\; \sum_{i \in V_k, j \in V_k^c}^{N} A_{ij}$ is the number of edges going from set V_k to set V_k^c. Directly minimising this quantity would lead to greatly unbalanced partitions, *e.g.*, partitions in which all or almost all the nodes are in the same cluster. To alleviate this issue, we first define the volume of a set by

$$\mathrm{vol}\,(V_k) \;=\; \sum_{i \in V_k} d_i.$$

We will then aim to find a partition $V_1, \dots, V_K$ that minimises

$$\mathrm{Ncut}(A, V_1, \dots, V_K) \;=\; \sum_{k=1}^{K} \frac{\mathrm{Cut}(A, V_k)}{\mathrm{vol}(V_k)}. \tag{1.3.1}$$

The *Normalized-Cut* corresponds to a Cut penalized with respect to the volume of the sets V_k: small sets bear a large penalty. Hence, we can expect that the solutions minimising the Normalized-Cut lead to clusters of similar sizes.

As minimising the Normalized-Cut is NP-hard [WW93], we instead perform a *continuous relaxation*. Let us define the matrix $N = (n_{ik}) \in \mathbb{R}^{n \times K}$ by

$$\forall\, i \in [n],\ \forall\, k \in [K]\colon \quad n_{ik} \;=\; \begin{cases} \dfrac{1}{\sqrt{\mathrm{vol}(V_k)}} & \text{if} \quad v_i \in V_k, \\ 0 & \text{otherwise.} \end{cases} \tag{1.3.2}$$

N is a matrix containing K indicators vectors as columns, where the size of each set V_k is used as a penalization term. We notice that

$$\mathrm{Ncut}(A, V_1, \dots, V_K) \;=\; \mathrm{Tr}\left(N^T L N\right),$$

while $N^T D N \;=\; I_K$. Therefore, minimising the Ncut can be rewritten as

$$\underset{(V_1, \dots, V_k)}{\arg\min}\, \mathrm{Tr}\left(X^T \mathcal{L} X\right), \tag{1.3.3}$$

where $X := D^{1/2} N$ and N is defined in equation (1.3.2). The final step is to relax (1.3.3) by keeping only the constraint $X^T X = I_K$. This leads to the following relaxed problem

$$\hat{X} \;=\; \underset{X \in \mathbb{R}^{n \times K}\colon\, X^T X = I_K}{\arg\min}\; \mathrm{Tr}\left(X^T \mathcal{L} X\right).$$

The solution of this relaxed problem is given by the matrix whose columns are given by the K orthonormal eigenvectors of $\mathcal{L}$ associated with the K smallest eigenvalues.

Once the relaxed problem is solved, we are left with a n-by-K matrix whose columns are the K first eigenvectors of $\mathcal{L}$. To reconvert this real-valued matrix to a discrete partition, a standard method is to apply k-means algorithm on the n rows of K (seen as n data points in $\mathbb{R}^K$). More precisely, k-means consists in the following minimisation problem

$$\left(\hat{Z}, \hat{X}\right) = \underset{\substack{Z \in \mathcal{Z}_{n,K} \\ X \in \mathbb{R}^{K \times K}}}{\arg\min} \|ZX - V\|_F^2 \tag{1.3.4}$$

where $\mathcal{Z}_{n,K}$ denotes the space of *membership matrices*, that is $n \times K$ matrices with entries in $\{0, 1\}$ for which each row i has only one non-zero element. While solving the minimisation problem (1.3.4) is NP-hard, there exists (see [KSS04]) a polynomial time procedure finding

$$\begin{aligned} &\left(\hat{Z}, \hat{X}\right) \in \mathcal{Z}_{n,K} \times \mathbb{R}^{K \times K} \\ \text{s.t.} \quad &\left\|\hat{Z}\hat{X} - V\right\|_F^2 \leq (1+\epsilon) \min_{\substack{Z \in \mathcal{Z}_{n,K} \\ X \in \mathbb{R}^{K \times K}}} \|ZX - V\|_F^2. \end{aligned} \tag{1.3.5}$$

Once $\hat{Z}$ is found, we return the predicted clusters: node i is in cluster k if $\hat{Z}_{ik} = 1$. We summarize this in Algorithm 1.

Algorithm 1: (Normalized) Spectral Clustering.

Input: Graph Laplacian L (resp. normalized Laplacian $\mathcal{L}$), number of clusters K.
Output: Predicted node labelling vector $\hat{z} \in [K]^n$.
Spectral Step:

- Compute $v_1, \ldots, v_K$ the K orthonormal eigenvectors of L (resp. $\mathcal{L}$) associated with the K smallest eigenvalues;
- Let $V \in \mathbb{R}^{n \times K}$ be the matrix whose column k is v_k.

Clustering Step:

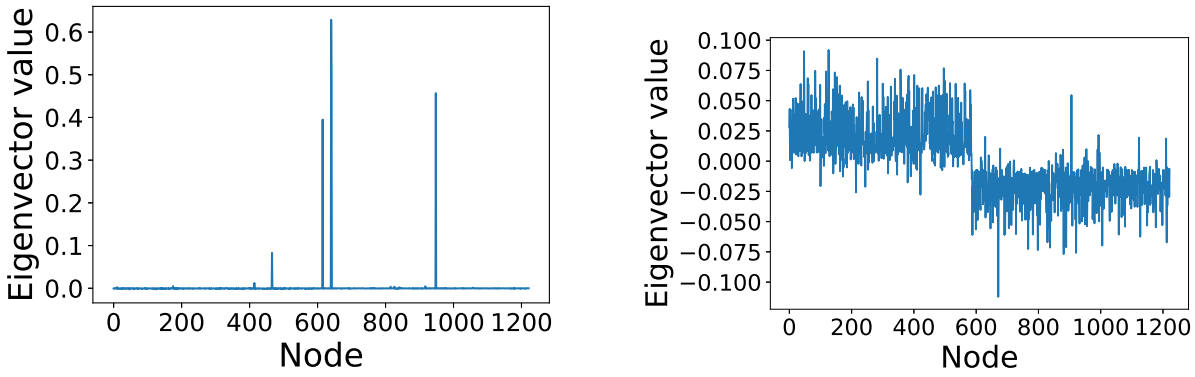
- Let $\left(\hat{Z}, \hat{X}\right)$ be an $(1+\epsilon)$ approximate solution to the k-means problem (1.3.5);
- For every node $i = 1 \cdots n$, let $z_i = k$ if $Z_{ik} = 1$.

In the particular case of two clusters, one can simply look at the second eigenvector of $\mathcal{L}$ (sometimes referred to as the *Fiedler vector* [Fie75]). The nodes are then clustered according to the sign of their entries.

Limitations of spectral clustering

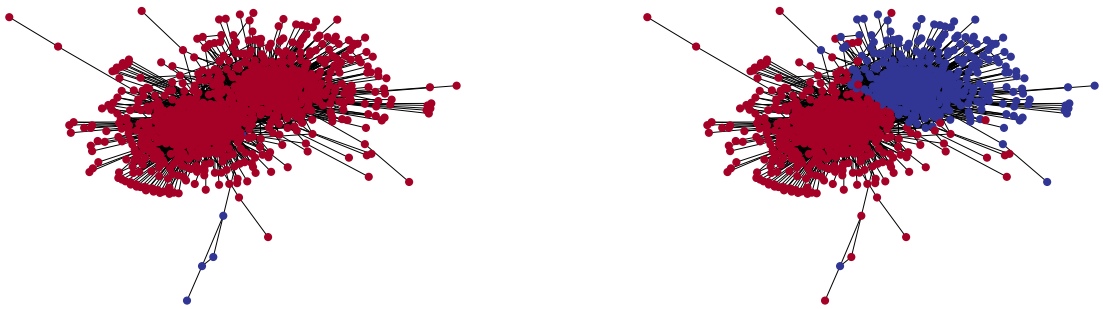
Dangling trees In this part, we will analyze the failure of spectral clustering on the *political blogs* data set. Figure 1.5 shows the value of eigenvectors of $\mathcal{L}$ associated with the second and third smallest eigenvalues. We see that the entries of the second eigenvector are localised

over a few nodes. Moreover, those nodes are associated with a *dangling tree* and do not represent a meaningful community structure (see Figure 1.5c). On the contrary, the entries of the third eigenvector lead to the correct community structure. In fact, using this eigenvector for clustering would lead to an accuracy of 95%.



(a) $k = 2$.

(b) $k = 3$.



(c) $k = 2$.

(d) $k = 3$.

Figure 1.5: Analysis of the failure of Spectral Clustering on the *political blogs* data set. Top: value of the eigenvector of $\mathcal{L}$ associated to the k-th smallest eigenvalue, for $k = 2$ and $k = 3$. Bottom: graph where the node colors show the prediction made using the sign of the entries of the k-th eigenvector.

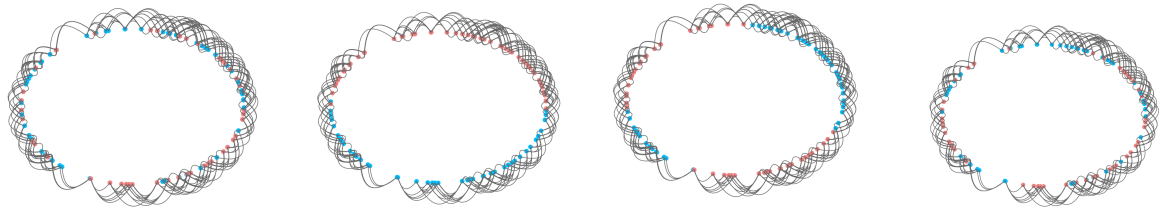
Figure 1.5 shows that the good eigenvector for clustering is the third one, while the second eigenvector is concentrated around low degree nodes, forming a dangling tree[1]. Since it results in a partition of the graph into one large community, with almost all the nodes, and a small one with only a few nodes, it is easy to spot in practice. To solve this issue, one simple solution would be to look at a higher-order eigenvector. But, how to determine the correct eigenvector? Indeed, this might not always be an easy task. First of all, it could happen that the correct eigenvector is in a lower position, say 5th or 7th, and localising it among noisy eigenvectors might be non-trivial. Besides, it is difficult to extend this reasoning for more than two clusters.

The *regularization technique* aims to solve this issue [ZR18]. It consists in performing Spectral Clustering on $\mathcal{L}_\tau := I - D_\tau^{-1/2} A_\tau D_\tau^{-1/2}$, where $A_\tau := A + \frac{\tau}{n} 1_n^T 1_n$ and D_τ is the

[1]Note that if one were to use the standard Laplacian, we would also observe an analogous phenomenon, with noisy eigenvectors concentrated around high degree nodes.

associated degree matrix. The matrix A_τ is a perturbed version of the initial adjacency matrix A, where was added an edge of weight $\frac{\tau}{n}$ between all nodes pairs. The perturbation parameter τ is typically taken as $\tau = 1$ or $\tau = \bar{d}$ where $\bar{d}$ is the average degree of the graph.

Spectral methods and geometric data In many situations, nodes can have geometric attributes (for example, a position in a metric space). The geometric structure of this model handicaps cut-based clustering methods. Indeed, in this case, the Fiedler vector might be associated with a geometric configuration, hence bearing no information about the latent community labelling. Figure 1.6 highlights this issue in GBM. While the second and fourth eigenvectors give configurations based on node positions, recovering the node labels is, in this example, better done with the 10th eigenvector.



(a) True labels. (b) $k = 2$. (c) $k = 4$. (d) $k = 10$.

Figure 1.6: Analysis of the failure of Spectral Clustering on a Geometric Block Model, with 100 nodes, inter-distance and intra-distance cutoffs $r_{\text{in}} = 0.07$, $r_{\text{out}} = 0.02$.

We show that a higher-order eigenvector can lead to better clustering in real data sets. We select 1000 images from MNIST, representing digits 4 and 9, and construct a k-nearest neighbours ($k = 8$) interaction graph with Gaussian weights. The digits 4 and 9 form the hardest digit pair to distinguish. We plot in Figure 1.7 the accuracy obtained by spectral clustering as a function of the eigenvector order. We emphasise the fact that, unlike the *political blog* data set, this is not an artefact due to dangling trees. We plot in Figure 1.8 the predicted clusters using the eigenvectors associated with the second and smallest eigenvalues of the graph's normalized Laplacian and compare them with the true clusters. We notice that the predicted clusters are of balanced sizes. We also note that the Ncut of the true labels is 3.8, while the Ncut of the predicted labels associated with the prediction using the second (resp. third) eigenvector is 2.7 (resp. 3.7). Therefore for this graph, the correct labels do not match with the smallest normalized cut.

Graph semi-supervised clustering

Semi-supervised learning (SSL) aims at achieving superior learning performance by combining unlabelled and labelled data. Since typically the amount of unlabelled data is large compared to the amount of labelled data, SSL methods are relevant when the performance of unsupervised learning is low, or when the cost of getting a large amount of labelled data for supervised learning is too high. Unfortunately, many standard semi-supervised learning techniques have been shown to not efficiently use the unlabelled data, leading to unsatisfactory

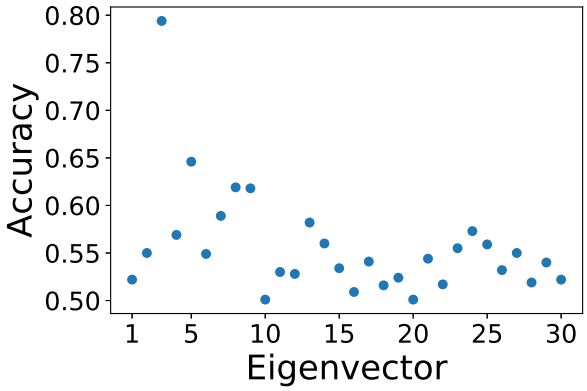
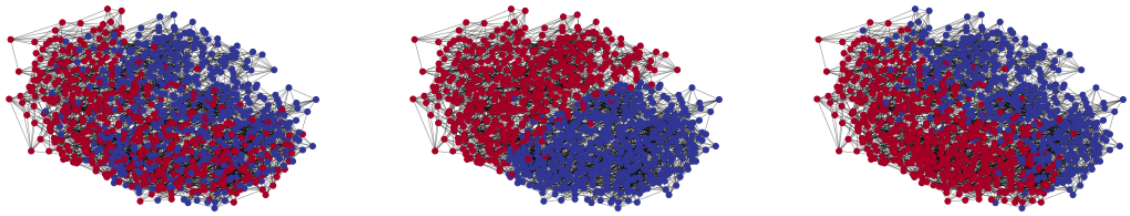


Figure 1.7: Accuracy obtained on weighted graph build using a subset of the MNIST data set ($n = 1000$ images representing digits 4 and 9) using the different eigenvectors of the normalized Laplacian matrix $\mathcal{L}$. The eigenvector of index k is the eigenvector associated with the k-th smallest eigenvalue of $\mathcal{L}$.



(a) True labels (b) Labels using v_2. (c) Labels using v_3.

Figure 1.8: Different clusterings on the same graph as Figure 1.7. The colours in Figure 1.8a shows the true labels, while the colours in Figures 1.8b and 1.8c represent the predicted labels using respectively the eigenvector associated with the second and third smallest eigenvalues of the normalized Laplacian.

or unstable performances ([CSZ06, Chapter 4]; [BLP08; CCC02]). Furthermore, the presence of noise in the labelled data can further degrade the performance. In practice, the noise can come from a tired or non-diligent expert carrying out the labelling task.

Let us consider a graph $G = (V, E)$ whose node-set is partitioned into K latent communities. For the simplicity of the current exposition, assume $K = 2$. The communities are thus represented by a vector $\sigma \in \{-1, 1\}^n$. In addition to the observation of the graph, an *oracle* gives us extra information about the cluster assignment of some nodes. We call those nodes *labelled nodes*, and we denote by ℓ the set of labelled nodes. Among those nodes, some are correctly labelled by the oracle, and some are mislabelled by the oracle. We denote by ℓ_0 the set of mislabelled nodes and ℓ_1 the set of correctly labelled nodes. In particular, $\ell = \ell_0 \sqcup \ell_1$. The oracle can be represented as a vector $s \in \{0, \dots, K\}^n$ such that

$$s_i = \begin{cases} \sigma_i & \text{if} \quad i \in \ell_1, \\ -\sigma_i & \text{if} \quad i \in \ell_0, \\ 0 & \text{if} \quad i \notin \ell. \end{cases} \tag{1.3.6}$$

In other words, the oracle (1.3.6) reveals the correct cluster assignment of $|\ell_1|$ nodes and a false cluster assignment for $|\ell_0|$ nodes. It reveals nothing for $n - |\ell|$ nodes. The quantity $|\ell_0|/|\ell|$ is the rate of mistake of the oracle (*i.e.,* the probability that the oracle reveals false information given that it reveals something). The oracle is informative if this quantity is less than $1/2$, which is equivalent to the intuitive condition $|\ell_1| > |\ell_0|$. In the following, we will always assume that the oracle is informative.

Assumption 1. The oracle is informative, that is $|\ell_1| > |\ell_0|$.

Given the oracle s and the graph G, our strategy is to find a vector $\hat{x} \in \mathbb{R}^n$ from which we could predict the node's labels: node i will be classified in cluster $\hat{\sigma}_i$ according to the sign of $\hat{x}_i$. A standard framework is to let $\hat{x}$ be the solution of an optimization problem of the type

$$\hat{x} \;=\; \arg\min_{x \in \mathcal{X}} C(x, s),$$

where $C(x, s)$ is a cost function, and $\mathcal{X}$ a subset of $\mathbb{R}^n$.

For instance, the paper [Avr+12] suggests the following optimization formulation

$$\hat{x} \;=\; \arg\min x^T D^{\sigma-1} L D^{\sigma-1} x + \lambda \left(x - s\right)^T D^{2\sigma-1} \left(x - s\right).$$

particular choices of σ lead to different methods (namely $\sigma = 1$ is the Standard Laplacian method [ZGL03], $\sigma = \frac{1}{2}$ is the Normalized Laplacian method [Zho+04] and $\sigma = 0$ is a Page Rank based method). Since the problem is equivalent to

$$\hat{x} \;=\; \arg\min_{x} \sum_{i<j} a_{ij} \left(d_i^{\sigma-1} x_i - d_j^{\sigma-1} x_j\right)^2 + \lambda \sum_{i=1}^{n} \left(x_i - s_i\right)^2,$$

we observe that the first term forces the smoothness of the solution $\hat{x}$, while the second term pushes the solution towards the oracle value s. This trade-off is governed by a parameter λ.

Clustering temporal networks

In this section, we consider a temporal network represented by a tuple of adjacency matrices $\left(A^1, \ldots, A^T\right)$, and we consider a setting in which the node labels are static through time. If we assume that the temporal snapshots A^t are independent of each other, one could simply generalize the classical $\min \mathrm{Cut}$ problem by considering

$$\arg\min_{z \in [K]^n} \sum_{t=1}^{T} \mathrm{Cut}\left(A^t, z\right).$$

Since $\sum_{t=1}^{T} \mathrm{Cut}\left(A^t, z\right) \;=\; \mathrm{Cut}\left(\sum_{t=1}^{T} A^t, z\right)$, we would then apply a spectral method on the time-aggregated graph (that is, the weighted graph represented by the adjacency matrix $\sum_{t=1}^{T} A^t$).

Unfortunately, this fails at integrating the time-correlation in the interaction patterns between nodes. As an example, consider a network in which the inter-community interactions

are sparse and temporally independent (hence forming spikes), while the intra-community interactions are strongly correlated in time. Consider two node-pairs whose interaction patterns are given by $x_1 = (0,1,0,0,0,0,1,0,0,1)$ and $x_2 = (0,0,1,1,1,0,0,0,0,0)$. Since $\|x_1\|_1 = \|x_2\|_1 = 3$, we see that simple time-aggregation is agnostic to the different time patterns between time series x_1 and x_2.

A possible correction is to account for the persistent links. Indeed, since x_1 (resp. x_2) has zero (resp. two) transitions $1 \to 1$, keeping the endpoints of x_2 in the same cluster could be preferred. This can be done by considering

$$\arg\min_z \sum_{t=1}^{T} \mathrm{Cut}\left(A^t, z\right) + \alpha \sum_{t=2}^{T} \mathrm{PerCut}\left(A^{t-1}, A^t, z\right)$$

where

$$\mathrm{PerCut}\left(A^{t-1}, A^t, z\right) \;=\; \sum_{i,j\colon z_i \neq z_j} A^{t-1}_{ij} A^t_{ij}$$

counts the number of persistent links from time $t-1$ to time t between nodes with different label. We further notice that

$$\mathrm{PerCut}\left(A^{t-1}, A^t, z\right) \;=\; \mathrm{Cut}\left(A^{t-1} \odot A^t, z\right)$$

where $\odot$ denotes the matrix element-wise product. Introducing

$$W \;=\; A^1 + \sum_{t=2}^{T} A^t + \alpha A^{t-1} \odot A^t$$

we can rewrite the minimisation problem as

$$\arg\min_z \mathrm{Cut}\left(W, z\right).$$

This problem can be approximately solved by Spectral clustering on the weighted graph whose adjacency matrix is W.

1.4 Contributions and open problems

The contributions of this Thesis are the following.

- We derive information-theoretic conditions for consistent and strongly consistent recovery for homogeneous SBM with non-binary interactions. This extends and unifies in a single framework existing results for the binary SBM, and for various extensions such as edge-labelled SBM or weighted SBM. We can emphasise the following methodological improvements from current literature:
  - In order to lower bound the error made by *any* algorithm, we use a change of measure argument, similar to the one made by [YP16] in edge-labelled networks. This lower bound is in particular derived for a non-homogeneous model.

  - We show that the Maximum Likelihood Estimator attains the lower bound for homogeneous models. This result was shown in a minimax framework for the binary SBM [ZZ16] and multiplex models with independent layers [PC16]; we extend it to non-binary models.
  - Under some additional technical assumptions on the Rényi divergence, we provide an algorithm attaining the desired lower bound in a polynomial time in the number of nodes. This also extends [ZZ16, Theorem 3.1] and [XJL20, Proposition 6.1] to a larger set of interaction spaces.
  - We explicit the particular case of a temporal SBM where interactions are Markovian, by computing the Rényi-divergence between two sparse Markov chains, which could be of independent interest. In particular, we show that consistent recovery can be achieved even if the interactions are very sparse, given a large enough number of snapshots. Moreover, we compare the recovery bounds obtained to those obtained if one aggregates the temporal data.

- We have established error bounds on the recovery made by semi-supervised extensions of spectral methods. Numerical experiments further show promising performance on synthetic and real data sets, even in the case of very noisy labelled data.

- We derived a framework for spectral clustering in temporal networks, using the persisting and freshly appearing edges.

- By characterising the spectrum of the adjacency matrix of geometric graphs, we show that a higher-order eigenvector can recover the community structure of clustered geometric networks.

Several problems remain open to future work.

- Studying the Stochastic Block Model (SBM) with non-uniform community labels and non-homogeneous, non-binary interactions is a natural follow-up. Existing results for the non-homogeneous binary SBM show that consistency thresholds are governed by the Chernoff-Hellinger divergence [AS15]. We refer to the discussion in Section 3.5.2 for some insights.

- The study of recovery thresholds in SBM with a growing number of communities, for example $K = N^\alpha$ for some $\alpha \in (0, 1)$. We note that this question is still under study in the binary SBM. While some of our results extend to allow infinitely many communities, most do not, and several interesting research directions are left open (see the discussion in Section 3.5.1).

- The derivation of consistency conditions for spectral methods (using persistent edges) in temporal networks. In the binary SBM, consistency of Spectral Clustering is done by showing that the normalized Laplacian concentrates around its expectation, and concluding using the Davis-Kahan theorem [LR15; Abb+20]. Showing strong consistency requires a finer entry-wise analysis [Abb+20; Che+21]. Since obtaining tight bounds as in [LLV17] on the concentration of the normalized Laplacian in binary networks is not trivial (for example using matrix-Bernstein inequalities lead to sub-optimal concentrations [Tro15]), extending such results to temporal networks is interesting.

- Clustering temporal networks with time-varying communities. A weak recovery threshold was conjectured in [Gha+16], alongside with a belief propagation algorithm on a space-time graph. Dealing with temporally correlated interactions was partially tackled in [Bar+18], where the authors showed that the persistence of communities makes the clustering problem easier, while the persistence of the edges makes it harder. Pursuing these directions, both from an information-theoretic and an algorithmic point of view is an important open question.

- The study of clustering in geometric graphs is still in its infancy. In particular, our result is valid only in a dense regime where the expected node degrees scale with the number of nodes, while the information-theoretic results derived in [ABS17] assume that the geometric position of the nodes is known.

1.5 Publications

Journal publication

- [ABD20] Avrachenkov, K., Bobu, A., & Dreveton, M. (2021). Higher-Order Spectral Clustering for Geometric Graphs. *Journal of Fourier Analysis and Applications, 27(2)*, 1-29.

Conference papers

- [ADL21] Avrachenkov, K., Dreveton, M., & Leskelä, L. (2021, November). Recovering Communities in Temporal Networks Using Persistent Edges. In *International Conference on Computational Data and Social Networks* (pp. 243-254).

- [AD19] Avrachenkov, K., & Dreveton, M. (2019). Almost Exact Recovery in Label Spreading. In *International Workshop on Algorithms and Models for the Web-Graph* (pp. 30-43).

Preprints

- [ADL22] Avrachenkov, K., Dreveton, M., & Leskelä, L. (2022). Community recovery in non-binary and temporal stochastic block models. *arXiv preprint* arXiv:2008.04790.

- [AD19] Avrachenkov, K., & Dreveton, M. (2020). Almost exact recovery in noisy semi-supervised learning. *arXiv preprint* arXiv:2007.14717.

Part I

Recovery bounds in non-binary stochastic block models

Chapter 2

Preliminaries

This chapter contains some preliminaries results helping the motivation and the understanding of this thesis. We first present a quick reminder about the Rényi divergences between probability distributions, including some basic properties and examples. Then, we introduce the $\mathcal{S}$-valued stochastic block model (SBM) as a straightforward extension of the standard (binary) SBM. The chapter ends on some reminders about current litterature results of information-theoretic conditions for recovery in SBM.

2.1 Divergences between probability measures

Notations

For a measurable set $(\mathcal{S}, \mathcal{F})$, we denote by $\mathcal{P}(\mathcal{S})$ the set of probability measures on $(\mathcal{S}, \mathcal{F})$. For two probability measures f and g, we denote by $f \ll g$ if f is absolutely continuous with respect to g and by $f \perp g$ if f and g are singular. The negations of those statements

are denoted $f \not\ll g$ and $f \not\perp g$. We recall that if $f \ll \mu$, then $\frac{df}{d\mu}$ denote the Radon-Nikodym derivative of f with respect to μ, and that Radon-Nikodym derivatives differ only on a μ-null set.

2.1.1 Rényi and Kullback-Leibler divergences

The *Rényi divergence* of positive order $\alpha \neq 1$ of a probability distributions f on a state space $\mathcal{S}$ over another distribution g is defined as

$$\mathrm{D}_\alpha(f\|g) \;=\; \frac{1}{\alpha-1}\log\int\left(\frac{df}{d\mu}\right)^{\alpha}\left(\frac{dg}{d\mu}\right)^{1-\alpha}d\mu, \tag{2.1.1}$$

where μ is an arbitrary measure which dominates f and g. We use the convenient conventions $\log 0 = -\infty$, $0/0 = 0$ and $x/0 = \infty$ for $x > 0$. In particular, if $f \not\perp g$ and $\alpha \in (0,1)$ then $\mathrm{D}_\alpha(f\|g) < \infty$. The quantity $Z_\alpha(f\|g) = \int \left(\frac{df}{d\mu}\right)^{\alpha}\left(\frac{dg}{d\mu}\right)^{1-\alpha}d\mu$ is called *Hellinger integral* of order α.

Since the Rényi divergence of order $\alpha = \frac{1}{2}$ is symmetric in f and g, we will denote indifferently $\mathrm{D}_{1/2}(f\|g)$ as $\mathrm{D}_{1/2}(f,g)$. We also introduce the squared Hellinger distance $\mathrm{Hel}^2(f,g) = 1 - Z$ where $Z = Z_{1/2}(f,g)$. In particular, $\mathrm{D}_{1/2}(f,g) = -2\log Z_{1/2}(f,g)$.

The divergence of order $\alpha = 1$ cannot be defined using Formula (2.1.1). We instead let $\mathrm{D}_1(f\|g)$ to be the *Kullback–Leibler* divergence, given by

$$\mathrm{D}_{\mathrm{KL}}(f\|g) \;=\; \int \frac{df}{d\mu}\log\left(\frac{df/d\mu}{dg/d\mu}\right)d\mu$$

with the conventions that $0\ln(0/t) = 0$ and $t\ln(t/0) = \infty$ if $t > 0$. These definition is motivated by the continuity of $\alpha \mapsto \mathrm{D}_\alpha(f\|g)$ (see Proposition 1). We observe that $\mathrm{D}_{\mathrm{KL}}(f\|g) = \infty$ if $f \not\ll g$. If X is a f-distributed random variable, then $\mathrm{D}_{\mathrm{KL}}(f\|g) = \mathbb{E}_f\left[\log\frac{f}{g}(X)\right]$. We also define the centered Kullback–Leibler variation by $\mathrm{V}_{\mathrm{KL}}(f\|g) = \mathrm{Var}_f\left[\log\frac{f}{g}(X)\right]$. Thus

$$\mathrm{V}_{\mathrm{KL}}(f\|g) \;=\; \int \frac{df}{d\mu}\left(\log\frac{df/d\mu}{dg/d\mu}\right)^2 d\mu - \left(\mathrm{D}_{\mathrm{KL}}(f\|g)\right)^2.$$

The quantities $\mathrm{D}_\alpha, \mathrm{D}_{\mathrm{KL}}, \mathrm{V}_{\mathrm{KL}}$ are in general not symmetric in their arguments, with the notable exception of $\mathrm{D}_{1/2}$. We denote with a superscript s the symmetrized version, that is

$$\mathrm{D}^s_\alpha(f,g) \;=\; \mathrm{D}_\alpha(f\|g) + \mathrm{D}_\alpha(g\|f),$$

and similarly for $\mathrm{D}^s_{\mathrm{KL}}, \mathrm{V}^s_{\mathrm{KL}}$.

2.1.2 Basic properties of Rényi-divergences

This section lists some common properties of Rényi divergences that will be used in this thesis. The letters f and g refer to two probability distributions over a space $\mathcal{S}$. We refer the reader to [VH14] for additional information and proofs.

Proposition 1 (Varying the order). *The function* $\alpha \in (0,\infty) \mapsto \mathrm{D}_\alpha(f\|g)$ *is increasing, while* $\alpha \in [0,\infty) \mapsto (1-\alpha)\,\mathrm{D}_\alpha(f\|g)$ *is concave. Finally, the function* $\alpha \mapsto \mathrm{D}_\alpha(f\|g)$ *is continuous on* $(0,1] \cup \{\alpha > 1 \colon \mathrm{D}_\alpha(f\|g) < \infty\}$.

Proposition 2 (Positivity and skew symmetry). *For any* $\alpha > 0$ *it holds* $\mathrm{D}_\alpha(f\|g) \geq 0$. *Moreover, for* $\alpha \in (0,1)$ *we have* $(1-\alpha)\,\mathrm{D}_\alpha(f\|g) = \alpha\,\mathrm{D}_{1-\alpha}(g\|f)$.

Proposition 3 (Orders $\alpha \in (0,1)$ are all equivalent). *For any* $0 < \beta_1 \leq \beta_2 < 1$ *we have*

$$\frac{\beta_1}{\beta_2}\frac{1-\beta_2}{1-\beta_1}\,\mathrm{D}_{\beta_2} \leq \mathrm{D}_{\beta_1} \leq \mathrm{D}_{\beta_2}\,.$$

Proposition 4. *If* $\alpha > 0$ *and* $\alpha \neq 1$ *then*

$$(1-\alpha)\,\mathrm{D}_\alpha(f\|g) \;=\; \inf_h \left\{\alpha\,\mathrm{D}_{\mathrm{KL}}(h\|f) + (1-\alpha)\,\mathrm{D}_{\mathrm{KL}}(h\|g)\right\},$$

where the inf *is taken over all probability distributions on* $\mathcal{S}$ *and with the convention* $\alpha\,\mathrm{D}_{\mathrm{KL}}(h\|f) + (1-\alpha)\,\mathrm{D}_{\mathrm{KL}}(h\|g) = \infty$ *if it were otherwise be undefined. Moreover, if the distribution* $h_\alpha = \frac{f^\alpha g^{1-\alpha}}{\int f^\alpha g^{1-\alpha}}$ *is well defined and* $\mathrm{D}_\alpha(h_\alpha\|f) < \infty$ *or* $\alpha \in (0,1)$ *then the infimum is uniquely achieved by* $h = h_\alpha$.

Proposition 5. *Suppose that* $\mathrm{D}_{\mathrm{KL}}(f\|g) < \infty$. *Then*

$$\sup_{\alpha\in(0,1)} \inf_h \left\{\alpha\,\mathrm{D}_{\mathrm{KL}}(h\|f) + (1-\alpha)\,\mathrm{D}_{\mathrm{KL}}(h\|g)\right\} \;=\; \inf_h \sup_{\alpha\in(0,1)} \left\{\alpha\,\mathrm{D}_{\mathrm{KL}}(h\|f) + (1-\alpha)\,\mathrm{D}_{\mathrm{KL}}(h\|g)\right\}.$$

Moreover, if there exists $\alpha^* \in (0,1)$ *such that* $\mathrm{D}_{\mathrm{KL}}(h_{\alpha^*}\|f) = \mathrm{D}_{\mathrm{KL}}(h_{\alpha^*}\|g)$ *where* $h_\alpha = \frac{f^\alpha g^{1-\alpha}}{\int f^\alpha g^{1-\alpha}}$ *then we have*

$$\sup_{\alpha\in(0,1)} (1-\alpha)\,\mathrm{D}_\alpha(f\|g) \;=\; (1-\alpha^*)\,\mathrm{D}_{\alpha^*}(f\|g) \;=\; \mathrm{D}_{\mathrm{KL}}(h_{\alpha^*}\|f) \;=\; \mathrm{D}_{\mathrm{KL}}(h_{\alpha^*}\|g).$$

The quantity $\sup_{\alpha\in(0,1)} (1-\alpha)\,\mathrm{D}_\alpha(f\|g)$ is called the *Chernoff information* between f and g.

2.1.3 Examples

Bernoulli distributions

Consider two Bernoulli distributions on $\{0,1\}$ defined by $f(x) = (1-p)^{1-x}p^x$ and $g(x) = (1-q)^{1-x}q^x$ for some $p, q \in (0,1)$. For any $\alpha \in (0,\infty)\backslash\{1\}$ we find that

$$\mathrm{D}_\alpha\left(f\|g\right) \;=\; \frac{1}{\alpha-1}\log\left((1-p)^\alpha(1-q)^{1-\alpha} + p^\alpha q^{1-\alpha}\right).$$

Remark 1. In an asymptotic regime with $p, q = o(1)$, Taylor's approximations $\log(1+t) = t + O(t^2)$ and $(1-t)^\alpha = 1 - \alpha t + O(t^2)$ for $t = o(1)$, together with the fact that $p^\alpha q^{1-\alpha} = O(p \vee q)$ for $\alpha \in (0,1)$ imply that

$$\mathrm{D}_\alpha(f\|g) \;=\; \frac{1}{1-\alpha}\left(\alpha p + (1-\alpha)q - p^\alpha q^{1-\alpha}\right) + O\left((p\vee q)^2\right)$$

for $\alpha \in (0,1)$. Especially $\mathrm{D}_{1/2}(f\|g) = \left(\sqrt{p} - \sqrt{q}\right)^2 + O\left((p\vee q)^2\right)$. In an asymptotic regime with $p, q = o(1)$ and $p \asymp q$, the above approximation also holds for $\alpha > 1$.

Observing that $\log \frac{f}{g}(x) = \log \frac{1-p}{1-q} + x\left(\log \frac{p}{q} - \log \frac{1-p}{1-q}\right)$, and noting that $\mathbb{E}_f X = p$ and $\mathrm{Var}_f\, X = p(1-p)$ for a f-distributed random variable X, the Kullback-Leibler divergence between two Bernoulli distributions equals

$$\mathrm{D}_{\mathrm{KL}}(f\|g) \;=\; \mathbb{E}_f\left(\log \frac{f}{g}(X)\right) \;=\; \log\frac{1-p}{1-q} + p\left(\log\frac{p}{q} - \log\frac{1-p}{1-q}\right),$$

and the Kullback-Leibler variation is

$$\mathrm{V}_{\mathrm{KL}}(f\|g) \;=\; \mathrm{Var}_f\left(\log \frac{f}{g}(X)\right) \;=\; p(1-p)\left(\log\frac{p}{q} - \log\frac{1-p}{1-q}\right)^2.$$

Especially, symmetrised versions of the above quantities equal

$$\begin{aligned}
\mathrm{D}^s_{\mathrm{KL}}(f,g) \;&=\; (p-q)\left(\log\frac{p}{q} - \log\frac{1-p}{1-q}\right),\\
\mathrm{V}^s_{\mathrm{KL}}(f,g) \;&=\; (p(1-p) + q(1-q))\left(\log\frac{p}{q} - \log\frac{1-p}{1-q}\right)^2.
\end{aligned}$$

Remark 2. In an asymptotic regime in which $p, q \ll 1$ and $|p-q| \ll p \vee q$, by noting that $\frac{p}{q} = \left(1 - \frac{|p-q|}{p\vee q}\right)^{\mathrm{sgn}(p-q)}$ and $\log(1+t) = t + O(t^2)$ for $t = o(1)$, we find that $\log\frac{p}{q} = \mathrm{sgn}(p-q)\log\left(1 - \frac{|p-q|}{p\vee q}\right) = \frac{p-q}{p\vee q} + O\left(\left(\frac{p-q}{p\vee q}\right)^2\right)$. Therefore $\mathrm{D}^s_{\mathrm{KL}}(f,g) = (1+o(1))\frac{(p-q)^2}{p\vee q}$ and $\mathrm{V}^s_{\mathrm{KL}}(f,g) = (1+o(1))(p+q)\left(\frac{p-q}{p\vee q}\right)^2 = (1+o(1))\frac{(p-q)^2}{p\vee q}$.

Remark 3. In an asymptotic regime in which $p, q \ll 1$ and $p \asymp q$, we find that $\mathrm{D}^s_{\mathrm{KL}}(f,g) = (1+o(1))(p-q)\log\frac{p}{q}$ and $\mathrm{V}^s_{\mathrm{KL}}(f,g) = (1+o(1))(p+q)\log^2\frac{p}{q}$.

Geometric distributions

We will investigate geometric distributions on the strictly positive integers with densities $f(x) = (1-a)a^{x-1}$ and $g(x) = (1-b)b^{x-1}$ for $a, b \in (0,1)$.

For any $\alpha \in (0,1)\cup(1,\infty)$, by applying the geometric sum formula we find that

$$\mathrm{D}_\alpha(f\|g) \;=\; \begin{cases} \frac{1}{\alpha-1}\log\frac{(1-a)^\alpha(1-b)^{1-\alpha}}{1-a^\alpha b^{1-\alpha}} & \text{if } a^\alpha b^{1-\alpha} < 1,\\ \infty & \text{otherwise.}\end{cases}$$

In particular, the Rényi divergence is finite for any $\alpha \in (0,1)$. Furthermore,

$$\mathrm{D}^s_\alpha(f,g) \;=\; \begin{cases} \frac{1}{\alpha-1}\left(\log\frac{1-a}{1-a^\alpha b^{1-\alpha}} + \log\frac{1-b}{1-b^\alpha a^{1-\alpha}}\right) & \text{if } \max\{a^\alpha b^{1-\alpha}, a^{1-\alpha}b^\alpha\} < 1, \\ \infty & \text{otherwise.} \end{cases}$$

Observing that $\log\frac{f}{g}(x) = \log\frac{1-a}{1-b} + (x-1)\log\frac{a}{b}$ and noting that $\mathbb{E}_f X = \frac{1}{1-a}$ and $\mathrm{Var}_f\, X = \frac{a}{(1-a)^2}$, it follows that

$$\begin{aligned} \mathrm{D}_{\mathrm{KL}}(f\|g) \;&=\; \mathbb{E}_f\left(\log\frac{1-a}{1-b} + (X-1)\log\frac{a}{b}\right) \;=\; \log\frac{1-a}{1-b} + \frac{a}{1-a}\log\frac{a}{b}, \\ \mathrm{V}_{\mathrm{KL}}(f\|g) \;&=\; \mathrm{Var}_f\left(\log\frac{1-a}{1-b} + (X-1)\log\frac{a}{b}\right) \;=\; \frac{a}{(1-a)^2}\log^2\frac{a}{b}. \end{aligned}$$

Especially, symmetrised versions of the above quantities equal

$$\mathrm{D}^s_{\mathrm{KL}}(f,g) \;=\; \left(\frac{a}{1-a} - \frac{b}{1-b}\right)\log\frac{a}{b} \quad \text{and} \quad \mathrm{V}^s_{\mathrm{KL}}(f,g) \;=\; \left(\frac{a}{(1-a)^2} + \frac{b}{(1-b)^2}\right)\log^2\frac{a}{b}.$$

Zero-inflated geometric distributions

For $a \in [0,1)$ and $p \in [0,1]$ we define a *zero-inflated geometric distribution* as the probability measure

$$\mathrm{Geo}^*_{a,p} = (1-p)\delta_0 + p\,\mathrm{Geo}_a$$

where Geo_a is geometric on $\{1,2,\cdots\}$ with density $\mathrm{Geo}_a(x) = (1-a)a^{x-1}$ and δ_0 is the Dirac measure at zero. The zero-inflated geometric distribution contains as special cases the Bernoulli distribution $\mathrm{Ber}_p = \mathrm{Geo}^*_{0,p}$ and the geometric distribution $\mathrm{Geo}_a = \mathrm{Geo}^*_{a,1}$.

For parameters $a,b,p,q \in (0,1)$ and an exponent $\alpha \in (0,\infty)\backslash\{1\}$, simple computations show that

$$\mathrm{D}_\alpha\left(\mathrm{Geo}^*_{a,p}\,\|\,\mathrm{Geo}^*_{b,q}\right) \;=\; \frac{1}{\alpha-1}\log\left((1-p)^\alpha(1-q)^{1-\alpha} + p^\alpha q^{1-\alpha} Z_\alpha\left(\mathrm{Geo}_a\,\|\,\mathrm{Geo}_b\right)\right),$$

where

$$Z_\alpha(\mathrm{Geo}_a\,\|\,\mathrm{Geo}_b) \;=\; \begin{cases} \frac{(1-a)^\alpha(1-b)^{1-\alpha}}{1-a^\alpha b^{1-\alpha}} & \text{if } 1-a^\alpha b^{1-\alpha} < \infty, \\ \infty & \text{otherwise.} \end{cases}$$

Remark 4. Consider a regime where $a,b = \Theta(1)$ and $p,q = \Theta(\delta)$ for $\delta = o(1)$. Taylor's expansions $(1-p)^\alpha(1-q)^{1-\alpha} = 1-\alpha p-(1-\alpha)q+O(\delta^2)$ and $p^\alpha q^{1-\alpha} = \Theta(\delta)$ ensures that

$$\begin{aligned} \mathrm{D}_\alpha\left(\mathrm{Geo}^*_{a,p}\,\|\,\mathrm{Geo}^*_{b,q}\right) \;&=\; \frac{\alpha p + (1-\alpha)q}{1-\alpha} - \frac{p^\alpha q^{1-\alpha}}{1-\alpha} Z_\alpha(\mathrm{Geo}_a\,\|\,\mathrm{Geo}_b) + O\left(\delta^2\right) \\ &=\; \mathrm{D}_\alpha\left(\mathrm{Ber}_p\,\|\,\mathrm{Ber}_q\right) + \frac{p^\alpha q^{1-\alpha}}{1-\alpha}\left(1 - Z_\alpha\left(\mathrm{Geo}_a\,\|\,\mathrm{Geo}_b\right)\right) + O\left(\delta^2\right) \end{aligned}$$

where we used the result of Remark 1.

Normal and doubly exponential distribution

This paragraph presents two real-valued distributions f and g such that $\mathrm{D}_{\mathrm{KL}}(f\|g) < \infty$ but $\mathrm{D}_{1+r}(f\|g) = \infty$ for all $r > 0$. This counter-example is taken from [LV06]. Define $f(x) = \mathrm{e}^{-2|x|}$ as the density of a doubly-exponential random variable and let $g(x) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-x^2/2}$ be the standard normal density. Then $Z_{1+r}(f\|g) = \infty$ and hence $\mathrm{D}_{1+r}(f\|g) = \infty$, while $\mathrm{D}_{\mathrm{KL}}(f\|g) < \infty$.

2.2 Graph data and clustered random graph models

2.2.1 Graph data and clustering

A *graph* G is a pair (V, E), where V is a (finite) set, whose elements are called *nodes* (or vertices, or points) and E is a set of ordered node pairs called *edges* (or links, lines, bonds). An edge (i, j) represents a link between two nodes, and we denote $i \sim j$ if $(i, j) \in E$. We will assume that the edges are undirected (that is $i \sim j$ implies $j \sim i$), and that graphs do not possess any self-loops (a link connecting a node with itself), unless otherwise stated.

The *adjacency matrix* A of a graph contains all the information from (V, E). Asuming that each node of V is assigned an arbitrary unique index in $[N]$ where $N = |V|$, we defined $A \in \{0, 1\}^{N\times N}$ such that

$$A_{ij} = \begin{cases} 1 & \text{if } i \sim j, \\ 0 & \text{otherwise.} \end{cases}$$

In particular, A is symmetric when the graph is undirected.

For binary networks, the *degree* d_i of a node i is the number of nodes linked to i, and is given by $d_i = \sum_{j=i}^{N} A_{ij}$. The *degree matrix* D is the diagonal matrix whose diagonal is given by $d = (d_1, \cdots, d_N)$. The graph *standard Laplacian* L (respectively the *normalized Laplacian* $\mathcal{L}$) is defined by $L = D - A$ (respectively by $\mathcal{L} = D^{-1/2}LD^{-1/2}$).

Graph clustering consists in partitioning the node set V into K non-overlapping communities (or clusters) $V_1, \cdots, V_K$ such that $\cup_{k=1}^{K} V_k = V$ and $V_k \cap V_\ell = \emptyset$ for $k \neq \ell$. *Spectral clustering* [SM00; NJW02; Von07] is one of the most popular graph clustering algorithms. It consists in selecting the K eigenvectors $v_1, \cdots, v_K$ of the normalized Laplacian $\mathcal{L}$ (or another graph matrix) associated to the K leading eigenvalues of $\mathcal{L}$, and performing k-means on the row of the matrix $\big(v_1, \cdots, v_K\big)$. In this Thesis, we refer to *spectral method* any graph clustering algorithm that uses the eigenvectors of some graph matrix.

A networks whose edges take value in a interaction space $\mathcal{S}$ is naturally represented by a $\mathcal{S}$-valued array $A \in \mathcal{S}^{N\times N}$. Examples of such networks include edge-labelled networks ($\mathcal{S} = \{0, 1 \ldots, L\}$), weighted networks ($\mathcal{S} = \mathbb{R}_+$), temporal networks with T snapshots ($\mathcal{S} = \{0, 1\}^T$), and multiplex networks with M layers ($\mathcal{S} = \{0, 1\}^M$).

2.2.2 Non-binary stochastic block models

The objective of study is a population of $N \geq 1$ mutually interacting nodes partitioned into $K \geq 1$ disjoint sets called blocks. The partition is represented by a node labelling $\sigma : [N] \to [K]$, so that $\sigma(i)$ indicates the block which contains node i. In line with the classical definition of a stochastic block model [HLL83], we assume that interactions between node pairs can be of arbitrary type, and the set of possible interaction types is a measurable space $\mathcal{S}$. This general setup allows to model usual random graphs ($\mathcal{S} = \{0, 1\}$), edge-labelled random graphs ($\mathcal{S} = \{0, \dots, L\}$, $\mathcal{S} = \mathbb{R}$), multiplex and temporal networks ($\mathcal{S} = \{0, 1\}^T$, $\mathcal{S} = \{0, 1\}^\infty$), and many other settings such as nodes interacting over a continuous time interval. In full generality, such a *stochastic block model (SBM)* is parameterised by a node labelling $\sigma : [N] \to [K]$ and an interaction kernel $(f_{k\ell})$ which is a collection of probability density functions with respect to a common sigma-finite reference measure μ on $\mathcal{S}$, such that $f_{k\ell} = f_{\ell k}$ for all $k, \ell = 1, \dots, K$. These parameters specify a probability measure on a space of observations

$$\mathcal{X} \;=\; \Big\{ x : [N] \times [N] \to \mathcal{S} : \; x_{ij} = x_{ji}, \; x_{ii} = 0 \text{ for all } i, j \Big\}$$

with probability density function

$$P_\sigma(x) \;=\; \prod_{1 \leq i < j \leq N} f_{\sigma(i)\sigma(j)}\,(x_{ij}) \tag{2.2.1}$$

with respect to the $N(N-1)/2$-fold product of the reference measure μ.

Our main focus is on homogeneous models in which the interaction kernel can be represented as

$$f_{k\ell} \;=\; \begin{cases} f & \text{if } k = \ell \\ g & \text{otherwise} \end{cases} \tag{2.2.2}$$

for some probability densities f and g on $\mathcal{S}$, called the intra-block and inter-block interaction distribution, respectively. A *homogeneous SBM* is hence a probability density P_σ on $\mathcal{X}$ specified by (2.2.1)–(2.2.2) and parameterised by a 5-tuple (N, K, σ, f, g). For an observation X distributed according to such P_σ, the entries X_{ij}, $1 \leq i < j \leq N$, are mutually independent, and X_{ij} is distributed according to f when $\sigma(i) = \sigma(j)$, and according to g otherwise.

The node labelling σ representing the block membership structure is considered an unknown parameter to be estimated. When studying the average error rate of estimators, it is natural to regard the node labelling as a random variable distributed according to the uniform distribution $\pi(\sigma) = K^{-N}$ on parameter space

$$\mathcal{Z} \;=\; \Big\{ \sigma : [N] \to [K] \Big\}.$$

In this case the joint distribution of the node labelling and the observed data is characterised by a probability density

$$\mathbb{P}(\sigma, x) \;=\; \pi(\sigma) P_\sigma(x) \tag{2.2.3}$$

on $\mathcal{Z} \times \mathcal{X}$ with respect to $\mathrm{card}_\mathcal{Z} \times \mu$, where $\mathrm{card}_\mathcal{Z}$ is the counting measure on $\mathcal{Z}$.

2.2.3 Community recovery and error indices

Community recovery refers to the estimation of an unknown node labelling $\sigma = (\sigma_i)_{1\le i\le N}$ based on an observed interaction array $X = (X_{ij})_{1\le i<j\le N}$. The community recovery problem then becomes the problem of developing an algorithm $\phi : \mathcal{X} \to \mathcal{Z}$ which maps an observed data array $X = (X_{ij})$ into an estimated node labelling $\hat{\sigma} = \phi(X)$. Stated like this, the recovery problem is ill-posed because the map $\sigma \mapsto P_\sigma$ defined by (2.2.1) is in general non-injective. Therefore, we adopt the common approach in which the goal is to recover the unlabelled block structure, that is, the partition $[\sigma] = \{\sigma^{-1}(k) : k \in [K]\}$, and the estimation error is considered small when $[\hat{\sigma}]$ is close to $[\sigma]$. Accordingly, we define for node labellings $\sigma_1, \sigma_2 : [N] \to [K]$ an error quantity by

$$\mathrm{d}^*_{\mathrm{Ham}}(\sigma_1, \sigma_2) \;=\; \min_{\rho\in \mathrm{Sym}(K)} \mathrm{d}_{\mathrm{Ham}}(\rho \circ \sigma_1, \sigma_2)$$

where $\mathrm{Sym}(K)$ denotes the group of permutations on $[K]$ and $\mathrm{d}_{\mathrm{Ham}}$ refers to the Hamming distance. The above error takes values in $\{0, \dots, N\}$ and depends on its inputs only via the partitions $[\sigma_1]$ and $[\sigma_2]$. The normalised error quantity $N^{-1}\,\mathrm{d}^*_{\mathrm{Ham}}(\sigma_1, \sigma_2)$ is known as the classification error in [For10; MH01].

When analysing the average performance of an estimator, we can view $\hat{\sigma}$ as a $\mathcal{Z}$-valued random variable defined on the observation space $\mathcal{X}$. Then $E_\sigma\,\mathrm{d}^*_{\mathrm{Ham}}(\hat{\sigma}, \sigma)$ equals the expected clustering error given a true parameter σ, and

$$\mathbb{E}\,\mathrm{d}^*_{\mathrm{Ham}}(\hat{\sigma}) \;=\; \sum_{\sigma\in\mathcal{Z}} \pi(\sigma) E_\sigma\,\mathrm{d}^*_{\mathrm{Ham}}(\hat{\sigma}, \sigma)$$

is the average clustering error with respect to the uniform distribution $\pi(\sigma) = K^{-N}$ on the parameter space.

2.2.4 Large-scale models

A large-scale network is represented as a sequence of models $P^{(\eta)}_\sigma$ indexed by a scale parameter $\eta = 1, 2, \dots$ In this setting the model dimensions $N^{(\eta)}, K^{(\eta)}$, the node labelling $\sigma^{(\eta)}$, the interaction densities $f^{(\eta)}, g^{(\eta)}$, as well as the spaces $\mathcal{S}^{(\eta)}, \mathcal{X}^{(\eta)}, \mathcal{Z}^{(\eta)}$ all depend on the scale parameter η. In this setup, an estimator is viewed as a map $\phi^{(\eta)} : \mathcal{X}^{(n)} \to \mathcal{Z}^{(\eta)}$. For nonnegative sequences $a = a^{(\eta)}$ and $b = b^{(\eta)}$ we denote $a = o(b)$ when $\limsup_{\eta\to\infty} a^{(\eta)}/b^{(\eta)} = 0$, and $a = O(b)$ when $\limsup_{\eta\to\infty} a^{(\eta)}/b^{(\eta)} < \infty$. We write $a = \omega(b)$ when $b = o(a)$, $a = \Omega(b)$ when $b = O(a)$, and $a = \Theta(b)$ when $a = O(b)$ and $b = O(a)$. We also denote $a \ll b$ for $a = o(b)$, $a \lesssim b$ for $a = O(b)$, and $a \asymp b$ for $a = \Theta(b)$. To avoid overburdening the notation, the scale parameter is mostly omitted from the notation in what follows.

2.2.5 Consistent estimators

For a large-scale model with $N \gg 1$ nodes, an estimator $\hat{\sigma} = \hat{\sigma}^{(\eta)}$ is called:

(i) *consistent* if $\mathbb{E}\, \mathrm{d}^*_{\mathrm{Ham}}(\hat{\sigma}) = o(N)$;

(ii) *strongly consistent* if $\mathbb{E}\, \mathrm{d}^*_{\mathrm{Ham}}(\hat{\sigma}) = o(1)$;

where we recall that $\mathbb{E}$ refers to the expectation with respect to the probability measure $\mathbb{P}$ defined by (2.2.3). A strongly consistent estimator is also said to achieve *exact recovery*, and a consistent estimator is said to achieve *almost exact recovery* [Abb18].

2.2.6 Information theoretic limits block model clustering

Theorem 1 ([MNS16; ABH16]). *Consider a homogeneous binary SBM, with $K = 2$ and where π is the uniform distribution over the set of equal size blocks. Let $f = \mathrm{Ber}\,(p_{\mathrm{in}})$ and $g = \mathrm{Ber}\,(p_{\mathrm{out}})$ and suppose $N \gg 1$. The following hold.*

- *Suppose $p_{\mathrm{in}} = aN^{-1}$ and $p_{\mathrm{out}} = bN^{-1}$ where a and b depend on the scale parameter η. Then a consistent estimator exists if and only if $\frac{(a-b)^2}{2(a+b)} \gg 1$.*
- *Suppose $p_{\mathrm{in}} = aN^{-1}\log N$ and $p_{\mathrm{out}} = bN^{-1}\log N$ with a, b constants. Then a strongly consistent estimator exist if $\left(\sqrt{a} - \sqrt{b}\right)^2 > 2$, and does not exist if $\left(\sqrt{a} - \sqrt{b}\right)^2 < 2$.*

The result of Theorem 1 has been extended to more than two blocks [AS15], as well as models considering $\mathcal{S} = \{0, 1, \cdots, L\}$ [JL15] or $\mathcal{S} = \mathbb{R}$ [XJL20]. This Thesis extends the study to homogeneous SBM with uniform node labels and general pairwise interaction space $\mathcal{S}$, which eventually might depend on the scale parameter η. Example naturally occurs by considering a temporal network with T_η binary snapshots, where both the number of nodes N_η and the number of snapshots T_η grow unbounded.

The connection between consistency conditions in binary SBM and $\mathcal{S}$-valued SBM is naturally done by considering the Rényi divergences between the probability distributions f, g. Indeed, using the Taylor's expansion of the Rényi divergence between Bernoulli distributions (see Section 2.1.3), we have

$$\mathrm{D}_{1/2}\left(\mathrm{Ber}\left(a\frac{\log N}{N}\right), \mathrm{Ber}\left(b\frac{\log N}{N}\right)\right) = \frac{\log N}{N}\left(\sqrt{a} - \sqrt{b}\right)^2 + O\left(\left(\frac{\log N}{N}\right)^2\right).$$

Thus, the strong consistency threshold can be rewritten as $\mathrm{D}_{1/2}(f, g) \geq (1 + \Omega(1))\frac{K \log N}{N}$. Similarly, the condition for consistent recovery is $\mathrm{D}_{1/2}\,(f, g) \gg N^{-1}$. The role of the Rényi divergence is further amplified by the following mini-max result.

Theorem 2 ([ZZ16]). *Consider a homogeneous binary SBM with connectivity distributions f, g. Let $\Sigma = \left\{\sigma \in [K]^N : \forall k \in [K], |\sigma^{-1}(k)| \in \left[\beta^{-1}\frac{N}{K}, \beta\frac{N}{K}\right]\right\}$ be the space of admissible distributions, where $\beta \in \left[1, \sqrt{5/3}\right]$ is an imbalanced parameter. Let $I = \mathrm{D}_{1/2}(f, g)$, and assume that $\frac{NI}{K \log K} \gg 1$. Then*

$$\inf_{\hat{\sigma}} \sup_{\sigma \in \Sigma} N^{-1} E_\sigma \, \mathrm{d}^*_{\mathrm{Ham}}(\hat{\sigma}, \sigma) \;=\; \begin{cases} \exp\left(-(1-o(1))\frac{NI}{2}\right) & \text{if } K = 2, \\ \exp\left(-(1-o(1))\frac{NI}{\beta K}\right) & \text{if } K \geq 3. \end{cases}$$

Assuming $K \asymp 1$ and $\beta \in [1, \sqrt{5/3}]$, the main implications of Theorem 2 are:

1. *Consistency is impossible* (in a worst-case sense) when $I \lesssim N^{-1}$. In this case, for any estimator $\hat{\sigma}$ there exists a worst-case node labelling σ with imbalance factor at most β, for which the relative error remains bounded away from zero, regardless of the network size. In other words, for any estimator $\hat{\sigma}$ (taking also scale as input), as the scale grows larger and larger, there always exist worst-case node labellings for which the relative error remains bounded away from zero.

2. *Strong consistency is impossible* (in a worst-case sense) when $I \leq (1 - \Omega(1))K\frac{\log N}{N}$ and $K = 2$, or $I \leq (1 - \Omega(1))\beta K\frac{\log N}{N}$ and $K \geq 3$. This means that for any estimator $\hat{\sigma}$ there exists a worst-case node labelling σ with imbalance factor at most β, for which the absolute error remains bounded away from zero.

Chapter 3

Recovery thresholds

In this section, we derive information-theoretic thresholds for consistent recovery in Section 3.1 and specialise in sparse networks in Section 3.2. We propose a polynomial-time algorithm for sparse networks and study its consistency in Section 3.3. To keep the exposition clean, we relegate the proofs in Chapter 4.

3.1 Error bounds and recovery conditions

The following theorem characterises fundamental information-theoretic limits for the recovery of block memberships from data generated by a homogeneous $\mathcal{S}$-valued SBM. It does not make any scaling assumptions on the model dimensions N and K, or on the space of interaction types $\mathcal{S}$, and its proof indicates that maximum likelihood estimators achieve the upper bound.

Theorem 3. *For a homogeneous SBM with N nodes, K blocks, and interaction distributions f, g on a general measurable space $\mathcal{S}$ having Rényi divergence $I = \mathrm{D}_{1/2}(f,g)$, the minimum average classification error among all estimators $\hat{\sigma} : \mathcal{X} \to \mathcal{Z}$ is bounded from below by*

$$\min_{\hat{\sigma}} \mathbb{E}\left(\frac{\mathrm{d}^*_{\mathrm{Ham}}(\hat{\sigma})}{N}\right) \geq \frac{1}{84}K^{-3}e^{-\frac{N}{K}I-\sqrt{8NI_{21}}} - \frac{1}{6}e^{-\frac{N}{8K}}$$

and from above by

$$\min_{\hat{\sigma}} \mathbb{E}\left(\frac{\mathrm{d}^*_{\mathrm{Ham}}(\hat{\sigma})}{N}\right) \leq 8e(K-1)e^{-(1-\zeta-\kappa)\frac{N}{K}I} + K^N e^{-\frac{1}{4}(\frac{\zeta}{K-1}-\epsilon)(N/K)^2 I} + 2Ke^{-\frac{1}{3}\epsilon^2\frac{N}{K}},$$

for all $0 \leq \epsilon \leq \zeta \leq \frac{1}{21}$, where $\kappa = 56\max\{K^2e^{-\frac{NI}{8K}},\, KN^{-1}\}$ and another auxiliary parameter is defined by $I_{21} = \left(\frac{1}{2} - K^{-1}\right)K^{-1}I + \frac{1}{2}K^{-1}J$ with $J = Z_{1/2}(f,g)^{-1}\int\sqrt{fg}\log^2\frac{f}{g}$.

Proof. The lower bound is established in Section 4.1, while the upper bound is analysed in Section 4.2 and follows from Proposition 11. □

Theorem 3 generalises [XJL20, Theorem 5.2] to a quantitative setting which requires neither regularity assumptions on f, g nor restrictions on the underlying space $\mathcal{S}$ of interaction types. This also allows equally sized blocks, unlike in [XJL20]. Note that [XJL20, Theorem 5.2] does not tell what happens for large T in case where $S = \{0,1\}^T$. Theorem 3 is also analogous to [ZZ16, Theorem 2.1 and 2.2] and [PC16, Theorem 6], who studied the binary SBM and multiplex SBM in a minimax framework.

The next key result characterises information-theoretic recovery conditions in large-scale networks, for which we emphasise that the model dimensions $N = N^{(\eta)}$, $K = K^{(\eta)}$, the interaction distributions $f = f^{(\eta)}$, $g = g^{(\eta)}$, and also the interaction type space $\mathcal{S} = \mathcal{S}^{(\eta)}$, are allowed to depend on the scale parameter.

Theorem 4. *Consider a homogeneous SBM with $N \gg 1$ nodes, $K \asymp 1$ blocks, and interaction distributions f and g having Rényi divergence $I = \mathrm{D}_{1/2}(f,g)$. The following hold:*

(i) a consistent estimator exists if $I \gg N^{-1}$, and does not exist if $I \lesssim N^{-1}$;

(ii) a strongly consistent estimator exists if $I \geq (1+\Omega(1))\frac{K\log N}{N}$, and does not exist if $I \leq (1-\Omega(1))\frac{K\log N}{N}$.

Proof. The nonexistence statements are a direct consequence of the lower bound in Theorem 3 combined with Lemma 8. The existence results follow by analysing the upper bound of Theorem 3, which is done in Proposition 12 in Section 4.2. □

Example 1 (Binary interactions)**.** The Rényi divergence of order $\frac{1}{2}$ for Bernoulli distributions with means p and q equals $I = -2\log((1-p)^{1/2}(1-q)^{1/2} + p^{1/2}q^{1/2})$. In a regime where $p = p_0\frac{\log N}{N}$ and $q = q_0\frac{\log N}{N}$ for scale-independent constants $p_0, q_0 > 0$, this is approximated by $I = (\sqrt{p_0}-\sqrt{q_0})^2\frac{\log N}{N} + o\left(\frac{\log N}{N}\right)$. Theorem 4 tells that a strongly consistent estimator exists if $(\sqrt{p_0}-\sqrt{q_0})^2 > K$ and does not exist if $(\sqrt{p_0}-\sqrt{q_0})^2 < K$. This is the well-known threshold for strong consistency in binary SBMs [ABH16; MNS16].

Example 2 (Poisson interactions)**.** The Rényi divergence of order $\frac{1}{2}$ for Poisson distributions with means λ and μ equals $I = (\sqrt{\lambda} - \sqrt{\mu})^2$. In a regime where $\lambda = a\frac{\log N}{N}$ and $\mu = b\frac{\log N}{N}$ for scale-independent constants $a, b > 0$, Theorem 4 tells that a strongly consistent estimator exists for $(\sqrt{a}-\sqrt{b})^2 > K$ and does not exist for $(\sqrt{a}-\sqrt{b})^2 < K$. This is a similar condition as in Example 1, and is due to the fact that Bernoulli distributions with small mean are well approximated by Poisson distributions.

Example 3 (Censored block model)**.** Consider a binary censored block model as in [Dha+22]. A latent graph is generated from a binary SBM with $f = \mathrm{Ber}(p_0)$ and $g = \mathrm{Ber}(q_0)$ and then each pair interaction is revealed independently with probability $r = r_0\frac{\log N}{N}$, where we assume that p_0, q_0, r_0 are scale-independent constants. The resulting observed network is a SBM with $\mathcal{S} = \{\text{present}, \text{absent}, \text{censored}\}$ (where censored denotes the non-observed interactions) and with intra-block and inter-block probability distributions $\tilde{f}$ and $\tilde{g}$. We have $\tilde{f}(\text{present}) = rp_0$, $\tilde{f}(\text{absent}) = r(1-p_0)$ and $\tilde{f}(\text{censored}) = 1-r$, and similarly for $\tilde{g}$. From $\mathrm{D}_{1/2}(\tilde{f}, \tilde{g}) = r\left(\left(\sqrt{p_0} - \sqrt{q_0}\right)^2 + \left(\sqrt{1-p_0} - \sqrt{1-q_0}\right)^2\right) + O(r^2)$ it follows that a strongly consistent estimator exists if $r_0 > r_0^{\mathrm{crit}}$ and does not exist if $r_0 < r_0^{\mathrm{crit}}$, where $r_0^{\mathrm{crit}} = K\left(\left(\sqrt{p_0} - \sqrt{q_0}\right)^2 + \left(\sqrt{1-p_0} - \sqrt{1-q_0}\right)^2\right)^{-1}$. For $K = 2$, this coincides with the critical threshold obtained in [Dha+22].

Example 4 (Multiplex networks with independent layers)**.** The interaction space of a multiplex network with M independent layers is $\mathcal{S} = \{0,1\}^M$, and for $x = (x_1, \dots, x_M) \in \mathcal{S}$ we have $f(x) = \prod_{m=1}^M f_m(x_m)$ and $g = \prod_{m=1}^M g_m(x_m)$. Then $\mathrm{D}_{1/2}(f, g) = \sum_{m=1}^M \mathrm{D}_{1/2}(f_m, g_m)$. Assuming $K \asymp 1$, strong consistency in expectation is possible if

$$\sum_{m=1}^{M} \mathrm{D}_{1/2}(f_m, g_m) \;\geq\; (1 + \Omega(1))\frac{K \log N}{N},$$

and impossible if

$$\sum_{m=1}^{M} \mathrm{D}_{1/2}(f_m, g_m) \;\leq\; (1 - \Omega(1))\frac{K \log N}{N}.$$

Similarly, consistency is possible if $\sum_{m=1}^M \mathrm{D}_{1/2}(f_m, g_m) \gg N^{-1}$ and impossible otherwise. These thresholds are similar to the minimax error rates reported in [PC16, Theorem 6]. An important extension here is that the above characterisation remains valid if $\{0,1\}^M$ is replaced by $\mathcal{S}_0^M$, where $\mathcal{S}_0$ is an arbitrary measurable space.

3.2 Recovery conditions for sparse non-binary SBM

Sparse networks can be modelled using interaction distributions f and g for which $p = 1 - f(0)$ and $q = 1 - g(0)$ are close to zero, where $0 \in \mathcal{S}$ is an element representing no-interaction. We can represent the interaction distributions as

$$f \;=\; (1-p)\delta_0 + p\tilde{f} \qquad \text{and} \qquad g \;=\; (1-q)\delta_0 + q\tilde{g}, \tag{3.2.1}$$

where the probability distributions $\tilde{f}$ and $\tilde{g}$ may depend on the scale parameter η and $\tilde{f}(0) = \tilde{g}(0) = 0$. Here $\tilde{f}$ and $\tilde{g}$ represent conditional distributions of intra-block and inter-block interactions given that there is an interaction. In a sparse setting where $p, q = O(\rho)$ for some $\rho \ll 1$, Taylor's approximations show that the Rényi divergence of order $\frac{1}{2}$ is given by

$$D_{1/2}(f, g) = (\sqrt{p} - \sqrt{q})^2 + 2\sqrt{pq}\,\mathrm{Hel}^2(\tilde{f}, \tilde{g}) + O(\rho^2). \tag{3.2.2}$$

Since $\mathrm{D}_{1/2}(\mathrm{Ber}_p, \mathrm{Ber}_q) = (\sqrt{p} - \sqrt{q})^2 + O(\rho^2)$, the Hellinger distance $\mathrm{Hel}(\tilde{f}, \tilde{g})$ characterises the information gained by observing types of interactions, compared to the binary data corresponding to just observing whether or not there is an interaction.

The following proposition states consistency thresholds when the interactions distributions are given by (3.2.1).

Proposition 6. *Consider a homogeneous SBM with $N \gg 1$ nodes, $K \asymp 1$ blocks, and with f, g given by (3.2.1). Suppose that $p = p_0\rho$ and $q = q_0\rho$ with p_0, q_0 constants and $\rho \ll 1$. Let $\tilde{I} = \left(\sqrt{p_0} - \sqrt{q_0}\right)^2 + 2\sqrt{p_0 q_0}\,\mathrm{Hel}^2(\tilde{f}, \tilde{g})$. Then the following hold:*

(i) a consistent estimator exists if $\rho\tilde{I} \gg N^{-1}$, and does not exist if $\rho\tilde{I} \lesssim N^{-1}$;

(ii) a strongly consistent estimator exist if $\rho\tilde{I} \geq (1 + \Omega(1))\frac{K \log N}{N}$, and does not exist if $\rho\tilde{I} \leq (1 - \Omega(1))\frac{K \log N}{N}$.

Proof. Equation (3.2.2) shows that $\mathrm{D}_{1/2}(f, g) = \rho\tilde{I} + O(\rho^2)$, and hence points (i) and (ii) are straightforward consequences of Corollary 4. □

Remark 5. Proposition 6 generalizes [JL15, Theorems 3.3 and 3.4], which only study strong consistency in probability and further restrict $\mathcal{S} = \{0, 1, \cdots, L\}$ while imposing $\tilde{f}(\ell) > 0$ and $\tilde{g}(\ell) > 0$ for all $\ell \in \{1, \ldots, L\}$.

3.3 Clustering sparse SBMs in polynomial time

To cluster sparse SBM in a polynomial time in N, we propose Algorithm 2 which employs spectral clustering as a subroutine to produce a moderately accurate initial clustering, and then performs a refinement step through node-wise likelihood maximisation [Gao+17; XJL20]. Similarly to [Gao+17; XJL20], for technical reasons related to the proofs, the initialisation step of Algorithm 2 involves N separate spectral clustering steps. A consensus step is therefore needed at the end, to correctly permute the individual predictions. Numerical experiments indicate that in practice it often suffices to do one spectral clustering on a binary matrix, and remove this consensus step.

Algorithm 2: Clustering using general $\mathcal{S}$-valued interaction data

Input: $\mathcal{S}$-valued interaction matrix A_{ij}; interaction distributions f, g.
Output: Estimated node labelling $\hat{\sigma}$.

Step 1: Coarse clustering using binary interaction data

Compute a binary matrix $\tilde{A}$ by setting $\tilde{A}_{ij} = 1(A_{ij} \neq 0)$ for all i, j
for $i = 1, \dots, N$ **do**
    Let $\tilde{A}_{-i}$ be the submatrix of $\tilde{A}$ with row i and column i removed.
    Compute a node labelling $\tilde{\sigma}_i$ on $[N] \setminus \{i\}$ by applying a standard graph clustering algorithm with adjacency matrix $\tilde{A}_{-i}$.

Step 2: Refined clustering using full interaction data

for $i = 1, \dots, N$ **do**
    Compute $h_i(k) = \sum_{j:\tilde{\sigma}_i(j)=k} \log \frac{f(A_{ij})}{g(A_{ij})}$ for all $k \in [K]$.
    Set $\hat{\sigma}_i(i) = \arg\max_{k\in[K]} h_i(k)$ with arbitrary tie breaks.
    Set $\hat{\sigma}_i(j) = \tilde{\sigma}_i(j)$ for $j \neq i$.

Step 3: Consensus

Select $\hat{\sigma}_1$ as a baseline node labelling and set $\hat{\sigma}(1) = \hat{\sigma}_1(1)$.
for $i = 2, \dots, N$ **do**
    Set $\hat{\sigma}(i) = \arg\max_\ell |\hat{\sigma}_i^{-1}(\hat{\sigma}_i(i)) \cap \hat{\sigma}_1^{-1}(\ell)|$ with arbitrary tie breaks.

The following theorem characterises the accuracy of Algorithm 2 for sparse large-scale models, and implies that under some technical conditions this algorithm achieves the exponential error rate in Theorem 3. The proof of Theorem 5 is given in Section 4.3.

Theorem 5. *Consider a homogeneous SBM with $N \gg 1$ nodes, $K \asymp 1$ blocks, and sparse interaction distributions f and g given by (3.2.1), having Rényi divergence $I = \mathrm{D}_{1/2}(f, g)$. Assume that $p \vee q \gtrsim N^{-1}$ and $\frac{(p-q)^2}{p\vee q} \gg N^{-1} \frac{\mathrm{D}^s_{1+r}(f,g)}{\mathrm{D}^s_r(f,g)}$ for some $r \in \left(0, \frac{1}{2}\right]$. Then the classification error of Algorithm 2 is bounded by*

$$\mathbb{E}\left(\frac{\mathrm{d}^*_{\mathrm{Ham}}(\hat{\sigma})}{N}\right) \leq Ke^{-(1-o(1))2r\frac{NI}{K}} + o(1).$$

Theorem 5 is similar in spirit to upper bounds in [XJL20] and [YP16] but is fundamentally different in that it makes neither assumptions about truncating the label space $\mathcal{S}$ nor any assumptions about the regularity of the interaction distributions f, g. Moreover, for temporal binary interactions with $\mathcal{S} = \{0, 1\}^T$, the algorithms in [XJL20] are of exponential complexity in T and the analytical techniques do not allow to consider regimes with $T \gg 1$. An open future research problem is to derive an upper bound in the case with only one initial clustering step.

When f, g are given by (3.2.1), a Taylor expansion of $\mathrm{D}_\alpha(f\|g)$ for $\alpha \in (0, \infty)\backslash\{1\}$ leads to

$$\mathrm{D}_\alpha(f\|g) = \frac{p^\alpha q^{1-\alpha}}{\alpha - 1} Z_\alpha(\tilde{f}\|\tilde{g}) - \frac{\alpha p + (1-\alpha)q}{\alpha - 1} + O\left(\rho^2\right),$$

where $Z_\alpha(\tilde{f}\|\tilde{g}) = \int \tilde{f}^\alpha \tilde{g}^{1-\alpha}$. If $p = p_0\rho$ and $q = q_0\rho$ where p_0, q_0 are constants, and $Z_\alpha(\tilde{f}\|\tilde{g}) = O(1)$, then $\mathrm{D}_\alpha(f\|g) \asymp \rho$. Hence, $\mathrm{D}^s_{1+r}(f,g) \asymp \mathrm{D}^s_r(f,g)$ for every r such that $\mathrm{D}^s_{1+r}(\tilde{f},\tilde{g}) = O(1)$.

Proposition 7. *Consider a homogeneous SBM with $N \gg 1$ nodes, $K \asymp 1$ blocks, and interaction distributions f and g having Rényi divergence $I = \mathrm{D}_{1/2}(f,g)$ and form (3.2.1). Suppose that $p = p_0\rho$ and $q = q_0\rho$ with $p_0 \neq q_0$ constants and $\rho \ll 1$. Let*

$$\tilde{I} = (\sqrt{p_0} - \sqrt{q_0})^2 + 2\sqrt{p_0 q_0}\,\mathrm{Hel}^2(\tilde{f},\tilde{g}).$$

Then, the following hold:

(i) if $\mathrm{D}^s_{1+r}(\tilde{f},\tilde{g}) = O(1)$ for some constant $r \in \left(0, \frac{1}{2}\right]$, then Algorithm 2 is consistent for $\rho\tilde{I} \gg N^{-1}$;

(ii) if $\mathrm{D}^s_{3/2}(\tilde{f},\tilde{g}) = O(1)$, then Algorithm 2 is strongly consistent for $\rho\tilde{I} \geq (1+\Omega(1))\frac{K\log N}{N}$.

Proof. Firstly, if $\rho \lesssim N^{-1}$ then consistency is never possible (see e.g., point (i) of Proposition 6). We can thus assume $\rho \gg N^{-1}$. Let $r \in \left(0, \frac{1}{2}\right]$ such that $\mathrm{D}^s_{1+r}(\tilde{f},\tilde{g}) = O(1)$. Then $\mathrm{D}^s_{1+r}(f,g) \asymp \mathrm{D}^s_r(f,g)$. Combined with $p_0 \neq q_0$, this implies that $\frac{(p-q)^2}{p\vee q} = \rho\frac{(p_0-q_0)^2}{p_0\vee q_0} \asymp \rho \gg N^{-1}\frac{\mathrm{D}_{1+r}(f,g)}{\mathrm{D}_r(f,g)}$, and we can apply Theorem 5 to prove (i). To prove (ii), we need to apply Theorem 5 with $r = \frac{1}{2}$, so that Algorithm 2 achieves the threshold for strong consistency. This is possible under the assumption $\mathrm{D}^s_{3/2}(\tilde{f},\tilde{g}) = O(1)$. □

We will next illustrate by examples the applicability of Theorem 5 and Proposition 7 in various contexts involving discrete labels and continuous weights.

Example 5 (Sparse categorical interactions)**.** Consider a categorical stochastic block model with intra- and inter-block interactions distributed according to

$$f = (1-\rho p_0)\delta_0 + \rho p_0 \tilde{f} \quad \text{and} \quad g = (1-\rho q_0)\delta_0 + \rho q_0 \tilde{g},$$

in which $\tilde{f}$ and $\tilde{g}$ are probability distributions on $\{1,\dots,L\}$, not depending on scale, such that $\tilde{f}(\ell), \tilde{g}(\ell) > 0$ for all ℓ. This is the model studied in [JL15]. In this case $\mathrm{D}^s_{3/2}(\tilde{f},\tilde{g})$ is finite, and the critical information quantity defined in Proposition 7 equals

$$\tilde{I} = (\sqrt{p_0} - \sqrt{q_0})^2 + \sqrt{p_0 q_0}\sum_{\ell=1}^{L}\left(\sqrt{\tilde{f}(\ell)} - \sqrt{\tilde{g}(\ell)}\right)^2.$$

Proposition 7 then tells that Algorithm 2 is consistent for $\rho\tilde{I} \gg N^{-1}$ and strongly consistent for $\rho\tilde{I} \geq (1+\Omega(1))\frac{K\log N}{N}$. In the degenerate case with $L = 1$, we see that $\tilde{I} = \left(\sqrt{p_0} - \sqrt{q_0}\right)^2$ and for $\rho = \frac{\log N}{N}$ we recover the well-known condition $\left(\sqrt{p_0} - \sqrt{q_0}\right)^2 > K$ in Example 1.

Example 6 (Sparse geometric interactions)**.** Consider an integer-valued stochastic block model with intra- and inter-block interactions distributed according to

$$f \;=\; (1-\rho p_0)\delta_0 + \rho p_0\,\mathrm{Geo}(a) \quad \text{and} \quad g \;=\; (1-\rho q_0)\delta_0 + \rho q_0\,\mathrm{Geo}(b),$$

for some scale-independent constants $a, b \in (0,1)$, where $\mathrm{Geo}(a)$ denotes a geometric distribution on $\{1,2,\dots\}$ with probability mass function $x \mapsto (1-a)a^{x-1}$. Basic computations show that the critical information quantity defined in Proposition 7 equals

$$\tilde{I} \;=\; \left(\sqrt{p_0}-\sqrt{q_0}\right)^2 + 2\sqrt{p_0 q_0}\left(1 - \frac{(1-a)^{1/2}(1-b)^{1/2}}{1-a^{1/2}b^{1/2}}\right).$$

Furthermore, $\mathrm{D}^s_{1+r}(\mathrm{Geo}(a),\mathrm{Geo}(b)) < \infty$ if and only if $b^{\frac{1+r}{r}} < a < b^{\frac{r}{1+r}}$, see Figure 3.1. For any $a, b \in (0,1)$, this condition holds for a small enough $r > 0$. Proposition 7 then tells that Algorithm 2 is consistent when $\rho\tilde{I} \gg 1$, and strongly consistent when $\rho\tilde{I} \geq (1+\Omega(1))\frac{K\log N}{N}$ and $b^3 < a < b^{1/3}$. The more a and b differ from each other, the easier it is to distinguish samples from the geometric distributions. On the other hand, very large differences between a and b might imply divergences in the likelihood ratios used by Algorithm 2. Such cases are ruled out by the extra condition $b^3 < a < b^{1/3}$.

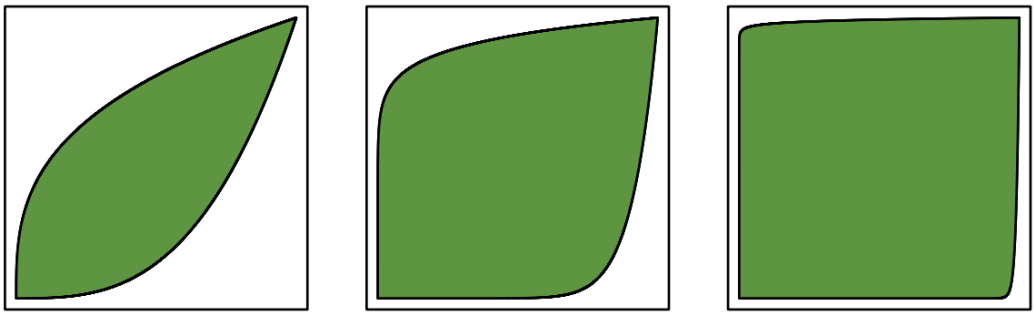


Figure 3.1: Shaded areas display (a,b) pairs in the unit rectangle satisfying $b^{\frac{1+r}{r}} < a < b^{\frac{r}{1+r}}$ for $r = 0.5$ (left), $r = 0.1$ (center) and $r = 0.01$ (right).

Example 7 (Sparse normal interactions)**.** Consider a real-valued stochastic block model with intra- and inter-block interactions distributed according to

$$f \;=\; (1-\rho p_0)\delta_0 + \rho p_0 \tilde{f} \quad \text{and} \quad g \;=\; (1-\rho q_0)\delta_0 + \rho q_0 \tilde{g},$$

in which $\tilde{f} = \mathrm{Nor}(0,\sigma^2)$ and $\tilde{g} = \mathrm{Nor}(0,\tau^2)$ are zero-mean normal distributions, and we assume that $p_0 \neq q_0$ and $\sigma \neq \tau$ are scale-independent constants. The critical information quantity defined in Proposition 7 then equals

$$\tilde{I} \;=\; \left(\sqrt{p_0}-\sqrt{q_0}\right)^2 + 2\sqrt{p_0 q_0}\left(1 - \sqrt{\frac{\sigma\tau}{\frac{1}{2}(\sigma^2+\tau^2)}}\right).$$

Simple computations show that the Rényi divergences between $\tilde{f}$ and $\tilde{g}$ are given by $\mathrm{D}_\alpha(\tilde{f}\|\tilde{g}) = \frac{1}{\alpha-1}\log\frac{\sigma^{1-\alpha}\tau^\alpha}{\xi_\alpha}$ when $\xi_\alpha^2 = \alpha\sigma^2 + (1-\alpha)\tau^2 > 0$, and by $\mathrm{D}_\alpha(\tilde{f}\|\tilde{g}) = \infty$ otherwise. Thus, the

symmetric Rényi divergence $\mathrm{D}^s_{1+r}(\tilde{f},\tilde{g})$ is finite if and only if $\frac{\max\{\sigma,\tau\}}{\min\{\sigma,\tau\}} < \sqrt{1+1/r}$. Proposition 7 then implies that Algorithm 2 is consistent when $\rho\tilde{I} \gg N^{-1}$, and strongly consistent when $\rho\tilde{I} \geq (1+\Omega(1))\frac{K\log N}{N}$ and $\frac{\sigma\vee\tau}{\sigma\wedge\tau} < \sqrt{3}$. The more σ and τ differ from each other, the easier it is to distinguish samples from the normal distributions $\tilde{f}$ and $\tilde{g}$. On the other hand, very large differences between σ and τ might imply divergences in the likelihood ratios used in Algorithm 2. Such cases are ruled out by the extra condition $\frac{\max\{\sigma,\tau\}}{\min\{\sigma,\tau\}} < \sqrt{3}$.

Example 8. Consider a real-valued stochastic block model with intra- and inter-block interactions distributed according to

$$f \;=\; (1-\rho p_0)\delta_0 + \rho p_0 \tilde{f} \quad \text{and} \quad g \;=\; (1-\rho q_0)\delta_0 + \rho q_0 \tilde{g},$$

in which $\tilde{f}(x) = e^{-2|x|}$ and $\tilde{g}(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$ are probability densities over $\mathbb{R}$. We notice that $D_{1+r}(f\|g) = \infty$ for every $r > 0$. Hence in that case, Proposition 7 cannot say if Algorithm 2 is consistent or not.

3.4 Temporal networks with sparse Markov interactions

As an instance of a network where interactions are correlated over time, we investigate a *Markov SBM* with interaction space $\mathcal{S} = \{0,1\}^T$ in which intra- and inter-block distributions are given by

$$f \;=\; \mu_{x_1} P_{x_1,x_2} \cdots P_{x_{T-1},x_T}, \qquad \text{and} \qquad g \;=\; \nu_{x_1} Q_{x_1,x_2} \cdots Q_{x_{T-1},x_T}, \tag{3.4.1}$$

where μ,ν are initial probability distributions on $\{0,1\}$ and P,Q are stochastic matrices on $\{0,1\}$. The following model instances deserve special attention.

Example 9 (Static SBM)**.** If P and Q are identity matrices, then there is no temporal activity, so that $A^1 = \cdots = A^T$ almost surely. Hence the model reduces to a static homogeneous SBM.

Example 10 (Independent SBM samples)**.** If the rows of P are equal, and the rows of Q are equal, then there is no temporal dependence, and $A^1,\ldots,A^T$ are mutually independent. This corresponds to observing T independent samples from a static SBM.

Example 11 (Static intra-block linkage)**.** If P is the identity matrix, and the rows of Q are identical, then the link states within blocks remain constant over time, whereas inter-block interactions may be considered as temporally uncorrelated noise.

A Markov SBM is an instance of the general SBM model in which the symmetric Rényi divergence between interaction distributions, a key quantity in Theorem 4, equals

$$D_{1/2}(f,g) \;=\; -2\log\left(\sum_{x\in\{0,1\}^T} (\mu_{x_1}\nu_{x_1})^{1/2} \prod_{t=2}^{T} (P_{x_{t-1}x_t} Q_{x_{t-1}x_t})^{1/2}\right). \tag{3.4.2}$$

In a sparse regime, the probability of observing a non-zero interaction between any particular pair of nodes is small, *i.e.,*

$$\max\{\mu_1, \nu_1, P_{01}, Q_{01}\} \;\le\; \rho \tag{3.4.3}$$

where $\rho = \rho_\eta \to 0$ as $\eta \to \infty$. One particular case is to assume that for some constants $u, v, p_{01}, q_{01} \in (0, \infty)$,

$$\mu_1 = u\rho, \quad \nu_1 = v\rho, \quad P_{01} = p_{01}\rho, \quad Q_{01} = q_{01}\rho. \tag{3.4.4}$$

Under this assumption, the expected number of 1's in a f-distributed signal is $\mathbb{E}\sum_{t=1}^{T} X_t \le \mu_1 + (T-1)P_{01} = O(\rho T)$. Hence when $\rho T = o(1)$, the probability of observing an interaction between any particular node pair is small. We will often make the assumption that the chains are indexed by a scale parameter η and verify

$$\rho_\eta T_\eta \ll 1. \tag{3.4.5}$$

3.4.1 Rényi divergence between sparse chains

The following result presents a key approximation formula with proof provided in Section 4.4.

Proposition 8. *Consider binary Markov chains with initial distributions μ, ν and transition probability matrices P, Q. Assume that $\mu_1, \nu_1, P_{01}, Q_{01} \le \rho$ for some ρ such that $\rho T \le 0.01$. Then the Rényi divergence (3.4.2) is approximated by $|D_{1/2}(f, g) - I| \le 92(\rho T)^2$, where*

$$\begin{aligned} I \;=\; & (\sqrt{\mu_1} - \sqrt{\nu_1})^2 + \left(\left(\sqrt{P_{01}} - \sqrt{Q_{01}}\right)^2 + 2H_{11}^2\sqrt{P_{01}Q_{01}}\right)(T-1) \\ & + 2\left(\Gamma\sqrt{\mu_1\nu_1} - \sqrt{P_{01}Q_{01}}\right)H_{11}^2 \sum_{t=0}^{T-2}(1-\Gamma)^t \end{aligned} \tag{3.4.6}$$

is defined in terms of $H_{11}^2 = 1 - \frac{\sqrt{1-P_{11}}\sqrt{1-Q_{11}}}{1-\sqrt{P_{11}Q_{11}}}$, and $\Gamma = 1 - \sqrt{P_{11}Q_{11}}$.

The quantities in (3.4.6) can be understood as follows. With the help of Taylor's approximations we see that

$$\begin{aligned} (\sqrt{\mu_1} - \sqrt{\nu_1})^2 \;&=\; \mathrm{D}_{1/2}(\mathrm{Ber}(\mu_1) \| \,\mathrm{Ber}(\nu_1)) + O(\rho^2), \\ (\sqrt{P_{01}} - \sqrt{Q_{01}})^2 \;&=\; \mathrm{D}_{1/2}(\mathrm{Ber}(P_{01}) \| \,\mathrm{Ber}(Q_{01})) + O(\rho^2), \\ H_{11} \;&=\; \mathrm{Hel}(\mathrm{Geo}(P_{11}), \mathrm{Geo}(Q_{11})). \end{aligned}$$

We also note that $\Gamma = 1 - \sqrt{P_{11}Q_{11}}$ may be interpreted as an effective spectral gap averaged over the two Markov chains[1].

[1]The nontrivial eigenvalues of transition matrices P and Q can be written as $\Lambda_P = P_{11} - P_{01}$ and $\Lambda_Q = Q_{11} - Q_{01}$. These are nonnegative when $P_{01} \le P_{11}$ and $Q_{01} \le Q_{11}$. The absolute spectral gaps characterising the mixing rates of these chains [LPW08] are then $\Gamma_P = 1 - \Lambda_P$ and $\Gamma_Q = 1 - \Lambda_Q$. When $P_{01} \ll P_{11}$ and $Q_{01} \ll Q_{11}$, we find that $\Gamma = 1 - \sqrt{P_{11}Q_{11}} = 1 - (\Lambda_P\Lambda_Q)^{1/2} + o(1)$.

3.4.2 Bounded time horizon

Consider a Markov SBM in which $T = O(1)$ is a scale-independent constant, and

$$\begin{aligned} \mu_1 &= u\rho + o(\rho), & P_{01} &= p_{01}\rho + o(\rho), & H_{11} &= h_{11} + o(1), \\ \nu_1 &= v\rho + o(\rho), & Q_{01} &= q_{01}\rho + o(\rho), & \Gamma &= \gamma + o(1), \end{aligned} \tag{3.4.7}$$

for some constants $u, v, p_{01}, q_{01}, h_{11}, \gamma$, and define a constant $\tilde{I}$ by

$$\begin{aligned} \tilde{I} &= \left(\sqrt{u} - \sqrt{v}\right)^2 + \left(\left(\sqrt{p_{01}} - \sqrt{q_{01}}\right)^2 + 2h_{11}^2\sqrt{p_{01}q_{01}}\right)(T-1) \\ &\quad + 2h_{11}^2\left(\gamma\sqrt{uv} - \sqrt{p_{01}q_{01}}\right)\sum_{t=0}^{T-2}(1-\gamma)^t. \end{aligned} \tag{3.4.8}$$

Theorem 6. *Consider a Markov SBM with $N \gg 1$ nodes, $K = O(1)$ blocks, and $T = O(1)$ snapshots, and assume that (3.4.7) holds for some constants $u, v, p_{01}, q_{01}, h_{11}, \gamma \geq 0$ such that $\tilde{I} \neq 0$, and some $\rho \ll 1$. Then:*

(i) a consistent estimator does not exist for $\rho \lesssim \frac{1}{N}$ and does exist for $\rho \gg \frac{1}{N}$;

(ii) a strongly consistent estimator does not exist for $\rho \ll \frac{\log N}{N}$ and does exist for $\rho \gg \frac{\log N}{N}$;

(iii) in a critical regime with $\rho = \frac{\log N}{N}$, a strongly consistent estimator does not exist for $\tilde{I} < K$ and does exist for $\tilde{I} > K$.

If we further assume that $u, v, p_{01}, q_{01} > 0$, $u + (T-1)p_{01} \neq v + (T-1)q_{01}$, $P_{10} \asymp Q_{10}$, and $P_{11} \asymp Q_{11}$, then Algorithm 2 is consistent when $\rho \gg \frac{1}{N}$; and strongly consistent when $\rho \gg \frac{\log N}{N}$, or when $\rho = \frac{\log N}{N}$ and $\tilde{I} > K$.

Proof. By Proposition 8, we find that

$$D_{1/2}(f, g) = (1 + o(1))\tilde{I}\rho + O\left(\rho^2\right).$$

The assumption that $\tilde{I} \neq 0$ now implies that $D_{1/2}(f, g) = (1 + o(1))\tilde{I}\rho$. The claims (i)–(iii) now follow Theorem 4.

Let us now impose the further extra assumptions of the theorem. In this case may fix a constant $M \geq 1$ such that $M^{-1} \leq \frac{\mu_1}{\nu_1}, \frac{P_{01}}{Q_{01}}, \frac{P_{10}}{Q_{10}} \leq M$. Moreover, the assumption $\gamma > 0$ implies that P_{11} and Q_{11} cannot both go to one. Thus, we may choose a $\beta \in [0, 1]$ such that $P_{11}^{3/2}Q_{11}^{\beta-3/2} \neq 1 + o(1)$. Denote $\Lambda = P_{11}^{3/2}Q_{11}^{-1/2}$. Because $P_{11} \asymp Q_{11}$, we find that $\Lambda \lesssim 1$. Proposition 14 then implies that $D_{3/2}(f\|g) \lesssim \rho$. A similar argument shows that $D_{3/2}(g\|f) \lesssim \rho$ as well. Therefore, $\frac{\mathrm{D}^s_{3/2}(f,g)}{\mathrm{D}_{1/2}(f,g)} \lesssim 1$. Taylor's approximations further show that the intra- and inter-block probabilities $p = 1 - (1-\mu_1)(1-P_{01})^{T-1}$ and $q = 1 - (1-\nu_1)(1-Q_{01})^{T-1}$ of observing a nonzero interaction pattern satisfy $p = (u + (T-1)p_{01})\rho + o(\rho)$ and

$q = (v + (T-1)q_{01})\rho + o(\rho)$. It follows that $p, q \asymp \rho$ and $p - q \asymp \rho$. When we assume that $\rho \gg \frac{1}{N}$, it follows that $p \vee q \gg N^{-1}$ and $\frac{(p-q)^2}{p \vee q} \asymp \rho$. We will apply Theorem 5 to conclude that Algorithm 2 is consistent when $\rho \gg \frac{1}{N}$, and strongly consistent when $\rho \gg \frac{\log N}{N}$, or when $\rho = \frac{\log N}{N}$ and $\tilde{I} > K$. □

Remark 6. The point (i) of Theorem 6 shows that consistency criteria is the same as for a binary SBM. Hence observing a finite number of snapshots does not change the consistency threshold.

Remark 7. Theorem 6 shows that the critical network density for strong consistency is $\rho = \frac{\log N}{N}$. In this regime, the the existence of a strongly consistent estimator is determined by $\tilde{I}$ defined in (3.4.8). The first term of $\tilde{I}$ equals $(\sqrt{u} - \sqrt{v})^2$ and accounts for the first snapshot: for $T = 1$ we recover the known threshold for strong consistency in the binary SBM [ABH16; MNS16]. Each additional snapshot adds to $\tilde{I}$ an extra term of size $\tilde{I}_t$ defined by

$$\tilde{I}_t \;=\; (\sqrt{p_{01}} - \sqrt{q_{01}})^2 + 2h_{11}^2\sqrt{p_{01}q_{01}} + 2h_{11}^2\Big(\gamma\sqrt{uv} - \sqrt{p_{01}q_{01}}\Big)(1-\gamma)^{t-2}.$$

In particular, $\tilde{I}_t$ is bounded by

$$(\sqrt{p_{01}} - \sqrt{q_{01}})^2 + 2c_1h_{11}^2 \;\le\; \tilde{I}_t \;\le\; (\sqrt{p_{01}} - \sqrt{q_{01}})^2 + 2c_2h_{11}^2$$

with $c_1 = \min\{\sqrt{p_{01}q_{01}}, \gamma\sqrt{uv}\}$ and $c_2 = \max\{\sqrt{p_{01}q_{01}}, \gamma\sqrt{uv}\}$. The extra term is zero when $p_{01} = q_{01}$ and $h_{11} = 0$. Notably, if the left side above is nonzero, then there exists a finite threshold T^* such that strong consistency is possible for $T \ge T^*$.

Remark 8. In a special case of (3.4.7) with $p_{01} = u$, $q_{01} = v$, $h_{11} = 0$, and $\gamma = 1$, the critical information quantity in (3.4.8) equals $\tilde{I} = T(\sqrt{u} - \sqrt{v})^2$. This coincides with multiplex networks composed of T independent layers studied in Example 4. This is also what we would obtain when studying transition matrices P and Q corresponding to independent Bernoulli sequences with means $\mu_1 = u\rho + o(\rho)$ and $\nu_1 = v\rho + o(\rho)$, because in this case $\Gamma = 1 - \sqrt{P_{11}Q_{11}} = 1 - O(\rho)$ and $H_{11} = \mathrm{Hel}(\mathrm{Geo}(\mu_1), \mathrm{Geo}(\nu_1)) = O(\rho)$.

3.4.3 Long time horizon

Consider a Markov SBM with $T \gg 1$ snapshots in which

$$P_{01} \;=\; p_{01}\rho + o(\rho), \qquad Q_{01} \;=\; q_{01}\rho + o(\rho), \qquad H_{11} \;=\; h_{11} + o(1), \tag{3.4.9}$$

for some constants p_{01}, q_{01}, h_{11}, and define

$$\tilde{I} \;=\; (\sqrt{p_{01}} - \sqrt{q_{01}})^2 + 2h_{11}^2\sqrt{p_{01}q_{01}}. \tag{3.4.10}$$

In the following result we assume that the effective spectral gap $\Gamma = 1 - \sqrt{P_{11}Q_{11}}$ satisfies $\Gamma \gg T^{-1}$ which guarantees that both Markov chains mix fast enough, and we may ignore the role of initial states. Indeed, in this case, expression (3.4.6) shows that

$$\mathrm{D}_{1/2}(f, g) \;=\; \tilde{I}\rho T + o(\rho T). \tag{3.4.11}$$

Theorem 7. *Consider a Markov SBM with $N \gg 1$ nodes, $K = O(1)$ blocks, and $T \gg 1$ snapshots, and assume that $\mu_1, \nu_1 \lesssim \rho$ and (3.4.9) holds for some constants $p_{01}, q_{01}, h_{11} \geq 0$ such that $\tilde{I} \neq 0$. Assume also that $\rho \ll T^{-1} \ll 1 - \sqrt{P_{11}Q_{11}}$. Then:*

(i) a consistent estimator does not exist for $\rho \lesssim \frac{1}{NT}$ and does exist for $\rho \gg \frac{1}{NT}$;

(ii) a strongly consistent estimator does not exist for $\rho \ll \frac{\log N}{NT}$ and does exist for $\rho \gg \frac{\log N}{NT}$;

(iii) in a critical regime with $\rho = \frac{\log N}{NT}$, a strongly consistent estimator does not exist for $\tilde{I} < K$ and does exist for $\tilde{I} > K$.

If we further assume that $p_{01}, q_{01} > 0$ and $p_{01} \neq q_{01}$, $\mu_1 \asymp \nu_1$, $P_{10} \asymp Q_{10}$, and that

$$(1 + \Omega(1))P_{11}^3 \;\leq\; Q_{11} \;\leq\; (1 - \Omega(1))P_{11}^{1/3}, \tag{3.4.12}$$

then Algorithm 2 is consistent when $\rho \gg \frac{1}{NT}$; and strongly consistent when $\rho \gg \frac{\log N}{NT}$, or when $\rho = \frac{\log N}{NT}$ and $\tilde{I} > K$.

Proof. By Proposition 8, we find that

$$D_{1/2}(f, g) \;=\; (1 + o(1))\tilde{I}\rho T + 2\Big(\Gamma\sqrt{\mu_1\nu_1} - \sqrt{P_{01}Q_{01}}\Big)H_{11}^2\Gamma_T + O((\rho T)^2),$$

where $\Gamma_T = \sum_{t=0}^{T-2}(1-\Gamma)^t$. Because $\Gamma_T \leq \Gamma^{-1}$ and $H_{11} \leq 1$, we see that the middle term on the right is bounded in absolute value by $2\Gamma^{-1}\rho$. The assumption that $\rho T \ll 1 \ll \Gamma T$, combined with the assumption that $\tilde{I} \neq 0$, now implies that $D_{1/2}(f, g) = (1 + o(1))\tilde{I}\rho T$. The claims (i)–(iii) now follow from Theorem 4.

Let us now impose the extra assumptions that $p_{01}, q_{01} > 0$ and $p_{01} \neq q_{01}$, $\mu_1 \asymp \nu_1$, $P_{10} \asymp Q_{10}$, and (3.4.12). In this case may fix a constant $M \geq 1$ such that $M^{-1} \leq \frac{\mu_1}{\nu_1}, \frac{P_{01}}{Q_{01}}, \frac{P_{10}}{Q_{10}} \leq M$. Furthermore, (3.4.12) implies that $\Lambda \leq 1 - \Omega(1)$. Proposition 14 then implies that

$$D_\alpha(f\|g) \;\leq\; 8C\rho T e^{5C\rho T} \qquad \text{with } C = \frac{M^3}{1-\Lambda}.$$

Because $C \lesssim 1$ and $\rho T \ll 1$, we conclude that $D_\alpha(f\|g) \lesssim \rho T$. A similar argument shows that $D_{3/2}(g\|f) \lesssim \rho T$ as well. Therefore, $\frac{\mathrm{D}^s_{3/2}(f,g)}{\mathrm{D}_{1/2}(f,g)} \lesssim 1$. Taylor's approximations further show that the intra- and inter-block probabilities $p = 1 - (1-\mu_1)(1-P_{01})^{T-1}$ and $q = 1-(1-\nu_1)(1-Q_{01})^{T-1}$ of observing a nonzero interaction pattern satisfy $p = p_{01}\rho T + o(\rho T)$ and $q = q_{01}\rho T + o(\rho T)$. It follows that $p, q \asymp \rho T$ and $p - q \asymp \rho T$. When we assume that $\rho \gg \frac{1}{NT}$, it follows that $p \vee q \gg N^{-1}$ and $\frac{(p-q)^2}{p\vee q} \asymp \rho T$. We will apply Theorem 5 to conclude that Algorithm 2 is consistent when $\rho \gg \frac{1}{NT}$, and strongly consistent when $\rho \gg \frac{\log N}{NT}$, or when $\rho = \frac{\log N}{NT}$ and $\tilde{I} > K$. □

Remark 9. Theorem 7 shows that consistent recovery may be possible even in cases where individual snapshots are very sparse, for example in regimes with $\rho \asymp \frac{1}{N}$ and $T \gg 1$. This is in stark contrast with standard binary SBMs, where in the constant-degree regime with $\rho \asymp \frac{1}{N}$, the best one can achieve is detection [Mas14; MNS15; MNS18].

Remark 10. We observe that the conditions derived in Theorem 7 are similar to those derived for an SBM with zero-inflated geometrically distributed edge weights. Indeed, the Rényi divergence between two sparse Markov chains with $T \gg 1$ (expression (3.4.11)) corresponds (up to second order terms) to the Rényi divergence between two zero-inflated geometric distributions (equation (3.2.2)).

Example 12. A temporal network model in [Bar+18] is characterised by link density $\rho \ll 1$ and parameters $0 \leq a, \xi, \eta \leq 1$ corresponding to assortativity, link persistence, and community persistence. For $\eta = 1$, the model corresponds to a Markov SBM with intra- and inter-block node pairs interacting according to stationary Markov chains having transition matrices $P = \xi \left[\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right] + (1-\xi) \left[\begin{smallmatrix} 1-\mu_1 & \mu_1 \\ 1-\mu_1 & \mu_1 \end{smallmatrix}\right]$ and $Q = \xi \left[\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right] + (1-\xi) \left[\begin{smallmatrix} 1-\nu_1 & \nu_1 \\ 1-\nu_1 & \nu_1 \end{smallmatrix}\right]$ and marginal link probabilities $\mu_1 = (1 - a + Ka)\rho$ and $\nu_1 = (1-a)\rho$, respectively. When $K = 2$ and $0 < a, \xi \leq 1$ are constants, conditions (3.4.9) are valid with $p_{01} = (1-\xi)(1+a)$, $q_{01} = (1-\xi)(1-a)$, and $h_{11} = 0$, and the critical information quantity in (3.4.10) equals

$$\tilde{I} \;=\; 2(1-\xi)\left(1 - \sqrt{(1-a)(1+a)}\right). \tag{3.4.13}$$

By Theorem 7, strong consistency in the critical regime with $\rho = \frac{\log N}{NT}$ is possible for $\tilde{I} > 2$ and impossible for $\tilde{I} < 2$. Formula (3.4.13) quantifies how higher link persistence ξ makes community recovery harder, whereas higher assortativity a makes it easier. The model in [Bar+18] assumes that intra-block and inter-block links have equal persistence ξ, leading to $h_{11} = 0$.

3.5 Discussions

3.5.1 Large number of blocks

While in some situations the assumption of a finite number of clusters is realistic (e.g., in the *political blog* data set, no matter the number of blogs considered there are only two clusters corresponding to political affiliations), some settings require an infinite number of clusters (or equivalently, communities of sub-linear size).

The lower bound of Theorem 3 tells nothing about impossibility of consistent recovery whenever $K \gg 1$, while it gives some information about the impossibility of strong consistency if $K \ll N^{-1/3}$. For example, for $K = N^{\gamma}$ with $\gamma < \frac{1}{3}$, Theorem 3 implies that a strongly consistent estimator does not exist if $\frac{NI}{K \log N} \leq (1 - \Omega(1))(1 - 3\gamma)$.

Furthermore, the form of the upper bound of Theorem 3 tempts to believe that the MLE is consistent whenever $I \gg \frac{K}{N}$. However, the proof requires the assumptions $I \gg \frac{K^3 \log K}{N}$ and $K \ll \left(\frac{N}{\log N}\right)^{1/3}$ (see Proposition 12).

Finally, a conjecture made in [ZZ16, Section 4.2] for binary SBMs in a minimax setting states that the exact rate for the smallest expected error might be $NKe^{-(1+o(1))\frac{NI}{K}}$. This conjecture implies that a strongly consistent estimator does not exist if $I \leq (1 - \Omega(1))\frac{K \log(NK)}{N}$

and exists if $I \geq (1+\Omega(1))\frac{K\log(NK)}{N}$. For a slowly growing K such that $\log K \ll \log N$ we recover the threshold stated in Theorem 4, while a regime in which $K = N^\gamma$ leads to the existence of a strongly consistent estimator if $I \geq (1+\gamma+\Omega(1))\frac{K\log N}{N}$.

3.5.2 Non-homogeneous models

Existing results for the non-homogeneous binary SBM show that the thresholds for almost exact and exact recovery are governed by the Chernoff-Hellinger divergence [AS15]. More precisely, let α_k be the probability that node $i \in [N]$ is in community $k \in [K]$ and $f_{k\ell}$ be the connection probability between a node in community k and a node in community ℓ. If $f_{k\ell}$ are Bernoulli probability distributions with parameters $p_{k\ell}$, then [AS15, Theorem 6] states that exact recovery of the community memberships is possible if $I_{CH}(\alpha, f) \geq (1+\Omega(1))\frac{K\log N}{N}$, where $I_{CH}(\alpha, f)$ is called *Chernoff-Hellinger* divergence and is defined by

$$I_{CH}(\alpha, f) = \min_{k_1 \neq k_2} \sup_{t\in(0,1)} \sum_{\ell\in[K]} \alpha_\ell \left(tp_{k_1\ell} + (1-t)p_{k_2\ell} - (p_{k_1\ell})^t (p_{k_2\ell})^{1-t}\right).$$

Noticing that the Rényi divergence of order $t \in (0,1)$ between two Poisson random variables of parameters λ and μ is given by

$$\mathrm{D}_t(\mathcal{P}(\lambda)\|\mathcal{P}(\mu)) = \frac{1}{1-t}\left(t\lambda + (1-t)\mu - \lambda^t\mu^{1-t}\right),$$

then $I_{CH}(\alpha, f) = \min_{k_1\neq k_2} \sup_{t\in(0,1)} \sum_{\ell\in[K]} (1-t)\,\mathrm{D}_t(\mathcal{P}(p_{k_1\ell})\|\mathcal{P}(p_{k_2\ell}))$. Approximating Poisson distributions by Bernoulli distributions leads to

$$I_{CH}(\alpha, f) \approx \min_{k_1\neq k_2} \sup_{t\in(0,1)} (1-t) \sum_{\ell\in[K]} \alpha_\ell\, \mathrm{D}_t(f_{k_1\ell}\|f_{k_2\ell}).$$

Theorem 8 provides a lower-bound for the expected loss made by any algorithm in clustering non-homogeneous and non-binary SBM. A careful analysis of this lower bound as well as deriving a matching upper bound is a challenging task.

Chapter 4

Proofs for Chapter 3

This section is devoted to the exposition of the proofs of the results of Chapter 3. Section 4.1 provides a lower bound on the estimation error made by any algorithm on non-binary SBM. Upper bound on the estimation error made by Maximum Likelihood Estimator and Algorithm 2 are given in Section 4.2 and Section 4.3 respectively. Rényi-divergences between sparse Markov chains are computed in Section 4.4.

4.1 Proof of the lower bound of Theorem 3

This section is devoted to proving the lower bound of Theorem 3 and is organised as follows: Section 4.1.1 describes a lower bound (Theorem 8) which is valid for general SBMs, not necessarily homogeneous or binary. Section 4.1.2 presents the proof of Theorem 8. Section 4.1.3 specialises the lower bound into homogeneous SBMs and leads to Proposition 9.

4.1.1 A quantitative lower bound

Let us recall the model. A general SBM has N nodes, K blocks, and interaction distributions $f_{k\ell}$ on a general interaction space $\mathcal{S}$ such that $f_{k\ell} = f_{\ell k}$ for all $k, \ell \in [K]$. The model has parameter space $\mathcal{Z} \subset [K]^{[N]}$, observation space

$$\mathcal{X} \;=\; \{x : [N] \times [N] \to S : x_{ij} = x_{ji},\; x_{ii} = 0 \text{ for all } i, j\}, \tag{4.1.1}$$

and probability kernel

$$P_\sigma(x) \;=\; \prod_{1 \le i < j \le N} f_{\sigma(i)\sigma(j)}(x_{ij}). \tag{4.1.2}$$

For a probability distribution π on $\mathcal{Z}$, define a probability measure on $\mathcal{Z} \times \mathcal{X}$ by

$$\mathbb{P}(\sigma, x) \;=\; \pi(\sigma) P_\sigma(x). \tag{4.1.3}$$

The SBM is called homogeneous if

$$f_{k\ell} \;=\; \delta_{k\ell} f + (1 - \delta_{k\ell}) g$$

for some probability distributions f and g on $\mathcal{S}$.

Define a probability measure $\pi = \alpha^{\otimes N}$ on $\mathcal{Z} = [K]^{[N]}$ as the N-fold product of a probability distribution α on $[K]$. This corresponds to the joint law of $\sigma = (\sigma(1), \dots, \sigma(N))$ in which all labels are mutually independent and α-distributed. For any $\mathcal{K} \subset [K]$ and any reference probability distributions $f_1^*, \dots, f_K^*$, denote

$$\begin{aligned} I_1 \;&=\; \sum_k \sum_\ell \alpha_k^* \alpha_\ell \, \mathrm{D}_{\mathrm{KL}}(f_\ell^* \| f_{k\ell}), \\ I_{21} \;&=\; \sum_k \sum_\ell \alpha_k^* \alpha_\ell \, \mathrm{V}_{\mathrm{KL}}(f_\ell^* \| f_{k\ell}) + \sum_k \alpha_k^* B_k, \\ I_{22} \;&=\; \sum_k \alpha_k^* A_k^2 - \Big(\sum_k \alpha_k^* A_k \Big)^2, \end{aligned} \tag{4.1.4}$$

with $A_k = \sum_\ell \alpha_\ell \, \mathrm{D}_{\mathrm{KL}}(f_\ell^* \| f_{k\ell})$ and $B_k = \sum_\ell \alpha_\ell \, \mathrm{D}_{\mathrm{KL}}(f_\ell^* \| f_{k\ell})^2 - (\sum_\ell \alpha_\ell \, \mathrm{D}_{\mathrm{KL}}(f_\ell^* \| f_{k\ell}))^2$, together with $\alpha_k^* = 1(k \in \mathcal{K}) \frac{\alpha_k}{\alpha_\mathcal{K}}$ and $\alpha_\mathcal{K} = \sum_{k \in \mathcal{K}} \alpha_k$.

Theorem 8. *Consider a stochastic block model defined by (4.1.1)–(4.1.3), and fix an arbitrary $\mathcal{K} \subset [K]$ and probability distributions $f_1^*, \dots, f_K^*$. Assume that $N \geq 8\alpha_{\min}^{-1} \log(K/\delta)$ for $\delta = \frac{1}{4}(\alpha_{\mathcal{K}} - \alpha_{\max,\mathcal{K}})$. Then for any estimator $\hat{\sigma} : \mathcal{X} \to \mathcal{Z}$, the expected error defined in Section 2.2.3 is lower bounded by*

$$\mathbb{E}\, \mathrm{d}_{\mathrm{Ham}}^*(\hat{\sigma}) \;\geq\; \frac{1}{21} N \alpha_{\min}^2 \delta e^{-NI_1 - \alpha_{\mathcal{K}}^{1/2} \delta^{-1/2} \sqrt{NI_{21} + N^2 I_{22}}} - \frac{1}{6} N \alpha_{\min} K e^{-\frac{1}{8} N \alpha_{\min}}. \tag{4.1.5}$$

Remark 11. The second term on the right side of (4.1.5) is $o(1)$ when $\alpha_{\min} \geq 9N^{-1} \log N$ and $2 \leq K \leq N$.

Remark 12. The lower bound of Theorem 8 is quantitative, and hence valid regardless of any scaling assumptions, and also for all finite models with fixed, not asymptotic, size. This one of the first explicit quantitative lower bounds in this context.

Remark 13. In homogeneous models with uniform node labels, one can specify the quantities I_1, I_{21} and I_{12} to obtain the lower bound stated in Theorem 3. This is done in Section 4.1.3.

4.1.2 Proof of Theorem 8

This section is devoted to proving Theorem 8 step by step.

Key result on block permutations

The following key result implies that when $L(\sigma_1, \sigma_2) = \min_\tau \mathrm{d}_{\mathrm{Ham}}(\sigma_1, \tau \circ \sigma_2) < \frac{1}{2} N_{\min}(\sigma_1)$, then the minimum Hamming distance is attained by a unique block permutation.

Lemma 1. *Let $\sigma_1, \sigma_2 : [N] \to [K]$ be such that $\mathrm{d}_{\mathrm{Ham}}(\sigma_1, \tau^* \circ \sigma_2) < \frac{1}{2} N_{\min}$ for some K-permutation τ^*, where $N_{\min} = \min_k |\sigma_1^{-1}(k)|$. Then τ^* is the unique minimiser of $\tau \mapsto \mathrm{d}_{\mathrm{Ham}}(\sigma_1, \tau \circ \sigma_2)$.*

This corresponds to [XJL20, Lemma B.6].

Proof. Assume that $\tau \in \mathrm{Sym}(K)$ satisfies $\mathrm{d}_{\mathrm{Ham}}(\tau \circ \sigma_1, \sigma_2) < \frac{s}{2}$, where $s = N_{\min}$. Fix $k \in [K]$ and let $U_k = \{i : \sigma_1(i) = k, \sigma_2(i) \neq \tau(k)\}$. Then every node i in U_k satisfies $\tau \circ \sigma_1(i) \neq \sigma_2(i)$, and therefore $|U_k| \leq \mathrm{d}_{\mathrm{Ham}}(\tau \circ \sigma_1, \sigma_2) < \frac{s}{2}$. Hence for any $\ell \neq \tau(k)$,

$$|\sigma_1^{-1}(k) \cap \sigma_2^{-1}(\ell)| \;\leq\; |U_k| \;<\; \frac{s}{2}.$$

On the other hand,

$$|\sigma_1^{-1}(k) \cap \sigma_2^{-1}(\tau(k))| \;=\; |\sigma_1^{-1}(k)| - |U_k| \;\geq\; s - \frac{s}{2} \;\geq\; \frac{s}{2}.$$

Hence $\tau(k)$ is the unique value which maximises $\ell \mapsto |\sigma_1^{-1}(k) \cap \sigma_2^{-1}(\ell)|$. Because this conclusion holds for all k, it follows that τ is uniquely defined. □

Lower bounding by critical node count

This method apparently originates from [ZZ16]. Let $\mathrm{Opt}(\sigma_1, \sigma_2)$ be the set of K-permutations τ for which $\mathrm{d}_{\mathrm{Ham}}(\sigma_1, \tau \circ \sigma_2)$ is minimised. Given an estimated node labelling $\hat{\sigma}_x$, we define a set of critical nodes by

$$\mathrm{Crit}(\sigma, \hat{\sigma}_x) \;=\; \{j \in [N] : \sigma(j) \neq \tau \circ \hat{\sigma}_x(j) \text{ for some } \tau \in \mathrm{Opt}(\sigma, \hat{\sigma}_x)\}.$$

We denote the number of critical nodes by

$$L^+(\sigma, x) \;=\; |\mathrm{Crit}(\sigma, \hat{\sigma}_x)|.$$

Lemma 2. *For any estimate $\hat{\sigma}_x$ obtained as a deterministic function of observed data. Then*

$$\mathbb{E}\, \mathrm{d}^*_{\mathrm{Ham}}(\sigma, \hat{\sigma}_x) \;\geq\; \frac{\alpha_{\min}}{6} \left(\mathbb{E}L^+ - NKe^{-\frac{1}{8}N\alpha_{\min}} \right). \tag{4.1.6}$$

Proof. Let $L = L(\sigma, \hat{\sigma}_x) = \mathrm{d}^*_{\mathrm{Ham}}(\sigma, \hat{\sigma}_x)$. We shall consider $L^+ = L^+(\sigma, \hat{\sigma}_x)$, and $N_{\min} = N_{\min}(\sigma)$ as random variables defined on $\mathcal{Z} \times \mathcal{X}$. By Lemma 1, $L = L^+$ on the event $L < c$ where $c = \frac{1}{2}N_{\min}$. Given a node labelling $\sigma \in \mathcal{Z}$, we consider $x \mapsto \hat{\sigma}_x$, $x \mapsto L(\sigma, x)$ and $x \mapsto L^+(\sigma, x)$ as random variables on $\mathcal{X}$. Consider the following two cases:

(i) If $P_\sigma(L \geq c) \geq \frac{1}{N+c} E_\sigma L^+$, then

$$E_\sigma L 1(L \geq c) \;\geq\; cP_\sigma(L \geq c) \;\geq\; \frac{c}{N+c} E_\sigma L^+.$$

(ii) If $P_\sigma(L \geq c) \leq \frac{1}{N+c} E_\sigma L^+$, then

$$E_\sigma L^+ 1(L \geq c) \;\leq\; NP_\sigma(L \geq c) \;\leq\; \frac{N}{N+c} E_\sigma L^+,$$

so that

$$E_\sigma L 1(L < c) \;=\; E_\sigma L^+ 1(L < c) \;=\; E_\sigma L^+ - E_\sigma L^+ 1(L \geq c) \;\geq\; \frac{c}{N+c} E_\sigma L^+.$$

In both cases, $E_\sigma L \geq \frac{c}{N+c} E_\sigma L^+$, so that

$$E_\sigma L \;\geq\; \frac{N_{\min}}{2N + N_{\min}} E_\sigma L^+ \;\geq\; \frac{N_{\min}}{3N} E_\sigma L^+.$$

By taking expectations with respect to the prior, we find that

$$\mathbb{E}L \;\geq\; \frac{1}{3N} \mathbb{E} N_{\min} Y, \tag{4.1.7}$$

where $Y = E_\sigma L^+$ is viewed as a random variable on probability space $\mathcal{S}$ equipped with probability measure π. Let $t = \frac{1}{2}N\alpha_{\min}$. We note that $0 \leq Y \leq N$ surely, and that $N_{\min} > t$ with high probability. Observe that

$$\mathbb{E} N_{\min} Y \;\geq\; \mathbb{E} N_{\min} Y\, 1(N_{\min} > t) \;\geq\; t\, \mathbb{E} Y 1(N_{\min} > t),$$

and, due to $Y \le N$,

$$\mathbb{E}Y1(N_{\min} > t) \;=\; \mathbb{E}Y - \mathbb{E}Y1(N_{\min} \le t) \;\ge\; \mathbb{E}Y - N\mathbb{P}(N_{\min} \le t).$$

By noting that $\mathbb{E}Y = \mathbb{E}L^+$ and applying Lemma 10, we find that

$$\begin{aligned}\mathbb{E}N_{\min}Y \;&\ge\; t\left(\mathbb{E}Y - N\mathbb{P}(N_{\min} \le t)\right)\\ &=\; \frac{1}{2}N\alpha_{\min}\left(\mathbb{E}L^+ - N\mathbb{P}(N_{\min} \le \frac{1}{2}N\alpha_{\min})\right)\\ &\ge\; \frac{1}{2}N\alpha_{\min}\left(\mathbb{E}L^+ - NKe^{-\frac{1}{8}N\alpha_{\min}}\right).\end{aligned}$$

Together with (4.1.7), the claim now follows. □

Change of measure

Fix a reference node i, a set $\mathcal{K} \subset [K]$, and some probability distributions $f_1^*, \dots, f_K^*$ on the interaction space S. We define an alternative statistical model for σ and x by modifying $P_\sigma(x)$ defined in (4.1.2) according to

$$P_\sigma^{*i}(x) \;=\; \left(1_{\mathcal{K}}(\sigma(i)) \prod_{j \ne i} \frac{f^*_{\sigma(j)}(x_{ij})}{f_{\sigma(i)\sigma(j)}(x_{ij})} + 1_{\mathcal{K}^c}(\sigma(i))\right) P_\sigma(x), \tag{4.1.8}$$

and defining a modified probability measure on $\mathcal{Z} \times \mathcal{X}$ by

$$\mathbb{P}^{*i}(\sigma, x) \;=\; \pi(\sigma)P_\sigma^{*i}(x). \tag{4.1.9}$$

In the modified model, node labels are sampled independently as before, and all interactions not involving node i are sampled just as in the original model. If the label of node i belongs to $\mathcal{K}$, then we sample all i-interactions from $f_1^*, \dots, f_K^*$. The following lemma confirms that under the alternative model, σ_i is conditionally independent of observed data x and other labels σ_{-i} given $\sigma_i \in \mathcal{K}$.

Lemma 3. *For (σ, x) sampled from model (4.1.9), the conditional distribution of the label σ_i given that $\sigma_i \in \mathcal{K}$, the other labels are σ_{-i}, and the observed interactions are x, equals*

$$\mathbb{P}^{*i}(\sigma_i = k \,|\, \sigma_i \in \mathcal{K}, \sigma_{-i}, x) \;=\; \alpha_k^* \qquad \textit{for all } \sigma_{-i}, x,$$

where $\alpha_k^* = 1(k \in \mathcal{K})\frac{\alpha_k}{\alpha_\mathcal{K}}$.

Proof. Observe that $P_\sigma^{*i}(x) = Q_{\sigma_{-i}}(x)$ for all σ such that $\sigma_i \in \mathcal{K}$, where

$$Q_{\sigma_{-i}}(x) \;=\; \left(\prod_{j \ne i} f^*_{\sigma_{-i}(j)}(x_{ij})\right)\left(\prod_{uv \in E_{-i}} f_{\sigma_{-i}(u)\sigma_{-i}(v)}(x_{uv})\right),$$

and E_{-i} is the set of unordered node pairs not incident to i. Especially,

$$\mathbb{P}^{*i}(\sigma, x) = \alpha(\sigma_i)\, \pi_{-i}(\sigma_{-i})\, Q_{\sigma_{-i}}(x)$$

whenever $\sigma_i \in \mathcal{K}$. Hence the conditional probability distribution of σ_i given (σ_{-i}, x) satisfies $\mathbb{P}^{*i}(\sigma_i \,|\, \sigma_{-i}, x) = \alpha(\sigma_i)$ for all $\sigma_i \in \mathcal{K}$. The claim follows by summing this equality with respect to $\sigma_i \in \mathcal{K}$. □

To analyse how much the alternative model differs from the original model, we will investigate the associated log-likelihood ratio

$$\Lambda_i(\sigma, x) \;=\; \log \frac{\mathbb{P}^{*i}(\sigma, x)}{\mathbb{P}(\sigma, x)}.$$

Lemma 4. *The mean and variance of the log-likelihood ratio given $\sigma(i) \in \mathcal{K}$ are equal to $\mathbb{E}^{*i}(\Lambda_i \,|\, \sigma_i \in \mathcal{K}) = (N-1)I_1$ and $\mathbb{V}^{*i}(\Lambda_i \,|\, \sigma_i \in \mathcal{K}) = (N-1)I_{21} + (N-1)^2 I_{22}$, where I_1, I_{21}, I_{22} are given by (4.1.4).*

Proof. The conditional distribution of (σ, x) sampled from $\mathbb{P}^{*i}$ given $\sigma(i) \in \mathcal{K}$ can be represented as

$$\tilde{\mathbb{P}}^{*i}(\sigma, x) \;=\; \tilde{\pi}^{*i}(\sigma) P^{*i}_{\sigma}(x),$$

where $\tilde{\pi}^{*i}(\sigma) = \alpha^*_{\sigma(i)} \prod_{j \neq i} \alpha_{\sigma(j)}$ and $\alpha^*_k = 1(k \in \mathcal{K}) \frac{\alpha_k}{\alpha_{\mathcal{K}}}$, and P^{*i}_{σ} is defined by (4.1.8). Furthermore, the log-likelihood ratio can be written as

$$\Lambda_i(\sigma, x) \;=\; 1(\sigma(i) \in \mathcal{K}) \sum_{j \neq i} \log \frac{f^*_{\sigma(j)}(x_{ij})}{f_{\sigma(i)\sigma(j)}(x_{ij})}.$$

The conditional expectation $A(\sigma) = E^{*i}_{\sigma} \Lambda_i$ of the log-likelihood ratio given σ hence equals

$$A(\sigma) \;=\; 1(\sigma(i) \in \mathcal{K}) \sum_{j \neq i} m_{\sigma(i)\sigma(j)}$$

where $m_{k\ell} = \mathrm{D}_{\mathrm{KL}}(f^*_{\ell} \| f_{k\ell})$. Hence, treating $(\sigma, x) \mapsto \sigma(i)$, $(\sigma, x) \mapsto \sigma(j)$, and $(\sigma, x) \mapsto A(\sigma)$, as random variables on probability space $(\mathcal{Z} \times \mathcal{X}, \tilde{\mathbb{P}}^{*i})$, and noting that $\sigma(i) \in \mathcal{K}$ with $\tilde{\mathbb{P}}^{*i}$-probability one, we find that

$$\tilde{\mathbb{E}}^{*i} \Lambda_i \;=\; \tilde{\mathbb{E}}^{*i} A \;=\; \sum_{j \neq i} \tilde{\mathbb{E}}^{*i} m_{\sigma(i)\sigma(j)} \;=\; (N-1) \sum_{k} \sum_{\ell} m_{k\ell} \alpha^*_k \alpha_{\ell},$$

which implies the first claim.

To compute the variance, we observe that

$$\tilde{\mathbb{V}}^{*i} \Lambda_i \;=\; \tilde{\mathbb{E}}^{*i} B + \tilde{\mathbb{V}}^{*i} A, \tag{4.1.10}$$

where $B = V^{*i}_{\sigma} \Lambda_i$. We note that by the conditional independence of x_{ij}, $j \neq i$, given σ, it follows that

$$B \;=\; 1(\sigma(i) \in \mathcal{K}) \sum_{j \neq i} v_{\sigma(i)\sigma(j)},$$

where $v_{k\ell} = \mathrm{V}_{\mathrm{KL}}(f_\ell^* \| f_{k\ell})$. By taking expectations, we find that

$$\tilde{\mathbb{E}}^{*i} B \;=\; (N-1) \sum_k \sum_\ell \alpha_k^* \alpha_\ell v_{k\ell}. \tag{4.1.11}$$

We still need to compute the variance of A. To do this, we condition on the label of node i and observe that on the event $\sigma(i) \in \mathcal{K}$ of $\tilde{\mathbb{P}}^{*i}$-probability one,

$$\begin{aligned} \tilde{\mathbb{E}}^{*i}(A \,|\, \sigma(i)\,) &= (N-1) A_{\sigma(i)}, \\ \tilde{\mathbb{V}}^{*i}(A \,|\, \sigma(i)\,) &= (N-1) B_{\sigma(i)}, \end{aligned}$$

where $A_k = \sum_\ell \alpha_\ell m_{k\ell}$ and $B_k = \sum_\ell \alpha_\ell m_{k\ell}^2 - (\sum_\ell \alpha_\ell m_{k\ell})^2$. Therefore,

$$\begin{aligned} \tilde{\mathbb{V}}^{*i} A &= \tilde{\mathbb{E}}^{*i} \tilde{\mathbb{V}}^{*i}(A \,|\, \sigma(i)\,) + \mathbb{V}\tilde{\mathbb{E}}^{*i}(A \,|\, \sigma(i)\,) \\ &= (N-1) \tilde{\mathbb{E}}^{*i} B_{\sigma(i)} + (N-1)^2 \tilde{\mathbb{V}}^{*i} A_{\sigma(i)} \\ &= (N-1) \sum_k \alpha_k^* B_k + (N-1)^2 \Big\{ \sum_k \alpha_k^* A_k^2 - \Big(\sum_k \alpha_k^* A_k \Big)^2 \Big\}. \end{aligned}$$

By combining this with (4.1.10) and (4.1.11), we find that

$$\begin{aligned} \tilde{\mathbb{V}}^{*i} \Lambda_i \;=\; & (N-1) \sum_k \sum_\ell \alpha_k^* \alpha_\ell v_{k\ell} + (N-1) \sum_k \alpha_k^* B_k \\ & + (N-1)^2 \Big\{ \sum_k \alpha_k^* A_k^2 - \Big(\sum_k \alpha_k^* A_k \Big)^2 \Big\}, \end{aligned}$$

and the second claim follows. □

Lower bound of critical node count

The following is key to proving the lower bound, and rigorously handling stochastic dependencies implied by optimal K-permutations in the definition of L. Recall that $\alpha_{\mathcal{K}} = \sum_{k \in \mathcal{K}} \alpha_k$ together with $\alpha_{\min} = \min_{k \in [K]} \alpha_k$ and $\alpha_{\max,\mathcal{K}} = \max_{k \in \mathcal{K}} \alpha_k$.

Lemma 5. *Assume that $N \geq 8 \alpha_{\min}^{-1} \log(K/\delta)$ for $\delta = \frac{1}{4}(\alpha_{\mathcal{K}} - \alpha_{\max,\mathcal{K}})$. Then for any estimator $x \mapsto \hat{\sigma}_x$, the expected number of critical nodes is bounded by*

$$\mathbb{E} L^+ \;\geq\; \frac{2}{7} \alpha_{\min} \delta N e^{-t} \tag{4.1.12}$$

for $t = \max_i \Big(\mathbb{E}^{*i}(\Lambda_i | \sigma_i \in \mathcal{K}) + \alpha_{\mathcal{K}}^{1/2} \delta^{-1/2} \sqrt{\mathbb{V}^{*i}(\Lambda_i | \sigma_i \in \mathcal{K})} \Big)$.

Proof. Denote $\epsilon = \frac{1}{6} \alpha_{\min}$. The proof contains four steps which are treated one by one in what follows.

(i) Denote the event that node i is critical by

$$\mathcal{C}_i \;=\; \big\{ (\sigma, x) : \sigma(i) \neq \tau(\hat{\sigma}_x(i)) \text{ for some } \tau \in \mathrm{Opt}(\sigma, \hat{\sigma}_x) \big\},$$

and let

$$\mathcal{E}_i \;=\; \mathcal{C}_i \cup \big\{(\sigma, x) : L^+(\sigma, \hat{\sigma}_x) > \epsilon N\big\}.$$

Recall that $\mathbb{E}L^+ = \sum_i \mathbb{P}(\mathcal{C}_i)$. Markov's inequality then implies that

$$\sum_i \mathbb{P}(\mathcal{E}_i) \;\le\; \sum_i \left(\mathbb{P}(\mathcal{C}_i) + (\epsilon N)^{-1}\mathbb{E}L^+\right).$$

By noting that the right side above equals $(1+\epsilon^{-1})\mathbb{E}L^+$, we obtain a lower bound

$$\mathbb{E}L^+ \;\ge\; \frac{\epsilon}{1+\epsilon}\sum_i \mathbb{P}(\mathcal{E}_i). \tag{4.1.13}$$

(ii) We will now focus on a particular node i, and derive a lower bound for the probability of event $\mathcal{E}_i$ under the perturbed model $\mathbb{P}^{*i}$ defined by (4.1.8). We start by deriving an upper bound for the probability of the event

$$\mathbb{P}^{*i}(\mathcal{E}_i^c,\, N_{\min} > 3\epsilon N,\, \sigma(i) \in \mathcal{K}) \;=\; \mathbb{P}^{*i}(\mathcal{C}_i^c,\, \mathcal{B},\, \sigma(i) \in \mathcal{K}),$$

where

$$\mathcal{B} \;=\; \{(\sigma, x) : L^+(\sigma, \hat{\sigma}_x) \le \epsilon N,\, N_{\min}(\sigma) > 3\epsilon N\}$$

and $N_{\min}(\sigma) = \min_k |\sigma^{-1}(k)|$. On the event $\mathcal{B}$, we see that $L^+(\sigma, \hat{\sigma}_x) < \frac{1}{3}N_{\min}(\sigma)$, and Lemma 1 implies that $L^+(\sigma, \hat{\sigma}_x) = \min_\tau \mathrm{Ham}(\sigma, \tau\circ\hat{\sigma}_x)$ is attained by a unique K-permutation τ. This is why we may split the above probability into

$$\mathbb{P}^{*i}(\mathcal{C}_i^c,\, \mathcal{B},\, \sigma(i) \in \mathcal{K}) \;=\; \sum_\tau \mathbb{P}^{*i}(\mathcal{C}_i^c,\, \mathcal{B}_\tau,\, \sigma(i) \in \mathcal{K}) \tag{4.1.14}$$

where

$$\mathcal{B}_\tau \;=\; \{(\sigma, x) : \mathrm{Ham}(\sigma, \tau \circ \hat{\sigma}_x) \le \epsilon N,\, N_{\min}(\sigma) > 3\epsilon N\}.$$

To analyse events associated with $\mathcal{B}_\tau$, we consider an event

$$\mathcal{B}_\tau^{-i} \;=\; \big\{(\sigma, x) : \mathrm{Ham}_{-i}(\sigma, \tau \circ \hat{\sigma}_x) \le \epsilon N,\, N_{-i}^{\min}(\sigma) > 3\epsilon N - 1\big\}$$

where $\mathrm{Ham}_{-i}(\sigma_1, \sigma_2) = \sum_{j\neq i} 1(\sigma_1(j) \neq \sigma_2(j))$ and $N_{-i}^{\min}(\sigma) = \min_k |\sigma^{-1}(k) \setminus \{i\}|$. Then,

$$\begin{aligned}\mathcal{C}_i^c \cap \mathcal{B}_\tau \;&=\; \{\sigma(i) = \tau(\hat{\sigma}_x(i))\} \cap \mathcal{B}_\tau \\ &\subset\; \{\sigma(i) = \tau(\hat{\sigma}_x(i))\} \cap \mathcal{B}_\tau^{-i},\end{aligned}$$

so that, under the conditional distribution $\tilde{\mathbb{P}}^{*i}(\cdot) = \mathbb{P}^{*i}(\cdot \,|\, \sigma(i) \in \mathcal{K})$,

$$\tilde{\mathbb{P}}^{*i}(\mathcal{C}_i^c,\, \mathcal{B}_\tau) \;\le\; \tilde{\mathbb{P}}^{*i}(\sigma(i) = \tau(\hat{\sigma}_x(i)),\, \mathcal{B}_\tau^{-i}). \tag{4.1.15}$$

We note that the event $\mathcal{B}_\tau^{-i}$ is completely determined by (σ_{-i}, x), and according to Lemma 3, we know that when (σ, x) is sampled from $\mathbb{P}^{*i}$, then $\sigma(i)$ is α^*-distributed and conditionally

independent of (σ_{-i}, x) given $\sigma(i) \in \mathcal{K}$. Therefore, under the conditional distribution $\tilde{\mathbb{P}}^{*i}(\cdot) = \mathbb{P}^{*i}(\cdot \,|\, \sigma(i) \in \mathcal{K})$, we find that

$$\begin{aligned}\tilde{\mathbb{P}}^{*i}\big(\sigma(i) = \tau(\hat{\sigma}_x(i)),\, \mathcal{B}_\tau^{-i}\big) &= \sum_{k\in\mathcal{K}} \tilde{\mathbb{P}}^{*i}\big(\sigma(i) = k,\, \tau(\hat{\sigma}_x(i)) = k,\, \mathcal{B}_\tau^{-i}\big) \\ &= \sum_{k\in\mathcal{K}} \alpha_k^* \,\tilde{\mathbb{P}}^{*i}\big(\tau(\hat{\sigma}_x(i)) = k,\, \mathcal{B}_\tau^{-i}\big),\end{aligned}$$

from which we conclude together with (4.1.15) that

$$\tilde{\mathbb{P}}^{*i}(\mathcal{C}_i^c,\, \mathcal{B}_\tau) \;\le\; \frac{\alpha_{\max,\mathcal{K}}}{\alpha_{\mathcal{K}}}\, \tilde{\mathbb{P}}^{*i}\big(\mathcal{B}_\tau^{-i}\big).$$

Because $N \ge \epsilon^{-1}$ due to $\log(K/\delta) \ge \log(4K) \ge 1$ and $N \ge 8\alpha_{\min}^{-1}\log(K/\delta)$, we see that $\epsilon N < \frac{1}{2}(3\epsilon N - 1)$. Therefore, $\mathrm{Ham}_{-i}(\sigma, \tau\circ\hat{\sigma}_x) < \frac{1}{2}N_{-i}^{\min}(\sigma)$ on the event $\mathcal{B}_\tau^{-i}$. Then again by Lemma 1, the events $\mathcal{B}_\tau^{-i}$ are mutually exclusive, and in light of (4.1.14) it follows that

$$\tilde{\mathbb{P}}^{*i}\big(\mathcal{C}_i^c,\, \mathcal{B}\big) \;\le\; \frac{\alpha_{\max,\mathcal{K}}}{\alpha_{\mathcal{K}}}\, \tilde{\mathbb{P}}^{*i}(\cup_\tau \mathcal{B}_\tau^{-i}) \;\le\; \frac{\alpha_{\max,\mathcal{K}}}{\alpha_{\mathcal{K}}}.$$

By recalling the definitions of $\mathcal{C}_i, \mathcal{E}_i$, we now conclude that

$$\begin{aligned}\mathbb{P}^{*i}\big(\mathcal{E}_i^c,\, N_{\min} > 3\epsilon N,\, \sigma(i)\in\mathcal{K}\big) &= \mathbb{P}^{*i}\big(\mathcal{C}_i^c, L^+ \le \epsilon N,\, N_{\min} > 3\epsilon N,\, \sigma(i)\in\mathcal{K}\big) \\ &= \mathbb{P}^{*i}\big(\mathcal{C}_i^c,\, \mathcal{B},\, \sigma(i)\in\mathcal{K}\big) \\ &\le \alpha_{\max,\mathcal{K}},\end{aligned}$$

and therefore,

$$\mathbb{P}^{*i}(\mathcal{E}_i^c,\, \sigma(i)\in\mathcal{K}) \;\le\; \alpha_{\max,\mathcal{K}} + \mathbb{P}^{*i}(N_{\min} \le 3\epsilon N). \tag{4.1.16}$$

(iii) Next, by recalling our choice of $\epsilon = \frac{1}{6}\alpha_{\min}$ and applying Lemma 10, we see that $\mathbb{P}^{*i}(N_{\min} \le 3\epsilon N) = \mathbb{P}(N_{\min} \le \frac{1}{2}N\alpha_{\min}) \le Ke^{-\frac{1}{8}N\alpha_{\min}} \le \delta$ due to $N \ge 8\alpha_{\min}^{-1}\log(K/\delta)$. By combining this with (4.1.16), we see that $\mathbb{P}^{*i}(\mathcal{E}_i^c,\, \sigma(i)\in\mathcal{K}) \le \alpha_{\max,\mathcal{K}} + \delta$. Hence, by our choice of δ, it follows that

$$\begin{aligned}\mathbb{P}^{*i}(\mathcal{E}_i,\, \sigma(i)\in\mathcal{K}) &\ge \mathbb{P}(\sigma(i)\in\mathcal{K}) - \alpha_{\max,\mathcal{K}} - \delta \\ &= \alpha_{\mathcal{K}} - \alpha_{\max,\mathcal{K}} - \delta \\ &= 3\delta.\end{aligned} \tag{4.1.17}$$

(iv) Finally, we will transform the lower bound (4.1.17) into one involving the original probability distribution $\mathbb{P}$ instead of $\mathbb{P}^{*i}$. By writing

$$\mathbb{P}(\mathcal{E}_i,\, \sigma(i)\in\mathcal{K}) \;=\; \mathbb{E}^{*i} e^{-\Lambda_i} 1(\mathcal{E}_i,\, \sigma(i)\in\mathcal{K}),$$

and noting that $e^{-\Lambda_i}1(\mathcal{E}_i, \sigma_i \in \mathcal{K}) \ge e^{-t}1(\mathcal{E}_i, \sigma_i\in\mathcal{K}, \Lambda_i \le t)$, it follows that

$$\begin{aligned}\mathbb{P}(\mathcal{E}_i,\, \sigma(i)\in\mathcal{K}) &\ge e^{-t}\mathbb{P}^{*i}(\mathcal{E}_i,\, \sigma(i)\in\mathcal{K},\, \Lambda_i \le t) \\ &\ge e^{-t}\Big(\mathbb{P}^{*i}(\mathcal{E}_i,\, \sigma(i)\in\mathcal{K}) - \mathbb{P}^{*i}(\mathcal{E}_i,\, \sigma(i)\in\mathcal{K},\, \Lambda_i > t)\Big) \\ &\ge e^{-t}\Big(\mathbb{P}^{*i}(\mathcal{E}_i,\, \sigma(i)\in\mathcal{K}) - \mathbb{P}^{*i}(\sigma(i)\in\mathcal{K},\, \Lambda_i > t)\Big).\end{aligned}$$

For $t \geq \tilde{\mathbb{E}}^{*i}(\Lambda_i) + \left(\frac{\alpha_{\mathcal{K}}}{\delta}\tilde{\mathbb{V}}^{*i}(\Lambda_i)\right)^{1/2}$, Chebyshev's inequality implies that $\tilde{\mathbb{P}}^{*i}(\Lambda_i > t) \leq \frac{\delta}{\alpha_{\mathcal{K}}}$, and hence $\mathbb{P}^{*i}(\sigma(i) \in \mathcal{K},\, \Lambda_i > t) \leq \delta$. By substituting this bound and the bound (4.1.17) to the right side above, we see that

$$\mathbb{P}(\mathcal{E}_i,\, \sigma(i) \in \mathcal{K}) \;\geq\; e^{-t}(3\delta - \delta) \;=\; 2\delta e^{-t}.$$

By (4.1.13) it now follows that

$$\mathbb{E}L^+ \;\geq\; \frac{\epsilon}{1+\epsilon}\sum_i \mathbb{P}(\mathcal{E}_i) \;\geq\; \frac{\epsilon}{1+\epsilon}\sum_i \mathbb{P}(\mathcal{E}_i,\, \sigma(i) \in \mathcal{K}),$$

so that

$$\mathbb{E}L^+ \;\geq\; \frac{2N\delta e^{-t}}{1+\epsilon^{-1}}.$$

Because $1 + \epsilon^{-1} \leq \frac{7}{6}\epsilon^{-1} = 7\alpha_{\min}^{-1}$, the claim follows. □

Concluding the proof of Theorem 8

By Lemma 2, we find that

$$\mathbb{E}L \;\geq\; \frac{\alpha_{\min}}{6}\left(\mathbb{E}L^+ - NKe^{-\frac{1}{8}N\alpha_{\min}}\right).$$

By Lemma 5,

$$\mathbb{E}L^+ \;\geq\; \frac{2}{7}\alpha_{\min}\delta N e^{-t}$$

for $t = \max_i\left(\tilde{\mathbb{E}}^{*i}\Lambda_i + \alpha_{\mathcal{K}}^{1/2}\delta^{-1/2}\sqrt{\tilde{\mathbb{V}}^{*i}(\Lambda_i)}\right)$. By Lemma 4, $\tilde{\mathbb{E}}^{*i}\Lambda_i \leq NI_1$ and $\tilde{\mathbb{V}}^{*i}(\Lambda_i) \leq NI_{21} + N^2I_{22}$, so that $t \leq NI_1 + \alpha_{\mathcal{K}}^{1/2}\delta^{-1/2}\sqrt{NI_{21} + N^2I_{22}}$. By combining these facts, it follows that

$$\begin{aligned}
\mathbb{E}L \;&\geq\; \frac{\alpha_{\min}}{6}\left(\mathbb{E}L^+ - NKe^{-\frac{1}{8}N\alpha_{\min}}\right)\\
&\geq\; \frac{\alpha_{\min}}{6}\left(\frac{2}{7}\alpha_{\min}\delta N e^{-t} - NKe^{-\frac{1}{8}N\alpha_{\min}}\right)\\
&\geq\; \frac{\alpha_{\min}}{6}\left(\frac{2}{7}\alpha_{\min}\delta N e^{-NI_1 - \alpha_{\mathcal{K}}^{1/2}\delta^{-1/2}\sqrt{NI_{21}+N^2I_{22}}} - NKe^{-\frac{1}{8}N\alpha_{\min}}\right).
\end{aligned}$$

Hence the claim of Theorem 8 is valid. □

4.1.3 Application to homogeneous models

Log-likelihood ratio in homogeneous models

The expected log-likelihood ratio equals $(N-1)I_1$ where I_1 is given by (4.1.4). The following result shows how to minimise this in the homogeneous case with intra-block and inter-block interaction distributions f and g.

Lemma 6. *For any homogeneous SBM and for any $\mathcal{K} \subset [K]$ of size at least two such that $\alpha_k > 0$ for all $k \in \mathcal{K}$,*

$$\min_{f_1^*,\dots,f_K^*} I_1 \;=\; \sum_{k\in\mathcal{K}} \alpha_k^* \alpha_k D_{1-\alpha_k^*}(g\|f), \tag{4.1.18}$$

with $\alpha_k^* = \alpha_k/(\sum_{k\in\mathcal{K}} \alpha_k)$, and the minimum is attained by setting

$$f_k^* \;=\; \begin{cases} Z_{\alpha_k^*}^{-1} f^{\alpha_k^*} g^{1-\alpha_k^*} & \text{for } k \in \mathcal{K}, \\ g, & \text{otherwise.} \end{cases} \tag{4.1.19}$$

Furthermore, when α is the uniform distribution on $[K]$,

$$\min_{\mathcal{K}:|\mathcal{K}|\geq 2} \min_{f_1^*,\dots,f_K^*} I_1 \;=\; K^{-1} D_{1/2}(f\|g). \tag{4.1.20}$$

Proof. Observe that $I_1 = I_{11} + I_{12}$ where

$$I_{11} = \sum_{\ell\in\mathcal{K}} \alpha_\ell \sum_{k\in\mathcal{K}} \alpha_k^* \,\mathrm{D_{KL}}(f_\ell^*\|f_{k\ell}) \quad \text{and} \quad I_{12} = \sum_{\ell\in\mathcal{K}^c} \alpha_\ell \sum_{k\in\mathcal{K}} \alpha_k^* \,\mathrm{D_{KL}}(f_\ell^*\|f_{k\ell}).$$

We see that

$$I_{11} \;=\; \sum_{\ell\in\mathcal{K}} \alpha_\ell \Big(\alpha_\ell^* \,\mathrm{D_{KL}}(f_\ell^*\|f) + (1-\alpha_\ell^*)\,\mathrm{D_{KL}}(f_\ell^*\|g) \Big)$$

and

$$I_{12} \;=\; \sum_{\ell\in\mathcal{K}^c} \alpha_\ell \,\mathrm{D_{KL}}(f_\ell^*\|g).$$

Because each f_ℓ^* appears only once in the sums above, we minimise I_{11} and I_{12} separately. To minimise I_{12}, we set $f_\ell^* = g$ for all $\ell \in \mathcal{K}^c$, leading to $I_{12} = 0$. To minimise I_{11}, we see by applying [VH14, Theorem 30] that for all $\ell \in \mathcal{K}$,

$$\min_{f_\ell^*} \Big(\alpha_\ell^* \,\mathrm{D_{KL}}(f_\ell^*\|f) + (1-\alpha_\ell^*)\,\mathrm{D_{KL}}(f_\ell^*\|g) \Big) \;=\; (1-\alpha_\ell^*) D_{\alpha_\ell^*}(f\|g),$$

and the minimum is attained by setting f_ℓ^* as in (4.1.19). Hence the minimum value of I_1 equals

$$I_1 \;=\; \sum_{\ell\in\mathcal{K}} \alpha_\ell (1-\alpha_\ell^*) D_{\alpha_\ell^*}(f\|g).$$

Finally, by skew symmetry of Rényi divergences, we know that $(1-\alpha_\ell^*) D_{\alpha_\ell^*}(f\|g) = \alpha_\ell^* D_{1-\alpha_\ell^*}(g\|f)$, so that we can also write the minimum as

$$I_1 \;=\; \sum_{\ell\in\mathcal{K}} \alpha_\ell \alpha_\ell^* D_{1-\alpha_\ell^*}(g\|f) \;=\; \alpha_{\mathcal{K}}^{-1} \sum_{\ell\in\mathcal{K}} \alpha_\ell^2 D_{1-\alpha_\ell^*}(g\|f).$$

Assume now that α is the uniform distribution on $[K]$. Then the minimum above equals $I_1 = (K/r)K^{-2} r D_{1-1/r}(g\|f) = K^{-1} D_{1-1/r}(g\|f)$ for $r = |\mathcal{K}|$. Because $r \mapsto D_{1-\frac{1}{r}}(g\|f)$ is increasing in r, we see that I_1 is increasing as a function of $|\mathcal{K}|$. The minimum with respect to $\mathcal{K}$ is hence attained at an arbitrary $\mathcal{K}$ with $|\mathcal{K}| = 2$, confirming (4.1.20). □

The following result describes the variance terms I_{21} and I_{22} given by (4.1.4) for a uniform homogeneous SBM, when the reference distributions $f_1^*, \dots, f_K^*$ are selected to minimise I_1 according to Lemma 6.

Lemma 7. *Consider a homogeneous SBM with intra-block and inter-block interaction distributions f and g, and uniform α on $[K]$. Fix $\mathcal{K} \subset [K]$ of size 2, and define f_ℓ^* as in (4.1.19). Then*

$$\begin{aligned} I_{21} &= \left(\frac{1}{2} - K^{-1}\right) K^{-1} I^2 + \frac{1}{2} K^{-1} J, \\ I_{22} &= 0, \end{aligned}$$

where $I = D_{1/2}(f\|g)$ and $J = \int h \log^2 \frac{f}{g}$ with $h = Z_{1/2}^{-1}(fg)^{1/2}$.

Proof. When α is uniform on $[K]$ and $|\mathcal{K}| = 2$, we see that the distributions in (4.1.19) are given by $f_\ell^* = h$ for $\ell \in \mathcal{K}$, $f_\ell^* = g$ otherwise. Recall that

$$\begin{aligned} I_{21} &= \sum_{k\in\mathcal{K}} \sum_{\ell} \alpha_k^* \alpha_\ell \, \mathrm{V_{KL}}(f_\ell^* \| f_{k\ell}) + \sum_{k\in\mathcal{K}} \alpha_k^* B_k, \\ I_{22} &= \sum_{k\in\mathcal{K}} \alpha_k^* A_k^2 - \Big(\sum_{k\in\mathcal{K}} \alpha_k^* A_k\Big)^2, \end{aligned}$$

with $A_k = \sum_\ell \alpha_\ell \, \mathrm{D_{KL}}(f_\ell^* \| f_{k\ell})$ and $B_k = \sum_\ell \alpha_\ell \, \mathrm{D_{KL}}(f_\ell^* \| f_{k\ell})^2 - (\sum_\ell \alpha_\ell \, \mathrm{D_{KL}}(f_\ell^* \| f_{k\ell}))^2$. Now, using Proposition 4 of Chapter 2, we obtain that for any $k \in \mathcal{K}$,

$$A_k = K^{-1}\Big(\mathrm{D_{KL}}(h\|f) + \mathrm{D_{KL}}(h\|g)\Big) = K^{-1} I.$$

This implies that $I_{22} = 0$.

Observe next that for $k \in \mathcal{K}$,

$$\begin{aligned} B_k &= \sum_\ell \alpha_\ell \, \mathrm{D_{KL}}(f_\ell^* \| f_{k\ell})^2 - A_k^2 \\ &= K^{-1}\Big(\mathrm{D_{KL}}(h\|f)^2 + \mathrm{D_{KL}}(h\|g)^2\Big) - K^{-2} I^2. \end{aligned}$$

Because $\log Z = -\frac{1}{2} I$, we find that $\log \frac{h}{f} = \frac{1}{2} I - \frac{1}{2} \log \frac{f}{g}$ and $\log \frac{h}{g} = \frac{1}{2} I + \frac{1}{2} \log \frac{f}{g}$. By squaring these equalities and integrating against h, we find that

$$\mathrm{V_{KL}}(h\|f) + \mathrm{V_{KL}}(h\|g) = \frac{1}{2} I^2 + \frac{1}{2} J - \mathrm{D_{KL}}(h\|f)^2 - \mathrm{D_{KL}}(h\|g)^2.$$

It follows that

$$\begin{aligned} \sum_{k\in\mathcal{K}} \sum_{\ell} \alpha_k^* \alpha_\ell \, \mathrm{V_{KL}}(f_\ell^* \| f_{k\ell}) &= \sum_{k\in\mathcal{K}} \sum_{\ell\in\mathcal{K}} \alpha_k^* \alpha_\ell \, \mathrm{V_{KL}}(h\|f_{k\ell}) \\ &= \frac{1}{2} K^{-1} \sum_{k\in\mathcal{K}} (\mathrm{V_{KL}}(h\|f) + \mathrm{V_{KL}}(h\|g)) \\ &= K^{-1}\Big(\mathrm{V_{KL}}(h\|f) + \mathrm{V_{KL}}(h\|g)\Big) \\ &= K^{-1}\Big(\frac{1}{2} I^2 + \frac{1}{2} J - \mathrm{D_{KL}}(h\|f)^2 - \mathrm{D_{KL}}(h\|g)^2\Big). \end{aligned}$$

Therefore,

$$\begin{aligned} I_{21} &= \sum_{k\in\mathcal{K}}\sum_{\ell}\alpha_k^*\alpha_\ell \,\mathrm{V}_{\mathrm{KL}}(f_\ell^*\|f_{k\ell}) + \sum_{k\in\mathcal{K}}\alpha_k^* B_k \\ &= K^{-1}\Big(\frac{1}{2}I^2 + \frac{1}{2}J - \mathrm{D}_{\mathrm{KL}}(h\|f)^2 - \mathrm{D}_{\mathrm{KL}}(h\|g)^2\Big) \\ &\quad + K^{-1}\Big(\mathrm{D}_{\mathrm{KL}}(h\|f)^2 + \mathrm{D}_{\mathrm{KL}}(h\|g)^2\Big) - K^{-2}I^2 \\ &= \left(\frac{1}{2} - K^{-1}\right)K^{-1}I^2 + \frac{1}{2}K^{-1}J. \end{aligned}$$

□

Lemma 8. *Let $I = \mathrm{D}_{1/2}(f,g) = -2\log Z$ and $J = Z^{-1}\int \log^2(f/g)\sqrt{fg}$, where $Z = \int\sqrt{fg}$. Assume that $f, g > 0$ on S, and that $Z > 0$. Then*

$$J \;\le\; 8(e^{I/2} - 1).$$

Especially, $J \le 14I$ whenever $I \le 1$.

Proof. Let us fix some $x \in S$ for which $f(x) \neq g(x)$. At this point, for $t = \sqrt{\frac{f}{g}}$,

$$\frac{(\log f - \log g)^2}{(\sqrt{f}-\sqrt{g})^2}\sqrt{fg} \;=\; 4\frac{(\log\sqrt{f} - \log\sqrt{g})^2}{(\sqrt{f}-\sqrt{g})^2}\sqrt{fg} \;=\; 4\phi(t)$$

where $\phi(t) = \frac{(\log t)^2}{(t-1)^2}\,t$. Assume that $t > 1$, and let $u = \frac{1}{2}\log t$. Then $t = e^{2u}$ and

$$\phi(t) \;=\; \left(\frac{2u}{e^{2u}-1}\right)^2 e^{2u} \;=\; \left(\frac{2u}{e^u - e^{-u}}\right)^2 \;=\; \left(\frac{u}{\sinh u}\right)^2,$$

where

$$\sinh u \;=\; \frac{1}{2}(e^u - e^{-u}) \;=\; \sum_{k>0,\mathrm{odd}}\frac{u^k}{k!} \;\ge\; u.$$

Hence $\phi(t) \le 1$ for all $t > 1$. Next, by noting that $\phi(t) = \phi(1/t)$ for all $0 < t$, we conclude that $\phi(t) \le 1$ for all $t > 0$ such that $t \neq 1$. We conclude that

$$(\log f - \log g)^2\sqrt{fg} \;\le\; 4(\sqrt{f}-\sqrt{g})^2$$

whenever $f \neq g$. Obviously the same inequality holds also when $f = g$. By integrating both sides, it follows that

$$ZJ \;\le\; 4\int(\sqrt{f}-\sqrt{g})^2 \;=\; 4(2-2Z) \;=\; 8(1-Z).$$

Hence $J \le 8(Z^{-1}-1)$. The first claim follows because $Z = e^{-I/2}$. The second claim follows by noting that $e^{t/2} - 1 = \int_0^{t/2} e^s ds \le e^{1/2}t$ for $t \le 1$, and $8e^{1/2} \le 14$. □

Proposition 9. *Consider a stochastic block model defined by (2.2.1)–(2.2.3). Suppose that α is the uniform distribution over $[K]$, and that the interactions are homogeneous. Then for any estimator $\hat{\sigma} : \mathcal{X} \to \mathcal{Z}$, the expected error is lower bounded by*

$$\mathbb{E}\left(\frac{\mathrm{d}^*_{\mathrm{Ham}}(\hat{\sigma})}{N}\right) \geq \frac{1}{84}K^{-3}e^{-\frac{N}{K}I-\sqrt{8NI_{21}}} - \frac{1}{6}e^{-\frac{N}{8K}}$$

where $I_{21} = \left(\frac{1}{2} - K^{-1}\right)K^{-1}I^2 + \frac{1}{2}K^{-1}J$.

Proof. Theorem 8 states that

$$\mathbb{E}\,\mathrm{Ham}^*(\hat{\sigma}) \geq \frac{1}{21}N\alpha_{\min}^2\delta e^{-NI_1-\alpha_{\mathcal{K}}^{1/2}\delta^{-1/2}\sqrt{NI_{21}+N^2I_{22}}} - \frac{1}{6}N\alpha_{\min}Ke^{-\frac{1}{8}N\alpha_{\min}}. \tag{4.1.21}$$

Lemma 6 implies that $\min_{\mathcal{K}:|\mathcal{K}|\geq 2}\min_{f_1^*,\ldots,f_K^*} I_1 = K^{-1}D_{1/2}(f\|g)$. When the minimum is achieved, Lemmas 7 and 8 ensure that $I_{22} = 0$ and $I_{21} = \left(\frac{1}{2} - K^{-1}\right)K^{-1}I^2 + \frac{1}{2}K^{-1}J$. Furthermore, we have $\alpha_{\mathcal{K}} = \frac{2}{K}$ and $\delta = \frac{1}{4}\left(\frac{2}{K} - \frac{1}{K}\right) = \frac{1}{4K}$ since α is uniform. □

4.1.4 Miscellaneous result: multinomial concentration

Fix integers $N, K \geq 1$, and consider the space $[K]^N$ of mappings $\sigma : [N] \to [K]$. For any such mapping, we denote the frequencies of output values by $N_k(\sigma) = \sum_{i=1}^N \delta_{\sigma(i)k}$ for $k = 1, \ldots, K$. When the space $[K]^N$ is equipped with a probability measure $\mathbb{P}$, then $\sigma \mapsto (N_1(\sigma), \ldots, N_K(\sigma))$ is considered as a random variable. Given $\epsilon > 0$ and $\alpha_1, \ldots, \alpha_K \in [0, 1]$, we shall be interested in probabilities of events of the form

$$\mathcal{A}_\epsilon = \Big\{\sigma : |N_k(\sigma) - \alpha_k N| \leq \epsilon\alpha_k N \text{ for all } k \in [K]\Big\}, \tag{4.1.22}$$

$$\mathcal{A}_{\epsilon,+} = \Big\{\sigma : N_k(\sigma) \geq (1-\epsilon)\alpha_k N \text{ for all } k \in [K]\Big\}. \tag{4.1.23}$$

Lemma 9. *Let* $0 < \epsilon \leq 1$.

(i) If $\mathbb{P} = \alpha^{\otimes N}$ *for a probability measure* α *on* $[K]$, *then* $\mathbb{P}(\mathcal{A}_\epsilon^c) \leq 2\sum_{k=1}^K e^{-(\epsilon^2/3)\alpha_k N}$ *and* $\mathbb{P}(\mathcal{A}_{\epsilon,+}^c) \leq \sum_{k=1}^K e^{-(\epsilon^2/2)\alpha_k N}$.

(ii) If $\mathbb{P}$ *is the uniform distribution on* $[K]^N$, *then* $\mathbb{P}(\mathcal{A}_\epsilon^c) \leq 2e^{\log K - \epsilon^2 N/(3K)}$ *and* $\mathbb{P}(\mathcal{A}_{\epsilon,+}^c) \leq e^{\log K - \epsilon^2 N/(2K)}$.

Proof. (i) Because N_k is $\mathrm{Bin}(N, \alpha_k)$-distributed, a Chernoff bound [JŁR00, Corollary 2.3] implies that

$$\mathbb{P}(|N_k(\sigma) - \alpha_k N| > \epsilon\alpha_k N) \leq 2e^{-(\epsilon^2/3)\alpha_k N}.$$

Similarly, another Chernoff bound [JŁR00, Theorem 2.1] implies that

$$\mathbb{P}(N_k(\sigma) \le (1-\epsilon)\alpha_k N) \;\le\; e^{-(\epsilon^2/2)\alpha_k N}.$$

Hence the first claim follows by the union bound.

(ii) The second claim follows immediately from (i) after noting that the uniform distribution on $[K]^N$ can be represented as $\pi = \alpha^{\otimes N}$ where $\alpha_k = K^{-1}$ for all k. □

We shall also be interested in random variables defined by $N_{\min}(\sigma) = \min_k N_k(\sigma)$ and $\Delta N(\sigma) = \max_{k,\ell} |N_k(\sigma) - N_\ell(\sigma)|$. The following result implies that for large-scale uniformly distributed settings with $N \gg K \log K$, these random variables are bounded by $N_{\min} \ge (1-\epsilon)K^{-1}N$ and $\Delta N \le 2\epsilon K^{-1}N$ with high probability for $(\frac{K\log K}{N})^{1/2} \ll \epsilon \le 1$. For example, we may select $\epsilon = (\frac{K\log K}{N})^{0.499}$.

Lemma 10. *Let $0 < \epsilon \le 1$. (i) If $\mathbb{P} = \alpha^{\otimes N}$ for a probability measure α on $[K]$, then*

$$\begin{aligned} \mathbb{P}\Big(N_{\min} \ge (1-\epsilon)\alpha_{\min} N\Big) &\ge 1-\delta_1, \\ \mathbb{P}\Big(\Delta N \le (2\epsilon\alpha_{\max} + \Delta\alpha) N\Big) &\ge 1-\delta_2, \end{aligned} \tag{4.1.24}$$

where $\delta_1 = Ke^{-(\epsilon^2/2)\alpha_{\min}N}$ and $\delta_2 = 2Ke^{-(\epsilon^2/3)\alpha_{\min}N}$, together with $\alpha_{\min} = \min_k \alpha_k$, $\alpha_{\max} = \max_k \alpha_k$, and $\Delta\alpha = \max_{k,\ell} |\alpha_k - \alpha_\ell|$.

(ii) If $\mathbb{P}$ is the uniform distribution on $[K]^N$, then

$$\begin{aligned} \mathbb{P}\Big(N_{\min} \ge (1-\epsilon)K^{-1}N\Big) &\ge 1-\delta_1, \\ \mathbb{P}\Big(\Delta N \le 2\epsilon K^{-1}N\Big) &\ge 1-\delta_2, \end{aligned}$$

with $\delta_1 = e^{\log K - \epsilon^2 N/(2K)}$ and $\delta_2 = 2e^{\log K - \epsilon^2 N/(3K)}$.

Proof. (i) By Lemma 9, then events $\mathcal{A}_\epsilon$ and $\mathcal{A}_{\epsilon,+}$ defined by (4.1.22)–(4.1.23) satisfy $\mathbb{P}(\mathcal{A}_{\epsilon_+}^c) \le \delta_1$ and $\mathbb{P}(\mathcal{A}_\epsilon^c) \le \delta_2$. On the event $\mathcal{A}_{\epsilon,+}$, $N_{\min} \ge (1-\epsilon)\alpha_{\min}N$. Hence the first inequality in (4.1.24) follows. For the second inequality, we note that on the event $\mathcal{A}_\epsilon$

$$\begin{aligned} |N_k - N_\ell| &\le |N_k - \alpha_k N| + |N_\ell - \alpha_\ell N| + |\alpha_k N - \alpha_\ell N| \\ &\le \epsilon\alpha_k N + \epsilon\alpha_\ell N + |\alpha_k - \alpha_\ell| N \\ &\le 2\epsilon\alpha_{\max} N + \Delta\alpha N \end{aligned}$$

for all k, ℓ. This confirms the second inequality in (4.1.24).

(ii) This follows immediately from (i) after noting that the uniform distribution on $[K]^N$ can be represented as $\pi = \alpha^{\otimes N}$ where $\alpha_k = K^{-1}$ for all k. □

4.2 Upper bound on ML estimation error

This section is devoted to analysing the accuracy of maximum-likelihood estimators. Section 4.2.1 describes how ML estimation error probabilities are characterised by Mirkin distances. Section 4.2.2 provides an upper bound on a worst-case ML estimation error among balanced block structures. Section 4.2.3 provides an upper bound (Proposition 11) on an average ML estimation error among all block structures, which confirms the upper bound of Theorem 3, and also shows that any maximum-likelihood estimator achieves the upper bound. Section 4.2.4 analyses the upper bound of Theorem 3 in a large-scale setting and yields a proof of the existence part of Theorem 4, summarised as Proposition 12.

4.2.1 Maximum likelihood estimators

We denote the set of all node labellings by $\mathcal{Z} = [K]^{[N]}$. A maximum likelihood estimator of σ is a map $\hat{\sigma} : \mathcal{X} \to \mathcal{Z}$ such that

$$P_{\hat{\sigma}_x}(x) \geq P_{\sigma'}(x) \qquad \text{for all } \sigma' \in \mathcal{Z} \text{ and } x \in \mathcal{X}. \tag{4.2.1}$$

The following results helps to analyse situations in which a maximum likelihood estimator produces outputs diverging from the correct value. The result is stated using the Mirkin distance $d_{\mathrm{Mir}}(\sigma, \sigma')$ defined in Section 4.2.5.

Lemma 11. *For a homogeneous SBM with N nodes, K blocks, and interaction distributions f and g with $I = \mathrm{D}_{1/2}(f, g)$*

$$P_\sigma\{x : P_{\sigma'}(x) \geq P_\sigma(x)\} \;\leq\; e^{-\frac{1}{4} d_{\mathrm{Mir}}(\sigma,\sigma') I},$$

for all node labellings σ, σ'.

Proof. Observe that $P_\sigma\{x : P_{\sigma'}(x) \geq P_\sigma(x)\} = P_\sigma(\ell \geq 0)$, where the log-likelihood ratio $\ell(x) = \log \frac{P_{\sigma'}(x)}{P_\sigma(x)}$ is viewed as a random variable on probability space $(\mathcal{X}, P_\sigma)$. Also denote by E (resp. E') the set of node pairs $\{i, j\}$ for which $\sigma(i) = \sigma(j)$ (resp. $\sigma'(i) = \sigma'(j)$). Then we find that

$$\ell(x) \;=\; \sum_{ij \in E' \setminus E} \log \frac{f}{g}(x_{ij}) \;-\; \sum_{ij \in E \setminus E'} \log \frac{f}{g}(x_{ij}).$$

Therefore, the distribution of $x \mapsto \ell(x)$ on the probability space $(\mathcal{X}, P_\sigma)$ is the same as the law of

$$\sum_{j=1}^{|E' \setminus E|} \log \frac{f}{g}(Y_j) - \sum_{i=1}^{|E \setminus E'|} \log \frac{f}{g}(X_i),$$

in which the random variables X_i, Y_j are mutually independent and distributed according to $\mathrm{Law}(X_i) = f$ and $\mathrm{Law}(Y_j) = g$. By applying Markov's inequality and the above representation, we find that

$$P_\sigma(\ell \geq 0) \;=\; P_\sigma(e^{\frac{1}{2}\ell} \geq 1) \;\leq\; E_\sigma e^{\frac{1}{2}\ell} \;=\; e^{-\frac{1}{2}(|E' \setminus E| + |E \setminus E'|) I},$$

where $I = \mathrm{D}_{1/2}(f, g)$. Hence the claim follows. □

4.2.2 Upper bound among balanced node labellings

The following result is key minimax upper bound characterising the worst-case estimation accuracy among block structures which are balanced according to $\sigma \in \mathcal{Z}_{1-\epsilon,1+\epsilon}$, where

$$\mathcal{Z}_{a,b} \;=\; \left\{\sigma \in \mathcal{Z} :\; a\frac{N}{K} \le |\sigma^{-1}(k)| \le b\frac{N}{K}\right\}, \tag{4.2.2}$$

and we recall that $\mathcal{Z} = [K]^{[N]}$. Similar upper bounds in the context of binary SBMs have been derived in [ZZ16].

Proposition 10. *For a homogeneous SBM with N nodes and K blocks, any estimator $\hat{\sigma} : \mathcal{X} \to \mathcal{Z}$ satisfying the MLE property (4.2.1) has classification error bounded by*

$$\max_{\sigma \in \mathcal{Z}_{1-\epsilon,1+\epsilon}} E_\sigma \, \mathrm{d}^*_{\mathrm{Ham}}(\sigma, \hat{\sigma}) \;\le\; 8eN(K-1)e^{-(1-\zeta-\kappa)\frac{NI}{K}} + NK^N e^{-\frac{1}{4}(\frac{\zeta}{K-1}-\epsilon)(N/K)^2 I}$$

for all $0 \le \epsilon \le \zeta \le \frac{1}{21}$, where $\kappa = 56\max\{K^2 e^{-\frac{1}{8}\frac{NI}{K}},\, KN^{-1}\}$ and $I = \mathrm{D}_{1/2}(f, g)$.

Proof. We note that due to homogeneity, $P_\sigma = P_{[\sigma]}$ depends on σ only via the partition $[\sigma] = \{\sigma^{-1}(k) : k \in [K]\}$. A similar observation also holds for the absolute classification error $\mathrm{d}^*_{\mathrm{Ham}}(\sigma_1, \sigma_2) = \mathrm{d}^*_{\mathrm{Ham}}([\sigma_1], [\sigma_2])$. In the proof we denote by $\mathcal{P}_{1-\epsilon,1+\epsilon} = \{[\sigma] : \sigma \in \mathcal{Z}_{1-\epsilon,1+\epsilon}\}$ the collection of partitions corresponding to node labellings in $\mathcal{Z}_{1-\epsilon,1+\epsilon}$. We select a node labelling $\sigma \in \mathcal{Z}_{1-\epsilon,1+\epsilon}$, and split the error according to

$$E_\sigma L \;=\; E_\sigma L1(\hat{\sigma} \in \mathcal{Z}_{1-\zeta,1+\zeta}) + E_\sigma L1(\hat{\sigma} \notin \mathcal{Z}_{1-\zeta,1+\zeta}). \tag{4.2.3}$$

The remainder of the proof consists of two parts, where we derive upper bounds for both terms on the right side above.

(i) For analysing the first term on the right side of (4.2.3), we note that $\hat{\sigma} \in \mathcal{Z}_{1-\zeta,1+\zeta}$ if and only if $[\hat{\sigma}] \in \mathcal{P}_{1-\zeta,1+\zeta}$, and therefore,

$$E_\sigma L1(\hat{\sigma} \in \mathcal{Z}_{1-\zeta,1+\zeta}) \;=\; \sum_{m=1}^{N} m p_m \tag{4.2.4}$$

where $p_m = P_\sigma\{x : [\hat{\sigma}_x] \in \mathcal{P}_{1-\zeta,1+\zeta}(\sigma, m)\}$ is the probability of the event that the partition associated to $\hat{\sigma}_x$ belongs to the set

$$\mathcal{P}_{1-\zeta,1+\zeta}(\sigma, m) \;=\; \{\theta \in \mathcal{P}_{1-\zeta,1+\zeta} : \mathrm{d}^*_{\mathrm{Ham}}([\sigma], \theta) = m\}.$$

On such event there exists a partition $\theta \in \mathcal{P}_{1-\zeta,1+\zeta}(\sigma, m)$ such that $P_\theta(x) \ge P_{[\sigma]}(x)$. Hence by the union bound,

$$p_m \;\le \sum_{\theta \in \mathcal{P}_{1-\zeta,1+\zeta}(\sigma,m)} P_\sigma\{x : P_\theta(x) \ge P_{[\sigma]}(x)\}.$$

Observe next that to every partition $\theta \in \mathcal{P}_{1-\zeta,1+\zeta}(\sigma, m)$ there corresponds exactly $K!$ node labellings σ' belonging to the set

$$\mathcal{Z}_{1-\zeta,1+\zeta}(\sigma, m) \;=\; \{\sigma' \in \mathcal{Z}_{1-\zeta,1+\zeta} : \mathrm{d}^*_{\mathrm{Ham}}(\sigma, \sigma') = m\}.$$

Therefore, the above upper bound can be rewritten as

$$p_m \;\leq\; (K!)^{-1} \sum_{\sigma' \in \mathcal{Z}_{1-\zeta,1+\zeta}(\sigma,m)} P_\sigma\{x : P_{\sigma'}(x) \geq P_\sigma(x)\}. \tag{4.2.5}$$

Let us next analyse the probabilities on the right side of (4.2.5). By Lemma 11, we find that

$$P_\sigma\{x : P_{\sigma'}(x) \geq P_\sigma(x)\} \;\leq\; e^{-\frac{1}{4} d_{\mathrm{Mir}}(\sigma,\sigma') I}.$$

Because $\epsilon \leq \zeta$, it follows that $\mathcal{Z}_{1-\epsilon,1+\epsilon} \subset \mathcal{Z}_{1-\zeta,1+\zeta}$. We note that $\frac{1}{4} d_{\mathrm{Mir}}(\sigma, \sigma') = \frac{1}{2}(|E \setminus E'| + |E' \setminus E| \geq \min\{|E \setminus E'|, |E' \setminus E|\}$, where E (resp. E') denotes the set of node pairs for which σ (resp. σ') assigns the same label. With the help of Lemma 14 we then find that for all $\sigma, \sigma' \in \mathcal{Z}_{1-\zeta,1+\zeta}$, such that $\mathrm{d}^*_{\mathrm{Ham}}(\sigma, \sigma') = m$,

$$\frac{1}{4} d_{\mathrm{Mir}}(\sigma, \sigma') \;\geq\; \max\left\{(1-\zeta)\frac{N}{K} - m,\; \frac{1}{3}(1-\zeta)\frac{N}{K} - \frac{1}{6}(1+\zeta)\frac{N}{K}\right\} m.$$

We note that $\frac{1}{3}(1-\zeta) - \frac{1}{6}(1+\zeta) = \frac{1}{6} - \frac{1}{2}\zeta \geq \frac{1}{7}$ when $\zeta \leq \frac{1}{21}$. Hence,

$$\frac{1}{4} d_{\mathrm{Mir}}(\sigma, \sigma') \;\geq\; \max\left\{(1-\zeta)\frac{N}{K} - m,\; \frac{1}{7}\frac{N}{K}\right\} m,$$

and we conclude that for all $\sigma \in \mathcal{Z}_{1-\epsilon,1+\epsilon}$ and $\sigma' \in \mathcal{Z}_{1-\zeta,1+\zeta}$,

$$P_\sigma\{x : P_{\sigma'}(x) \geq P_\sigma(x)\} \;\leq\; \min\left\{e^{-(1-\zeta)\frac{NI}{K} + mI},\; e^{-\frac{1}{7}\frac{NI}{K}}\right\}^m. \tag{4.2.6}$$

Furthermore, let us analyse the cardinality of the sum on (4.2.5). Because $\mathrm{d}^*_{\mathrm{Ham}}(\sigma, \sigma') = m$ if and only if $\mathrm{Ham}(\tau \circ \sigma, \sigma') = m$ for some $\tau \in \mathrm{Sym}(K)$, a union bound combined with Lemma 15 implies that

$$|\mathcal{Z}_{1-\zeta,1+\zeta}(\sigma, m)| \;\leq\; K!\,|\{\sigma' \in \mathcal{Z} : \mathrm{Ham}(\sigma, \sigma') = m\}| \;\leq\; K!\left(\frac{eN(K-1)}{m}\right)^m.$$

By combining this bound with (4.2.5) and (4.2.6), we may now conclude that

$$p_m \;\leq\; \min\left\{\frac{eN(K-1)}{m} e^{-(1-\zeta)\frac{NI}{K} + mI},\; \frac{eN(K-1)}{m} e^{-\frac{1}{7}\frac{NI}{K}}\right\}^m. \tag{4.2.7}$$

We will now apply the bounds in (4.2.7) to derive an upper bound for the sum in (4.2.4) which we will split according to

$$\sum_{m=1}^{N} m p_m \;=\; \sum_{m \leq m_1} m p_m \;+\; \sum_{m_1 < m \leq N} m p_m \tag{4.2.8}$$

using a threshold parameter m_1. We will also select another threshold parameter $0 < m_0 \leq m_1$. Using these, the probabilities p_m are bounded by $p_m \leq s_1^m$ for $m_0 \leq m \leq m_1$, and $p_m \leq s_2^m$ for $m \geq m_1$, where

$$s_1 = \frac{eN(K-1)}{m_0} e^{-(1-\zeta)\frac{NI}{K}+m_1 I} \qquad \text{and} \qquad s_2 = \frac{eN(K-1)}{m_1} e^{-\frac{1}{7}\frac{NI}{K}}.$$

To obtain a good upper bound, m_1 should be small enough to keep the exponent in s_1 small, and large enough so that $s_2 < 1$. From the latter point of view, we see that $s_2 \leq \frac{1}{2}$ when $m_1 \geq 2eN(K-1)e^{-\frac{1}{7}\frac{NI}{K}}$. To leave some headroom, we set a slightly larger m_1 corresponding to $\frac{1}{7}$ replaced by $\frac{1}{8}$. For later use purposes, we also require that $m_1 \geq 56$ which guarantees that $\frac{1}{m_1} \leq \frac{1}{7} - \frac{1}{8}$. Therefore, we set

$$m_1 \;=\; 2eN(K-1)e^{-\frac{1}{8}\frac{NI}{K}} \vee 56.$$

With this choice, we find that $s_2 \leq \frac{1}{2} e^{-\frac{1}{56}\frac{NI}{K}} \leq \frac{1}{2}$. Hence,

$$\sum_{m_1 < m \leq N} m p_m \;\leq\; N \sum_{m \geq m_1} s_2^m \;=\; N \frac{s_2^{\lceil m_1 \rceil}}{1-s_2} \;\leq\; N \frac{s_2^{m_1}}{1-s_2} \;\leq\; 2N s_2^{m_1}.$$

Furthermore, $m_1 \geq 56$ implies that $s_2^{m_1} \leq e^{-\frac{m_1}{56}\frac{NI}{K}} \leq e^{-\frac{NI}{K}}$. It follows that the second term on the right side of (4.2.8) is bounded by

$$\sum_{m_1 < m \leq N} m p_m \;\leq\; 2N e^{-\frac{NI}{K}}. \tag{4.2.9}$$

Let us next derive an upper bound for the first term on the right side of (4.2.8). We define $B = eN(K-1)e^{-(1-\zeta)\frac{NI}{K}+m_1 I}$, and consider the following two cases.
(a) If $B \leq \frac{1}{2}$, we set $m_0 = 1$, which implies that $s_1 = B$, and we find that

$$\sum_{1 \leq m \leq m_1} m p_m \;\leq\; \sum_{m=1}^{\infty} m s_1^m \;=\; \sum_{m=1}^{\infty} m B^m \;=\; \frac{B}{(1-B)^2} \;\leq\; 4B. \tag{4.2.10}$$

(b) If $B > \frac{1}{2}$, we set $m_0 = 2B$, so that $s_1 = \frac{1}{2}$, and we find that

$$\sum_{1 \leq m \leq m_1} m p_m \;=\; \sum_{1 \leq m \leq m_0} m p_m + \sum_{m_0 < m \leq m_1} m p_m \;\leq\; m_0 \;+ \sum_{m > m_0} m s_1^m.$$

By noting that $m_0 > 1$, we find that $2 \leq \lfloor m_0 \rfloor + 1 \leq 2m_0$. Then by applying Lemma 16 it follows that

$$\sum_{m > m_0} m s_1^m \;=\; \sum_{m=\lfloor m_0 \rfloor + 1}^{\infty} m 2^{-m} \;\leq\; 4(\lfloor m_0 \rfloor + 1) 2^{-(\lfloor m_0 \rfloor + 1)} \;\leq\; 2m_0.$$

Hence, $\sum_{1 \leq m \leq m_1} m p_m \leq 3m_0 = 6B$. In light of (4.2.10), we conclude that the latter conclusion holds for both $B \leq \frac{1}{2}$ and $B > \frac{1}{2}$. By combining these observations with (4.2.9), and noting that $B \geq N e^{-\frac{NI}{K}}$, it follows that

$$\sum_{1 \leq m \leq N} m p_m \;\leq\; 2N e^{-\frac{NI}{K}} + 6B \;\leq\; 8B \;=\; 8eN(K-1)e^{-(1-\zeta)\frac{NI}{K}+m_1 I}.$$

After noting that $m_1 I = \frac{NI}{K} \max\{2eK(K-1)e^{-\frac{1}{8}\frac{NI}{K}}, 56\frac{K}{N}\}$, we see that $m_1 I \leq \kappa \frac{NI}{K}$ for $\kappa = 56 \max\{K^2 e^{-\frac{1}{8}\frac{NI}{K}}, KN^{-1}\}$. Then we conclude that the first term on the right side of (4.2.3) is bounded by

$$E_\sigma L1(\hat{\sigma} \in \mathcal{Z}_{1-\zeta,1+\zeta}) \leq 8eN(K-1)e^{-(1-\zeta-\kappa)\frac{NI}{K}}. \tag{4.2.11}$$

(ii) Finally, it remains to derive an upper bound for the second term on the right side of (4.2.3). Denote $\gamma = (K-1)^{-1}\zeta$. Then the generic bound $N \leq N_{\min}(\sigma') + (K-1)N_{\max}(\sigma')$ implies that $N_{\min}(\sigma') \geq N - (K-1)(1+\gamma)\frac{N}{K} = (1-\zeta)\frac{N}{K}$ for all $\sigma' \in \mathcal{Z}_{0,1+\gamma}$. Therefore, $\mathcal{Z}_{0,1+\gamma} \subset \mathcal{Z}_{1-\zeta,1+\gamma} \subset \mathcal{Z}_{1-\zeta,1+\zeta}$. Especially,

$$P_\sigma(\hat{\sigma} \notin \mathcal{Z}_{1-\zeta,1+\zeta}) \leq P_\sigma(\hat{\sigma} \notin \mathcal{Z}_{0,1+\gamma}).$$

On the event that $\hat{\sigma} \notin \mathcal{Z}_{0,1+\gamma}$, the MLE property (4.2.1) implies that there exists σ' with $N_{\max}(\sigma') > (1+\gamma)\frac{N}{K}$ for which $P_{\sigma'}(x) \geq P_\sigma(x)$. For any such σ', $N_{\max}(\sigma') - N_{\max}(\sigma) \geq (\gamma - \epsilon)\frac{N}{K}$, so that by Lemma 12, we see that

$$d_{\mathrm{Mir}}(\sigma, \sigma') = 2(|E \setminus E'| + |E' \setminus E|) \geq 2|E' \setminus E| \geq (\gamma - \epsilon)(N/K)^2.$$

By Lemma 11, we conclude that for all $\sigma' \notin \mathcal{Z}_{0,1+\gamma}$,

$$P_\sigma\{x : P_{\sigma'}(x) \geq P_\sigma(x)\} \leq e^{-\frac{1}{4}(\gamma-\epsilon)(N/K)^2 I}.$$

Hence, by a crude union bound it follows that

$$P_\sigma(\hat{\sigma} \notin \mathcal{Z}_{1-\zeta,1+\zeta}) \leq P_\sigma(\hat{\sigma} \notin \mathcal{Z}_{0,1+\gamma}) \leq K^N e^{-\frac{1}{4}(\gamma-\epsilon)(N/K)^2 I}, \tag{4.2.12}$$

and we conclude that second term on the right side of (4.2.3) is bounded by

$$E_\sigma L1(\hat{\sigma} \notin \mathcal{Z}_{1-\zeta,1+\zeta}) \leq NP_\sigma(\hat{\sigma} \notin \mathcal{Z}_{1-\zeta,1+\zeta}) \leq NK^N e^{-\frac{1}{4}(\gamma-\epsilon)(N/K)^2 I}. \tag{4.2.13}$$

The claim now follows by combining (4.2.11)–(4.2.13). □

4.2.3 Upper bound on average error among all node labellings

The following result is the upper bound of Theorem 3.

Proposition 11. *For a homogeneous SBM with N nodes and K blocks, any estimator $\hat{\sigma} : \mathcal{X} \to \mathcal{Z}$ satisfying the MLE property (4.2.1) has classification error bounded by*

$$\mathbb{E}\, \mathrm{d}^*_{\mathrm{Ham}}(\sigma, \hat{\sigma}) \leq 8eN(K-1)e^{-(1-\zeta-\kappa)\frac{NI}{K}} + NK^N e^{-\frac{1}{4}(\frac{\zeta}{K-1}-\epsilon)(N/K)^2 I} + 2NKe^{-\frac{1}{3}\epsilon^2 \frac{N}{K}},$$

for all $0 \leq \epsilon \leq \zeta \leq \frac{1}{21}$, where $\kappa = 56 \max\{K^2 e^{-\frac{1}{8}\frac{NI}{K}}, KN^{-1}\}$ and $I = \mathrm{D}_{1/2}(f, g)$.

Proof. Denote $L = \mathrm{d}^*_{\mathrm{Ham}}(\sigma, \hat{\sigma})$. By noting that the classification error is bounded by $L \leq N$ with probability one, it follows that

$$\begin{aligned}\mathbb{E}L \;&\leq \sum_{\sigma \in \mathcal{Z}_{1-\epsilon,1+\epsilon}} \pi_\sigma E_\sigma L \;+ \sum_{\sigma \in \mathcal{Z}^c_{1-\epsilon,1+\epsilon}} \pi_\sigma E_\sigma L \\ &\leq \max_{\sigma \in \mathcal{Z}_{1-\epsilon,1+\epsilon}} E_\sigma L \;+ N\pi(\mathcal{Z}^c_{1-\epsilon,1+\epsilon}).\end{aligned}$$

For a random node labelling $\sigma = (\sigma_1, \ldots, \sigma_N)$ sampled from the uniform distribution π on $\mathcal{Z}$, we see that coordinates are mutually independent and uniformly distributed on in $[K]$. A multinomial concentration inequality (Lemma 10) then implies that

$$\pi(\mathcal{Z}^c_{1-\epsilon,1+\epsilon}) \;\leq\; 2Ke^{-\frac{1}{3}\epsilon^2 \frac{N}{K}}.$$

The claim follows by Proposition 10. □

4.2.4 Upper bound for large-scale settings

The following result implies the existence statements of Theorem 4.

Proposition 12. *Consider a large-scale homogeneous SBM with $N \gg 1$ nodes and $K = O(1)$ blocks, and interaction distributions f, g such that $I = \mathrm{D}_{1/2}(f, g)$, and let $\hat{\sigma}$ be any estimator having the MLE property (4.2.1). Then the following hold:*

(i) if $I \gg N^{-1}$, then the estimator $\hat{\sigma}$ is consistent;

(ii) if $I \geq (1 + \Omega(1))\frac{K \log N}{N}$, then the estimator $\hat{\sigma}$ is strongly consistent.

Proof. Denote $L = \mathrm{d}^*_{\mathrm{Ham}}(\hat{\sigma})$. By Proposition 11, we see that

$$\mathbb{E}L \;\leq\; 8eNKe^{-(1-\zeta-\kappa)\frac{NI}{K}} + NK^N e^{-\frac{1}{4}(\frac{\zeta}{K-1}-\epsilon)(N/K)^2 I} + 2NKe^{-\frac{1}{3}\epsilon^2\frac{N}{K}}, \tag{4.2.14}$$

where $\kappa = 56 \max\{K^2 e^{-\frac{1}{8}\frac{NI}{K}},\, KN^{-1}\}$ and $I = \mathrm{D}_{1/2}(f, g)$, and where we are free to choose any $0 \leq \epsilon \leq \zeta \leq \frac{1}{21}$,

(i) Recall that we have assumed that $K \ll (\frac{N}{\log N})^{1/3}$ and $I \gg \frac{K^3 \log K}{N}$. Let us define $\epsilon = 3(\frac{K \log N}{N})^{1/2}$ and $\zeta = \epsilon K + 5\frac{K^3 \log K}{NI}$. Furthermore, $\epsilon \ll K^{-1}$ due to our assumption $K \ll (\frac{N}{\log N})^{1/3}$. Then it also follows that $\zeta \ll 1$. Furthermore, we find that $e^{-\frac{1}{3}\epsilon^2\frac{N}{K}} = N^{-3}$, and the last term on the right side of (4.2.14) equals $2KN^{-2}$. We also find that

$$\left(\frac{\zeta}{K-1} - \epsilon\right)(N/K)^2 I \;\geq\; \left(\frac{\zeta}{K} - \epsilon\right)(N/K)^2 I \;=\; 5N \log K,$$

so that the middle term on the right side of (4.2.14) is bounded by

$$NK^N e^{-\frac{1}{4}(\frac{\zeta}{K-1}-\epsilon)(N/K)^2 I} \;\leq\; NK^N K^{-\frac{5}{4}N} \;=\; NK^{-\frac{1}{4}N}.$$

We conclude that

$$\mathbb{E}L \;\leq\; 8eNKe^{-(1-\zeta-\kappa)\frac{NI}{K}} + NK^{-\frac{1}{4}N} + 2N^{-2}K.$$

We note that $\kappa \ll 1$ and $\log(8eK) \ll \frac{NI}{K}$ when $I \gg \frac{K\log K}{N}$. We note that $NK^{-\frac{1}{4}N} \leq N2^{-\frac{1}{4}N} = o(1)$. We conclude that

$$\mathbb{E}L \;\leq\; Ne^{-(1-o(1))\frac{NI}{K}} + o(1)$$

for homogeneous SBMs with $2 \leq K \ll (\frac{N}{\log N})^{1/3}$ and $I \gg \frac{K^3\log K}{N}$.

(ii) The condition for strong consistency follows immediately from the above bounds. □

4.2.5 Comparing partitions

Classification error

The absolute classification error between node labellings $\sigma, \sigma' : [N] \to [K]$ is defined by

$$\mathrm{d}^*_{\mathrm{Ham}}(\sigma, \sigma') \;=\; \min_{\rho \in \mathrm{Sym}(K)} \mathrm{Ham}(\sigma, \rho \circ \sigma'),$$

where $\mathrm{Ham}(\sigma, \sigma') = \sum_{i=1}^N 1(\sigma(i) \neq \sigma'(i))$ denotes the Hamming distance and $\mathrm{Sym}(K)$ denotes the group of permutations on $[K]$. We note that $\mathrm{d}^*_{\mathrm{Ham}}(\sigma, \sigma') = \mathrm{d}^*_{\mathrm{Ham}}(\rho\circ\sigma, \rho'\circ\sigma')$ for all $\rho, \rho' \in \mathrm{Sym}(K)$, which confirms that the classification error depends on its inputs only via the partitions induced by the preimages of the node labellings. The relative error $N^{-1}\,\mathrm{d}^*_{\mathrm{Ham}}(\sigma, \sigma')$ is called usually just called the classification error [MH01; Mei07].

Mirkin distance

The Mirkin distance is one of the common pair-counting based cluster validity indices [GTP21; Lei+17], and it is related to the Rand index by $d_{\mathrm{Mir}}(\sigma, \sigma') = N(N-1)(1 - d_{\mathrm{Rand}}(\sigma, \sigma'))$. The Mirkin distance between node labellings $\sigma, \sigma' : [N] \to [K]$ is defined by

$$d_{\mathrm{Mir}}(\sigma, \sigma') \;=\; 2 \sum_{1 \leq i < j \leq N} \big(e_{ij}(1 - e'_{ij}) + (1 - e_{ij})e'_{ij}\big)$$

where $e_{ij} = 1(\sigma(i) = \sigma(j))$ and $e'_{ij} = 1(\sigma'(i) = \sigma'(j))$.

For any node labelling $\sigma : [N] \to [K]$, we denote by $E(\sigma)$ the set of unordered node pairs $\{i, j\}$ such that $\sigma(i) = \sigma(j)$, by $N^\sigma_{\min} = \min_k |C_k|$ and $N^\sigma_{\max} = \max_k |C_k|$ where $C_k = \{i : \sigma(i) = k\}$. Then we note that the Mirkin distance can be written as

$$d_{\mathrm{Mir}}(\sigma, \sigma') \;=\; 2\left(|E(\sigma) \setminus E(\sigma')| + |E(\sigma') \setminus E(\sigma)|\right). \tag{4.2.15}$$

The following result shows that when the Mirkin metric is small, then the maximum set sizes in two partitions cannot differ arbitrarily much.

Lemma 12. *For any node labellings $\sigma, \sigma' : [N] \to [K]$,*

$$|E(\sigma) \setminus E(\sigma')| \;\geq\; \frac{1}{2}(N_{\max}^{\sigma} - N_{\max}^{\sigma'})N_{\max}^{\sigma}.$$

Proof. For any k, denote by $E(C_k)$ the set of unordered node pairs in C_k. Also denote $N_k = |E(C_k)|$, $N'_\ell = |E(C'_\ell)|$, and $N_{k\ell} = |C_k \cap C'_\ell|$. Then we find that

$$|E(C_k) \setminus E(\sigma')| \;=\; \binom{N_k}{2} - \sum_{\ell} \binom{N_{k\ell}}{2}.$$

By applying the bound $N_{k\ell} \leq N_{\max}^{\sigma'}$, we see that

$$\sum_{\ell} \binom{N_{k\ell}}{2} \;\leq\; \frac{1}{2}(N_{\max}^{\sigma'} - 1) \sum_{\ell} N_{k\ell} \;=\; \frac{1}{2}(N_{\max}^{\sigma'} - 1)N_k.$$

Therefore,

$$|E(C_k) \setminus E(\sigma')| \;\geq\; \binom{N_k}{2} - \frac{1}{2}(N_{\max}^{\sigma'} - 1)N_k \;\geq\; \frac{1}{2}(N_k - N_{\max}^{\sigma'})N_k.$$

The claim now follows after noting that

$$|E(\sigma) \setminus E(\sigma')| \;=\; \sum_{k} |E(C_k) \setminus E(\sigma')| \;\geq\; \max_{k} |E(C_k) \setminus E(\sigma')|.$$

□

Optimal alignments

The confusion matrix of node labellings $\sigma, \sigma' : [N] \to [K]$ is the K-by-K matrix having entries

$$N_{k\ell} \;=\; |C_k \cap C'_\ell|,$$

where $C_k = \sigma^{-1}(k)$ and $C'_\ell = (\sigma')^{-1}(\ell)$. We say that node labellings $\sigma, \sigma' : [N] \to [K]$ are optimally aligned if

$$\mathrm{d}^{*}_{\mathrm{Ham}}(\sigma, \sigma') \;=\; \mathrm{Ham}(\sigma, \sigma'). \tag{4.2.16}$$

The following result provides an entrywise upper bound for the confusion matrix of optimally aligned node labellings.

Lemma 13. *If σ and σ' are optimally aligned, then the associated confusion matrix is bounded by*

$$N_{k\ell} + N_{\ell k} \;\leq\; N_{kk} + N_{\ell\ell} \tag{4.2.17}$$

and

$$N_{k\ell} \;\leq\; \frac{1}{3}(N_k + N'_\ell) \tag{4.2.18}$$

for all $k \neq \ell$, where $N_k = \sum_\ell N_{k\ell}$ and $N'_\ell = \sum_k N_{k\ell}$.

Proof. Fix some distinct $k, \ell \in [K]$. Define $\sigma'' = \tau \circ \sigma'$ where τ is the K-permutation which swaps k and ℓ and leaves other elements of $[K]$ intact. Denote $C''_j = (\sigma'')^{-1}(j)$. Then we see that $C''_j = C'_\ell$ for $j = k$, $C''_j = C'_k$ for $j = \ell$, and $C''_j = C'_j$ otherwise. Using the formulas

$$\mathrm{Ham}(\sigma, \sigma') = \sum_j |C_j \setminus C'_j| \qquad \text{and} \qquad \mathrm{Ham}(\sigma, \sigma'') = \sum_j |C_j \setminus C''_j|$$

we find that

$$\mathrm{Ham}(\sigma, \sigma'') - \mathrm{Ham}(\sigma, \sigma') \;=\; |C_k \setminus C'_\ell| - |C_k \setminus C'_k| + |C_\ell \setminus C'_k| - |C_\ell \setminus C'_\ell|.$$

Because

$$|C_k \setminus C'_\ell| - |C_k \setminus C'_k| \;=\; (N_k - N_{k\ell}) - (N_k - N_{kk}) \;=\; N_{kk} - N_{k\ell},$$

and the same formula holds also with the roles of k and ℓ swapped, it follows that

$$\mathrm{Ham}(\sigma, \sigma'') - \mathrm{Ham}(\sigma, \sigma') \;=\; N_{kk} - N_{k\ell} + N_{\ell\ell} - N_{\ell k}.$$

Because σ and σ' are optimally aligned, we see that $\mathrm{Ham}(\sigma, \sigma') \leq \mathrm{Ham}(\sigma, \sigma'')$. Therefore, the left side of the above equality is nonnegative, and (4.2.17) follows.

Next, by applying the bounds $N_{kk} \leq N_k - N_{k\ell}$ and $N_{\ell\ell} \leq N'_\ell - N_{k\ell}$, we may conclude that

$$0 \;\leq\; N_{kk} - N_{k\ell} + N_{\ell\ell} - N_{\ell k} \;\leq\; N_k + N'_\ell - 3N_{k\ell} - N_{\ell k}.$$

The inequality (4.2.18) now follows by noting that

$$N_{k\ell} \;\leq\; \frac{1}{3}(N_k + N'_\ell - N_{\ell k}) \;\leq\; \frac{1}{3}(N_k + N'_\ell).$$

□

Relating the classification error and the Mirkin distance

The next result provides a way to bound the absolute classification error $\mathrm{d}^*_{\mathrm{Ham}}(\sigma, \sigma')$ using the Mirkin metric $d_{\mathrm{Mir}}(\sigma, \sigma')$. For any node labelling $\sigma : [N] \to [K]$, we denote by $E(\sigma)$ the set of unordered node pairs $\{i, j\}$ such that $\sigma(i) = \sigma(j)$, by $N^\sigma_{\min} = \min_k |C_k|$ and $N^\sigma_{\max} = \max_k |C_k|$ where $C_k = \{i : \sigma(i) = k\}$. Then we note that the Mirkin distance can be written using $E(\sigma), E(\sigma')$ according to (4.2.15).

Lemma 14. *For any node labellings $\sigma, \sigma' : [N] \to [K]$,*

$$|E(\sigma) \setminus E(\sigma')| \;\geq\; \max\left\{ N^\sigma_{\min} - \mathrm{d}^*_{\mathrm{Ham}}(\sigma, \sigma'),\; \frac{1}{3}N^\sigma_{\min} - \frac{1}{6}N^{\sigma'}_{\max} \right\} \mathrm{d}^*_{\mathrm{Ham}}(\sigma, \sigma').$$

Proof. Let us note that all quantities appearing in the statement of the lemma remain invariant if we replace σ' by $\rho \circ \sigma'$, where $\rho \in \mathrm{Sym}(K)$ is an arbitrary permutation. Therefore, we may without loss of generality assume that σ and σ' are optimally aligned according to (4.2.16).

For sets $C, D \subset [N]$, we denote by $E(C, D)$ the collection of unordered pairs which can be written as $e = \{i, j\}$ with $i \in C$ and $j \in D$, and we denote the set of node pairs internal to C by $E(C) = E(C, C)$. We observe that the set $\Gamma = E(\sigma) \setminus E(\sigma')$ can be partitioned into $\Gamma = \cup_k \Gamma_k$, where $\Gamma_k = E(C_k) \setminus E(\sigma')$. We may further split this set according to $\Gamma_k = \Gamma_{k1} \cup \Gamma_{k2}$, where

$$\begin{aligned} \Gamma_{k1} &= E(C_k \cap C'_k, C_k \setminus C'_k), \\ \Gamma_{k2} &= E(C_k \setminus C'_k) \setminus E(\sigma'). \end{aligned}$$

Therefore, it follows that $|\Gamma| = \sum_k (|\Gamma_{k1}| + |\Gamma_{k2}|)$.

To analyse the sizes of Γ_{k1} and Γ_{k2}, denote $N_{k\ell} = |C_k \cap C'_\ell|$ and $D_k = |C_k \setminus C'_k|$. Then we immediately see that

$$|\Gamma_{k1}| = N_{kk} D_k. \tag{4.2.19}$$

Furthermore, we see that $E(C_k \setminus C'_k) \cap E(\sigma') = \cup_{\ell \neq k} E(C_k \cap C'_\ell)$, and it follows that

$$|\Gamma_{k2}| = |E(C_k \setminus C'_k)| - \sum_{\ell \neq k} |E(C_k \cap C'_\ell)| = \binom{D_k}{2} - \sum_{\ell \neq k} \binom{N_{k\ell}}{2}. \tag{4.2.20}$$

By combining (4.2.19)–(4.2.20) we conclude that

$$|\Gamma| = \sum_k (|\Gamma_{k1}| + |\Gamma_{k2}|) = \sum_k \left\{ N_{kk} D_k + \binom{D_k}{2} - \sum_{\ell \neq k} \binom{N_{k\ell}}{2} \right\}.$$

Let us derive a lower bound for $|\Gamma|$. Denote $B_k = \max_{\ell \neq k} N_{k\ell}$. Then by noting that $\sum_{\ell \neq k} N_{k\ell} = D_k$, we see that

$$\sum_{\ell \neq k} \binom{N_{k\ell}}{2} = \frac{1}{2} \sum_{\ell \neq k} N_{k\ell}(N_{k\ell} - 1) \leq \frac{1}{2} D_k (B_k - 1),$$

and by applying (4.2.20), it follows that

$$|\Gamma_{k2}| \geq \frac{1}{2} D_k (D_k - 1) - \frac{1}{2} D_k (B_k - 1) = \frac{1}{2} D_k (D_k - B_k).$$

By applying (4.2.19) and noting that $N_{kk} = N_k - D_k$, it now follows that

$$|\Gamma_k| \geq D_k (N_k - D_k) + \frac{1}{2} D_k (D_k - B_k). \tag{4.2.21}$$

We shall apply (4.2.21) to derive two lower bounds for $|\Gamma|$. First, by Lemma 13, we find that $B_k \leq \frac{1}{3}(N_k + N^{\sigma'}_{\max})$, and hence

$$\begin{aligned} |\Gamma_k| &\geq (N_k - D_k) D_k + \frac{1}{2} \left(D_k - \frac{1}{3} N_k - \frac{1}{3} N^{\sigma'}_{\max} \right) D_k \\ &= \left(\frac{5}{6} N_k - \frac{1}{2} D_k - \frac{1}{6} N^{\sigma'}_{\max} \right) D_k. \end{aligned}$$

Because $D_k \leq N_k$, we conclude that

$$|\Gamma_k| \;\geq\; \left(\frac{1}{3}N_k - \frac{1}{6}N_{\max}^{\sigma'}\right)D_k \;\geq\; \left(\frac{1}{3}N_{\min}^{\sigma} - \frac{1}{6}N_{\max}^{\sigma'}\right)D_k$$

By summing the above inequality over k and noting that $\sum_k D_k = \mathrm{Ham}(\sigma, \sigma') = L$ for optimally aligned σ and σ', we conclude that

$$|\Gamma| \;\geq\; \left(\frac{1}{3}N_{\min}^{\sigma} - \frac{1}{6}N_{\max}^{\sigma'}\right)L. \tag{4.2.22}$$

Second, by noting that $B_k \leq D_k$, we see that (4.2.21) implies

$$|\Gamma_k| \;\geq\; D_k(N_k - D_k) \;\geq\; D_k(N_{\min}^{\sigma} - D_k).$$

By summing the above inequality over k, we find that

$$|\Gamma| \;\geq\; N_{\min}^{\sigma}\sum_k D_k - \sum_k D_k^2 \;\geq\; N_{\min}^{\sigma}\sum_k D_k - (\sum_k D_k)^2.$$

By recalling that $\sum_k D_k = L$, we conclude that

$$|\Gamma| \;\geq\; N_{\min}^{\sigma}L - L^2. \tag{4.2.23}$$

By combining (4.2.22)–(4.2.23), the claim follows. □

4.2.6 Additional lemmas

Lemma 15. *For any node labelling $\sigma : [N] \to [K]$, the number $Z_{\sigma,m}$ of node labellings $\sigma' : [N] \to [K]$ such that $\mathrm{Ham}(\sigma, \sigma') = m$ satisfies*

$$Z_{\sigma,m} \;=\; \binom{N}{m}(K-1)^m \;\leq\; \left(\frac{eN(K-1)}{m}\right)^m.$$

Proof. Any node labelling $\sigma' : [N] \to [K]$ which differs from a particular σ at exactly m input values can be constructed as follows. First choose a set of m input values out of N; there are $\binom{N}{m}$ ways to do this. Then for each i of the chosen m input values, select a new output value from the of $K-1$ values excluding $\sigma(i)$; there are $(K-1)^m$ ways to do this. Hence the equality follows.

To verify the inequality, we note that $\frac{m^m}{m!} \leq \sum_{s=0}^{\infty}\frac{m^s}{s!} = e^m$. Therefore, we see that $\binom{N}{m} \leq \frac{N^m}{m!} \leq (\frac{eN}{m})^m$, and the inequality follows. □

Lemma 16. *For any integer $M \geq 1$ and any number $0 \leq s < 1$,*

$$Ms^M \;\leq\; \sum_{m=M}^{\infty} ms^m \;\leq\; (1-s)^{-2}Ms^M.$$

Proof. Denote $S = \sum_{m=M}^{\infty} m s^m$. By differentiating $\sum_{m=M}^{\infty} s^m = (1-s)^{-1} s^M$, we find that

$$s^{-1} S \;=\; \sum_{m=M}^{\infty} m s^{m-1} \;=\; (1-s)^{-2} s^M + (1-s)^{-1} M s^{M-1},$$

from which we see that

$$S \;=\; s(1-s)^{-2}\Big(s^M + (1-s) M s^{M-1}\Big) \;=\; \frac{M s^M}{(1-s)^2}\Big(1 - s(1-1/M)\Big)$$

The upper bound now follows from $1 - s(1-1/M) \le 1$. The lower bound is immediate, corresponding to the first term of the nonnegative series. □

4.3 Consistency of Algorithm 2

4.3.1 Single node label estimation

For $r > 0$ we define the ratio

$$\beta_r(f,g) \;=\; \frac{\mathrm{D}^s_{1+r}(f,g)}{\mathrm{D}^s_r(f,g)}. \tag{4.3.1}$$

Given a reference node i and a node labelling[1] $\tilde{\sigma}_i$ on $[N] \setminus \{i\}$, define an estimator for the label of i by $\hat{\sigma}_i(i) = \operatorname*{arg\,max}_{k \in [K]} h_i(k)$, with arbitrary tie breaks, where

$$h_i(k) \;=\; \sum_{j\colon \tilde{\sigma}_i(j) = k} \log \frac{f(X_{ij})}{g(X_{ij})}. \tag{4.3.2}$$

This is a maximum likelihood estimator in the special case where $\tilde{\sigma}_i$ assigns a correct label to all $j \neq i$. When this is not the case, we need to account for errors caused by corrupted likelihoods due to misclassified nodes in $\tilde{\sigma}_i$. The error in such a setting is given by the following lemma.

Lemma 17. *Let $\sigma : [N] \to [K]$ and assume that X_{ij}, $j \neq i$, are mutually independent $\mathcal{S}$-valued random variables such that $\mathrm{Law}(X_{ij}) = f$ for $\sigma(i) = \sigma(j)$ and $\mathrm{Law}(X_{ij}) = g$ otherwise. The error probability when estimating the label of node i as a maximiser of (4.3.2) is bounded by*

$$\mathbb{P}\left(\tau \circ \hat{\sigma}_i(i) \neq \sigma(i)\right) \;\le\; K e^{-(N_{\min} - 1 - 2d_i^* - \frac{2r}{1-r} d_i^* \beta_r) 2(1-r)\, \mathrm{D}^s_r(f,g)} \qquad \forall r \in [0,1],$$

where $\beta_r = \beta_r(f,g)$ is defined by (4.3.1), $d_i^* = \mathrm{d}^*_{\mathrm{Ham}}(\tilde{\sigma}_i, \sigma_{-i})$ is the symmetrised Hamming distance from $\tilde{\sigma}_i$ to the restriction σ_{-i} of the true node labelling σ to $[N] \setminus \{i\}$, and τ is an arbitrary K-permutation such that $\mathrm{d}_{\mathrm{Ham}}(\tau \circ \tilde{\sigma}_i, \sigma_{-i}) = d_i^*$.

[1] In this section we assume that $\tilde{\sigma}_i$ is nonrandom.

Proof. Denote $k^* = \tau^{-1}(\sigma(i))$. Observe that $\tau \circ \hat{\sigma}_i(i) \neq \sigma(i)$ if and only if $\hat{\sigma}_i(i) \neq k^*$, and the latter is possible only if $L_k = h_i(k) - h_i(k^*) \geq 0$ for some $k \neq k^*$. After noting that $\mathbb{P}(L_k \geq 0) = \mathbb{P}\left(e^{sL_k} \geq 1\right) \leq \mathbb{E}e^{sL_k}$, it follows that

$$\mathbb{P}(\tau \circ \hat{\sigma}_i(i) \neq \sigma(i)) \;\leq\; \sum_{k \neq k^*} \mathbb{P}(L_k \geq 0) \;\leq\; \sum_{k \neq k^*} \mathbb{E}e^{sL_k}. \tag{4.3.3}$$

Denote by $C_k = \{j \neq i : \sigma(j) = k\}$ the peers of i with true label k, and by $\tilde{C}_k = \{j \neq i : \tilde{\sigma}_i(j) = k\}$ the set of peers labelled k by $\tilde{\sigma}_i$. Denote $Z_\alpha(f\|g) = \int f^\alpha g^{1-\alpha}$. By noting that for any $j \neq i$,

$$\mathbb{E}\left(\frac{f(X_{ij})}{g(X_{ij})}\right)^r \;=\; \begin{cases} Z_{1+r}(f\|g), & \sigma(j) = \sigma(i), \\ Z_r(f\|g), & \text{else}, \end{cases}$$

and

$$\mathbb{E}\left(\frac{f(X_{ij})}{g(X_{ij})}\right)^{-r} \;=\; \begin{cases} Z_r(g\|f), & \sigma(j) = \sigma(i), \\ Z_{1+r}(g\|f), & \text{else}, \end{cases}$$

we find that for all k, the log-likelihood ratio $h_i(k)$ defined in (4.3.2) satisfies

$$\mathbb{E}e^{rh_i(k)} \;=\; Z_{1+r}(f\|g)^{v_k^{\text{in}}} Z_r(f\|g)^{v_k^{\text{out}}} \quad \text{and} \quad \mathbb{E}e^{-rh_i(k)} \;=\; Z_{1+r}(g\|f)^{v_k^{\text{out}}} Z_r(g\|f)^{v_k^{\text{in}}},$$

where $v_k^{\text{in}} = |\tilde{C}_k \cap C_{\sigma(i)}|$ and $v_k^{\text{out}} = |\tilde{C}_k \setminus C_{\sigma(i)}|$. Because $h_i(k)$ and $h_i(\ell)$ are mutually independent for $k \neq \ell$, it follows that L_k for $k \neq k^*$ satisfies

$$\mathbb{E}e^{rL_k} \;=\; Z_{1+r}(f\|g)^{v_k^{\text{in}}} Z_{1+r}(g\|f)^{v_{k^*}^{\text{out}}} Z_r(f\|g)^{v_k^{\text{out}}} Z_r(g\|f)^{v_{k^*}^{\text{in}}}.$$

Because $Z_r = e^{-(1-r)\,\mathrm{D}_r}$ and $Z_{1+r} = e^{r\,\mathrm{D}_{1+r}}$, we may rephrase the above equality as $\mathbb{E}e^{rL_k} = e^t$, where

$$t \;=\; s_1 v_k^{\text{in}} + s_2 v_{k^*}^{\text{out}} - u_1 v_k^{\text{out}} - u_2 v_{k^*}^{\text{in}},$$

with $s_1 = rD_{1+r}(f\|g)$, $s_2 = rD_{1+r}(g\|f)$, $u_1 = (1-r)\,\mathrm{D}_r(f\|g)$, and $u_2 = (1-r)\,\mathrm{D}_r(g\|f)$. By noting that $v_k^{\text{in}} + v_k^{\text{out}} = |\tilde{C}_k|$, we see that

$$t \;=\; (u_1 + s_1)\, v_k^{\text{in}} + (u_2 + s_2)\, v_{k^*}^{\text{out}} - u_1 |\tilde{C}_k| - u_2 |\tilde{C}_{k^*}|.$$

One may verify that $\tau \circ \tilde{\sigma}_i(j) \neq \sigma(j)$ for all $j \in \tilde{C}_k \cap C_{\sigma(i)}$ and all $k \neq k^*$. Therefore, $v_k^{\text{in}} = |C_k \cap C_{\sigma(i)}| \leq \mathrm{d}_{\text{Ham}}(\tau \circ \tilde{\sigma}_i, \sigma_{-i})$. Similarly, $\tau \circ \tilde{\sigma}_i(j) \neq \sigma(j)$ for all $j \in \tilde{C}_{k^*} \setminus C_{\sigma(i)}$ implies that $v_{k^*}^{\text{out}} = |\tilde{C}_{k^*} \setminus C_{\sigma(i)}| \leq \mathrm{d}_{\text{Ham}}(\tau \circ \tilde{\sigma}_i, \sigma_{-i})$. Next, by noting that $\tau \circ \tilde{\sigma}_i(j) \neq \tau(k)$ and $\sigma(j) = \tau(k)$ for $j \in C_{\tau(k)} \setminus \tilde{C}_k$, it follows that $|C_{\tau(k)} \setminus \tilde{C}_k| \leq \mathrm{d}_{\text{Ham}}(\tau \circ \tilde{\sigma}_i, \sigma_{-i})$. Therefore,

$$|\tilde{C}_k| \;\geq\; |\tilde{C}_k \cap C_{\tau(k)}| \;=\; |C_{\tau(k)}| - |C_{\tau(k)} \setminus \tilde{C}_k| \;\geq\; N_{\min} - 1 - \mathrm{d}_{\text{Ham}}(\tau \circ \tilde{\sigma}_i, \sigma_{-i}),$$

and the above inequality also holds for $k = k^*$. By collecting the above inequalities and recalling that $\mathrm{d}_{\text{Ham}}(\tau \circ \tilde{\sigma}_i, \sigma_{-i}) = d_i^*$, we conclude that

$$\begin{aligned} t \;&\leq\; d_i^*\,(u_1 + u_2 + s_1 + s_2) - (N_{\min} - 1 - d_i^*)\,(u_1 + u_2) \\ &\leq\; -(u_1 + u_2)\left(N_{\min} - 1 - 2d_i^* - d_i^* \frac{s_1 + s_2}{u_1 + u_2}\right). \end{aligned}$$

The claim follows by observing that $u_1 + u_2 = 2(1-r)\,\mathrm{D}_r^s(f, g)$ and $s_1 + s_2 = 2r\,\mathrm{D}_{1+r}^s(f, g)$. □

4.3.2 Analysis of refinement and consensus procedures

Let us start with a Lemma bounding difference between the block sizes given by two node labelling σ_1, σ_2 as a function of the Hamming distance.

Lemma 18. *For any $\sigma_1, \sigma_2 : [N] \to [K]$, (i) $\big||\sigma_1^{-1}(k)| - |\sigma_2^{-1}(k)|\big| \leq \mathrm{d}_{\mathrm{Ham}}(\sigma_1, \sigma_2)$ for all k, and (ii) $|N_{\min}(\sigma_1) - N_{\min}(\sigma_2)| \leq \mathrm{d}^*_{\mathrm{Ham}}(\sigma_1, \sigma_2)$, where $N_{\min}(\sigma_1) = \min_k |\sigma_1^{-1}(k)|$ and $N_{\min}(\sigma_2) = \min_k |\sigma_2^{-1}(k)|$.*

Proof. (i) Because $|\sigma_1^{-1}(k) \setminus \sigma_2^{-1}(k)| \leq \mathrm{d}_{\mathrm{Ham}}(\sigma_1, \sigma_2)$, we find that

$$\begin{aligned} |\sigma_1^{-1}(k)| &= |\sigma_1^{-1}(k) \cap \sigma_2^{-1}(k)| + |\sigma_1^{-1}(k) \setminus \sigma_2^{-1}(k)| \\ &\leq |\sigma_2^{-1}(k)| + \mathrm{d}_{\mathrm{Ham}}(\sigma_1, \sigma_2). \end{aligned}$$

By symmetry, the same inequality is true also with σ_1, σ_2 swapped.

(ii) Let π be a K-permutation for which $\mathrm{d}_{\mathrm{Ham}}(\pi \circ \sigma_1, \sigma_2) = \mathrm{d}^*_{\mathrm{Ham}}(\sigma_1, \sigma_2)$. Then by (i),

$$|\sigma_2^{-1}(k)| \geq |(\pi \circ \sigma_1)^{-1}(k)| - \mathrm{d}_{\mathrm{Ham}}(\pi \circ \sigma_1, \sigma_2) \geq N_{\min}(\sigma_1) - \mathrm{d}^*_{\mathrm{Ham}}(\sigma_1, \sigma_2).$$

This implies that $N_{\min}(\sigma_2) \geq N_{\min}(\sigma_1) - \mathrm{d}^*_{\mathrm{Ham}}(\sigma_1, \sigma_2)$. The second claim hence follows by symmetry. □

The following result describes the behaviour of Steps 2 and 3 in Algorithm 2 on the event that Step 1 achieves moderate accuracy.

Lemma 19. *Assume that the outputs $\tilde{\sigma}_i$ of Step 1 in Algorithm 2 satisfy $\mathrm{d}^*_{\mathrm{Ham}}(\tilde{\sigma}_i, \sigma_{-i}) < \frac{1}{5} N_{\min} - 1$ for all i. Then there exist unique K-permutations $\pi_1, \ldots, \pi_N$ such that for all i:*

(i) The outputs $\tilde{\sigma}_i$ of Step 1 satisfy $\mathrm{d}^*_{\mathrm{Ham}}(\tilde{\sigma}_i, \sigma_{-i}) = \mathrm{d}_{\mathrm{Ham}}(\pi_i \circ \tilde{\sigma}_i, \sigma_{-i})$.

(ii) The outputs $\hat{\sigma}_i$ of Step 2 satisfy $\mathrm{d}^*_{\mathrm{Ham}}(\hat{\sigma}_i, \sigma) = \mathrm{d}_{\mathrm{Ham}}(\pi_i \circ \hat{\sigma}_i, \sigma)$.

(iii) The final output $\hat{\sigma}$ from Step 3 satisfies $\hat{\sigma}(i) = (\pi_1^{-1} \circ \pi_i)(\hat{\sigma}_i(i))$.

Proof. (i) Denote $\epsilon = \max_i \mathrm{d}^*_{\mathrm{Ham}}(\tilde{\sigma}_i, \sigma_{-i})$. Because the smallest block size of σ_{-i} is bounded by $\frac{1}{2} N_{\min}(\sigma_{-i}) \geq \frac{1}{2}(N_{\min} - 1) > \epsilon$, it follows by Lemma 1 that for every i there exists a unique K-permutation π_i such that $\mathrm{d}_{\mathrm{Ham}}(\pi_i \circ \tilde{\sigma}_i, \sigma_{-i}) = \mathrm{d}^*_{\mathrm{Ham}}(\tilde{\sigma}_i, \sigma_{-i})$.

(ii) Observe that

$$\mathrm{d}_{\mathrm{Ham}}(\pi_i \circ \hat{\sigma}_i, \sigma) = \mathrm{d}_{\mathrm{Ham}}(\pi_i \circ \tilde{\sigma}_i, \sigma_{-i}) + 1(\pi_i(\hat{\sigma}_i(i)) = \sigma(i)) \leq \epsilon + 1. \tag{4.3.4}$$

Because $\epsilon + 1 < \frac{1}{2} N_{\min}$, Lemma 1 implies that $\mathrm{d}^*_{\mathrm{Ham}}(\hat{\sigma}_i, \sigma) = \mathrm{d}_{\mathrm{Ham}}(\pi_i \circ \hat{\sigma}_i, \sigma)$.

(iii) By Lemma 18 and (4.3.4), the minimum block size of $\hat{\sigma}_1$ is bounded by $N_{\min}(\hat{\sigma}_1) \geq N_{\min} - \mathrm{d}^*_{\mathrm{Ham}}(\hat{\sigma}_1, \sigma) \geq N_{\min} - (\epsilon + 1)$. Inequality (4.3.4) also implies that

$$\mathrm{d}_{\mathrm{Ham}}(\pi_i \circ \hat{\sigma}_i, \pi_1 \circ \hat{\sigma}_1) \;\leq\; \mathrm{d}_{\mathrm{Ham}}(\pi_i \circ \hat{\sigma}_i, \sigma) + \mathrm{d}_{\mathrm{Ham}}(\sigma, \pi_1 \circ \hat{\sigma}_1) \;\leq\; 2(\epsilon + 1).$$

Therefore, $\mathrm{d}_{\mathrm{Ham}}(\pi_1^{-1} \circ \pi_i \circ \hat{\sigma}_i, \hat{\sigma}_1) \leq 2(\epsilon + 1)$ as well. Furthermore, because $2(\epsilon + 1) < \frac{1}{2}(N_{\min} - (\epsilon+1)) \leq \frac{1}{2} N_{\min}(\hat{\sigma}_1)$, we conclude by Lemma 1 that $\pi_1^{-1} \circ \pi_i$ is the unique minimiser of $\pi \mapsto \mathrm{d}_{\mathrm{Ham}}(\pi \circ \hat{\sigma}_i, \hat{\sigma}_1)$, and

$$\pi_1^{-1} \circ \pi_i(k) \;=\; \arg\max_{\ell} |\hat{\sigma}_i^{-1}(k) \cap \hat{\sigma}_1^{-1}(\ell)| \quad \text{for all } k.$$

Hence, the output value $\hat{\sigma}(i)$ satisfies $\hat{\sigma}(i) = (\pi_1^{-1} \circ \pi_i)(\hat{\sigma}_i(i))$. □

4.3.3 Proof of Theorem 5

In Algorithm 2, the outputs of Step 1 are denoted by $\tilde{\sigma}_1, \ldots, \tilde{\sigma}_N$, the outputs of Step 2 by $\hat{\sigma}_1, \ldots, \hat{\sigma}_N$, and the final output from Step 3 by $\hat{\sigma}$. Recall that σ denotes the unknown true node labelling, and σ_{-i} its restriction to $[N] \setminus \{i\}$. As a standard graph clustering algorithm for Step 1, we will employ a spectral clustering algorithm described in [XJL20, Algorithm 4] with tuning parameter $\mu = 4K(\alpha_{\min} N)^{-1} N = 4K\alpha_{\min}^{-1}$ and trim threshold $\tau = 40K\bar{d}$, where $\bar{d}$ is the average degree of $\tilde{A}_{-i}$, which is a modified version of [Gao+17, Algorithm 2] with explicitly known error bounds.

Denote by B the event that $N_{\min} \geq N\alpha_{\min}\left(1 - \sqrt{\frac{8 \log N}{\alpha_{\min} N}}\right)$. Lemma 10 shows that

$$\mathbb{P}\left(B^c\right) \;\leq\; KN^{-4}.$$

Since $N \gg 1$ we have $N \geq 2000$ when the scale-parameter η is large enough. Moreover, $\alpha_{\min} N \geq 6N^{1/2}$. Thus, we have $\sqrt{\frac{8 \log N}{\alpha_{\min} N}} \leq \frac{1}{2}$, and hence event B implies $N_{\min} \geq \frac{1}{2} N\alpha_{\min}$.

Let $J_0 = \frac{(p-q)^2}{p \vee q}$ and denote by E_i the event that Step 1 for node i succeeds with accuracy $\mathrm{d}^*_{\mathrm{Ham}}(\tilde{\sigma}_i, \sigma_{-i}) \leq \epsilon N$, where $\epsilon = 2^{30} K^3 \left(\alpha_{\min} N\right)^{-1} J_0^{-1}$. The matrix $\tilde{A}_{-i}$ computed in Step 1 of Algorithm 2 is the adjacency matrix of a standard binary SBM with intra-block link probability $p = 1 - f(0)$, inter-block link probability $q = 1 - g(0)$, and node labelling σ_{-i}, and by [XJL20, Proposition B.3] it follows that

$$\mathbb{P}\left(E_i^c \cap B\right) \;\leq\; N^{-5}. \tag{4.3.5}$$

Since $\mathrm{D}_{1+r}(f\|g) \geq \mathrm{D}_r(f\|g)$, the assumption $J_0 \gg N^{-1} \frac{\mathrm{D}^s_{1+r}(f\|g)}{\mathrm{D}^s_r(f\|g)}$ implies, for η large enough, that $\epsilon N \leq 2^{-4} K^{-1} N_{\min} \leq \frac{1}{32} N_{\min}$, and the event B implies $N_{\min} \geq 135$. Therefore, $\epsilon N \leq \frac{1}{32} N_{\min} < \frac{1}{5} N_{\min} - 1$, and we see by applying Lemma 19 that on the event $E \cap B$ where $E = \cap_i E_i$ there exist unique K-permutations $\tau_1, \ldots, \tau_N$ such that $\hat{\sigma}(i) = (\tau_1^{-1} \circ \tau_i)(\hat{\sigma}_i(i))$ for all i. Especially,

$$\mathrm{d}^*_{\mathrm{Ham}}(\hat{\sigma}, \sigma) \;\leq\; \mathrm{d}_{\mathrm{Ham}}(\tau_1 \circ \hat{\sigma}, \sigma) \;=\; \sum_i 1(\tau_i(\hat{\sigma}_i(i)) \neq \sigma(i)) \qquad \text{on } E \cap B,$$

so it follows that $\mathbb{E}\,\mathrm{d}^*_{\mathrm{Ham}}(\hat{\sigma},\sigma)1_{E\cap B} \le \sum_i \mathbb{P}(\tau_i(\hat{\sigma}_i(i)) \neq \sigma(i), E_i, B)$. In light of (4.3.5), by using $d^*_H = d^*_H 1_E 1_B + d^*_H 1_{E^c} 1_B + d^*_H 1_{B^C}$ and applying the bounds $\mathrm{d}^*_{\mathrm{Ham}}(\hat{\sigma},\sigma) \le N$ and $\mathbb{P}(E^c \cap B) \le \sum_i \mathbb{P}(E_i^c \cap B) \le N^{-4}$, we conclude that

$$\mathbb{E}\,\mathrm{d}^*_{\mathrm{Ham}}(\hat{\sigma},\sigma) \;\le\; \sum_i \mathbb{P}\left(\tau_i(\hat{\sigma}_i(i)) \neq \sigma(i), E_i, B\right) + (K+1)N^{-3}. \tag{4.3.6}$$

Let us analyse the sum on the right side of (4.3.6). Note that $\tilde{\sigma}_i$ and the K-permutation τ_i are fully determined by the entries of the sub-array $A_{-i} = (A^t_{j,j'} : j,j' \in [N]\setminus\{i\})$. Conditionally on A_{-i}, we may hence treat $\tilde{\sigma}_i$ and τ_i as nonrandom, and apply Lemma 17 to conclude that on the event $E_i \cap B$,

$$\mathbb{P}\left(\pi_i \circ \hat{\sigma}_i(i) \neq \sigma(i) \,|\, A_{-i}\right) \;\le\; K e^{-(N\alpha_{\min}\left(1-\sqrt{\frac{8\log N}{\alpha_{\min}N}}\right)-1-2d^*_i-\frac{2r}{1-r}d^*_i\beta_r)2(1-r)\,\mathrm{D}^s_r(f,g)},$$

Because $\mathrm{d}^*_{\mathrm{Ham}}(\tilde{\sigma}_i,\sigma_{-i}) \le \epsilon N$ on E_i, and the event $E_i \cap B$ is measurable with respect to the sigma-algebra generated by A_{-i}, we conclude that

$$\mathbb{P}\left(\tau_i \circ \hat{\sigma}_i(i) \neq \sigma(i), E_i, B\right) \;\le\; K e^{-\left(N\alpha_{\min}\left(1-\sqrt{\frac{8\log N}{\alpha_{\min}N}}\right)-1-3\frac{r}{1-r}\beta_r\epsilon N\right)2(1-r)\,\mathrm{D}^s_r(f,g)}.$$

By combining this with (4.3.6), we conclude that

$$\mathbb{E}\,\mathrm{d}^*_{\mathrm{Ham}}(\hat{\sigma}) \;\le\; KN e^{-\left(N\alpha_{\min}\left(1-\sqrt{\frac{8\log N}{\alpha_{\min}N}}\right)-1-2\left(1+\frac{r}{1-r}\beta_r\right)\epsilon N\right)2(1-r)\,\mathrm{D}^s_r(f,g)} + (K+1)N^{-3}.$$

We finish the proof of Theorem 5 by using the inequality $(1-r)\,\mathrm{D}_r \ge r\,\mathrm{D}_{1/2}$, valid for any $r \in \left(0,\frac{1}{2}\right]$ [VH14, Theorem 16]. □

4.4 Rényi divergences of sparse binary Markov chains

This section discusses binary Markov chains with initial distributions μ,ν and transition probability matrices P,Q. In this case the Rényi divergence of order $\alpha \in (0,\infty)\setminus\{1\}$ for the associated path probability distributions f,g on $\{0,1\}^T$ equals

$$D_\alpha(f\|g) \;=\; \frac{1}{\alpha-1}\log\left(\sum_{x\in\{0,1\}^T} \mu_{x_1}^\alpha \nu_{x_1}^{1-\alpha} \prod_{t=2}^T P^\alpha_{x_{t-1}x_t} Q^{1-\alpha}_{x_{t-1}x_t}\right). \tag{4.4.1}$$

Such divergences will be analysed using weighted geometric and arithmetic averages of transition parameters defined by

$$\begin{aligned} r_a &= \mu_a^\alpha \nu_a^{1-\alpha}, & R_{ab} &= P^\alpha_{ab} Q^{1-\alpha}_{ab}, \\ \hat{r}_a &= \alpha\mu_a + (1-\alpha)\nu_a, & \hat{R}_{ab} &= \alpha P_{ab} + (1-\alpha)Q_{ab}. \end{aligned} \tag{4.4.2}$$

We note that $D_\alpha(f\|g) = \frac{1}{\alpha-1}\log Z$, where $Z = \sum_{x\in\{0,1\}^T} r_{x_1}\prod_{t=2}^T R_{x_{t-1}x_t}$. Moreover, $r_1 = 1-\hat{r}_1+O(\rho^2)$ and $R_{01} = 1-\hat{R}_{01}+O(\rho^2)$ when $\mu_1,\nu_1,P_{01},Q_{01} \lesssim \rho$.

4.4.1 Rényi divergences of order $\alpha \in (0,1)$

The following proposition provides an expression for the Rényi divergence between sparse Markov chains. In particular, Proposition 8 follows by substituting $\alpha = \frac{1}{2}$ in the following result.

Proposition 13. *Consider binary Markov chains with initial distributions μ, ν and transition probability matrices P, Q. Assume that $\mu_1, \nu_1, P_{01}, Q_{01} \le \rho$ for some ρ such that $\rho T \le 0.01$. Then the Rényi divergence of order $\alpha \in (0,1)$ between the associated path probability distributions defined by (4.4.1) is approximated by*

$$\mathrm{D}_\alpha(f\|g) \;=\; \frac{1}{1-\alpha}\Bigg(\hat{r}_1 - r_1 + \sum_{t=2}^{T} J_t + \epsilon\Bigg), \tag{4.4.3}$$

where the error term satisfies $|\epsilon| \le 46(\rho T)^2$,

$$J_t \;=\; \begin{cases} \hat{R}_{01} - R_{01} + \Big(1 - \frac{R_{10}}{1-R_{11}}\Big)\Big(R_{01} + \Big(r_1(1-R_{11}) - R_{01}\Big)R_{11}^{t-2}\Big), & R_{11} < 1, \\ \hat{R}_{01} - R_{01}, & R_{11} = 1, \end{cases}$$

and the parameters $r_a, \hat{r}_a, R_{ab}, \hat{R}_{ab}$ are given by (4.4.2).

The rest of this section is devoted to proving Proposition 13.

Basic results on binary sequences

For a path $x = (x_1, \dots, x_T)$ in $\{0,1\}^T$, by $x_{ij} = \sum_{t=2}^{T} 1(x_{t-1} = i, x_t = j)$ the number of ij-transitions. Then the path probability of a binary Markov chain with initial distribution μ and transition matrix P can be written as $f(x) = \mu_{x_1} \prod_{ij} P_{ij}^{x_{ij}}$. For sparse Markov chains, we will analyse path probabilities by focusing on the total number of 1's $\|x\| = \sum_t x_t$, and the number of on-periods $x_{\mathrm{on}} = x_1 + x_{01} = x_{10} + x_T$, in which the count in middle is the number of on-period start times, and the count on the right is the number of on-period end times. We also note that $x_{01} + x_{11} = \sum_{t=2}^{T} x_t$ implies that $\|x\| = x_{\mathrm{on}} + x_{11}$. The data $(x_{\mathrm{on}}, \|x\|, x_1, x_T)$ suffices to determine the path probability of x because the transition counts can be recovered using the formulas $x_{01} = x_{\mathrm{on}} - x_1$, $x_{10} = x_{\mathrm{on}} - x_T$, $x_{11} = \|x\| - x_{\mathrm{on}}$, together with $x_{00} + x_{01} + x_{10} + x_{11} = T - 1$. Especially, the probability of a path with $(x_{\mathrm{on}}, \|x\|, x_1, x_T) = (j, t, a, b)$ equals

$$f(x) \;=\; \mu_0^{1-a}\mu_1^a P_{00}^{T-1-(t+j-a-b)} P_{01}^{j-a} P_{10}^{j-b} P_{11}^{t-j}. \tag{4.4.4}$$

The number of such paths is summarised in the next result.

Lemma 20. *Denote by $c_{jt}(ab)$ the number of paths $x \in \{0,1\}^T$ such that $x_{\mathrm{on}} = j$, $\|x\| = t$, $x_1 = a$, and $x_T = b$. Then the nonzero values of $c_{jt}(ab)$ are given by $c_{00}(00) = 1$,*

$$c_{1t}(ab) \;=\; \begin{cases} T - t - 1, & (a,b) = (0,0),\ 1 \le t \le T-2, \\ 1, & (a,b) = (0,1), (1,0),\ 1 \le t \le T-1, \\ 1, & (a,b) = (1,1),\ t = T, \end{cases}$$

and $c_{jt}(ab) = \binom{t-1}{j-1}\binom{T-t-1}{j-a-b}$ *for* $2 \le j \le \lceil T/2 \rceil$, *and* $j \le t \le T-1-j+a+b$.

Proof. We compute the cardinalities separately for the three cases in which the number of on-periods equals $j = 0$, $j = 1$, and $j \ge 2$.

(i) Case $j = 0$. The only path with no on-periods is the path of all zeros. Therefore, $c_{0t}(ab) = 1$ for $t = 0$ and $(a, b) = (0, 0)$, and $c_{0t}(a, b) = 0$ otherwise.

(ii) Case $j = 1$. In this case $c_{1t}(00) = T - t - 1$ for $1 \le t \le T - 2$ and zero otherwise. Furthermore, $c_{1t}(01) = c_{1t}(10) = 1$ for $1 \le t \le T - 1$, and both are zero otherwise. Finally, $c_{1t}(11) = 1$ for $t = T$ and zero otherwise.

(iii) Case $j \ge 2$. Now we proceed as follows. First, given a series of t ones, we choose $j-1$ places to break the series: there are $\binom{t-1}{j-1}$ ways of doing so. Then, we need to fill those breaks with zeros chosen among the $T-t$ zeros of the chain. Note that when $a = b = 0$, we also need to put zeros before and after the chain of ones. There are $j-1+(1-a)+(1-b) = j+1-a-b$ places to fill with $T - t$ zeros, and we need to put at least one zero in each place: there are $\binom{T-t-1}{j-a-b}$ ways of doing so.[2] Therefore, we conclude that

$$c_{jt}(ab) \;=\; \binom{t-1}{j-1}\binom{T-t-1}{j-a-b}.$$

□

Useful Taylor expansions

Lemma 21. *Assume that* $\alpha \in (0, 1)$ *and* $\max\{\mu_1, \nu_1, P_{01}, Q_{01}\} \le \rho$ *for some* $\rho \le \frac{1}{3}$. *Then the geometric and arithmetic means defined by* (4.4.2) *are related according to*

$$\begin{aligned} r_0 &= 1 - \hat{r}_1 + \epsilon_1, \\ R_{00} &= 1 - \hat{R}_{01} + \epsilon_2, \\ R_{00}^{T-1} &= 1 - (T-1)\hat{R}_{01} + \epsilon_3, \\ r_0 R_{00}^{T-1} &= 1 - \hat{r}_1 - (T-1)\hat{R}_{01} + \epsilon_4, \end{aligned}$$

where the error terms are bounded by $|\epsilon_1|, |\epsilon_2| \le (1+\rho)\rho^2$, $|\epsilon_3| \le 2(1+\rho)(\rho T)^2$, *and* $|\epsilon_4| \le 4(1+2\rho)(\rho T)^2$.

Proof. Note that $\hat{r}_1 \le \rho$ and $\hat{R}_{01} \le \rho$. Taylor's approximation (Lemma 25) implies that $(1-\mu_1)^\alpha = 1-\alpha\mu_1+\epsilon_{11}$ and $(1-\nu_1)^{1-\alpha} = 1-(1-\alpha)\nu_1+\epsilon_{12}$ for $|\epsilon_{11}|, |\epsilon_{12}| \le \frac{1}{2}\rho^2$. By multiplying these, we find that

$$r_0 \;=\; (1-\mu_1)^\alpha (1-\nu_1)^{1-\alpha} \;=\; 1 - \hat{r}_1 + \epsilon_1,$$

[2] A combinatorial fact, often referred as the *stars and bars* method, is that the number of ways in which n identical balls can be divided into m distinct bins is $\binom{n+m-1}{m-1}$, and $\binom{n-1}{m-1}$ if bins cannot be empty.

where the error term is bounded by $|\epsilon_1| \le (1+\frac{1}{4}\rho^2)\rho^2 \le (1+\rho)\rho^2$. Because $R_{00} = (1-P_{01})^\alpha(1-Q_{01})^{1-\alpha}$, repeating the same argument yields $|\epsilon_2| \le (1+\rho)\rho^2$.

Assume next that $T \ge 2$ (otherwise the third claim is trivial). Note that $0 \le 1 - R_{00} = \hat{R}_{01} - \epsilon_2 \le \rho + (1+\rho)\rho^2 \le \frac{1}{2}$ due to $\hat{R}_{01} \le \rho$ and $\rho \le \frac{1}{3}$. By applying Lemma 25, we then see that

$$R_{00}^{T-1} \;=\; (1-\hat{R}_{01}+\epsilon_2)^{T-1} \;=\; 1-(T-1)(\hat{R}_{01}-\epsilon_2)+\epsilon_{31},$$

where $|\epsilon_{31}| \le T^2(\hat{R}_{01}-\epsilon_2)^2 \le 2T^2(\hat{R}_{01}^2+\epsilon_2^2)$. It follows that $R_{00}^{T-1} = 1-(T-1)\hat{R}_{01}+\epsilon_3$ with $\epsilon_3 = (T-1)\epsilon_2+\epsilon_{31}$ bounded by $|\epsilon_3| \le T|\epsilon_2| + |\epsilon_{31}| \le T|\epsilon_2| + 2(T|\epsilon_2|)^2 + 2(\rho T)^2$, so that $|\epsilon_3| \le 2(1+\rho)\rho^2T^2$.

Finally, by multiplying the approximation formulas of r_0 and R_{00}^{T-1}, we find that

$$\epsilon_4 \;=\; \epsilon_1(1-(T-1)\hat{R}_{01}) + \epsilon_3(1-\hat{r}_1) + \epsilon_1\epsilon_3 + (T-1)\hat{r}_1\hat{R}_{01}.$$

By the triangle inequality, we find that for $T \ge 2$, $|\epsilon_4| \le (1+\rho T)|\epsilon_1| + |\epsilon_3| + |\epsilon_1\epsilon_3| + \rho^2 T$, from which one may check that $|\epsilon_4| \le 4(1+2\rho)(\rho T)^2$. □

Analysing paths with two or more on-periods

Lemma 22. *For any $\alpha \in (0,1)$ and any Markov chain path distributions f, g with transition matrices P, Q satisfying $P_{11}Q_{11} < 1$,*

$$\sum_{x:x_{\mathrm{on}}\ge 2} f_x^\alpha g_x^{1-\alpha} \;\le\; R_{01}(r_1+R_{01})T^2W(1+W)e^{WR_{01}T},$$

where $W = \frac{R_{10}}{1-R_{11}}$ and the weighted geometric means r_a, R_{ab} are defined by (4.4.2).

Proof. Fix an integer $2 \le j \le \lceil T/2 \rceil$, and denote $Z_j = \sum_{x:x_{\mathrm{on}}=j} f_x^\alpha g_x^{1-\alpha}$. By (4.4.4), we see that for any path x with j on-periods, t ones, initial state a, and final state b,

$$f_x^\alpha g_x^{1-\alpha} \;=\; r_0^{1-a}r_1^a R_{00}^{T-1-(t+j-a-b)} R_{01}^{j-a}R_{10}^{j-b}R_{11}^{t-j} \;\le\; r_1^a R_{01}^{j-a}R_{10}^{j-b}R_{11}^{t-j}.$$

By Lemma 20, the number of such paths equals

$$c_{jt}(ab) \;=\; \binom{t-1}{j-1}\binom{T-t-1}{j-a-b}.$$

To obtain an upper bound for the path count, we note that $\binom{T-t-1}{j-a-b} \le \frac{T^{j-a-b}}{(j-a-b)!}$. Furthermore, we also see that $\binom{t-1}{j-1} = \frac{t-1}{j-1}\binom{t-2}{j-2} \le T\binom{t-2}{j-2}$. The latter bound implies that $\binom{t-1}{j-1} \le T^b\binom{t-b-1}{j-b-1}$ for all $b \in \{0,1\}$. As a consequence, we conclude that

$$c_{jt}(ab) \;\le\; \frac{T^j}{(j-2)!}\binom{t-b-1}{j-b-1}$$

holds for all $a, b \in \{0,1\}$. Hence,

$$Z_j \;\le\; \frac{T^j}{(j-2)!}\sum_{a,b=0}^{1}\sum_{t\ge j}\binom{t-b-1}{j-b-1} r_1^a R_{01}^{j-a} R_{10}^{j-b} R_{11}^{t-j}.$$

Using a geometric moment formula (Lemma 23), we find that

$$\sum_{t=j}^{\infty}\binom{t-b-1}{j-b-1} R_{11}^{t-j} \;=\; (1-R_{11})^{-(j-b)} \;=\; R_{10}^{b-j} W^{j-b},$$

and it follows that

$$Z_j \;\le\; \frac{T^j}{(j-2)!}\sum_{a,b=0}^{1} r_1^a R_{01}^{j-a} W^{j-b} \;=\; T^2\frac{(R_{01}WT)^{j-2}}{(j-2)!}\sum_{a,b=0}^{1} r_1^a R_{01}^{2-a} W^{2-b}.$$

By noting that $\sum_{a,b=0}^{1} r_1^a R_{01}^{2-a} W^{2-b} \;=\; R_{01}(r_1+R_{01})W(1+W)$ and summing the above inequality with respect to $j \ge 2$, the claim follows. □

Proof of Theorem 8

By definition, the Rényi divergence of order α can be written as $\mathrm{D}_\alpha(f\|g) \;=\; \frac{1}{\alpha-1} Z_\alpha(f\|g)$, where $Z_\alpha(f\|g) = \sum_x f_x^\alpha g_x^{1-\alpha}$. We will split this Hellinger sum into

$$Z_\alpha(f\|g) \;=\; Z_0 + Z_1 + \sum_{j=2}^{\lceil T/2\rceil} Z_j$$

where $Z_j = \sum_{x:x_{\mathrm{on}}=j} f_x^\alpha g_x^{1-\alpha}$ indicates a Hellinger sum over paths with j on-periods. We will approximate the first two terms on the right by $Z_0 = \hat{Z}_0 + \epsilon_0$, $Z_1 = \hat{Z}_1 + \epsilon_1$, where

$$\hat{Z}_0 \;=\; 1 - \hat{r}_1 - (T-1)\hat{R}_{01}$$

and

$$\hat{Z}_1 \;=\; R_{01}R_{10}\sum_{t=1}^{T-2}(T-t-1)R_{11}^{t-1} + (R_{01}+r_1R_{10})\sum_{t=1}^{T-1} R_{11}^{t-1} + r_1 R_{11}^{T-1}.$$

Then is follows that

$$Z_\alpha(f\|g) \;=\; \hat{Z}_0 + \hat{Z}_1 + \epsilon_0 + \epsilon_1 + \epsilon_2,$$

where $\epsilon_2 \;=\; \sum_{j=2}^{\lceil T/2\rceil} Z_j$. We see that $\epsilon_2 \;=\; 0$ for $R_{11} \;=\; 1$, whereas for $R_{11} \;<\; 1$, Lemma 22 shows that $0 \le \epsilon_2 \le 4(\rho T)^2 e^{\rho T} \le 5(\rho T)^2$.

Let analyse the error term ϵ_0. Because the only path with $x_{\mathrm{on}} = 0$ is the identically zero path, we find that $Z_0 = r_0 R_{00}^{T-1}$. By Lemma 21 we have $|\epsilon_0| \le 4(1+2\rho)(\rho T)^2 \le 5(\rho T)^2$.

For the error term ϵ_1, with the help of formula (4.4.4) and Lemma 20, we see that

$$Z_1 \;=\; r_0R_{01}R_{10}\sum_{t=1}^{T-2}(T-t-1)R_{11}^{t-1}R_{00}^{T-2-t} + (r_0R_{01}+r_1R_{10})\sum_{t=1}^{T-1} R_{11}^{t-1}R_{00}^{T-1-t} + r_1R_{11}^{T-1}.$$

Because $r_0, R_{00} \leq 1$, it follows that $Z_1 \leq \hat{Z}_1$, and hence $\epsilon_1 \leq 0$. Furthermore, Lemma 21 implies that $r_0, R_{00} \geq 1 - 2\rho$. By noting that $R_{00}^{T-t} \geq R_{00}^{T-1}$ for $t \geq 1$, it follows that

$$Z_1 \;\geq\; (1-2\rho)^T \hat{Z}_1 \;\geq\; (1-2\rho T)\hat{Z}_1.$$

For $R_{11} = 1$ we have $R_{10} = 0$ and $\hat{Z}_1 = r_1 + (T-1)R_{01}$. For $R_{11} < 1$, we observe that

$$\begin{aligned}\hat{Z}_1 \;&\leq\; (T-1)R_{01}\frac{R_{10}}{1-R_{11}} + (T-1)R_{01} + r_1\frac{R_{10}}{1-R_{11}} + r_1 \\ &\leq\; (1+W)\Big(r_1 + (T-1)R_{01}\Big),\end{aligned}$$

where $W = \frac{R_{10}}{1-R_{11}}$. We note that $W = Z_\alpha(\mathrm{Geo}(P_{11}) \| \,\mathrm{Geo}(Q_{11}))$ equals the Hellinger sum of two geometric distributions, and therefore, $W \in (0, 1]$. Therefore, $\hat{Z}_1 \leq 2\rho T$, and it follows that

$$Z_1 \;\geq\; \hat{Z}_1 - 2\rho T \hat{Z}_1 \;\geq\; \hat{Z}_1 - 4(\rho T)^2.$$

Hence $|\epsilon_1| \leq 4(\rho T)^2$ for both $R_{11} < 1$ and $R_{11} = 1$.

We may now conclude that

$$Z_\alpha(f\|g) \;=\; \hat{Z}_0 + \hat{Z}_1 + \epsilon',$$

where

$$|\epsilon'| \;\leq\; |\epsilon_0| + |\epsilon_1| + |\epsilon_2| \;\leq\; 14(\rho T)^2.$$

(iv) Next, Taylor's approximation (Lemma 24) shows that $\log(1-t) = -t - \epsilon''$ where $0 \leq \epsilon'' \leq 2t^2$ for $0 \leq t \leq \frac{1}{2}$. By applying this with $t = 1 - Z_\alpha(f\|g)$, and noting that $|J_t| \leq \rho$ implies $0 \leq t \leq 3\rho T + |\epsilon'| \leq 4\rho T$, we find that

$$D_\alpha(f\|g) \;=\; \frac{1}{1-\alpha}\left(1 - Z_\alpha(f\|g) + \epsilon''\right) \;=\; \frac{1}{1-\alpha}\left(1 - \hat{Z}_0 - \hat{Z}_1 - \epsilon' + \epsilon''\right).$$

The claim about the error bound follows after noting that

$$|\epsilon'| + |\epsilon''| \;\leq\; 14(\rho T)^2 + 2(4\rho T)^2 \;\leq\; 46(\rho T)^2.$$

(v) Finally, let us simplify $\hat{Z}_0 + \hat{Z}_1$. When $R_{11} < 1$, by applying formulas $\sum_{t=1}^{T-2}(T-t-1)R_{11}^{t-1} = (1-R_{11})^{-1}\left((T-1) - \sum_{t=2}^{T} R_{11}^{t-2}\right)$ and $R_{11}^{T-1} = 1 - (1-R_{11})\sum_{t=2}^{T} R_{11}^{t-2}$ we find that

$$\hat{Z}_1 \;=\; r_1 + WR_{01}(T-1) - (1-W)\left(r_1(1-R_{11}) - R_{01}\right)\sum_{t=2}^{T} R_{11}^{t-2}.$$

Hence,

$$\hat{Z}_0 + \hat{Z}_1 \;=\; 1 - \left(\hat{r}_1 - r_1 + \sum_{t=2}^{T} J_t\right), \tag{4.4.5}$$

where the expression of J_t coincides with the one in the statement of the theorem. When $R_{11} = 1$, we find that $J_t = \hat{R}_{01} - R_{01}$. □

4.4.2 High-order Rényi divergences

The following result provides an upper bound on the Rényi divergence of order $\alpha > 1$ between path probability distributions of binary Markov chains defined by (4.4.1).

Proposition 14. *Assume that* $\frac{\mu_1}{\nu_1}, \frac{P_{01}}{Q_{01}}, \frac{P_{10}}{Q_{10}} \le M$ *for some* $M \ge 1$, $Q_{11} > 0$, *and* $\nu_1, Q_{01} \le \rho$ *for some* $\rho \le \frac{1}{2}$. *Then the Rényi divergence of order* $1 < \alpha < \infty$ *is bounded by*

$$\begin{aligned} D_\alpha(f\|g) \;\le\; & \frac{2\alpha}{\alpha-1}\rho T + \frac{M^{2\alpha}}{\alpha-1}\rho T \sum_{t=0}^{T-1} \Lambda^t \\ & + \frac{4}{\alpha-1} \sum_{j=2}^{\lceil T/2 \rceil} \frac{(M^{2\alpha}\rho T)^j}{(j-2)!} \sum_{t=j}^{T} \binom{t-1}{j-1} \Lambda^{t-j}, \end{aligned} \tag{4.4.6}$$

where $\Lambda = P_{11}^\alpha Q_{11}^{1-\alpha}$. *Furthermore, when* $\Lambda < 1$,

$$D_\alpha(f\|g) \;\le\; \frac{2\alpha+1}{\alpha-1} C\rho T e^{5C\rho T} \qquad \textit{with } C = \frac{M^{2\alpha}}{1-\Lambda}. \tag{4.4.7}$$

Proof. Recall that $D_\alpha(f\|g) = \frac{1}{\alpha-1}\log Z$ where $Z = \sum_x g_x (f_x/g_x)^\alpha$. Because $\nu_1 \le \rho$ with $\rho \le \frac{1}{2}$, we find that $\frac{\mu_0}{\nu_0} \le \frac{1}{1-\nu_1} = 1 + \frac{\nu_1}{1-\nu_1} \le 1 + 2\rho$. Because $Q_{01} \le \rho$, the same argument shows that $\frac{P_{00}}{Q_{00}} \le 1 + 2\rho$. Because $1 - x_1 + x_{00} \le T$, it follows that

$$\begin{aligned} \frac{f_x}{g_x} \;&=\; \left(\frac{\mu_0}{\nu_0}\right)^{1-x_1} \left(\frac{\mu_1}{\nu_1}\right)^{x_1} \left(\frac{P_{00}}{Q_{00}}\right)^{x_{00}} \left(\frac{P_{01}}{Q_{01}}\right)^{x_{01}} \left(\frac{P_{10}}{Q_{10}}\right)^{x_{10}} \left(\frac{P_{11}}{Q_{11}}\right)^{x_{11}} \\ &\le\; (1+2\rho)^T M^{x_1+x_{01}+x_{10}} \left(\frac{P_{11}}{Q_{11}}\right)^{x_{11}}. \end{aligned}$$

Observe also that $g_x \le \nu_1^{x_1} Q_{01}^{x_{01}} Q_{11}^{x_{11}} \le \rho^{x_1+x_{01}} Q_{11}^{x_{11}}$. Therefore,

$$Z \;\le\; (1+2\rho)^{\alpha T} \sum_x \rho^{x_1+x_{01}} M^{\alpha(x_1+x_{01}+x_{10})} \Lambda^{x_{11}},$$

where $\Lambda = P_{11}^\alpha Q_{11}^{1-\alpha}$. By recalling that $x_{\mathrm{on}} = x_1 + x_{01} = x_{10} + x_T$ and $\|x\| = x_1 + x_{01} + x_{11} = x_{10} + x_{11} + x_T$, we find that $x_1 + x_{01} + x_{10} = 2x_{\mathrm{on}} - x_T \le 2x_{\mathrm{on}}$ and $x_{11} = \|x\| - x_{\mathrm{on}}$. Hence

$$Z \;\le\; (1+2\rho)^{\alpha T} \sum_x \rho^{x_{\mathrm{on}}} M^{2\alpha x_{\mathrm{on}}} \Lambda^{\|x\|-x_{\mathrm{on}}} \;=\; (1+2\rho)^{\alpha T} \sum_{j=0}^{\lceil T/2 \rceil} S_j \tag{4.4.8}$$

where

$$S_j \;=\; (M^{2\alpha}\rho)^j \sum_{t=j}^{T} c_{jt} \Lambda^{t-j},$$

and c_{jt} is the number of paths $x \in \{0,1\}^T$ containing $x_{\mathrm{on}} = j$ on-periods and $\|x\| = t$ ones. Because there is only one path containing no ones, and this path has no on-periods, we find that $S_0 = 1$. By noting that $\log(1+t) \le t$, it follows from (4.4.8) that

$$D_\alpha(f\|g) \;\le\; \frac{2\alpha}{\alpha-1}\rho T + \frac{1}{\alpha-1} \sum_{j=1}^{\lceil T/2 \rceil} S_j. \tag{4.4.9}$$

Because $c_{1t} \le T$ for all t, we se that

$$S_1 \;\le\; M^{2\alpha}\rho T \sum_{t=0}^{T-1} \Lambda^t. \tag{4.4.10}$$

For $j \ge 2$, Lemma 20 implies that $c_{jt} = \sum_{a,b=0}^{1} \binom{t-1}{j-1}\binom{T-t-1}{j-a-b} \le 4\frac{T^j}{(j-2)!}\binom{t-1}{j-1}$, and we find that

$$S_j \;\le\; 4\frac{(M^{2\alpha}\rho T)^j}{(j-2)!} \sum_{t=j}^{T} \binom{t-1}{j-1}\Lambda^{t-j}. \tag{4.4.11}$$

Inequality (4.4.6) follows by substituting (4.4.10)–(4.4.11) into (4.4.9).

Assume next that $\Lambda < 1$, and denote $C = \frac{M^{2\alpha}}{1-\Lambda}$. By replacing $T-1$ by infinity on the right side of (4.4.10), it follows that $S_1 \le C\rho T$. By a geometric moment formula (Lemma 23), we find that

$$\sum_{t=j}^{T} \binom{t-1}{j-1}\Lambda^{t-j} \;\le\; \sum_{t=j}^{\infty} \binom{t-1}{j-1}\Lambda^{t-j} \;=\; (1-\Lambda)^{-j}.$$

Then (4.4.11) implies that

$$\sum_{j=2}^{\lceil T/2\rceil} S_j \;\le\; 4\sum_{j=2}^{\lceil T/2\rceil} \frac{(C\rho T)^j}{(j-2)!} \;\le\; 4\sum_{j=2}^{\infty} \frac{(C\rho T)^j}{(j-2)!} \;=\; 4(C\rho T)^2 e^{C\rho T}.$$

Now it follows by (4.4.9) that

$$D_\alpha(f\|g) \;\le\; \frac{2\alpha}{\alpha-1}\rho T + \frac{C\rho T}{\alpha-1} + \frac{4(C\rho T)^2}{\alpha-1}e^{C\rho T}.$$

Therefore,

$$\frac{(\alpha-1)D_\alpha(f\|g)}{C\rho T} \;\le\; \frac{2\alpha}{C} + 1 + 4C\rho T e^{C\rho T} \;\le\; \left(\frac{2\alpha}{C} + 1 + 4C\rho T\right) e^{C\rho T}.$$

Because $\frac{2\alpha}{C} + 1 + 4C\rho \le \left(\frac{2\alpha}{C}+1\right)(1+4C\rho T) \le \left(\frac{2\alpha}{C}+1\right) e^{4C\rho T}$, we conclude that

$$\frac{(\alpha-1)D_\alpha(f\|g)}{C\rho T} \;\le\; \left(\frac{2\alpha}{C}+1\right) e^{5C\rho T}$$

Because $C \ge 1$, we see that $\frac{2\alpha}{C} + 1 \le 2\alpha + 1$, and (4.4.7) follows. □

4.4.3 Additional lemmas

Lemma 23. *For any integer $j \geq 1$ and any real number $0 \leq q < 1$, we have*

$$\sum_{k=j}^{\infty} \binom{k}{j} q^{k-j} = (1-q)^{-(j+1)}.$$

Proof. Denote the falling factorial by $(x)_j = x(x-1)\cdots(x-j+1)$ and let $f(q) = (1-q)^{-1}$. Then the j-th derivative of f equals $f^{(j)}(q) = j!(1-q)^{-(j+1)}$. Because $f(q) = \sum_{k=0}^{\infty} q^k$, we find that the j-th derivative of f also equals $\sum_{k=j}^{\infty} (k)_j q^{k-j}$. Hence the claim follows. □

Lemma 24. *(i) For $t \geq 0$, $\log(1+t) = t - \epsilon_1$ where $0 \leq \epsilon_1 \leq \frac{1}{2}t^2$. (ii) For $0 \leq t < 1$, $\log(1-t) = -t - \epsilon_2$ where $0 \leq \epsilon_2 \leq \frac{t^2}{2(1-t)^2}$, and especially, $0 \leq \epsilon_2 \leq 2t^2$ for $0 \leq t \leq \frac{1}{2}$.*

Proof. (i) By taking two derivatives of $t \mapsto \log(1+t)$, we find that $\log(1+t) = t - \epsilon_1$ with $\epsilon_1 = \int_0^t \int_0^s (1+u)^{-2} du ds$.

(ii) Similarly, we find that $\log(1-t) = -t - \epsilon_2$ with $\epsilon_2 = \int_0^t \int_0^s (1-u)^{-2} du ds$. □

Lemma 25. *For any $0 \leq x \leq \frac{1}{2}$ and $a > 0$, the error term in the approximation $(1-x)^a = 1 - ax - r(x)$ is bounded by $|r(x)| \leq \frac{2|a-1|}{2^a} ax^2$. Moreover, $r(x) \geq 0$ when $a \geq 1$.*

Proof. The error term in the approximation $f(x) = f(0) + f'(0)x + r(x)$ equals $r(x) = \int_0^x \int_0^t f''(s) ds dt$ and is bounded by $|r(x)| \leq \frac{1}{2}cx^2$ with $c = \max_{0 \leq x \leq 1/2} |f''(x)|$. The function $f(x) = (1-x)^a$ satisfies $f(0) = 1$ and $f'(0) = -a$, together with $f''(x) = a(a-1)(1-x)^{a-2}$. The claims follow after noticing that

$$\max_{0 \leq x \leq 1/2} |f''(x)| = \begin{cases} |f''(\frac{1}{2})| = \frac{4}{2^a} a|a-1| & \text{for } 0 < a < 2, \\ f''(0) = a(a-1) & \text{for } a \geq 2. \end{cases}$$

□

Lemma 26. *Fix $0 \leq \delta < 1$. Then the error term in the approximation $\sqrt{1-x} = 1 - \frac{1}{2}x - \epsilon(x)$ satisfies $0 \leq \epsilon(x) \leq cx^2$ for all $0 \leq x \leq \delta$, where $c = \frac{1}{8}(1-\delta)^{-3/2}$.*

Proof. Consider Taylor's approximation $f(x) = f(0) + f'(0)x + r(x)$ where the quantity $r(x) = \int_0^x \int_0^t f''(s) ds dt$ is bounded by $\frac{1}{2}c_1 x^2 \leq r(x) \leq \frac{1}{2}c_2 x^2$ with $c_1 = \min_{0 \leq x \leq \delta} f''(x)$ and $c_2 = \max_{0 \leq x \leq \delta} f''(x)$. The function $f(x) = (1-x)^{1/2}$ satisfies $f(0) = 1$ and $f'(0) = -\frac{1}{2}$, together with $f''(x) = -\frac{1}{4}(1-x)^{-3/2}$. Now $c_1 = -\frac{1}{4}(1-x)^{-3/2}$ and $c_2 = -\frac{1}{4} \leq 0$. Hence the claim is true with $\epsilon(x) = -r(x)$. □

Part II

Temporal, geometric, and semi-supervised graph clustering

Chapter 5

Almost exact recovery in noisy semi supervised learning

Graph-based semi-supervised learning methods combine the graph structure and labelled data to classify unlabelled data. In this Chapter, we study the effect of a noisy oracle on classification. In particular, we derive in Section 5.1 the Maximum A Posteriori (MAP) estimator for clustering a Degree Corrected Stochastic Block Model (DC-SBM) when a noisy oracle reveals a fraction of the labels. We then propose an algorithm derived from a continuous relaxation of the MAP, and we establish its consistency in Section 5.2. Numerical experiments in Section 5.3 show that our approach achieves promising performance on synthetic and real data sets, even in the case of very noisy labelled data.

5.1 MAP estimator in a noisy semi-supervised setting

5.1.1 Problem formulation and notations

A homogeneous *degree corrected stochastic block model* (DC-SBM) is parametrized by the number of nodes n, two class-affinity parameters $p_{\text{in}}, p_{\text{out}}$, and a pair (θ, Z) where $\theta \in \mathbb{R}^n$ is a vector of intrinsic connection intensities and $Z \in \{-1, 1\}^n$ is the community labelling vector. Given $(p_{\text{in}}, p_{\text{out}}, \theta, Z)$, the graph adjacency matrix $A = (a_{ij})$ is generated as

$$A_{ij} = A_{ji} \sim \begin{cases} \text{Ber}\left(\theta_i \theta_j p_{\text{in}}\right) & \text{if} \quad Z_i = Z_j, \\ \text{Ber}\left(\theta_i \theta_j p_{\text{out}}\right) & \text{otherwise,} \end{cases} \tag{5.1.1}$$

for $i \neq j$, and $A_{ii} = 0$. We assume throughout the paper that $Z_i \sim \text{Uni}\left(\{-1, 1\}\right)$, and that the entries of θ are independent random variables satisfying $\theta_i \in [\theta_{\min}, \theta_{\max}]$ with $\mathbb{E}\theta_i = 1$, $\theta_{\min} > 0$, and $\theta_{\max}^2 \max(p_{\text{in}}, p_{\text{out}}) \leq 1$. In particular, when all the θ_i's are equal to one, the model reduces to the Stochastic Block Model (SBM):

$$A_{ij} = A_{ji} \sim \begin{cases} \text{Ber}\left(p_{\text{in}}\right) & \text{if} \quad Z_i = Z_j, \\ \text{Ber}\left(p_{\text{out}}\right) & \text{otherwise.} \end{cases} \tag{5.1.2}$$

In addition to the observation of the graph adjacency matrix A, an oracle gives us extra information about the cluster assignment of some nodes. This can be represented as a vector $s \in \{0, -1, 1\}^n$, whose entries s_i are independent and distributed as follows:

$$s_i \;=\; \begin{cases} Z_i & \text{with probability} \quad \eta_1, \\ -Z_i & \text{with probability} \quad \eta_0, \\ 0 & \text{otherwise.} \end{cases} \tag{5.1.3}$$

In words, the oracle (5.1.3) reveals the correct cluster assignment of node i with probability η_1 and gives a false cluster assignment with probability η_0. It reveals nothing with probability $1 - \eta_1 - \eta_0$. The quantity $\mathbb{P}\left(s_i \neq Z_i \,|\, s_i \neq 0\right)$ is the rate of mistakes of the oracle (*i.e.,* the probability that the oracle reveals a false information given that it reveals something), and is equal to $\eta_0/(\eta_1 + \eta_0)$. The oracle is informative if this quantity is less than $1/2$, which is equivalent to $\eta_1 > \eta_0$. In the following, we will always assume that the oracle is informative.

Assumption 2. The oracle is informative, that is, $\eta_1 > \eta_0$.

Given the observation of A and s, the goal of clustering is to recover the community labelling vector Z. For an estimator $\hat{Z} \in \{-1, 1\}^n$ of Z, the relative error is defined as the proportion of misclustered nodes

$$L\left(\hat{Z}, Z\right) \;=\; \frac{1}{n} \sum_{i=1}^{n} 1\left(\hat{Z}_i \neq Z_i\right). \tag{5.1.4}$$

Note that, unlike unsupervised clustering, we do not take a minimum over the permutations of the predicted labels since we should be able to learn the correct community labels from the informative oracle.

Notations Given an oracle s, we let ℓ be the set of labelled nodes, that is $\ell := \{i \in V : s_i \neq 0\}$, and denote $\mathcal{P}$ the diagonal matrix with entries $(\mathcal{P})_{ii} = 1$ if $i \in \ell$, and $(\mathcal{P})_{ii} = 0$ otherwise.

The notation I_n stands for the identity matrix of size $n \times n$, and 1_n (resp. 0_n) is the vector of size $n \times 1$ of all ones (resp. of all zeros).

For any matrix $A = (a_{ij})_{i\in[n],j\in[m]}$ and two sets $S \subset [n]$, $T \subset [m]$, we denote $A_{S,T} = (a_{ij})_{i\in S,j\in T}$ the matrix obtained from A by keeping elements whose row indices are in S and column indices are in T. We denote by $\|x\|$ the Euclidean norm of a vector x and by $\|A\|$ the spectral norm of a matrix $A \in \mathbb{R}^{n\times m}$. Finally, $A \odot B$ refers to the entry-wise matrix product between two matrices A and B of the same size.

5.1.2 MAP estimator of semi-supervised recovery in DC-SBM

Given a realization of a DC-SBM graph adjacency matrix A and the oracle information s, the Maximum A Posteriori (MAP) estimator is defined as

$$\hat{Z}^{\mathrm{MAP}} = \underset{z\in\{-1,1\}^n}{\arg\max}\, \mathbb{P}(z \mid A, s). \tag{5.1.5}$$

This estimator is known to be optimal (in the sense that if it fails then any other estimator would also fail, see *e.g.*, [Iba99]) for the exact recovery of all the community labels. Theorem 9 provides an expression of the MAP.

Theorem 9. *Let G be a graph drawn from DC-SBM as defined in (5.1.1) and s be the oracle information as defined in (5.1.3). Denote $M = (F_1 - F_0) \odot A + F_0$, where $F_0 = \left(f_{ij}^{(0)}\right)$ and $F_1 = \left(f_{ij}^{(1)}\right)$ such that $f_{ij}^{(a)} = \log \frac{\mathbb{P}(A_{ij}=a \mid z_i=z_j)}{\mathbb{P}(A_{ij}=a \mid z_i\neq z_j)}$ for $a \in \{0,1\}$. The MAP estimator defined in (5.1.5) is given by*

$$\hat{Z}^{\mathrm{MAP}} = \underset{z\in\{-1,1\}^n}{\arg\min} \left(z^T M z + \log\left(\frac{\eta_1}{\eta_0}\right) \|\mathcal{P}z - s\|^2 \right). \tag{5.1.6}$$

For a perfect oracle ($\eta_0 = 0$) this reduces to

$$\hat{Z}^{\mathrm{MAP}} = \underset{\substack{z\in\{-1,1\}^n \\ z_\ell = s_\ell}}{\arg\min}\; z^T M z. \tag{5.1.7}$$

The proof of Theorem 9 is standard and postponed to Section 5.4.1. We note that, despite being *a priori* standard, this result did not appear previously in the literature (neither for the standard SBM nor for the perfect oracle).

The minimisation problem (5.1.6) consists of a trade-off between minimising a quadratic function $z^T M z$ and a penalty term. This trade-off reads as follows: for each labelled node such that the prediction contradicts the oracle, a penalty $\log\left(\frac{\eta_1}{\eta_0}\right) > 0$ is added. In particular, when the oracle is uninformative, that is $\eta_1 = \eta_0$, then this term is null, and Expression (5.1.6) reduces to the MAP for unsupervised clustering.

The following Corollary 1, whose proof is in Section 5.4.1, provides the expression of the MAP estimator for a standard SBM.

Corollary 1. *The MAP estimator for semi-supervised clustering on SBM graph with $p_{\text{in}} > p_{\text{out}}$ and with an oracle s defined in (5.1.3) is given by*

$$\hat{Z}^{\text{MAP}} = \underset{z\in\{-1,1\}^n}{\arg\min} \left(-z^T \left(A - \tau 1_n 1_n^T\right) z + \lambda \left\|\mathcal{P}z - s\right\|_2^2\right), \tag{5.1.8}$$

where $\tau = \dfrac{\log\left(\dfrac{1-p_{\text{out}}}{1-p_{\text{in}}}\right)}{\log\left(\dfrac{p_{\text{in}}(1-p_{\text{out}})}{p_{\text{out}}(1-p_{\text{in}})}\right)}$ *and* $\lambda = \dfrac{\log\left(\dfrac{\eta_1}{\eta_0}\right)}{\log\left(\dfrac{p_{\text{in}}(1-p_{\text{out}})}{p_{\text{out}}(1-p_{\text{in}})}\right)}$. *For the perfect oracle, this reduces to*

$$\hat{Z}^{\text{MAP}} = \underset{\substack{z\in\{-1,1\}^n \\ z_\ell = s_\ell}}{\arg\min} \left(z^T \left(-A + \tau 1_n 1_n^T\right) z\right). \tag{5.1.9}$$

5.2 Almost exact recovery using a continuous relaxation

As finding the MAP estimate is NP-hard [WW93], we perform a continuous relaxation (Section 5.2.1). We then give an upper bound on the number of misclustered nodes in Section 5.2.2.

5.2.1 Continuous relaxation of the MAP

For the sake of simplicity, we focus on the MAP for SBM, *i.e.,* minimisation problem (5.1.8). We perform a continuous relaxation mirroring what is commonly done for spectral methods [New13], namely

$$\hat{X} = \underset{\substack{x\in\mathbb{R}^n \\ \sum_i \kappa_i x_i^2 = \sum_i \kappa_i}}{\arg\min} \left(-x^T A_\tau x + \lambda (s - \mathcal{P}x)^T (s - \mathcal{P}x)\right), \tag{5.2.1}$$

where $A_\tau = A - \tau 1_n 1_n^T$ and $\kappa = (\kappa_1, \ldots, \kappa_n)$ is a vector of positive entries. For the simplicity of the derivations, we choose to constrain x on the hyper-sphere $\|x\|^2 = n$ by letting $\kappa_i = 1$, but other choices would lead to a similar analysis. In particular, in the numerical Section 5.3 we will compare this choice with a degree-normalization approach ($\kappa_i = d_i$).

We further note that for the perfect oracle the corresponding relaxation of (5.1.9) is

$$\hat{X} = \underset{\substack{x\in\mathbb{R}^n \\ x_\ell = s_\ell \\ \|x\|^2 = n}}{\arg\min} \left(-x^T A_\tau x\right). \tag{5.2.2}$$

Given the classification vector $\hat{X} \in \mathbb{R}^n$, node i is classified into cluster $\hat{Z}_i \in \{-1, 1\}$ such that

$$\hat{Z}_i = \begin{cases} 1 & \text{if} \quad \hat{X}_i > 0, \\ -1 & \text{otherwise.} \end{cases} \tag{5.2.3}$$

Let us solve the minimisation problem (5.2.1). By letting $\gamma \in \mathbb{R}$ be the Lagrange multiplier associated with the constraint $\|x\|^2 = n$, the Lagrangian of the optimization problem (5.2.1) is

$$-x^T A_\tau x + \lambda (s - \mathcal{P}x)^T (s - \mathcal{P}x) - \gamma \left(x^T x - n\right).$$

This leads to the *constrained* linear system

$$\begin{cases} (-A_\tau + \lambda\mathcal{P} - \gamma I_n)\, x &= \lambda s, \\ x^T x &= n, \end{cases} \tag{5.2.4}$$

whose unknowns are γ and x.

While [MC21] let γ to be a hyper-parameter (hence the norm constraint $x^T x = n$ is no longer verified), the exact optimal value of γ can be found explicitly following [GGV89]. Firstly, we note that if (γ_1, x_1) and (γ_2, x_2) are solutions of the system (5.2.4), then

$$\mathcal{C}(x_1) - \mathcal{C}(x_2) = \frac{\gamma_1 - \gamma_2}{2} \|x_1 - x_2\|^2,$$

where $\mathcal{C}(x) = -x^T A_\tau x + \lambda(s - \mathcal{P}x)^T(s - \mathcal{P}x)$ is the cost function minimised in (5.2.1). Hence, among the solution pairs (γ, x) of the system (5.2.4), the solution of the minimisation problem (5.2.1) is the vector x associated with the smallest γ.

Secondly, the eigenvalue decomposition of $-A_\tau + \lambda\mathcal{P}$ reads as

$$-A_\tau + \lambda\mathcal{P} = Q\Delta Q^T,$$

where $\Delta = \mathrm{diag}(\delta_1, \dots, \delta_n)$ with $\delta_1 \leq \cdots \leq \delta_n$ and $Q^T Q = I_n$. Therefore, after the change of variables $u = Q^T x$ and $b = \lambda Q^T s$, the system (5.2.4) is transformed to

$$\begin{cases} \Delta u &= \gamma u + b, \\ u^T u &= n. \end{cases}$$

Thus, the solution $\hat{X}$ of the optimization problem (5.2.1) verifies

$$(-A_\tau + \lambda\mathcal{P} - \gamma_* I_n)\, \hat{X} = \lambda s, \tag{5.2.5}$$

where γ_* is the smallest solution of the *explicit secular equation* [GGV89]

$$\sum_{i=1}^{n} \left(\frac{b_i}{\delta_i - \gamma}\right)^2 - n = 0. \tag{5.2.6}$$

We summarize this in Algorithm 3. Note that for the sake of generality we let λ and τ be hyper-parameters of the algorithm. If the model parameters are known, we can use the expressions of λ and τ derived in Corollary 1. The choice of λ and τ is further discussed in Section 5.3.

5.2.2 Ratio of misclustered nodes

This section gives bounds on the number of unlabelled nodes misclassified by Algorithm 3. We then specialize the results for some particular cases.

Theorem 10. *Consider a DC-SBM with a noisy oracle as defined in* (5.1.1),(5.1.3). *Let* $\bar{d} = \frac{n}{2}(p_{\text{in}} + p_{\text{out}})$ *and* $\bar{\alpha} = \frac{n}{2}(p_{\text{in}} - p_{\text{out}})$. *Suppose that* $\tau > p_{\text{out}}$, *and let* $\hat{Z}$ *be the output of Algorithm 3. Then, the proportion of misclustered unlabelled nodes verifies*

$$L\left(\hat{Z}_u, Z_u\right) \leq C \left(\frac{p_{\text{in}} + p_{\text{out}}}{p_{\text{in}} - p_{\text{out}}}\right)^2 \left(\frac{\bar{\alpha} + \lambda}{\lambda}\right)^2 \frac{1}{(\eta_1 + \eta_0)\,(\eta_1 - \eta_0)^2\, \bar{d}}.$$

Algorithm 3: Semi-supervised learning with regularized adjacency matrix.

Input: Adjacency matrix A, oracle information s, parameters τ and λ.
Procedure:
Let γ^* be the smallest solution of Equation (5.2.6).
Compute $\hat{X}$ as the solution of Equation (5.2.5).
for $i = 1 \dots n$ **do**
  $\hat{Z}_i = \operatorname{sign}\left(\hat{X}_i\right)$.
end for
return $\hat{Z}$.

The core of the proof relies on the concentration of the adjacency matrix towards its expectation. This result, as presented in [LLV17], holds under loose assumptions: it is valid for any random graph whose edges are independent of each other. To use this result for $\bar{d} = o\big(\log n\big)$, one need to replace the matrix A_τ by $A'_\tau = A' - \tau 1_n 1_n^T$, where A' is the adjacency matrix of the graph obtained after reducing the weights on the edges incident to the high degree vertices. We refer to [LLV17, Section 1.4] for more details. This extra technical step is not necessary when $\bar{d} = \Omega(\log n)$. Moreover, concentration also occurs if we replace the adjacency matrix by the normalized Laplacian in Equation (5.2.5). In that case, we obtain a generalization of the Label Spreading algorithm [Zho+04], [CSZ06, Chapter 11].

In the following, the mean-field graph refers to the weighted graph formed by the expected adjacency matrix of a DC-SBM graph. Moreover, we assume without loss of generality that the first $\frac{n}{2}$ nodes are in the first cluster and the last $\frac{n}{2}$ are in the second cluster. Therefore, $\mathbb{E}A = ZBZ^T$ with $B = \begin{pmatrix} p_{\text{in}} & p_{\text{out}} \\ p_{\text{out}} & p_{\text{in}} \end{pmatrix}$ and $Z = \begin{pmatrix} 1_{n/2} & 0_{n/2} \\ 0_{n/2} & 1_{n/2} \end{pmatrix}$. In particular, the coefficients θ_i disappear because $\mathbb{E}\theta_i = 1$. We consider the setting where diagonal elements of $\mathbb{E}A$ are not zeros. This accounts for modifying the definition of DC-SBM, where we can have self-loops with probability p_{in}. Nonetheless, we could set the diagonal elements of $\mathbb{E}A$ to zeros and our results would still hold at the expense of cumbersome expressions. Note that the matrix $\mathbb{E}A$ has two non-zero eigenvalues: $\bar{d} = n\frac{p_{\text{in}}+p_{\text{out}}}{2}$ and $\bar{\alpha} = n\frac{p_{\text{in}}-p_{\text{out}}}{2}$.

Proof of Theorem 10. We prove the statement in three steps. We first show that the solution $\hat{X}$ of the constrained linear system (5.2.4) is concentrated around the solution $\bar{x}$ of the same system for the mean-field model. Then, we compute $\bar{x}$ and show that we can retrieve the correct cluster assignment from it. We finally conclude with the derivation of the bound.

(i) Similarly to [AKL18] and [AD19], let us rewrite equation (5.2.5) as a perturbation of a system of linear equations corresponding to the mean-field solution. We thus have

$$\big(\mathbb{E}\tilde{\mathcal{L}} + \Delta\tilde{\mathcal{L}}\big)\big(\bar{x} + \Delta x\big) = \lambda s,$$

where $\tilde{\mathcal{L}} = -A_\tau + \lambda\mathcal{P} - \gamma_* I_n$, $\Delta x := \hat{X} - \bar{x}$ and $\Delta\tilde{\mathcal{L}} := \tilde{\mathcal{L}} - \mathbb{E}\tilde{\mathcal{L}}$.

We recall that a perturbation of a system of linear equations $(A + \Delta A)(x + \Delta x) = b$ leads

to the following sensitivity inequality (see e.g., [HJ12, Section 5.8]):

$$\frac{\|\Delta x\|}{\|x\|} \leq \kappa(A)\frac{\|\Delta A\|}{\|A\|},$$

where $\|.\|$ is the operator norm associated to a vector norm $\|.\|$ (we use the same notations for simplicity) and $\kappa(A) := \|A^{-1}\| \cdot \|A\|$ is the condition number. In our case, the above inequality can be rewritten as follows:

$$\frac{\left\|\hat{X} - \bar{x}\right\|}{\|\bar{x}\|} \leq \left\|\left(\mathbb{E}\,\tilde{\mathcal{L}}\right)^{-1}\right\| \cdot \left\|\Delta\,\tilde{\mathcal{L}}\right\|, \tag{5.2.7}$$

employing the Euclidean vector norm and spectral operator norm. The spectral study of $\mathbb{E}\,\tilde{\mathcal{L}}$ (see Corollary 4 in Section 5.4.2) gives:

$$\left\|\left(\mathbb{E}\,\tilde{\mathcal{L}}\right)^{-1}\right\| = \frac{1}{\min\left\{|\lambda| : \lambda \in \mathrm{Sp}\left(\mathbb{E}\,\tilde{\mathcal{L}}\right)\right\}} = \frac{1}{-t_2^+ - \bar{\gamma}_*},$$

where t_2^+ is defined in Corollary 4 in Section 5.4.2 and $\bar{\gamma}_*$ is the solution of Equation (5.2.6) for the mean-field model. Lemma 28 in Section 5.4.2 leads to

$$\left\|\left(\mathbb{E}\,\tilde{\mathcal{L}}\right)^{-1}\right\| \leq \frac{1}{\lambda + \bar{\alpha}}. \tag{5.2.8}$$

The last ingredient needed is the concentration of the adjacency matrix around its expectation. We have

$$\left\|\tilde{\mathcal{L}} - \mathbb{E}\tilde{\mathcal{L}}\right\| \leq \|(\gamma_* - \bar{\gamma}_*)\, I_n\| + \|A - \mathbb{E}\,A\| \leq |\,\gamma_* - \bar{\gamma}_*\,| + \|A - \mathbb{E}\,A\|.$$

Proposition 16 in Section 5.4.2 shows that

$$|\,\gamma_* - \bar{\gamma}_*\,| \leq \left(1 + \frac{27\,(\bar{\alpha} + \lambda)^3}{\sqrt{2}\sqrt{\eta_1 + \eta_0}(\eta_1 - \eta_0)\bar{\alpha}^2\lambda}\right)\sqrt{\bar{d}}.$$

Moreover, when $d = \Omega(\log n)$, it is shown in [FO05] that $\|A - \mathbb{E}\,A\| = O\left(\sqrt{\bar{d}}\right)$. If $\bar{d} = o(\log n)$, the same result holds with a proper pre-processing on A, and we refer the reader to [LLV17] for more details. To keep notations short, we will omit this extra step in the proof. Using this concentration bound, we have

$$\begin{aligned}\left\|\tilde{\mathcal{L}} - \mathbb{E}\tilde{\mathcal{L}}\right\| &\leq \left(C' + \frac{27\,(\bar{\alpha} + \lambda)^3}{\sqrt{2}\sqrt{\eta_1 + \eta_0}(\eta_1 - \eta_0)\bar{\alpha}^2\lambda}\right)\sqrt{\bar{d}} \\ &\leq \left(C' + \frac{27}{\sqrt{2}}\right)\frac{(\lambda + \bar{\alpha})^3}{\bar{\alpha}^2\lambda}\frac{\sqrt{\bar{d}}}{\sqrt{\eta_1 + \eta_0}\,(\eta_1 - \eta_0)}.\end{aligned}$$

for some constant C'. Let $C = C' + \frac{27}{\sqrt{2}}$. By combining the above with inequality (5.2.8), the inequality (5.2.7) becomes

$$\frac{\left\|\hat{X} - \bar{x}\right\|}{\|\bar{x}\|} \leq C \frac{(\lambda + \bar{\alpha})^2}{\bar{\alpha}^2 \lambda} \frac{\sqrt{\bar{d}}}{\sqrt{\eta_1 + \eta_0}\,(\eta_1 - \eta_0)}. \tag{5.2.9}$$

(ii) Node i in the mean-field model is correctly classified by decision rule (5.2.3) if the sign of $\bar{x}_i$ equals the sign of Z_i. Corollary 5 in Section 5.4.3 shows that this is indeed the case for the unlabelled nodes.

(iii) Finally, for an unlabelled node i to be correctly classified, the node's value $\hat{X}_i$ should be close enough to its mean-field value $\bar{x}_i$. In particular, the part (ii) shows that if $|\hat{X}_i - \bar{x}_i|$ is smaller than some non-vanishing constant β, then an unlabelled node i will be correctly classified. An unlabelled node i is said to be β-bad if $\left|\hat{X}_i - \bar{x}_i\right| > \beta$. We denote by S_β the set of β-bad nodes. The nodes that are not β-bad are a.s. correctly classified, and thus $L\left(\hat{Z}_u, Z_u\right) \leq \frac{|S_\beta|}{n}$.

From $\left\|\hat{X} - \bar{x}\right\|^2 \geq \sum_{i \in S_\beta} \left|\hat{X}_i - \bar{x}_i\right|^2$, it follows that $\left\|\hat{X} - \bar{x}\right\|^2 \geq |S_\beta| \times \beta^2$. Thus, using inequality (5.2.9) and the norm constraint $\|\bar{x}\|^2 = n$, we have

$$|S_\beta| \leq \frac{1}{\beta^2} \left(\frac{C}{\eta_1 - \eta_0} \frac{\bar{\alpha} + \lambda}{\bar{\alpha}\lambda} \sqrt{\bar{d}} \right)^2 n,$$

for some constant C. We end the proof by noticing that $\frac{\bar{d}}{\bar{\alpha}} = \frac{p_{\text{in}} + p_{\text{out}}}{p_{\text{in}} - p_{\text{out}}}$. □

Corollary 2 (Almost exact recovery in the diverging degree regime)**.** *Consider a DC-SBM such that* $\bar{d} \gg 1$, $\frac{p_{\text{in}} + p_{\text{out}}}{p_{\text{in}} - p_{\text{out}}} = O(1)$, *and* $\sqrt{\eta_0 + \eta_1}(\eta_1 - \eta_0) \gg \frac{1}{\sqrt{\bar{d}}}$. *Suppose that* $\tau > p_{\text{out}}$ *and* $\lambda \gtrsim \bar{\alpha}$. *Then, Algorithm 3 correctly classifies almost all the unlabelled nodes.*

Proof. With the corollary's assumptions $(\eta_1 - \eta_0)^2 \bar{d} \to +\infty$ and $\frac{\bar{\alpha} + \lambda}{\lambda} = O(1)$, by Theorem 10 the fraction of misclustered nodes is of the order $o(1)$. □

The quantity $(\eta_1 - \eta_0)n$ is the expected difference between the number of nodes correctly labelled and the number of nodes wrongly labelled by the oracle. In particular, Corollary 2 allows for a sub-linear number of labelled nodes since η_0 and η_1 can go to zero.

Corollary 3 (Detection in the constant degree regime)**.** *Consider a DC-SBM such that* $p_{\text{in}} = \frac{c_{\text{in}}}{n}$ *and* $p_{\text{out}} = \frac{c_{\text{out}}}{n}$ *where* $c_{\text{in}}, c_{\text{out}}$ *are constants. Suppose that* $\sqrt{\eta_0 + \eta_1}(\eta_1 - \eta_0)$ *is a non-zero constant, and let* $\tau > 2p_{\text{out}}$ *and* $\lambda \gtrsim 1$. *Then, for* $\frac{(c_{\text{in}} - c_{\text{out}})^2}{c_{\text{in}} + c_{\text{out}}}$ *bigger than some constant, w.h.p. Algorithm 3 performs better than a random guess.*

Proof. According to Theorem 10, the fraction of misclustered nodes is smaller than $\frac{1}{2}$ when $\frac{(c_{\text{in}} - c_{\text{out}})^2}{c_{\text{in}} + c_{\text{out}}}$ is larger than $\frac{2C}{(\eta_1 - \eta_0)^2} \left(\frac{\bar{\alpha} + \lambda}{\lambda}\right)^2$, which is lower bounded by a constant. □

The quantity $\frac{(c_{\text{in}}-c_{\text{out}})^2}{c_{\text{in}}+c_{\text{out}}}$ can be interpreted as the signal-to-noise ratio. It is unfortunate that Corollary 3 does not allow us to control the constant in the statement of the corollary. This constant comes from concentration of the adjacency matrix. Similar remarks were made in [LLV17] for the analysis of spectral clustering in the constant degree regime for SBM graph.

5.3 Numerical experiments

This section presents numerical experiments both on simulated data sets generated from DC-SBMs and on real networks. In particular, we discuss the impact of the oracle mistakes (defined by the ratio $\frac{\eta_0}{\eta_0+\eta_1}$) on the performance of the algorithms. The code for the simulations is available on github at https://github.com/mdreveton/ssl-sbm.

Choice of λ and τ Let us denote by σ_1 and σ_2 the largest and second largest eigenvalues of A. We choose $\tau = \frac{4}{n}(\sigma_1+\sigma_2)$ and $\lambda = \frac{\log\frac{\eta_1}{\eta_0}}{\log\frac{\sigma_1+\sigma_2}{\sigma_1-\sigma_2}}$ if $\eta_0 \neq 0$, and $\lambda = \frac{\log(n\eta_1)}{\log\frac{\sigma_1+\sigma_2}{\sigma_1-\sigma_2}}$ otherwise. The heuristic for this choice is as follows. For a SBM graph, we have $\sigma_1 \approx \frac{n}{2}(p_{\text{in}}+p_{\text{out}})$ and $\sigma_2 \approx \frac{n}{2}(p_{\text{in}}-p_{\text{out}})$, hence $\frac{4}{n}(\sigma_1+\sigma_2) = 2p_{\text{in}} > p_{\text{out}}$, and τ verifies the condition of Theorem 10. For λ, we have $\frac{\log\frac{\eta_1}{\eta_0}}{\log\frac{\sigma_1+\sigma_2}{\sigma_1-\sigma_2}} \approx \frac{\log\frac{\eta_1}{\eta_0}}{\log\frac{p_{\text{in}}}{p_{\text{out}}}}$, which is indeed close to the expression of λ derived in Corollary 1 if $p_{\text{in}}, p_{\text{out}} = o(1)$.

Choice of relaxation We first compare the choice of the constraint in the continuous relaxation (5.2.1). Specifically, we compare the choice $\sum_i x_i^2 = n$ (we refer to as *standard relaxation*) versus $\sum_i d_i x_i^2 = 2|E|$ (we refer to as *degree-normalized relaxation*). This leads to two versions of Algorithm 3, whose cost obtained on SBM graph with a noisy oracle is presented in Figure 5.1. In particular, we observe that the normalized choice leads to a smaller cost. Therefore, in the following we will only consider the version of Algorithm 3 solving the relaxed problem (5.2.1) with constraint $\sum_i d_i x_i^2 = 2|E|$ instead of $\sum_i x_i^2 = n$, as it gives better numerical results.

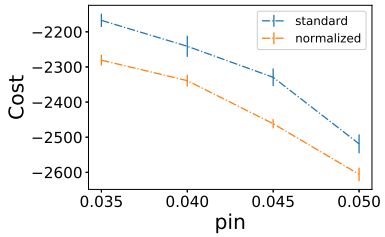
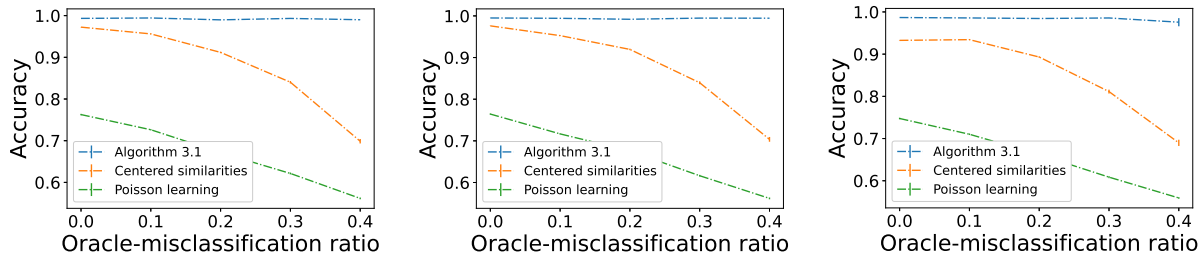


Figure 5.1: Cost in Algorithm 3 with the standard and normalized versions of the constraint, on 50 realizations of SBM with $n = 500$, $p_{\text{out}} = 0.03$ and 50 labelled nodes with 10% noise.

Experiments on synthetic graphs We first consider clustering on DC-SBM. We set $n = 2000$, $p_{\text{in}} = 0.04$ and $p_{\text{out}} = 0.02$. We consider three scenarios.

- In Figure 5.2a we consider a standard SBM ($\theta_i = 1$ for all i);
- In Figure 5.2b we generate θ_i according to $|\mathcal{N}(0, \sigma^2)| + 1 - \sigma\sqrt{2/\pi}$ where $|\mathcal{N}(0, \sigma^2)|$ denotes the absolute value of a normal random variable with mean 0 and variance σ^2. We take $\sigma = 0.25$. Note that this definition enforces $\mathbb{E}\theta_i = 1$.
- In Figure 5.2c we generate θ_i from Pareto distribution with density function $f(x) = \frac{am^a}{x^{a+1}} 1(x \geq m)$ with $a = 3$ and $m = 2/3$ (chosen such that $\mathbb{E}\theta_i = 1$).

We compare the performance of Algorithm 3 to the algorithm of [MC21] (referred to as *Centered Similarities*) and the *Poisson learning* algorithm described in [Cal+20]. Results are shown in Figure 5.2. We observe that when the oracle noise is low, the performance of Algorithm 3 is comparable to *Centered Similarities*. But, when the noise starts to be non-negligible, the performance of *Centered Similarities* deteriorates, while the accuracy of Algorithm 3 remains high. We notice that *Poisson learning* gives poor result on synthetic data sets.

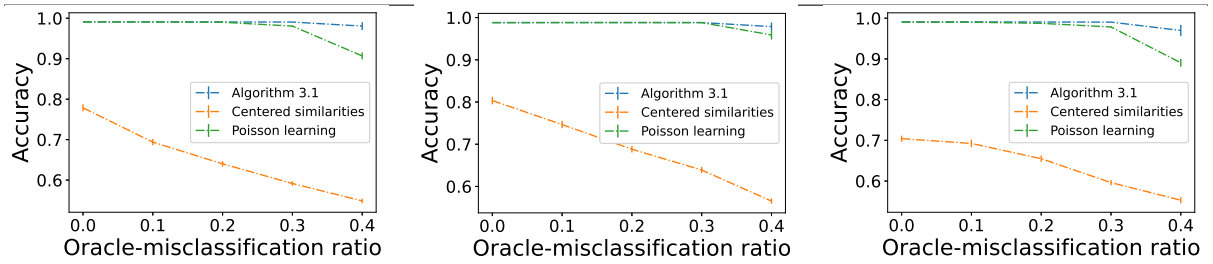


(a) SBM (b) Normal Degree (c) Pareto Degree

Figure 5.2: Average accuracy obtained by different semi-supervised clustering methods on DC-SBM graphs, with $n = 2000$, $p_{\text{in}} = 0.04$ and $p_{\text{out}} = 0.02$ with different distributions for θ. The number of labelled nodes is equal to 40. Accuracies are computed on the unlabelled nodes, and are averaged over 100 realisations; the error bars show the standard error.

Experiments on real data As a real-life example, we perform simulations on the standard MNIST data set [LCB98]. As preprocessing, we select 1000 images corresponding to two digits and compute the k-nearest-neighbors graph (we take $k = 8$) with gaussian weights $w_{ij} = \exp\left(-\|x_i - x_j\|^2/s_i^2\right)$ where x_i represents the data for image i and s_i is the average distance between x_i and its K-nearest neighbors. Accuracy for different digit pairs is given in Figure 5.3. While the performance of *Poisson learning* is excellent, it can suffer from the oracle noise, while the accuracy of Algorithm 3 remains unchanged.

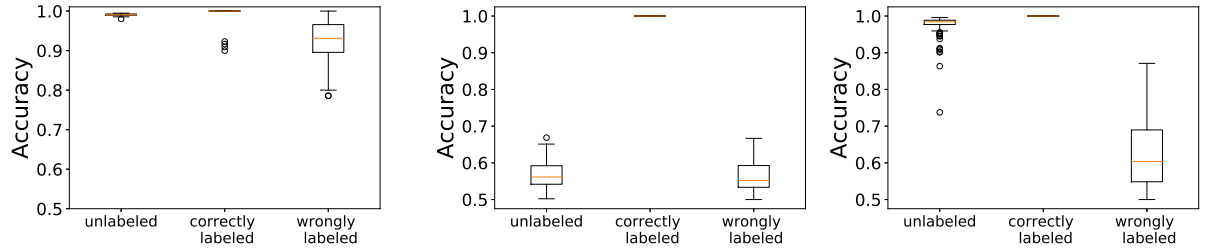
To further highlight the influence of the noise, we plot in Figure 5.4 the accuracy obtained by the three algorithms on the unlabelled nodes, the correctly labelled nodes, and the wrongly labelled nodes. We observe that the hard constraint $X_\ell = s_\ell$ imposed by *Centered Similarities* forces that the correctly labelled nodes to be correctly classified, while the wrongly labelled nodes are not classified much better than a random guess. In an extremely noisy setting, this heavily penalizes the unlabelled nodes' accuracy. On the contrary, Algorithm 3 allows for a smoother recovery: the unlabelled, correctly labelled, and wrongly labelled nodes have



(a) Digits (2,4). (b) Digits (3,6). (c) Digits (7,8).

Figure 5.3: Average accuracy obtained on a subset of the MNIST data set by different semi-supervised algorithms as a function of the oracle-misclassification ratio, when the number of labelled nodes is equal to 10. Accuracy is averaged over 100 random realizations, and the error bars show the standard error.

roughly the same classification accuracy. While some correctly labelled nodes are misclassified, many wrongly labelled nodes become correctly classified, and the unlabelled nodes are better recovered. Finally, *Poisson learning* lies between these two extreme cases: its accuracy on the unlabelled nodes is excellent, but it fails at correctly classifying the wrongly labelled nodes.



(a) Algorithm 3. (b) *Centered Similarities*. (c) *Poisson learning*.

Figure 5.4: Average accuracy obtained on the unlabelled, correctly labelled, and wrongly labelled nodes by the oracle. Simulations are done on the 1000 digits (2,4). The noisy oracle correctly classifies 24 nodes and misclassifies 16 nodes, and the boxplots show 100 realizations.

5.4 Additional proofs

5.4.1 Derivation of the MAP

Proof of Theorem 9. Bayes' formula gives $\mathbb{P}(z \mid A, s) \propto \mathbb{P}(A \mid z, s)\, \mathbb{P}(z \mid s)$, where the proportionality symbol hides $\mathbb{P}(A \mid s)$-term independent of z.

The likelihood term can be rewritten as follows:

$$\mathbb{P}(A \mid z, s) \;=\; \mathbb{P}(A \mid z) \;\propto\; \prod_{\substack{i<j \\ z_i = z_j}} \left(\frac{p_{\text{in}}}{p_{\text{out}}} \frac{1 - \theta_i \theta_j p_{\text{out}}}{1 - \theta_i \theta_j p_{\text{in}}} \right)^{a_{ij}} \left(\frac{1 - \theta_i \theta_j p_{\text{in}}}{1 - \theta_i \theta_j p_{\text{out}}} \right),$$

where the proportionality hides a constant $C = \prod_{i<j} \left(\frac{\theta_i \theta_j p_{\text{out}}}{1 - \theta_i \theta_j p_{\text{out}}} \right)^{a_{ij}} (1 - \theta_i \theta_j p_{\text{out}})$ independent of z. Hence,

$$\begin{aligned} \log \mathbb{P}\left(A \mid z, s\right) \;&=\; \log C + \frac{1}{2} \sum_{i,j} 1(z_i \neq z_j) \left(\left(f_{ij}^{(1)} - f_{ij}^{(0)} \right) a_{ij} + f_{ij}^{(0)} \right) \\ &=\; \log C + \frac{1}{2} \sum_{i,j=1}^{n} \frac{1 - z_i z_j}{2} \left(\left(f_{ij}^{(1)} - f_{ij}^{(0)} \right) a_{ij} + f_{ij}^{(0)} \right) \\ &=\; \log C' - \frac{1}{4} x^T M x. \end{aligned} \tag{5.4.1}$$

for some constant C' and $M = (F_1 - F_0) \odot A + F_0$.

The oracle information, given by the term $\mathbb{P}(z \mid s)$, is equal to

$$\begin{aligned} \mathbb{P}(z \mid s) \;&=\; \prod_{i=1}^{n} \frac{\mathbb{P}(s_i \mid z_i)}{\mathbb{P}(s_i)} \mathbb{P}(z_i) \\ &=\; \left(\frac{\eta_1}{\eta_1 + \eta_0} \right)^{\left| \{i \in \ell \colon z_i = s_i\} \right|} \left(\frac{\eta_0}{\eta_1 + \eta_0} \right)^{\left| \{i \in \ell \colon z_i \neq s_i\} \right|} \left(\frac{1}{2} \right)^{n} \\ &=\; \left(\frac{\eta_0}{\eta_1} \right)^{\left| \{i \in \ell \,:\, z_i \neq s_i\} \right|} \left(\frac{\eta_1}{\eta_1 + \eta_0} \right)^{\left| \ell \right|} \left(\frac{1}{2} \right)^{n}, \end{aligned} \tag{5.4.2}$$

where we used $\left| \{i \in \ell \colon z_i = s_i\} \right| + \left| \{i \in \ell \colon z_i \neq s_i\} \right| = \left| \ell \right|$ in the last line. Noticing that

$$\left| \{i \in \ell : z_i \neq s_i\} \right| \;=\; \frac{1}{4} \sum_{i=1}^{n} \left((\mathcal{P} z)_i - s_i \right)^2 \;=\; \frac{1}{4} \left(\mathcal{P} z - s \right)^T \left(\mathcal{P} z - s \right),$$

yields

$$\log \mathbb{P}\left(z \mid s\right) \;=\; -\frac{1}{4} \log \left(\frac{\eta_1}{\eta_0} \right) \cdot \left\| \mathcal{P} z - s \right\|^2 + C', \tag{5.4.3}$$

where C' is a term independent of z.

If $\eta_0 \neq 0$, the combination of Equations (5.4.1) and (5.4.3) with Bayes' formula gives Expression (5.1.6). If $\eta_0 = 0$, then from Equation (5.4.2) the term $\mathbb{P}(z \mid s)$ is non-zero (and constant) if and only if $z_i = s_i$ for every labelled node $i \in [\ell]$, and we obtain Expression (5.1.7). □

Proof of Corollary 1. The proof follows from Theorem 9 and the fact that $f_{ij}^{(0)} = \log \frac{1 - p_{\text{in}}}{1 - p_{\text{out}}}$ and $f_{ij}^{(1)} = \log \frac{p_{\text{in}}}{p_{\text{out}}}$. □

5.4.2 Lemmas related to mean-field solution of the secular equation

Spectral study of a perturbed rank-2 matrix

Lemma 27 (Matrix determinant lemma). *Suppose $A \in \mathbb{R}^n$ is invertible, and let U, V be two n by m matrices. Then* $\det(A + UV^T) = \det A \det(I_m + V^T A^{-1} U)$.

Proof. We take the determinant of $\begin{pmatrix} A & -U \\ V^T & I \end{pmatrix} = \begin{pmatrix} A & 0 \\ V^T & I \end{pmatrix} \cdot \begin{pmatrix} I & -A^{-1}U \\ 0 & I + V^T A^{-1} U \end{pmatrix}$ and by the Schur complement formula [HJ12, Section 0.8.5], $\det \begin{pmatrix} A & -U \\ V^T & I \end{pmatrix} = \det I \det \left(A + UV^T\right)$. □

Proposition 15. *Let $M = ZBZ^T$, where $B = \begin{pmatrix} a & b \\ b & a \end{pmatrix}$ is a 2×2 matrix, and $Z = \begin{pmatrix} 1_{n/2} & 0_{n/2} \\ 0_{n/2} & 1_{n/2} \end{pmatrix}$ is an $n \times 2$ matrix. Let m be an even number. We denote by $P_{\mathcal{L}}$ the $n \times n$ diagonal matrix whose first $\frac{m}{2}$ and last $\frac{m}{2}$ diagonal elements are ones, all other elements being zeros. Then,* $\det\left(tI_n + \lambda P_{\mathcal{L}} - M\right) = t^{n-m-2}(t+\lambda)^{m-2}(t - t_1^+)(t - t_1^-)(t - t_2^+)(t - t_2^-)$ *with*

$$t_1^{\pm} = \frac{1}{2}\left(\frac{n}{2}(a+b) - \lambda \pm \sqrt{\left(\lambda + \frac{n}{2}(a+b)\right)^2 - 2(a+b)\lambda m}\right),$$
$$t_2^{\pm} = \frac{1}{2}\left(\frac{n}{2}(a-b) - \lambda \pm \sqrt{\left(\lambda + \frac{n}{2}(a-b)\right)^2 - 2(a-b)\lambda m}\right).$$

Proof. For now, assume that $t \neq -\lambda$ and $t \neq 0$. Then, $tI_n + \lambda P_{\mathcal{L}}$ is invertible, and by Lemma 27,

$$\begin{aligned} \det\left(tI_n + \lambda P_{\mathcal{L}} - M\right) &= \det(tI_n + \lambda P_{\mathcal{L}}) \det\left(I_2 + Z^T (tI_n + \lambda P_{\mathcal{L}})^{-1}(-ZB)\right) \\ &= (t+\lambda)^m t^{n-m} \det\left(I_2 - Z^T (tI_n + \lambda P_{\mathcal{L}})^{-1} ZB\right). \end{aligned} \tag{5.4.4}$$

Moreover,

$$\left(tI_n + \lambda P_{\mathcal{L}}\right)^{-1} = \frac{1}{t}(I_n - P_{\mathcal{L}}) + \frac{1}{t+\lambda} P_{\mathcal{L}} = \frac{1}{t} I_n - \frac{\lambda}{t(t+\lambda)} P_{\mathcal{L}}.$$

Therefore, we can write

$$Z^T\left(tI_n + \lambda P_{\mathcal{L}}\right)^{-1} ZB = \frac{1}{t} Z^T ZB - \frac{\lambda}{t(t+\lambda)} Z^T P_{\mathcal{L}} ZB = \frac{1}{t}\frac{n}{2} B - \frac{\lambda}{t(t+\lambda)}\frac{m}{2} B = xB,$$

where $x := \frac{n}{2}\frac{1}{t(t+\lambda)}\left(t + \lambda\left(1 - \frac{m}{n}\right)\right)$. Thus, a direct computation of the determinant gives

$$\det\left(I_2 - Z^T\left(tI_n + \lambda P_{\mathcal{L}}\right)^{-1} ZB\right) = \left(1 - x(a+b)\right)\left(1 - x(a-b)\right).$$

Going back to equation (5.4.4), we can write

$$\det\Big(tI_n + \lambda P_{\mathcal{L}} - M\Big) \;=\; (t+\lambda)^{m-2}t^{n-m-2}P_1(t)P_2(t), \tag{5.4.5}$$

with $P_1(t) = t(t+\lambda) - \frac{n}{2}(a+b)\big(t+\lambda(1-\frac{m}{n})\big)$ and $P_2(t) = t(t+\lambda) - \frac{n}{2}(a-b)\big(t+\lambda(1-\frac{m}{n})\big)$. Since $t \in \mathbb{R} \mapsto \det(tI_n + \lambda P_{\mathcal{L}} - M)$ is continuous (even analytic), expression (5.4.5) is also valid for $t = 0$ and $t = -\lambda$ [AFH13]. We end the proof by observing that

$$P_1(t) \;=\; (t - t_1^+)(t - t_1^-) \qquad \text{and} \qquad P_2(t) \;=\; (t - t_2^+)(t - t_2^-),$$

where $t_1^\pm$ and $t_2^\pm$ are defined in the proposition's statement. □

Corollary 4. *Let A be the adjacency matrix of a DC-SBM with $p_{\text{in}} > p_{\text{out}} > 0$, and s be the oracle information. Let $\lambda, \tau > 0$, and $\bar{d}_\tau = \frac{n}{2}(p_{\text{in}} + p_{\text{out}}) - n\tau$, $\bar{\alpha} = \frac{n}{2}(p_{\text{in}} - p_{\text{out}})$. Let $A_\tau := A - \tau 1_n 1_n^T$ and $P_{\mathcal{L}}$ be the diagonal matrix whose element $(P_{\mathcal{L}})_{ii}$ is 1 if $s_i \neq 0$, and 0 otherwise. Then, the spectrum of $\mathbb{E}\tilde{\mathcal{L}} = -\mathbb{E}A_\tau + \lambda\mathcal{P} - \gamma I_n$ is $\big\{-\gamma - t_1^\pm; -\gamma - t_2^\pm; -\gamma; -\gamma + \lambda; 0\big\}$, where*

$$t_1^\pm \;=\; \frac{1}{2}\left(\bar{d}_\tau - \lambda \pm \sqrt{\big(\lambda + \bar{d}_\tau\big)^2 - 4\bar{d}_\tau\lambda\,(\eta_1 + \eta_0)}\right),$$
$$t_2^\pm \;=\; \frac{1}{2}\left(\bar{\alpha} - \lambda \pm \sqrt{\Big(\lambda + \bar{\alpha}\Big)^2 - 4\bar{\alpha}\lambda\,(\eta_1 + \eta_0)}\right).$$

Proof. Let $M = \begin{pmatrix} p_{\text{in}} - \tau & p_{\text{out}} - \tau \\ p_{\text{out}} - \tau & p_{\text{in}} - \tau \end{pmatrix}$ and $Z = \begin{pmatrix} 1_{n/2} & 0_{n/2} \\ 0_{n/2} & 1_{n/2} \end{pmatrix}$. Then, we notice that $\mathbb{E}A_\tau = ZMZ^T$ and we can apply Proposition 15 to compute the characteristic polynomial of $\mathbb{E}\tilde{\mathcal{L}}$. For $x \in \mathbb{R}$, $\det\Big(\mathbb{E}\tilde{\mathcal{L}} - xI_n\Big) \;=\; \det\Big((-\gamma - x)I_n - \mathbb{E}A_\tau + \lambda\mathcal{P}\Big)$, whose roots are $-\gamma - t_1^\pm$, $-\gamma - t_2^\pm$, $-\gamma$, and $-\gamma + \lambda$. □

Estimation of $\bar{\gamma}_*$

Lemma 28. *Let $\bar{\gamma}_*$ be the solution of Equation (5.2.6) for the mean-field model. Then,*

$$-\bar{\alpha}(1 - 2\eta_0) \;\leq\; \bar{\gamma}_* \;\leq\; -\bar{\alpha}.$$

Proof. For $\lambda \geq 0$, we denote by $(\bar{x}_\lambda, \bar{\gamma}_*(\lambda))$ the solution of the system (5.2.4) on a mean-field DC-SBM. The proof is in two steps. First, let us show that $\bar{\gamma}_*(0) = -\bar{\alpha}$ and $\bar{\gamma}_*(\infty) = -\bar{\alpha}(1 - 2\eta_0)$. For $\lambda = 0$, the constrained linear system (5.2.4) reduces to an eigenvector problem, and hence $\bar{\gamma}_*(0)$ equals $-\alpha$, the smallest eigenvalue of $-\mathbb{E}A_\tau$. Moreover, when $\lambda = \infty$, the hard constraint $x_\ell = \bar{s}_\ell$ is enforced, and the system (5.2.4) becomes

$$\begin{cases} (-\mathbb{E}A_\tau - \bar{\gamma}_*(\infty)I_n)_{uu}\bar{x}_u &= (\mathbb{E}A_\tau)_{u\ell}\bar{s}_\ell \\ \bar{x}_u^T\bar{x}_u &= n(1 - \eta_0 - \eta_1) \end{cases}$$

and we verify by hand that $\bar{\gamma}_*(\infty) = -\bar{\alpha}(1-2\eta_0)$ together with $\bar{x}_u = Z_u$ is indeed the solution.

Second, if we let $C_\lambda(x) = -x^T\mathbb{E}A_\tau x + \lambda(\bar{s}-\mathcal{P}x)^T(\bar{s}-\mathcal{P}x)$ be the cost function minimised in (5.2.1), then from equation (5.2.4) we have $\bar{\gamma}_*(\lambda_1) - \bar{\gamma}_*(\lambda_2) = C_{\lambda_1}(\bar{x}_1) - C_{\lambda_2}(\bar{x}_2) + \lambda_1 \bar{x}_1^T\bar{s} - \lambda_2\bar{x}_2^T\bar{s}$. Since $\lambda \mapsto C_\lambda(x)$ is increasing, then $\lambda_1 \leq \lambda_2$ implies $C_{\lambda_1}(\bar{x}_1) \leq C_{\lambda_2}(\bar{x}_2)$. Since $\bar{x}_\lambda^T\bar{s} \geq 0$ (if it was not the case, then $C_\lambda(-\bar{x}_\lambda) \leq C_\lambda(\bar{x}_\lambda)$, and hence $\bar{x}_\lambda \neq \arg\min_{x\in\mathbb{R}^n} C_\lambda(x)$), we can conclude that $\bar{\gamma}_*(0) \leq \bar{\gamma}_*(\lambda)$ and that $\bar{\gamma}_*(\lambda) \leq \bar{\gamma}_*(\infty)$. □

Concentration of γ_*

Proposition 16. *Let γ_* and $\bar{\gamma}_*$ be the solutions of Equation (5.2.4) for a DC-SBM and the mean-field DC-SBM, respectively. Then*

$$|\gamma_* - \bar{\gamma}_*| \leq \left(1 + \frac{27\,(\bar{\alpha}+\lambda)^3}{\sqrt{2}\sqrt{\eta_1+\eta_0}(\eta_1-\eta_0)\bar{\alpha}^2\lambda}\right)\sqrt{\bar{d}}.$$

Proof. The gradient with respect to $(\bar{\delta}_1, ..., \bar{\delta}_n, \bar{b}_1, ..., \bar{b}_n, \gamma)$ of the left-hand-side of Equation (5.2.6) is equal to

$$2\sum_{i=1}^n \frac{\bar{b}_i}{\bar{\delta}_i - \bar{\gamma}}\left[\frac{\Delta b_i}{\bar{\delta}_i - \bar{\gamma}_*} - \frac{\bar{b}_i\Delta\delta_i}{(\bar{\delta}_i-\bar{\gamma}_*)^2} + \frac{\bar{b}_i\Delta\gamma}{(\bar{\delta}_i-\bar{\gamma}_*)^2}\right].$$

Thus, we have

$$\Delta\gamma\sum_{i=1}^n \frac{\bar{b}_i^2}{(\bar{\delta}_i-\bar{\gamma}_*)^3} = \sum_{i=1}^n \frac{\bar{b}_i^2}{(\bar{\delta}_i-\bar{\gamma}_*)^3}\Delta\delta_i - \sum_{i=1}^n \frac{\bar{b}_i}{(\bar{\delta}_i-\bar{\gamma}_*)^2}\Delta b_i + o\left(\Delta\delta_i, \Delta b_i\right).$$

Firstly, we see that for all $i \in [n]$, $\Delta\delta_i = \left|\delta_i - \bar{\delta}_i\right| \leq \|A - \mathbb{E}A\| \leq \bar{d}$ by the concentration of the adjacency matrix of a DC-SBM graph. Therefore, using this fact and $\bar{\gamma}_* \leq \bar{\delta}_1 \leq \bar{\delta}_2 \leq \cdots \leq \bar{\delta}_n$,

$$\begin{aligned}\Delta\gamma = |\gamma_* - \bar{\gamma}_*| &\leq \max_i\left|\delta_i - \bar{\delta}_i\right| + \frac{\max_i \frac{1}{(\bar{\delta}_i-\bar{\gamma}_*)^2}}{\min_i \frac{1}{(\bar{\delta}_i-\bar{\gamma}_*)^3}}\frac{\sum_i |\bar{b}_i|\cdot|b_i-\bar{b}_i|}{\sum_i \bar{b}_i^2} \\ &\leq \sqrt{\bar{d}} + \frac{\max_i\left(\bar{\delta}_i-\bar{\gamma}_*\right)^3}{\min_i\left(\bar{\delta}_i-\bar{\gamma}_*\right)^2}\frac{\sum_i |\bar{b}_i|\cdot|b_i-\bar{b}_i|}{\sum_i \bar{b}_i^2}.\end{aligned}$$

We notice that $\min_i|\bar{\delta}_i - \bar{\gamma}_*| = \bar{\delta}_1 - \bar{\gamma}_*$. By using Lemma 28 and the expression of $\bar{\delta}_1$ given in Corollary 4, we have

$$\min_i|\bar{\delta}_i - \bar{\gamma}_*| \geq \bar{\alpha} + \lambda.$$

Similarly, $\max_i|\bar{\delta}_i - \bar{\gamma}_*| = \bar{\delta}_n - \bar{\gamma}_* = \bar{\delta}_n - \bar{\delta}_1 + \bar{\delta}_1 - \bar{\gamma}_*$. Corollary 4 implies $\bar{\delta}_n = \lambda$ and $\bar{\delta}_1 = \frac{1}{2}\left(\lambda - \bar{\alpha} - \sqrt{(\lambda+\bar{\alpha})^2 - 4\bar{\alpha}\lambda(\eta_0+\eta_1)}\right)$, thus $\bar{\delta}_n - \bar{\delta}_1 \leq \bar{\alpha}+\lambda$. Hence, using Lemma 28,

$$\max_i|\bar{\delta}_i - \bar{\gamma}_*| \leq \frac{3}{2}(\bar{\alpha}+\lambda).$$

Therefore, we have

$$|\gamma_* - \bar{\gamma}_*| \;\leq\; \sqrt{\bar{d}} + \frac{27}{8}(\bar{\alpha}+\lambda)\cdot\frac{\sum_i |\bar{b}_i|\cdot|b_i-\bar{b}_i|}{\sum_i \bar{b}_i^2}. \tag{5.4.6}$$

The term $\frac{\sum_i |\bar{b}_i|\cdot|b_i-\bar{b}_i|}{\sum_i \bar{b}_i^2}$ can be bounded as follow. Let $\mathcal{I} = \{i \in [n] \colon \bar{b}_i \neq 0\}$. Then

$$\sum_i |\bar{b}_i|\cdot|b_i-\bar{b}_i| \;\leq\; \max_{i\in\mathcal{I}} |b_i-\bar{b}_i| \cdot \sum_{i\in\mathcal{I}} |\bar{b}_i|\,.$$

Combining the Cauchy–Schwarz inequality

$$\left|b_i-\bar{b}_i\right| \;=\; \lambda\left|(Q_{\cdot i}-\bar{Q}_{\cdot i})^T\bar{s}\right| \;\leq\; \lambda\left\|Q_{\cdot i}-\bar{Q}_{\cdot i}\right\|_2\cdot\|\bar{s}\|,$$

with the Davis-Kahan theorem [YWS15]

$$\left\|Q_{\cdot i}-\bar{Q}_{\cdot i}\right\|_2 \;\leq\; \frac{2^{3/2}\,\|A-\mathbb{E}A\|}{\min\left\{\bar{\delta}_i-\bar{\delta}_{i-1},\bar{\delta}_{i+1}-\bar{\delta}_i\right\}},$$

$\|\bar{s}\| = \sqrt{(\eta_0+\eta_1)n}$, and the concentration of A towards $\mathbb{E}A$, yields

$$\max_{i\in\mathcal{I}} |b_i-\bar{b}_i| \;\leq\; \frac{\lambda\sqrt{(\eta_0+\eta_1)n}}{\min_{i\in\mathcal{I}}\left\{\bar{\delta}_i-\bar{\delta}_{i-1},\bar{\delta}_{i+1}-\bar{\delta}_i\right\}}\cdot 2^{3/2}\sqrt{\bar{d}}.$$

Using Lemma 29, we see that $\mathcal{I} = \{i \in [n] : \delta_i \notin \{0, t_1^-\}\}$. Combining it with Corollary 4, gives

$$\begin{aligned}
\min_{i\in\mathcal{I}}\left\{\bar{\delta}_i-\bar{\delta}_{i-1},\bar{\delta}_{i+1}-\bar{\delta}_i\right\} \;&=\; \lambda + t_2^+ \\
&=\; \frac{\alpha+\lambda}{2}\left(1-\sqrt{1-4\frac{\alpha\lambda}{(\alpha+\lambda)^2}(\eta_0+\eta_1)}\right) \\
&\geq\; \frac{\alpha\lambda}{\alpha+\lambda}(\eta_0+\eta_1),
\end{aligned}$$

where we used $\sqrt{1-x} \leq 1 - x/2$. Therefore,

$$\sum_i \left|\bar{b}_i\right|\cdot\left|b_i-\bar{b}_i\right| \;\leq\; 2^{3/2}\sqrt{\frac{n\bar{d}}{\eta_0+\eta_1}}\cdot\frac{\alpha+\lambda}{\alpha}\cdot\sum_i\left|\bar{b}_i\right|.$$

By noticing that $\sum_i \bar{b}_i^2 \geq \left(\sum_i |\bar{b}_i|\right)^2 \geq |\bar{b}_1|\cdot\sum_i|\bar{b}_i| \geq \sqrt{n}\,\frac{\eta_1-\eta_0}{2}\frac{\bar{\alpha}\lambda}{\lambda+\bar{\alpha}}\sum_i|\bar{b}_i|$ where we used $\bar{b}_1 \geq \sqrt{n}\frac{\eta_1-\eta_0}{2}\frac{\bar{\alpha}\lambda}{\lambda+\bar{\alpha}}$ (Lemma 29), we have

$$\frac{\sum_i \left|\bar{b}_i\right|\cdot\left|b_i-\bar{b}_i\right|}{\sum_i \bar{b}_i^2} \;\leq\; \frac{2^{5/2}}{(\eta_1-\eta_0)\sqrt{\eta_1+\eta_0}}\frac{(\alpha+\lambda)^2}{\alpha^2\lambda}\sqrt{\bar{d}}.$$

Going back to inequality (5.4.6), implies that $|\gamma_* - \bar{\gamma}_*| \;\leq\; \left(1+\frac{27(\bar{\alpha}+\lambda)^3}{\sqrt{2}\sqrt{\eta_1+\eta_0}(\eta_1-\eta_0)\bar{\alpha}^2\lambda}\right)\sqrt{\bar{d}}$. □

Lemma 29. *Let $-\mathbb{E}A\tau + \lambda\mathcal{P} = \bar{Q}\bar{\Delta}\bar{Q}^T$, where $\bar{\Delta} = \operatorname{diag}\left(\bar{\delta}_1, \dots, \bar{\delta}_n\right)$ and $\bar{Q}^T\bar{Q} = I_n$. Denote $\bar{b} = \lambda\bar{Q}^T s$. We have $\bar{b}_1 \geq \sqrt{n}\frac{\lambda(\eta_1-\eta_0)}{2}\frac{\bar{\alpha}}{\lambda+\bar{\alpha}}$. Moreover, $\bar{b}_i = 0$ if $\bar{\delta}_i = 0$ or if $\bar{\delta}_i = -t_1^-$.*

Proof. First, from Corollary 4, $\bar{\delta}_1 = -t_2^+ = -\frac{1}{2}\left(\bar{\alpha} - \lambda + \sqrt{\left(\lambda+\bar{\alpha}\right)^2 - 4\bar{\alpha}\lambda\left(\eta_1+\eta_0\right)}\right)$. By symmetry, the i-th component of the first eigenvector $\bar{Q}_{\cdot 1}$ (associated with $\bar{\delta}_1$) is equal to

$$\begin{cases} v_1\, Z_i & \text{if } i \in [\ell], \\ v_0\, Z_i & \text{if } i \notin [\ell], \end{cases}$$

where v_1 and v_0 are to be determined. Thus, the equation $\left(-\mathbb{E}A_\tau + \lambda\mathcal{P}\right)\bar{Q}_{\cdot 1} = \bar{\delta}_1\bar{Q}_{\cdot 1}$ leads to

$$\begin{cases} \bar{\alpha}\left((\eta_1+\eta_0)v_1 + (1-\eta_1-\eta_0)v_0\right) & = -t_2^+ v_0 \\ \bar{\alpha}\left((\eta_1+\eta_0)v_1 + (1-\eta_1-\eta_0)v_0\right) + \lambda v_1 & = -t_2^+ v_1, \end{cases}$$

which, given the norm constraint $\|v\|_2 = 1$, yields

$$\begin{cases} v_1 & = \dfrac{1}{\sqrt{n}}\dfrac{t_2^+}{\sqrt{(\eta_1+\eta_0)\left(t_2^+\right)^2 + (1-\eta_1-\eta_0)\left(t_2^+ + \lambda\right)^2}}, \\ v_0 & = \dfrac{1}{\sqrt{n}}\dfrac{+t_2^+ + \lambda}{\sqrt{(\eta_1+\eta_0)\left(t_2^+\right)^2 + (1-\eta_1-\eta_0)\left(t_2^+ + \lambda\right)^2}}. \end{cases}$$

Since $\bar{b}_1 = \lambda v^T\bar{s} = \lambda(\eta_1-\eta_0)nv_1$, we have

$$\frac{\bar{b}_1}{\sqrt{n}} = \lambda(\eta_1-\eta_0)\frac{t_2^+}{\sqrt{(\eta_1+\eta_0)\left(t_2^+\right)^2 + (1-\eta_1-\eta_0)\left(t_2^+ + \lambda\right)^2}}.$$

The proof ends by noticing that $t_2^+ \geq \frac{\bar{\alpha}}{2}$ and $t_2^+ \leq \bar{\alpha}$. Indeed,

$$\begin{aligned} \frac{\bar{b}_1}{\sqrt{n}} &\geq \lambda(\eta_1-\eta_0)\frac{\bar{\alpha}}{2\sqrt{(\eta_1+\eta_0)\bar{\alpha}^2 + (1-\eta_1-\eta_0)(\bar{\alpha}+\lambda)^2}} \\ &\geq \frac{\lambda(\eta_1-\eta_0)}{2}\frac{\bar{\alpha}}{(\bar{\alpha}+\lambda)\sqrt{(\eta_1+\eta_0)\left(\frac{\bar{\alpha}}{\bar{\alpha}+\lambda}\right)^2 + 1 - \eta_1 - \eta_0}} \\ &\geq \frac{\lambda(\eta_1-\eta_0)}{2}\frac{\bar{\alpha}}{\lambda+\bar{\alpha}}. \end{aligned}$$

This proves the first claim of the lemma.

Similarly, by symmetry the i-th component of the eigenvector v' associated with $-t_1^-$ equals v'_ℓ if $i \in \ell$, and v'_u otherwise, and therefore $(v')^T s = 0$.

Finally, let $I_0 := \{i \in [n] : \bar{\delta}_i = 0\}$. By Corollary 4, we have $|I_0| = n(1-\eta_1-\eta_0) - 2$. Since 0 is also eigenvalue of order $n(1-\eta_0-\eta_1)-2$ of the extracted sub-matrix $\left(-\mathbb{E}A_\tau + \lambda\mathcal{P}\right)_{u,u} = \left(-\mathbb{E}A_\tau\right)_{u,u}$, we have for all $k \in I_0$, $\bar{Q}_{ik} = 0$ for every $i \in [n]$. Therefore, for $k \in I_0$, $b_k = \lambda\bar{Q}_{\cdot k}^T s = 0$. □

5.4.3 Mean-field solution

In this section, we calculate the solution $\bar{x}$ to the mean-field model and deduce from it the conditions to recover the clusters.

Proposition 17. *Suppose that $\tau > p_{\text{out}}$. Then the solution of Equation (5.2.5) on the mean-field DC-SBM is the vector $\bar{x}$ whose element $\bar{x}_i$ is given by*

$$\bar{x}_i = \begin{cases} C\left(-1 + (\eta_1 - \eta_0)\bar{\alpha}B\right) Z_i, & \text{if } i \in \ell \text{ and } s_i \neq Z_i, \\ C\left(1 + (\eta_1 - \eta_0)\bar{\alpha}B\right) Z_i, & \text{if } i \in \ell \text{ and } s_i = Z_i, \\ \frac{-\bar{\alpha}C}{\bar{\alpha}(1-\eta_1-\eta_0)+\bar{\gamma}_*}(\eta_1 - \eta_0)\left(1 + (\eta_1 + \eta_0)\bar{\alpha}B\right) Z_i, & \text{if } i \notin \ell, \end{cases}$$

where $\bar{\alpha} = \frac{n}{2}(p_{\text{in}} - p_{\text{out}})$, $B = \frac{\bar{\alpha}\bar{\gamma}_*}{\lambda\bar{\alpha}(1-\eta_1-\eta_0)+\bar{\gamma}_*(\lambda-\bar{\alpha}-\bar{\gamma}_*)}$ and $C = \frac{\lambda}{\lambda-\bar{\gamma}_*}$.

Proof. Let $\bar{x}$ be a solution of Equation (5.2.5). By symmetry, we have

$$\bar{x}_i = \begin{cases} x_t\, Z_i, & \text{if } i \in [\ell] \text{ and } \bar{s}_i = Z_i, \\ x_f\, Z_i, & \text{if } i \in [\ell] \text{ and } \bar{s}_i = -Z_i, \\ x_0\, Z_i, & \text{if } i \notin [\ell], \end{cases}$$

where x_t, x_f and x_0 are unknowns to be determined. Since for every $i \in [n]$

$$\left(\mathbb{E}A_\tau \bar{x}\right)_i = \bar{\alpha}\left(x_0(1 - \eta_1 - \eta_0) + x_t\eta_1 + x_f\eta_0\right),$$

the linear system composed of the equations $\left((-\mathbb{E}A_\tau + \lambda\mathcal{P} - \bar{\gamma}_* I_n)\,\bar{x}\right)_i = \lambda s_i$ for all $i \in [n]$ leads to the system

$$\begin{cases} -\bar{\alpha}\left((1 - \eta_1 - \eta_0)x_0 + x_t\eta_1 + x_f\eta_0\right) - \bar{\gamma}_* x_0 &= 0, \\ -\bar{\alpha}\left((1 - \eta_1 - \eta_0)x_0 + x_t\eta_1 + x_f\eta_0\right) - \bar{\gamma}_* x_t + \lambda x_t &= \lambda, \\ -\bar{\alpha}\left((1 - \eta_1 - \eta_0)x_0 + x_t\eta_1 + x_f\eta_0\right) - \bar{\gamma}_* x_f + \lambda x_f &= -\lambda. \end{cases}$$

The rows of the latter system correspond to a node unlabelled by the oracle, correctly labelled and falsely labelled, respectively. This system can be rewritten as follows:

$$\begin{cases} x_0 &= \frac{-\bar{\alpha}}{\bar{\alpha}(1-\eta_1-\eta_0)+\bar{\gamma}_*}\left(\eta_1 x_t + \eta_0 x_f\right), \\ \bar{\gamma}_* x_0 + x_t(\lambda - \bar{\gamma}_*) &= \lambda, \\ \bar{\gamma}_* x_0 + x_f(\lambda - \bar{\gamma}_*) &= -\lambda. \end{cases}$$

In particular, we have $x_t - x_f = \frac{2\lambda}{\lambda-\bar{\gamma}_*}$. By subsequently eliminating x_0 and x_t in the equation $\bar{\gamma}_* x_0 + x_f(\lambda - \bar{\gamma}_*) = -\lambda$, we find

$$\begin{aligned} x_f &= \frac{\lambda}{\lambda - \bar{\gamma}_*}\left(-1 + \frac{\bar{\alpha}\bar{\gamma}_*\left(\eta_1 - \eta_0\right)}{\lambda\bar{\alpha}(1 - \eta_1 - \eta_0) + \lambda\bar{\gamma}_* - \bar{\gamma}_*(\bar{\alpha} + \bar{\gamma}_*)}\right), \\ x_t &= \frac{\lambda}{\lambda - \bar{\gamma}_*}\left(1 + \frac{\bar{\alpha}\bar{\gamma}_*\left(\eta_1 - \eta_0\right)}{\lambda\bar{\alpha}(1 - \eta_1 - \eta_0) + \lambda\bar{\gamma}_* - \bar{\gamma}_*(\bar{\alpha} + \bar{\gamma}_*)}\right), \end{aligned}$$

and finally

$$x_0 \;=\; \frac{-\bar{\alpha}}{\bar{\alpha}(1-\eta_1-\eta_0)+\bar{\gamma}_*} \cdot \frac{\lambda}{\lambda-\bar{\gamma}_*}\left(1+\frac{\bar{\alpha}\bar{\gamma}_*\,(\eta_1+\eta_0)}{\lambda\bar{\alpha}(1-\eta_1-\eta_0)+\lambda\bar{\gamma}_*-\bar{\gamma}_*(\bar{\alpha}+\bar{\gamma}_*)}\right).$$

□

Corollary 5. *Suppose that* $\tau > p_{\text{out}}$. *Then* $\operatorname{sign}(\bar{x}_i) = \operatorname{sign}(Z_i)$ *if*

- *node* i *is not labelled by the oracle;*
- *node* i *is correctly labelled by the oracle;*
- *node* i *is mislabelled by the oracle and* $\lambda < (1-2\eta_0)\bar{\alpha}\frac{\eta_1-\eta_0}{\eta_1+\eta_0}$.

Proof. A node i is correctly classified by decision rule (5.2.3) if the sign of $\bar{x}_i$ is equal to the sign of Z_i. Using Lemma 28 in Section 5.4.2, we have $-\bar{\alpha} \le \bar{\gamma}_* \le -\bar{\alpha}(1-2\eta_0)$. Therefore, the quantities B and C in Proposition 17 verify $C \ge 0$ and $\frac{1-2\eta_0}{\lambda(\eta_0+\eta_1)} \le B \le \frac{1}{\lambda(\eta_0+\eta_1)}$. The statement then follows from the expression of $\bar{x}_i$ computed in Proposition 17. □

Chapter 6

Clustering temporal networks

This chapter presents some algorithms for clustering temporal networks. Section 6.1 introduces various baseline algorithms for special cases, while Section 6.2 is devoted to online algorithms based on likelihood. Section 6.3 studies spectral methods applied to temporal networks with fixed communities, and Section 6.4 generalizes to the case of time-varying communities.

6.1 Baseline algorithms for special cases

This section provides some baseline algorithms to recover the blocks in some particular cases, without prior knowledge of the block interaction parameters. Section 6.1.1 concerns regimes with $N = O(1)$ and $T \gg 1$. An algorithm based on parameter estimations is proposed, and showed to converge to the true community structure. Section 6.1.2 describes tailored-made algorithms for a specific model instance with static intra-block interactions and uncorrelated inter-block noise.

6.1.1 Clustering using empirical transition rates

We consider a homogeneous Markov SBM, as defined in Section 3.4. The intra- and inter-block interactions are Markovian, and we denote by P and Q the respective transition probability matrices. We study the situation where the number of snapshots T goes to infinity while the number of nodes N remains bounded. The main idea is to use the ergodicity of the Markov chains to estimate the parameters using standard techniques, and then perform inference. For now, we will assume that the interaction parameters P and Q are known, but K is unknown. We refer to Remark 14 when P and Q are unknown as well.

Let $n_{ab}(i,j)$ be the observed number of transitions $a \to b$ in the interaction pattern between nodes i and j, and let $n_a(i,j) = \sum_b n_{ab}(i,j)$. Let $P(i,j)$ be the 2-by-2 matrix transition probabilities for the evolution of the pattern interaction between a node pair $\{i,j\}$. By the law of large numbers (for stationary and ergodic random processes), the empirical transition probabilities

$$\hat{P}_{ab}(i,j) \;=\; \frac{n_{ab}(i,j)}{n_a(i,j)} \tag{6.1.1}$$

are with high probability close to $P_{ab}(i,j)$ for $T \gg 1$.

Once all $P_{ab}(i,j)$ are known with a good precision, we can use our knowledge of P, Q to distinguish whether nodes i and j are in the same block or not, and use this data to construct a similarity graph on the set of nodes. This leads to Algorithm 4 which does not require a priori knowledge about the number of blocks, but instead estimates it as a byproduct. Note that this algorithm is tailor-made for homogeneous interaction tensors.

Proposition 18. *Consider a homogeneous Markov SBM with N nodes, K communities and T snapshots. Assume that N is fixed, and the interaction pattern probabilities f, g are known and ergodic. Then with high probability Algorithm 4 correctly classifies every node when T goes to infinity, as long as the evolution is not static and $P \neq Q$.*

Proof. For $a, b \in \{0,1\}$, let $n_a(i,j) = \sum_b n_{ab}(i,j)$ where $n_{ab}(i,j)$ counts the observed number of transitions $a \to b$ between a node pair ij. From [Bil61, Theorem 3.1 and Formula 3.13], the distribution of the random variable $\xi_{ab}(i,j) := \frac{n_{ab}(i,j) - n_a(i,j)P_{ab}(i,j)}{\sqrt{n_a(i,j)}}$ tends to a normal distribution with the zero mean and finite variance given by $\lambda_{(ab),(cd)} := \delta_{ac}\big(\delta_{bd}P_{ab}(i,j) -$

Algorithm 4: Clustering by empirical transition rates.

Input: Observed interaction array (X_{ij}^t); transition probability matrices P, Q.
Output: Estimated node labelling $\hat{\sigma} = (\hat{\sigma}_1, \dots, \hat{\sigma}_N)$; estimated number of communities $\hat{K}$.

Set $V \leftarrow \{1, \dots, N\}$ and $E \leftarrow \emptyset$.
for *all unordered node pairs* ij **do**
    Compute $\hat{P}_{ab}(i,j)$ for $a, b = 0, 1$
    if $|\hat{P}_{ab}(i,j) - P_{ab}| \leq \frac{1}{2}|P_{ab} - Q_{ab}|$ *for some* a, b **then**
        Set $E \leftarrow E \cup \{ij\}$.
Compute $\mathcal{C} \leftarrow$ set of connected components in $G = (V, E)$
Let $\hat{K} \leftarrow |\mathcal{C}|$ and $(C_1, \dots, C_{\hat{K}}) \leftarrow$ members of $\mathcal{C}$ listed in arbitrary order.
for $i = 1, \dots, N$ **do**
    $\hat{\sigma}_i \leftarrow$ unique k for which $C_k \ni i$.

$P_{ab}(i,j)P_{a,d}(i,j)\big)$. Therefore, for any $\alpha > 0$,

$$\mathbb{P}\Big(|\widehat{P}_{ab}(i,j) - P_{ab}(i,j)| \geq \alpha\Big) \;=\; \mathbb{P}\Big(|\xi_{ab}(i,j)| \geq \alpha\sqrt{n_a(i,j)}\Big) \tag{6.1.2}$$

and this quantity goes to zero as T goes to infinity.

From model identifiability, $P \neq Q$. Therefore, w.l.o.g. we can assume $P_{01} \neq Q_{01}$, and let α such that $0 < \alpha < \dfrac{|P_{01} - Q_{01}|}{2}$. The nodes i and j are predicted to be in the same community if $\widehat{P}_{ij}(0,1) > \dfrac{P_{01} + Q_{01}}{2}$, and the probability of making an error is

$$\mathbb{P}\left(\left|\widehat{P}_{01}(i,j) - P_{01}(i,j)\right| \geq \alpha\right).$$

By the union bound, the probability that all nodes are correctly classified is bounded by

$$\frac{N(N-1)}{2} \max_{ij} \mathbb{P}\left(\left|\widehat{P}_{01}(i,j) - P_{01}(i,j)\right| \geq \alpha\right),$$

where the maximum is taken over all nodes pair ij. By equation (6.1.2), for all node pairs ij we have $\mathbb{P}\Big(|\widehat{P}_{01}(i,j) - P_{01}(i,j)| \geq \alpha\Big) \to 0$. Therefore, all nodes are a.s. correctly classified as $T \to \infty$. □

Remark 14. If P and Q are unknown, we can add a step where the estimated transition matrices $(\hat{P}_{ij})$ are clustered into two classes (for example using k-means).

6.1.2 Algorithms for static and deterministic inter-block patterns

This section investigates special data tensors where the intra-block interactions are static and deterministic, and the inter-block interactions are considered as (non-static) random noise. For example, a Markov SBM with static intra-block interactions corresponds to $P = I_2$. For such data, we will first make two simple observations that greatly help recovering the underlying block structure. Those observations lead to two different algorithms, and we will study their performance in Section 9.

Description of the algorithms based on two simple observations

When the intra-block interactions are static, the two following observations hold.

Observation 1. If nodes i and j interact at time t but not at time $t+1$ (or vice versa), then i and j do not belong to the same block.

Observation 2. If nodes i and j interact at every time step, then i and j probably belong to the same block.

Observation 2 suggests a very simple and extremely fast clustering method (Algorithm 5) which tracks persistent interactions and disregards other information. Persistent interactions can be represented as an intersection graph $G = \cap_t G^t$, where G^t is the graph with adjacency matrix A^t. By noting that G can be computed by performing $O(\log T)$ graph intersections of complexity $O(\Delta_{\max} N)$, and that a breadth-first search finds the connected components in $O(N)$ time, we see that Algorithm 5 runs in $O(\Delta_{\max} N \log T)$ time, where $\Delta_{\max} = \max_t \max_i \sum_j |A^t_{ij}|$ is the maximum degree of the graphs G^t.

Algorithm 5: Best friends forever

Input: Observed interaction tensor (A^t_{ij})

Output: Estimated node labelling $\hat{\sigma} = (\hat{\sigma}_1, \dots, \hat{\sigma}_N)$; estimated number of communities $\hat{K}$.

1

2 Set $V \leftarrow \{1, \dots, N\}$.

3 Compute $E_T \leftarrow \cap_{t=1}^T E^t$ where $E^t = \{ij : A^t_{ij} = 1\}$

4 Compute $\mathcal{C} \leftarrow$ set of connected components in $G_T = (V, E_T)$ and set $\hat{K} \leftarrow$ number of members in $\mathcal{C}$ of size larger than $N^{1/2}$, and $(C_1, \dots, C_{\hat{K}}) \leftarrow$ list of $\hat{K}$ largest members in $\mathcal{C}$ in arbitrary order.

5 Set $V_1 \leftarrow \cup_{k=1}^{\hat{K}} C_k$.

6 For $i \in V_1$, set $\hat{\sigma}_i \leftarrow$ unique k for which $C_k \ni i$.

7 For $i \in V \setminus V_1$, set $\hat{\sigma}_i \leftarrow$ arbitrarily value $k \in \{1, \dots, \hat{K}\}$.

Similarly, we propose a clustering method based on Observation 1. We call *enemies* two nodes i and j such that there is a change in the interaction pattern between i and j. Then we

can group nodes that share a common enemy. Indeed, if $K = 2$, the fact that node i is enemy with h, and h is also enemy with j means that nodes i and j belong to the same cluster. This *enemies of my enemies are my friends* procedure leads to Algorithm 6.

Algorithm 6: Enemies of my enemy.

Input: Observed interaction tensor (A^t_{ij}).
Output: Estimated node labelling $\hat{\sigma} = (\hat{\sigma}_1, \dots, \hat{\sigma}_N)$.

Compute $E_\cap \leftarrow \cap_t E^t$ and $E_\cup \leftarrow \cup_t E^t$ where $E^t = \{ij : A^t_{ij} = 1\}$.
Compute $E' = E_\cup \setminus E_\cap$.
Set $V \leftarrow \{1, \dots, N\}$.
Set $G' \leftarrow (V, E')$.
Set $G'' = (V, E'')$ where $ij \in E''$ iff there is a 2-path $i \to h \to j$ in G'.
Compute $\mathcal{C} \leftarrow$ set of connected components in G'' and set $\hat{K} \leftarrow |\mathcal{C}|$ and $(C_1, \dots, C_{\hat{K}}) \leftarrow$ members of $\mathcal{C}$ listed in arbitrary order.
for $i = 1, \dots, N$ **do**
    $\hat{\sigma}_i \leftarrow$ unique k for which $C_k \ni i$.

Remark 15. The above description for Algorithm 6 runs in $O(\Delta_{\max} NT)$, where $\Delta_{\max}$ is the maximal degree over all single layers. A faster, but less transparent, implementation is possible, by first computing the union graph. Then, two nodes are marked as enemies if the weight between them in the union graph belongs to the interval $[1, T-1]$. This reduces the time complexity to $O(\Delta_{\max} N \log T)$.

Performance guarantees for Algorithms 5 and 6

Proposition 19 states the performance guarantees for Algorithm 5.

Proposition 19. *Consider a SBM with T snapshots and $K = O(1)$ blocks of size $N_1, \dots, N_K$. Assume that $N_k \asymp N$ for all k, and that*

$$\forall k \neq \ell \ : \ N^2 \max_{1 \le k < \ell \le K} f_{k\ell}(1, \dots, 1) \ll 1. \tag{6.1.3}$$

Then Algorithm 5 is consistent if

$$\forall k \in [K] \ : \ N_k f_{kk}(1, \dots, 1) \gg 1, \tag{6.1.4}$$

and is strongly consistent if

$$\forall k \in [K] \ : \ N_k f_{kk}(1, \dots, 1) \ \ge \ (1 + \Omega(1)) \log(K N_k). \tag{6.1.5}$$

Remark 16. Condition (6.1.3) ensures that the number of nodes in different community interacting at every time step remains small, making Observation 2 meaningful. The extra conditions (6.1.5) and (6.1.4) ensures that in each community, there is enough node pairs interacting at all time step.

Proof of Proposition 19. Assume that the true block membership structure σ contains K blocks $C_1, \dots, C_K$ of sizes $N_k = |C_k|$. Let G^t be the graph on node set $V = [N]$ and edge set $E^t = \{ij : A^t_{ij} = 1\}$. Let $G_T = \cap_t G^t$ be the intersection graph. We denote by $p^T_{k\ell} = f_{k\ell}(\underbrace{1, \dots, 1}_{T})$ the probability of a persistent interaction of duration T between a pair of nodes in blocks k and ℓ.

(a) Conditions for strong consistency. Algorithm 5 returns exactly the correct block membership structure if and only if each C_k forms a connected set of nodes in G_T, and for all blocks $k \neq \ell$, there are no links between C_k and C_ℓ in G_T.

The probability that the intersection graph G_T contains a link between some distinct blocks is bounded by

$$\sum_{1 \leq k < \ell \leq K} N_k N_\ell \, p^T_{k\ell}.$$

Hence, by the union bound, the probability that Algorithm 5 does not give exact recovery is bounded by

$$\sum_{k \in [K]} (1 - c^T_k) + \sum_{1 \leq k < \ell \leq K} N_k N_\ell \, p^T_{k\ell},$$

where c^T_k is the probability that the subgraph of G_T induced by C_k is connected. By classical results about Erdős–Rényi graph models [Les] we know that

$$c^T_k \;\geq\; 1 - 100 e^{-(N_k p^T_{kk} - \log N_k)},$$

whenever $N_k p^T_{kk} \geq \max\{9e, \log N_k\}$. Hence

$$\sum_{k \in [K]} (1 - c^T_k) \;\leq\; 100 \sum_{k \in [K]} e^{-(N_k p^T_{kk} - \log N_k)} \;\leq\; 100 \, e^{-\min_{k \in [K]} \log(KN_k) \left(\frac{N_k p^T_{kk}}{\log(KN_k)} - 1 \right)},$$

and this last term goes to zero under Condition (6.1.5). Moreover,

$$\sum_{1 \leq k < \ell \leq K} N_k N_\ell p^T_{k\ell} \leq \binom{N}{2} \max_{k \neq \ell} p^T_{k\ell}$$

which also goes to zero under Condition (6.1.3).

(b) Condition for consistency. We just saw that the probability that the intersection graph G_T contains a link between some distinct blocks is bounded by $\binom{N}{2} \max_{1 \leq k < \ell \leq K} p^T_{k\ell}$, and hence goes to zero if Condition (6.1.3) holds.

Let $G_T[C_k]$ be the subgraph of G_T induced by C_k. Let $\mathcal{A}_{kT}$ be the event that the largest connected component of $G_T[C_k]$ has size at least $N^{1/2}$, and all other components are smaller than $N^{1/2}$. Observe that $G_T[C_k]$ is an instance of a Bernoulli random graph with N_k nodes where all node pairs are independently linked with probability p^T_{kk}. When $N_k p^T_{kk} \gg 1$, classical Erdős–Rényi random graph theory tells that $\mathbb{P}(\mathcal{A}_{kT}) = 1 - o(1)$ for any fixed k as $N \gg 1$. For bounded $K = O(1)$ this implies that $\mathbb{P}(\cap_k \mathcal{A}_{kT}) = 1 - o(1)$.

On the event $\mathcal{A} = (\cap_k \mathcal{A}_{kT}) \cap \mathcal{B}$, the algorithm estimates $\hat{K} = K$ correctly, and (with the correct permutation), the number of misclustered nodes is at most

$$\sum_{k \in [K]} |C_k \setminus \hat{C}_{kT}| \ll 1,$$

where $\hat{C}_{kT}$ is the largest component of $G_T[C_k]$. □

The following Proposition 20 gives the guarantees of convergence of Algorithm 6 for a general temporal SBM with two communities.

Proposition 20. *Consider a dynamic SBM with $N \gg 1$ nodes and $K = 2$ blocks of sizes $N_1, N_2 \asymp N$. Assume that $\log(1/p_{11T}) + \log(1/p_{22T}) \ll N^{-2}$ and $1 - p_{12T} \gg N^{-1} \log N$, where*

$$p_{k\ell T} \;=\; f_{k\ell}(\underbrace{0, \dots, 0}_{T}) + f_{k\ell}(\underbrace{1, \dots, 1}_{T})$$

is the probability of observing a static interaction pattern of length T between any particular pair of nodes in blocks k and ℓ. Then Algorithm 6 is strongly consistent.

Proof. Denote the time-aggregated interaction tensor by $A^+_{ij} = \sum_t A^t_{ij}$. Let G' be the "enemy graph" with node set $\{1, \dots, N\}$ and adjacency matrix $A'_{ij} = 1(0 < A^+_{ij} < T)$. Let C_1, C_2 be blocks corresponding to the true labelling σ. The probability that all intra-block interactions are static is

$$p_{11T}^{\binom{N_1}{2}} p_{22T}^{\binom{N_2}{2}} \;\geq\; (p_{11T} p_{22T})^{N^2} \;\to\; 1.$$

Hence, it follows that G' is whp bipartite with respect to partition $\{C_1, C_2\}$.

Let us next analyze the probability that G' is connected. Let G'' be the graph on node set $\{1, \dots, N\}$ obtained by deleting all edges connecting pair of nodes within C_1 or within C_2. Then G'' is random bipartite graph with bipartition $\{C_1, C_2\}$ where each node pair ij with $i \in C_1$ and $j \in C_2$ is linked with probability $q = 1 - p_{12T}$, independently of other node pairs. Because blocks sizes are balanced according $N_1, N_2 \asymp N$ and $Nq \gg \log N$, it follows by applying [Sin95, Theorem 3.3] that G'' is connected with high probability. Because G'' is a subgraph of G', the same is true for G'.

We have now seen that G' is whp connected and bipartite with respect to partition $\{C_1, C_2\}$. Let $\tilde{G}$ be the graph on $[N]$, of which nodes i and j are linked if and only if there exists a 2-path in G' between i and j. Then the connected components of $\tilde{G}$ are C_1 and C_2. Hence Algorithm 6 estimates the correct block memberships on the high-probability event that G' is connected. □

6.2 Online likelihood based algorithms

Algorithms 5 and 6 are based on two Observations that rely a lot on the initial assumption of static inter-community edge, and i.i.d. evolution for the intra-community edges. If those assumptions are relaxed, both algorithms fail (see Section 6.2.3 for empirical evidence). Hence, we need to make softer decisions for clustering, e.g. based on the likelihood that node i is in cluster k.

To do so, we first derive the log-likelihood of a node i being in a cluster k given a clustering assignment of the other nodes. For homogeneous Markov SBM, we derive an online algorithm based on this likelihood.

6.2.1 Online algorithm when the model parameters are known

Let f and g be the intra- and inter-block interaction distributions of a temporal SBM. Given $A^{1:t} = (A^1, \dots, A^t)$, we define a log-likelihood ratio matrix by

$$M_{ij}^{(t)} \;=\; \log \frac{f(A_{ij}^{1:t})}{g(A_{ij}^{1:t})}. \tag{6.2.1}$$

Then the log of the probability of observing a graph sequence $A^{1:t}$ given node labelling σ equals $\frac{1}{2}\sum_i \sum_{j\neq i} M_{ij}^{(t)} \delta_{\sigma_j,\sigma_i} + \frac{1}{2}\sum_i \sum_{j\neq i} g(A_{ij}^{1:t})$. Therefore, given an assignment $\hat{\sigma}^{(t-1)}$ computed from the observation of the $t-1$ first snapshots, one can compute a new assignment $\hat{\sigma}^{(t)}$ such that node i is assigned to any block k which maximises

$$L_{i,k}^{(t)} \;=\; \sum_{j\neq i} M_{ij}^{(t)} \delta_{\hat{\sigma}_j^{(t-1)} k}.$$

This formula is interesting only if the computation of $M^{(t)}$ can be easily done from $M^{(t-1)}$. This is in particular the case for a Markov evolution. Indeed, if μ and ν are the initial probability distributions, and P, Q the transition matrices, then the cumulative log-likelihood matrices defined in Equation (6.2.1) can be computed recursively by $M^{(t)} = M^{(t-1)} + \Delta^{(t)}$ with $M_{ij}^{(1)} = \log \frac{\mu}{\nu}\left(A_{ij}^1\right)$ and $\Delta_{ij}^{(t)} = \log \frac{P}{Q}\left(A_{ij}^{t-1}, A_{ij}^t\right)$. We summarize this in Algorithm 7. Let us emphasize that this algorithm works in an online adaptive fashion.

The time complexity (worst case complexity) of Algorithm 7 is $O(KN^2T)$ plus the time complexity of the initial clustering. The space complexity is $O(N^2)$. In addition:

- Since at each time step, Δ can take only one of four values, these four different values of Δ can be precomputed and stored to avoid computing N^2T logarithms.
- The N-by-K matrix (L_{ik}) can be computed as a matrix product $L = M^0\Sigma$, where M^0 is the matrix obtained by zeroing out the diagonal of M, and Σ is the one-hot representation of $\hat{\sigma}$ such that $\Sigma_{ik} = 1$ if $\hat{\sigma}_i = k$ and zero otherwise.

Algorithm 7: Online clustering for homogeneous Markov dynamics when the block interaction parameters are known.

Input: Observed interaction tensor (A^t_{ij}); block interaction parameters μ, ν, P, Q; number of communities K; static graph clustering algorithm `algo`.
Output: Node labelling $\hat{\sigma} = (\hat{\sigma}_1, \dots, \hat{\sigma}_N) \in [N]^K$.

Initialize: Compute $\hat{\sigma} \leftarrow \texttt{algo}(A^1)$, and $M_{ij} \leftarrow \log \frac{\mu(A^1_{ij})}{\nu(A^1_{ij})}$ for $i, j = 1, \dots, N$.
for $t = 2, \dots, T$ **do**
  Compute $\Delta_{ij} \leftarrow \log \frac{P(A^{t-1}_{ij}, A^t_{ij})}{Q(A^{t-1}_{ij}, A^t_{ij})}$ for $i, j = 1, \dots, N$.
  Update $M \leftarrow M + \Delta$.
  for $i = 1, \dots, N$ **do**
    Set $L_{ik} \leftarrow \sum_{j \neq i} M_{ij}\, \delta_{\hat{\sigma}_j k}$ for $k = 1, \dots, K$.
    Set $\hat{\sigma}_i \leftarrow \arg\max_{1 \leq k \leq K} L_{ik}$.
Return: $\hat{\sigma}$

- For sparse networks the time and space complexity (average complexity) can be reduced by a factor of d/N where d is the average node degree, by neglecting the $0 \to 0$ transitions and only storing nonzero entries (similarly to what is often done for belief propagation in the static SBM [Moo17]).

6.2.2 Extension when the parameters are unknown

Algorithm 7 requires the *a priori* knowledge of the interaction parameters. This is often not the case in practice, and one has to learn the parameters during the process of recovering communities. In this section, we adapt Algorithm 7 to estimate the parameters on the fly.

An estimator of P is obtained by averaging the probabilities $\hat{P}_{ij}$ obtained using Formula 6.1.1 over the pairs of nodes predicted to belong to the same community. More precisely, after t snapshots observed ($t \geq 2$), given a predicted community assignment $\hat{\sigma}^{(t)}$, we define for $a, b \in \{0, 1\}$,

$$\hat{P}^{(t)}_{ab} = \frac{1}{\left|\left\{(i,j) \;:\; \hat{\sigma}^{(t)}_i = \hat{\sigma}^{(t)}_j\right\}\right|} \sum_{(i,j)\,:\,\hat{\sigma}^{(t)}_i = \hat{\sigma}^{(t)}_j} \frac{n^{(t)}_{ab}(i,j)}{n^{(t)}_a(i,j)}, \tag{6.2.2}$$

where $n^{(t)}_{ab}(i,j) = \sum_{t'=1}^{t-1} 1\big(A^{t'}_{ij} = a\big) 1\big(A^{t'+1}_{ij} = b\big)$ is the number of $a \to b$ transitions in the interaction pattern between nodes i and j seen during the t first snapshots and $n^{(t)}_a(i,j) = \sum_{b=0}^{1} n^{(t)}_{ab}(i,j)$. Similarly,

$$\hat{Q}^{(t)}_{ab} = \frac{1}{\left|\left\{(i,j) \;:\; \hat{\sigma}^{(t)}_i \neq \hat{\sigma}^{(t)}_j\right\}\right|} \sum_{(i,j)\,:\,\hat{\sigma}^{(t)}_i \neq \hat{\sigma}^{(t)}_j} \frac{n^{(t)}_{ab}(i,j)}{n^{(t)}_a(i,j)}, \tag{6.2.3}$$

is an estimator of Q_{ab}. Moreover, the quantities $n_{ab}^{(t)}(i,j)$ can be updated inductively. Indeed,

$$n_{ab}^{(t+1)}(i,j) \;=\; n_{ab}^{(t)}(i,j) + 1\big(A_{ij}^t = a\big)\, 1\big(A_{ij}^{t+1} = b\big). \tag{6.2.4}$$

Finally, the initial distribution can also be estimated by averaging:

$$\hat{\mu}^{(t)} \;=\; \frac{1}{\left|\left\{(i,j) \;:\; \hat{\sigma}_i^{(t)} = \hat{\sigma}_j^{(t)}\right\}\right|} \sum_{(i,j)\,:\,\hat{\sigma}_i^{(t)}=\hat{\sigma}_j^{(t)}} A_{ij}^t \tag{6.2.5}$$

and

$$\hat{\nu}^{(t)} \;=\; \frac{1}{\left|\left\{(i,j) \;:\; \hat{\sigma}_i^{(t)} \neq \hat{\sigma}_j^{(t)}\right\}\right|} \sum_{(i,j)\,:\,\hat{\sigma}_i^{(t)}\neq\hat{\sigma}_j^{(t)}} A_{ij}^t. \tag{6.2.6}$$

This leads to Algorithm 8, for clustering in a Markov SBM when only the number of communities K is known. Note that to save computation time, we can choose not to update the parameters at each time step.

Algorithm 8: Online clustering for homogeneous Markov dynamics when the block interaction parameters are unknown.

Input: Observed graph sequence $A^{1:T} = \big(A^1, \dots, A^T\big)$; number of communities K; static graph clustering algorithm `algo`.

Output: Node labelling $\hat{\sigma} = (\hat{\sigma}_1, \dots, \hat{\sigma}_n)$.

Initialize:

- Compute $\hat{\sigma} \leftarrow \texttt{algo}\,(A^1)$;
- Compute $\hat{\mu}, \hat{\nu}$ using formulas (6.2.5)-(6.2.6), and let $M_{ij} \leftarrow \log \frac{\hat{\mu}\left(A_{ij}^1\right)}{\hat{\nu}\left(A_{ij}^1\right)}$;
- Let $n_{ij}(a,b) \leftarrow 0$ for $i,j \in [N]$ and $a,b \in \{0,1\}$.

Update:

for $t = 2, \dots, T$ **do**
  Compute $\Delta_{ij} \leftarrow \log \frac{\hat{P}\left(A_{ij}^{t-1}, A_{ij}^t\right)}{\hat{Q}\left(A_{ij}^{t-1}, A_{ij}^t\right)}$ for $i,j = 1, \dots, N$;
  Set $M \leftarrow M + \Delta$.
  for $i = 1, \dots, n$ **do**
    Set $L_{i,k} \leftarrow \sum_{j \neq i} M_{ij} 1\,(\hat{\sigma}_j = k)$ for all $k = 1, \dots, K$
    Set $\hat{\sigma}_i \leftarrow \arg\max_{1 \le k \le K} L_{i,k}$
  For every node pair (ij), update $n_{ij}(a,b)$ using (6.2.4);
  Update $\hat{P}$, $\hat{Q}$ using (6.2.2) and (6.2.3).

6.2.3 Numerical illustrations and experiments

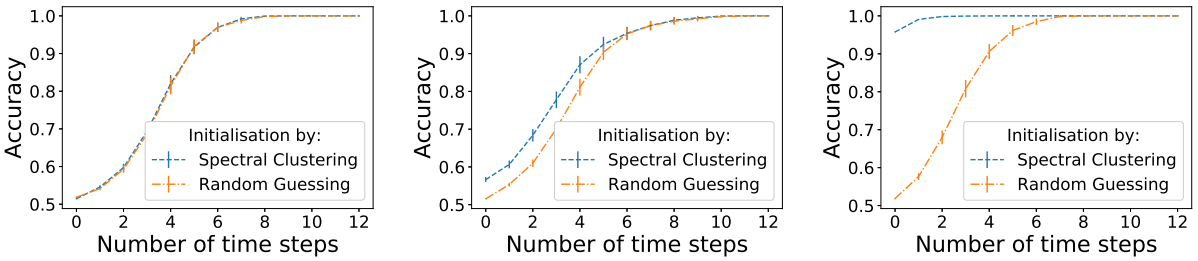
In the numerical simulations, we suppose that $\begin{pmatrix} 1-\mu_1 \\ \mu_1 \end{pmatrix}$ and $\begin{pmatrix} 1-\nu_1 \\ \nu_1 \end{pmatrix}$ are the stationary distributions of P and Q, respectively. Therefore,

$$P = \begin{pmatrix} 1-\mu_1\frac{1-P_{11}}{1-\mu_1} & \mu_1\frac{1-P_{11}}{1-\mu_1} \\ 1-P_{11} & P_{11} \end{pmatrix},$$

and Q has a similar expression.

Evolution of accuracy with the number of snapshots

Let us now study the effect of the initialization step. We plot in Figure 6.1 the evolution of the averaged accuracy obtained when we run Algorithm 7 on 50 realizations of a Markov SBM, where the initialization is done either using Spectral Clustering or Random Guessing. Obviously, when Spectral Clustering works well (see Figure 6.1c), it is preferable to use it than a random guess. Nonetheless, it is striking to see that when the initial Spectral Clustering gives a bad accuracy, then the likelihood method can overcome it. For example, in Figure 6.1a, the initial clustering with Spectral Clustering on the first snapshot is really bad (accuracy $\approx 50\%$, hence not much better than a random guessing), Algorithm 7 does overcome this and reaches a perfect clustering after a few snapshots. In that particular setting, there is no advantage in using Spectral Clustering rather than Random Guessing.



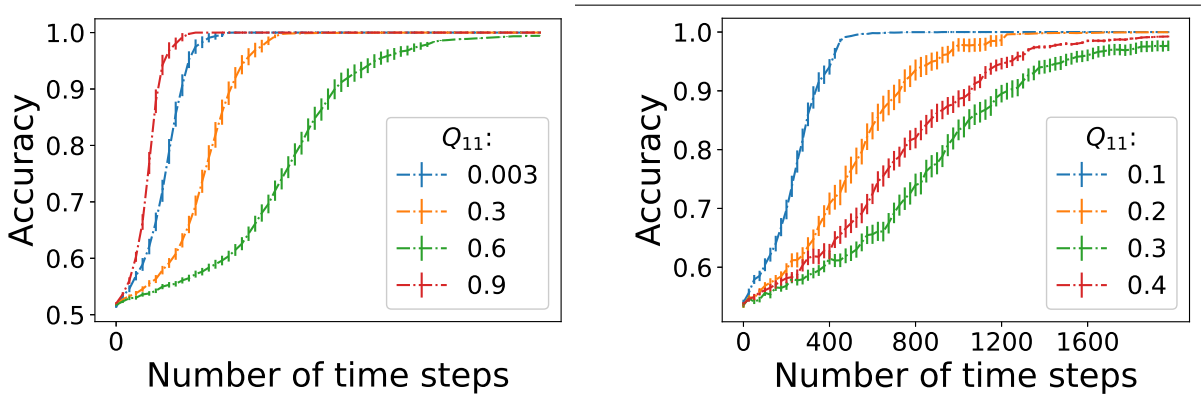
(a) $\mu_1 = 1.5\,\frac{\log N}{N}$ ($T^*_{\text{theo}} = 13$)

(b) $\mu_1 = 2.5\,\frac{\log N}{N}$ ($T^*_{\text{theo}} = 14$)

(c) $\mu_1 = 4.0\,\frac{\log N}{N}$ ($T^*_{\text{theo}} = 11$)

Figure 6.1: Evolution of the accuracy given by Algorithm 7 when the initialisation is done via Spectral Clustering or Random Guessing. The synthetic graphs are Markov SBM with $N = 500$ nodes (equally divided in two clusters), and parameters $\nu_1 = 1.5\,\frac{\log N}{N}$, $P_{11} = 0.7$ and $Q_{11} = 0.3$. Accuracy is averaged over 50 realisations, and the error bars represent the standard error. T^*_{theo} is the theoretical minimum number of time steps needed to get above the exact recovery threshold.

This is further strengthened by our numerical observations in the constant degree regime. As we see in Figure 6.2, our Algorithm performs well when $\mu_1 = \frac{c_{\text{in}}}{N}$ and $\nu = \frac{c_{\text{out}}}{N}$ ($c_{\text{in}}, c_{\text{out}}$ constants), even if $c_{\text{in}} \approx c_{\text{out}}$ (see Figure 6.2b). This is very similar to what we saw in the logarithmic degree regime (Figure 6.1), except that the number of snapshots needed to get excellent accuracy is higher since the graphs are sparser.



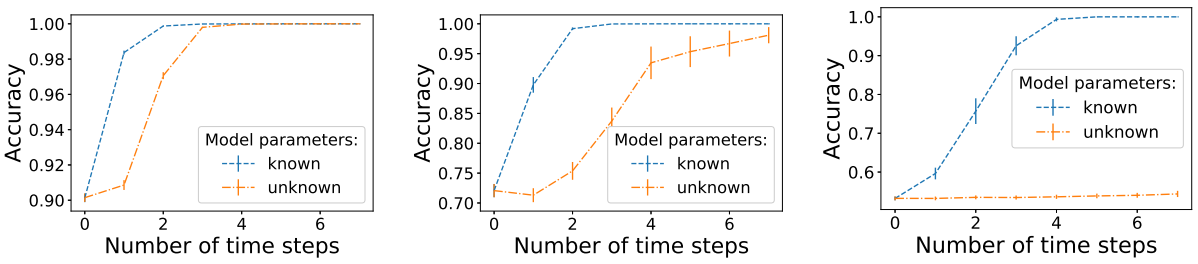
(a) $N = 500$, $\mu_1 = \frac{2.5}{N}$, $\nu_1 = \frac{1.5}{N}$ and $P_{11} = 0.6$.

(b) $N = 100$, $\mu_1 = \frac{0.15}{N}$, $\nu_1 = \frac{0.1}{N}$ and $P_{11} = 0.4$.

Figure 6.2: Evolution of the accuracy with the number of snapshots obtained by Algorithm 7 in a sparse setting, when the initialisation is done via Random Guessing. We draw 50 synthetic Markov SBM with two equal size communities. The choice of parameters in Figure (b) is much more challenging than Figure (a). The different curves show the averaged accuracy over 50 trials, and errors bars correspond to the empirical standard errors.

Unknown interaction parameters are unknown

We show in Figure 6.3 the comparison of the online Algorithm 7 (with known interaction parameters) with the online Algorithm 8 (with unknown interaction parameters). We see that, when the starting round of Spectral Clustering gives a decent accuracy (at least 75%), then Algorithm 8 can learn the model parameters as well as communities. However, when Spectral Clustering gives a bad accuracy, Algorithm 8 without the model parameters fails, whereas the version with the known interaction parameters succeeds.



(a) $\nu_1 = 0.03$.

(b) $\nu_1 = 0.035$.

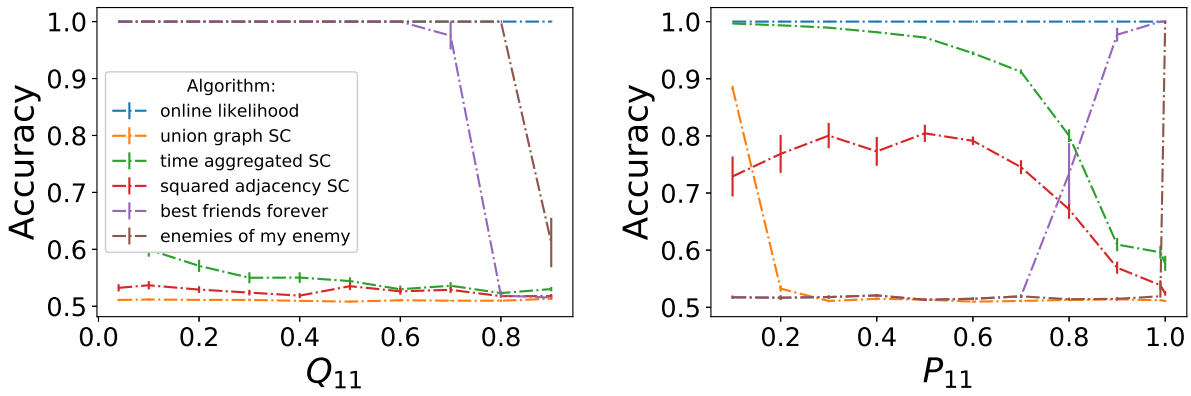
(c) $\nu_1 = 0.04$.

Figure 6.3: Comparison of the accuracy given by the online versions of the algorithm. The results are averaged on 20 realizations of Markov SBM with parameters $N = 1000$, $\mu_1 = 0.05$, $P_{11} = 0.6$, $Q_{11} = 0.3$, and for different ν_1.

Comparison with the baseline algorithms

In this section, we compare the performance of Algorithm 7 to the baseline methods proposed in Section 6.1.2. Results are shown in Figure 6.4. We draw the following observations:

- Algorithm 7 (called *online likelihood* in the plots) always achieves very high accuracy, and outperforms all other methods;
- Spectral Clustering on the union graph always performs very poorly, while Spectral Clustering on the time-aggregated graph can perform very well if the evolution of the pattern interactions are not too static (*i.e.*, P_{11} and Q_{11} are both away from 1);
- Spectral Clustering on $\sum_{t=1}^{T} A_t^2 - D_t$, where D_t is the degree matrix of layer t, is the method proposed and analysed in [LL20]. This method, called *squared adjacency SC* in the caption of Figure 6.4, is always outperformed by Spectral Clustering on the time-aggregated graph;
- Algorithms 5 and 6 are more sensitive to the hypothesis $P_{11} = 1$ than to $Q_{11} = \nu_1$. In particular, Algorithm 6 (*enemies of my enemy*) fails as soon as $P_{11} \neq 1$ (in Figure 6.4b, when $P_{11} = 0.99$, the accuracy of Algorithm 6 drops to 50%);
- Given its simplicity, Algorithm 5 (*best friends forever*) performs surprisingly well. Of course, when the parameter setting is too far from the ideal situation $P_{11} = 1$ and $Q_{11} = \nu_1$, the algorithm fails as expected. However, even at not too short distances from this ideal case, Algorithm 5 gives meaningful classification.



(a) $P_{11} = 1$.

(b) $Q_{11} = \nu_1$.

Figure 6.4: Comparison of the accuracy given by the different algorithms. The results are averaged on 50 realisations of Markov SBM with $N = 500$, $T = 30$ and $\mu_1 = 0.05$ and $\nu_1 = 0.04$. Figure (6.4a) shows the situation $P_{11} = 1$ (static intra-community interaction patterns) and Q_{11} varies, while Figure (6.4b) shows $Q_{11} = \nu_1$ (i.i.d. inter-community interaction pattern) and P_{11} varies. Colours correspond to the same algorithms in both plots.

6.3 Spectral methods for temporal networks with static communities

This Section studies the recovery of static communities in a temporal network. We introduce in Section 6.3.1 a temporal stochastic block model where dynamic interaction patterns between node pairs follow a Markov chain. We render this model versatile by adding degree correction parameters, describing the tendency of each node to start new interactions. We show that in some cases the likelihood of this model is approximated by the regularized modularity of a time-aggregated graph. This time-aggregated graph involves a trade-off between new edges and persistent edges. A continuous relaxation reduces the regularized modularity maximisation to a normalized spectral clustering. In Section 6.3.2, we illustrate by numerical experiments the importance of edge persistence, both on simulated and real data sets.

6.3.1 Degree-corrected temporal network model

Consider a population of N nodes partitioned into K static communities such that node i belongs to community $\sigma_i \in [K]$. We write $A_{ij}^t = 1$ if nodes i and j interact at time t, and $A_{ij}^t = 0$ otherwise. We investigate methods of recovering the community structure $\sigma = (\sigma_1, \dots, \sigma_N)$ from an observed adjacency tensor $A = \left(A_{ij}^t\right)$. The following section describes a versatile statistical model for this setting.

Model description

A degree-corrected temporal stochastic block model with N nodes, K blocks and T snapshots is a probability distribution

$$\mathbb{P}(A \,|\, \sigma, F, \theta) \;=\; \prod_{1 \le i < j \le N} F_{\sigma_i \sigma_j}^{\theta_i \theta_j}\left(A_{ij}^1, \dots, A_{ij}^T\right) \tag{6.3.1}$$

of a symmetric adjacency tensor $A \in \{0,1\}^{N \times N \times T}$ with zero diagonal entries, where $\sigma = (\sigma_1, \dots, \sigma_N)$ is a community assignment with $\sigma_i \in \{1, \dots, K\}$ indicating the community of node i, $F = (F_{k\ell}^{xy})$ is a collection of probability distributions over $\{0,1\}^T$, and $\theta = (\theta_1, \dots, \theta_N)$ is a vector of node-specific degree correction parameters, with $0 \le \theta_i < \infty$.

In the following, we will restrict ourselves to homogeneous inter-block interactions with Markov edge dynamics, for which the nodes' static community labellings are sampled uniformly at random from the set $[K]$ of all node labellings, and

$$F_{\sigma_i \sigma_j}^{\theta_i \theta_j}(x) \;=\; \begin{cases} \mu_{x_1}^{\theta_i \theta_j} \prod_{t=2}^{T} P_{x_{t-1}, x_t}^{\theta_i \theta_j} & \text{if } \sigma_i = \sigma_j, \\ \nu_{x_1}^{\theta_i \theta_j} \prod_{t=2}^{T} Q_{x_{t-1}, x_t}^{\theta_i \theta_j} & \text{otherwise,} \end{cases} \tag{6.3.2}$$

with initial distributions

$$\mu^{\theta_i \theta_j} = \begin{pmatrix} 1 - \theta_i \theta_j \mu_1 \\ \theta_i \theta_j \mu_1 \end{pmatrix} \quad \text{and} \quad \nu^{\theta_i \theta_j} = \begin{pmatrix} 1 - \theta_i \theta_j \nu_1 \\ \theta_i \theta_j \nu_1 \end{pmatrix}, \tag{6.3.3}$$

and transition probability matrices

$$P^{\theta_i\theta_j} = \begin{pmatrix} 1-\theta_i\theta_j P_{01} & \theta_i\theta_j P_{01} \\ 1-P_{11} & P_{11} \end{pmatrix} \quad \text{and} \quad Q^{\theta_i\theta_j} = \begin{pmatrix} 1-\theta_i\theta_j Q_{01} & \theta_i\theta_j Q_{01} \\ 1-Q_{11} & Q_{11} \end{pmatrix}. \tag{6.3.4}$$

The parameters θ_i account for the fact that some nodes might be more inclined than others to start new connections, similarly to the degree-corrected block model of [KN11]. To keep the model simple, we do not add degree correction parameters in front of P_{11}; hence once a connection started, the probability to keep it active is simply P_{11} or Q_{11}. Moreover, we assume that $\min_{i,j}\{\theta_i\theta_j\delta\} \leq 1$, where $\delta = \max\{\mu_1, \nu_1, P_{01}, Q_{01}\}$. Finally, we normalise the degree correction parameters so that $\sum_i 1(\sigma_i = k)\theta_i = \sum_i 1(\sigma_i = k)$ for all k.

Maximum likelihood estimator

Proposition 21. *A maximum likelihood estimator for the Markov block model defined by* (6.3.1)–(6.3.2) *is any community assignment* $\sigma \in [K]^N$ *that maximises*

$$\sum_{i,j} \delta(\sigma_i,\sigma_j)\left\{ A^1_{ij}\left(\rho_1^{\theta_i\theta_j} - \rho_0^{\theta_i\theta_j}\right) + \rho_0^{\theta_i\theta_j} + \left(A^1_{ij} - A^T_{ij}\right)\ell_{10}^{\theta_i\theta_j} \right\}$$
$$+ \sum_{i,j} \delta(\sigma_i,\sigma_j) \sum_{t=2}^{T} \left\{ \left(\ell_{01}^{\theta_i\theta_j} + \ell_{10}^{\theta_i\theta_j}\right)\left(A^t_{ij} - A^{t-1}_{ij}A^t_{ij}\right) + \ell_{11}^{\theta_i\theta_j} A^{t-1}_{ij}A^t_{ij} - \log\frac{Q_{00}^{\theta_i\theta_j}}{P_{00}^{\theta_i\theta_j}} \right\}$$

where $\rho_a^{\theta_i\theta_j} = \log \frac{\mu_a^{\theta_i\theta_j}}{\nu_a^{\theta_i\theta_j}}$ *and* $\ell_{ab}^{\theta_i\theta_j} = \log \frac{P_{ab}^{\theta_i\theta_j}}{Q_{ab}^{\theta_i\theta_j}} - \log \frac{P_{00}^{\theta_i\theta_j}}{Q_{00}^{\theta_i\theta_j}}$.

The MLE derived in Proposition 21 is more complex that summing all snapshots independently. In particular, the terms $A^{t-1}_{ij}A^t_{ij}$ account for *persistent edges* over two consecutive snapshots. Denote by $A^t_{\text{pers}} = A^{t-1} \odot A^t$ the entrywise product of adjacency matrices A^{t-1} and A^t. Then A^t_{pers} is the adjacency matrix of the graph containing the persistent edges between $t-1$ and t, and $A^t_{\text{new}} = A^t - A^t_{\text{pers}}$ corresponds to the graph containing the *edges freshly appearing* at time t.

Proof of Proposition 21. By the temporal Markov property, the log-likelihood of the model can be written as $\log \mathbb{P}(A \mid \sigma, \theta) = \log \mathbb{P}(A^1 \mid \sigma, \theta) + \sum_{t=2}^{T} \mathbb{P}(A^t \mid A^{t-1}, \sigma, \theta)$. By denoting $\rho_a^{\theta_i\theta_j} = \log \frac{\mu_a^{\theta_i\theta_j}}{\nu_a^{\theta_i\theta_j}}$, we find that

$$\begin{aligned} \log \mathbb{P}(A^1 \mid \sigma, \theta) &= \frac{1}{2}\sum_{i,j}\sum_a \delta(A^1_{ij}, a)\Big(\delta(\sigma_i,\sigma_j)\rho_a^{\theta_i\theta_j} + \log \nu_a^{\theta_i\theta_j}\Big) \\ &= \frac{1}{2}\sum_{i,j}\delta(\sigma_i,\sigma_j)\sum_a \delta(A^1_{ij}, a)\rho_a^{\theta_i\theta_j} + c_1(A), \end{aligned}$$

where $c_1(A) = \frac{1}{2}\sum_{i,j}\sum_a \delta(A^1_{ij},a)\log \nu_a^{\theta_i\theta_j}$ does not depend on the community structure. Similarly, by denoting $R^{\theta_i\theta_j}_{ab} = \log \frac{P^{\theta_i\theta_j}_{ab}}{Q^{\theta_i\theta_j}_{ab}}$ we find that

$$\begin{aligned}\log \mathbb{P}(A^t \mid A^{t-1},\sigma,\theta) &= \frac{1}{2}\sum_{i,j}\sum_{a,b}\delta(A^{t-1}_{ij},a)\delta(A^t_{ij},b)\Big(\delta(\sigma_i,\sigma_j)R^{\theta_i\theta_j}_{ab} + \log Q^{\theta_i\theta_j}_{ab}\Big)\\ &= \frac{1}{2}\sum_{i,j}\delta(\sigma_i,\sigma_j)\sum_{a,b}\delta(A^{t-1}_{ij},a)\delta(A^t_{ij},b)R^{\theta_i\theta_j}_{ab} + c_t(A),\end{aligned}$$

where $c_t(A) = \frac{1}{2}\sum_{i,j}\sum_{a,b}\delta(A^{t-1}_{ij},a)\delta(A^t_{ij},b)\log Q^{\theta_i\theta_j}_{ab}$ does not depend on the community structure. Simple computations show that

$$\sum_a \delta(A^1_{ij},a)\rho_a^{\theta_i\theta_j} = A^1_{ij}(\rho_1^{\theta_i\theta_j} - \rho_0^{\theta_i\theta_j}) + \rho_0^{\theta_i\theta_j}$$

and

$$\begin{aligned}\sum_{a,b}\delta(A^{t-1}_{ij},a)\delta(A^t_{ij},b)R^{\theta_i\theta_j}_{ab} &= R^{\theta_i\theta_j}_{00} + A^{t-1}_{ij}\big(R^{\theta_i\theta_j}_{10} - R^{\theta_i\theta_j}_{00}\big) + A^t_{ij}\big(R^{\theta_i\theta_j}_{01} - R^{\theta_i\theta_j}_{00}\big)\\ &\quad + A^{t-1}_{ij}A^t_{ij}\big(R^{\theta_i\theta_j}_{11} - R^{\theta_i\theta_j}_{01} - R^{\theta_i\theta_j}_{10} + R^{\theta_i\theta_j}_{00}\big)\\ &= R^{\theta_i\theta_j}_{00} + A^{t-1}_{ij}\ell^{\theta_i\theta_j}_{10} + A^t_{ij}\ell^{\theta_i\theta_j}_{01} + A^{t-1}_{ij}A^t_{ij}\big(\ell^{\theta_i\theta_j}_{11} - \ell^{\theta_i\theta_j}_{01} - \ell^{\theta_i\theta_j}_{10}\big).\end{aligned}$$

By collecting the above observations, we now find that $\log \mathbb{P}(A \mid \sigma,\theta)$ equals

$$\begin{aligned}&c(A) + \frac{1}{2}\sum_{i,j}\delta(\sigma_i,\sigma_j)\left\{A^1_{ij}(\rho_1^{\theta_i\theta_j} - \rho_0^{\theta_i\theta_j}) + \rho_0^{\theta_i\theta_j} + (A^1_{ij} - A^T_{ij})\ell^{\theta_i\theta_j}_{10}\right\}\\ &\quad + \frac{1}{2}\sum_{i,j}\delta(\sigma_i,\sigma_j)\sum_{t=2}^{T}\left\{(\ell^{\theta_i\theta_j}_{01} + \ell^{\theta_i\theta_j}_{10})\left(A^t_{ij} - A^{t-1}_{ij}A^t_{ij}\right) + \ell^{\theta_i\theta_j}_{11}A^{t-1}_{ij}A^t_{ij} - \log\frac{Q^{\theta_i\theta_j}_{00}}{P^{\theta_i\theta_j}_{00}}\right\},\end{aligned}$$

where $c(A) = \sum_t c_t(A)$ does not depend on σ. Hence the claim follows. □

Assuming that the number of snapshots T is large, we can ignore the boundary terms, and the MLE expressed in Proposition 21 reduces to maximising

$$\sum_{t=2}^{T}\sum_{\substack{i,j:\\ \sigma_i=\sigma_j}}\left(\left(\ell^{\theta_i\theta_j}_{01} + \ell^{\theta_i\theta_j}_{10}\right)\left(A^t_{ij} - A^{t-1}_{ij}A^t_{ij}\right) + \ell^{\theta_i\theta_j}_{11}A^{t-1}_{ij}A^t_{ij} - \log\frac{Q^{\theta_i\theta_j}_{00}}{P^{\theta_i\theta_j}_{00}}\right).$$

By utilising (6.3.3)–(6.3.4), we can further simplify it to express this as a modularity. Recall given a weighted graph W, a partition Z and a resolution parameter γ, the regularized modularity is defined as [NG04; RB06]

$$\mathcal{M}(W,\sigma,\gamma) = \sum_{i,j}\delta(\sigma_i,\sigma_j)\left(W_{ij} - \gamma\frac{d_i d_j}{2m}\right)$$

where $d_i = \sum_j W_{ij}$ and $m = \sum_i d_i$. Hence, suppose that $P^{\theta_i\theta_j}$ and $Q^{\theta_i\theta_j}$ are nondegenerate, and $\mu^{\theta_i\theta_j}$ (resp. $\nu^{\theta_i\theta_j}$) is the stationary distribution of $P^{\theta_i\theta_j}$ (resp. $Q^{\theta_i\theta_j}$). In a sparse setting, P_{01} and Q_{01} are small, and after a Taylor expansion (see Section 6.3.3 for the full derivations) the previous expression is approximately equal to $\mathcal{M}(W, \sigma, \gamma)$, where W is defined by

$$W \;=\; \sum_{t=2}^{T} \left(\alpha A^t_{\text{new}} + \beta A^t_{\text{pers}}\right) \tag{6.3.5}$$

with

$$\alpha \;=\; \log \frac{P_{01}}{Q_{01}} + \log \frac{1-P_{11}}{1-Q_{11}} \quad \text{and} \quad \beta \;=\; \log \frac{P_{11}}{Q_{11}}, \tag{6.3.6}$$

and $\gamma \;=\; (P_{01} - Q_{01})\, \frac{\alpha(\mu_1+(K-1)\nu_1)+(\beta-\alpha)(\mu_1 P_{11}+(K-1)\nu_1 Q_{11})}{K}$.

Comparison with previous work Correspondence between maximum likelihood estimator and modularity maximisation are long known in static block models [New16]. Analogously to the single-layer case, the modularity of a temporal network, with possibly time-dependent community structure, was previously defined in [Muc+10; Pam+19] by

$$\sum_{t=1}^{T} \mathcal{M}(A^t, \sigma^t, \gamma_t) + \sum_{t=1}^{T} \sum_{s\neq t} \sum_{i} \omega_i^{st}\, \delta\left(\sigma_i^s, \sigma_i^t\right) \tag{6.3.7}$$

where γ_t is the resolution parameter for layer t, σ_i^t is the community membership of node i at time step t, and w_i^{st} denotes a coupling between time instants s and t. For a static community structure, the second term in (6.3.7) is irrelevant. When the resolution is constant over time, the relevant term in (6.3.7) can be written as

$$\sum_{t=1}^{T} \mathcal{M}(A^t, \sigma, \gamma) \;=\; \mathcal{M}(A^{\text{agg}}, \sigma, \gamma),$$

where $A^{\text{agg}} \;=\; \sum_{t=1}^{T} A^t$ is the weighted adjacency matrix of the time-aggregated data. In contrast, the matrix W in (6.3.5) involves a trade-off between new edges and persistent edges. We notice that $W = A^{\text{agg}}$ only if $\alpha = \beta = 1$.

Temporal spectral clustering combining new and persistent edges

Following our analysis in Section 6.3.1, the community prediction should verify

$$\hat{\sigma} \;=\; \underset{\sigma\in[K]^N}{\arg\max}\, \mathcal{M}(W, \sigma, \gamma)$$

where W is defined in Equation (6.3.5) and γ is a proper resolution parameter. This optimisation problem is NP-complete in general [Bra+07], but can be approximately solved by continuous relaxation. We can choose the relaxation so that the optimization problem reduces to normalized spectral clustering algorithm on the weighted graph W (we refer to [New13] and to Section 6.3.3 for the full computations). We note that in order to compute the normalized Laplacian of W, we should restrict $\alpha, \beta \geq 0$, which is not necessarily guaranteed by Formula (6.3.6). We summarize this in Algorithm 9.

Algorithm 9: Spectral clustering for temporal networks with Markov edge dynamics and static node labelling.

Input: Adjacency matrices $A^1, \dots, A^T$, number of clusters K, parameter α, β.
Output: Predicted membership matrix $\hat{\sigma} \in [K]^N$

1 **Process:**

- Let $W = \sum_{t=2}^{T} \alpha A^t_{\text{new}} + \beta A^t_{\text{pers}}$ where $A^t_{\text{new}} = A^t - A^{t-1} \odot A^t$ and $A^t_{\text{pers}} = A^{t-1} \odot A^t$;
- Compute $\mathcal{L} = I_n - D^{-1/2} W D^{-1/2}$ where $D = \text{diag}(W 1_n)$;
- Compute the matrix $\hat{X} \in \mathbb{R}^{N \times K}$ whose columns consist of the K orthonormal eigenvectors of $\mathcal{L}$ associated to the K smallest eigenvalues.

2 **Return** $\hat{\sigma} \leftarrow \text{kmeans}\left(D^{-1/2}\hat{X}, K\right)$.

6.3.2 Numerical experiments

The Python source code for reproducing our results is available online[1].

Synthetic data

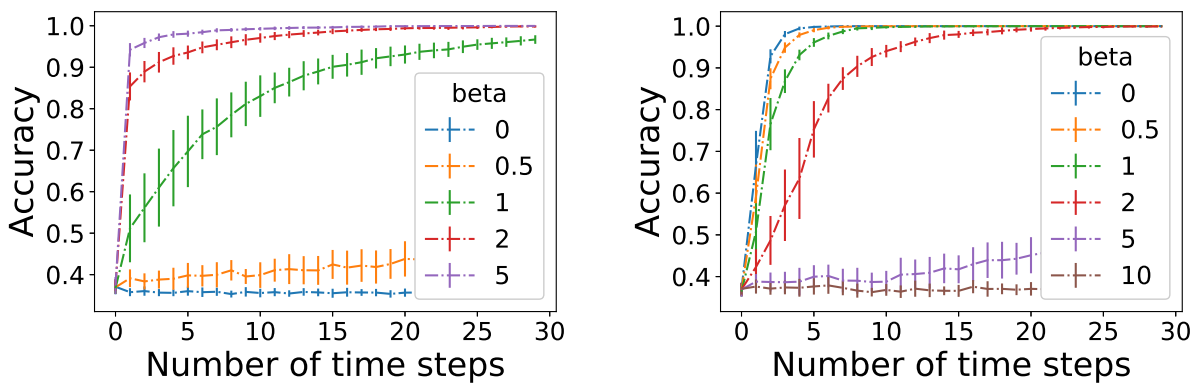
We first examine the effect of the choice of the parameters α and β in Algorithm 9. For this, we let $\alpha = 1$ and we plot in Figure 6.5 the averaged accuracy obtained on 25 realizations of stochastic block models with Markov edge dynamics for various β. While spectral clustering on the time-aggregated graph (corresponding to $\beta = 1$) works well, it is striking to notice that other values of β give better results. The choice of β depends on the probabilities of persistent interactions. For example, if $P_{11} > Q_{11}$ (Figure 6.5a), then $\beta > 1$ are preferred, while if $P_{11} < Q_{11}$ (Figure 6.5b) large choice of β are penalized. This is in accordance to the values of α, β derived in Formula (6.3.6) (albeit in Formula (6.3.6), α and β could be negative).

We show the robustness of Algorithm 9 on the degree correction parameters in Figure 6.6. More precisely:

- Figure 6.6a generates θ_i according to $|\mathcal{N}(0, \sigma^2)| + 1 - \sigma\sqrt{2/\pi}$ where $|\mathcal{N}(0, \sigma^2)|$ denotes the absolute value of a normal random variable with mean 0 and variance σ^2. We choose $\sigma = 0.25$.
- Figure 6.6b generates the θ_i from a Pareto distribution with density function $f(x) = \frac{am^a}{x^{a+1}} 1(x \geq m)$ with $a = 3$ and $m = 2/3$ (chosen such that $\mathbb{E}\theta_i = 1$).

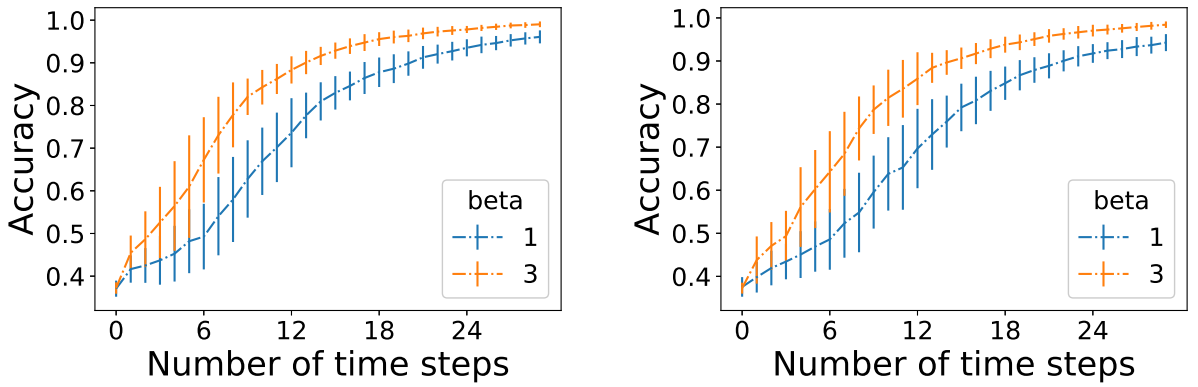
Note that the sampling of the θ_i's enforces $\mathbb{E}\theta_i = 1$ in both settings. We notice that in both cases, letting $\beta \neq 1$ improves the performance of Algorithm 9.

[1] https://github.com/mdreveton/Spectral-clustering-with-persistent-edges



(a) $P_{11} = 0.9$ (b) $P_{11} = 0.1$

Figure 6.5: Accuracy of Algorithm 9 on a SBM with 300 nodes in $K = 3$ blocks, degree correction parameters $\theta_1 = \cdots = \theta_n = 1$, and a stationary Markov edge evolution $\mu_1 = 0.04$, $\nu_1 = 0.02$ and $Q_{11} = 0.3$. The results are averaged over 25 synthetic graphs, and error bars show the standard deviation.



(a) Normal (b) Pareto

Figure 6.6: Accuracy of Algorithm 9 with $\alpha = 1$ and different β, on a SBM with 300 nodes and $K = 3$ blocks (with uniform prior), and a stationary Markov edge evolution $\mu_1 = 0.06$, $\nu_1 = 0.03$, $P_{11} = 0.7$ and $Q_{11} = 0.4$, for different generation of the degree correction parameters θ. The results are averaged over 25 synthetic graphs, and error bars show the standard deviation.

Social networks of high school students

We investigate three data sets collected during three consecutive years from the high school *Lycée Thiers* in Marseilles, France [FB14; MFB15]. Nodes correspond to students, interactions to close-proximity encounters, and communities to classes, with dimensions given in Table 6.1.

We make a hypothesis that the temporal characteristics of the interactions are similar each year. We then use the 2011 data set to estimate the transition probability matrices P and Q, and use these for clustering the 2012 and 2013 data sets. We assume that $\theta_i = 1$ (no degree

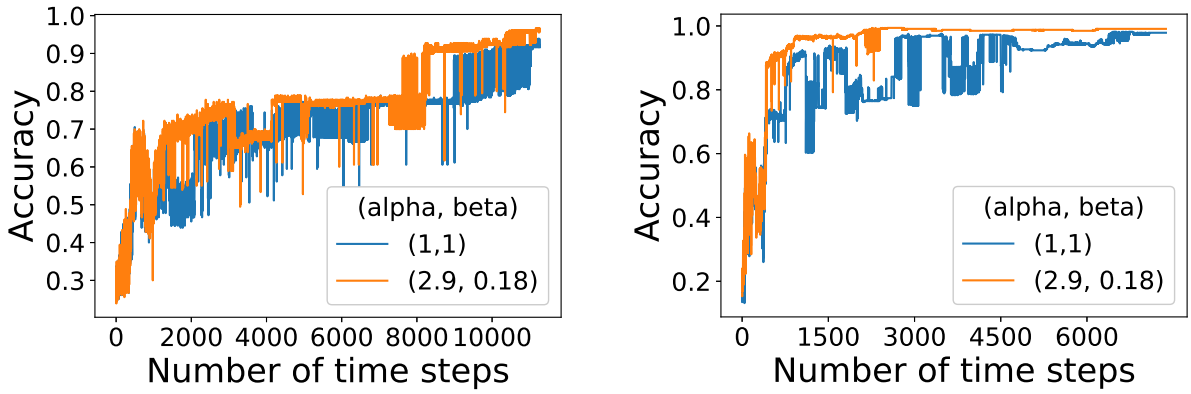
Year	N	K	T
2011	118	3	5609
2012	180	5	11273
2013	327	9	7375

Table 6.1: Dimensions of three data sets of interacting high school students.

correction). Assuming the ground truth clustering is known for the year 2011[2], a standard estimator of Markov chain transition probability matrices [Bil61] gives

$$\hat{P} = \begin{pmatrix} 0.9992 & 0.0008 \\ 0.37 & 0.63 \end{pmatrix} \quad \text{and} \quad \hat{Q} = \begin{pmatrix} 0.999967 & 3.3 \times 10^{-5} \\ 0.48 & 0.52 \end{pmatrix}.$$

Using (6.3.6), leads to $\hat{\alpha} = 2.9$ and $\hat{\beta} = 0.18$. We observe in Figure 6.7b that this choice of parameters gives a better accuracy on the 2013 data set than simply applying spectral clustering on the time-aggregated graph ($\alpha = \beta = 1$). For the 2012 data set (Figure 6.7a), this improvement is not so clearly visible.



(a) Year 2012

(b) Year 2013

Figure 6.7: Accuracy of Algorithm 9 on the 2012 and 2013 high school datasets, using uniform $\alpha = \beta = 1$ (blue) and adjusted α, β predicted using 2011 data (orange).

To understand why Algorithm 9 performs better for 2013 than for 2012, we have listed in Table 6.2 temporal transition probabilities and clustering weights $\hat{\alpha}, \hat{\beta}$ estimated separately for each dataset. For year 2012, the difference between intra-community edge persistence $\hat{P}_{11}$ and inter-community edge persistence $\hat{Q}_{11}$ is small, implying that persistent edges do not add much extra information for distinguishing communities ($\hat{\beta} \approx 0$). For years 2011 and 2013, this difference is larger, manifesting that edge persistence contains information that can be employed to recover communities with a higher accuracy.

[2]Alternatively, one could perform spectral clustering on the time-aggregated graph, which produces an excellent prediction of the ground truth community labels.

Dataset	$\hat{P}_{01}$	$\hat{Q}_{01}$	$\hat{P}_{11}$	$\hat{Q}_{11}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\beta}/\hat{\alpha}$
2011	0.00080	0.000033	0.63	0.52	2.9	0.58	0.060
2012	0.00050	0.000011	0.57	0.56	3.8	0.01	0.003
2013	0.00150	0.000014	0.64	0.40	4.5	0.07	0.015

Table 6.2: Markov chain transition probabilities and adjusted clustering weights estimated separately for each dataset.

6.3.3 Additional proofs

Approximation of the MLE

Recall the structural assumptions (6.3.3)–(6.3.4) about the degree correction parameters. Because $P_{01}, Q_{01} = o(1)$, a first-order Taylor expansion yields

$$\log \frac{1-\theta_i\theta_j Q_{01}}{1-\theta_i\theta_j P_{01}} \;=\; \theta_i\theta_j\left(P_{01}-Q_{01}\right) + o\left(P_{01}^2+Q_{01}^2\right),$$

as well as $\ell_{01}^{\theta_i\theta_j} \approx \log\frac{P_{01}}{Q_{01}}$, $\ell_{10}^{\theta_i\theta_j} \approx \log\frac{1-P_{11}}{1-Q_{11}}$ and $\ell_{11}^{\theta_i\theta_j} \approx \log\frac{P_{11}}{Q_{11}}$. Using these approximations in the MLE expression leads to the maximisation of

$$\sum_{t=2}^{T}\sum_{i,j\colon \sigma_i=\sigma_j}\left(\tilde{a}_{ij}^t - \theta_i\theta_j\left(P_{01}-Q_{01}\right)\right), \tag{6.3.8}$$

where $\tilde{a}_{ij}^t = \alpha\left(A_{\text{new}}^t\right)_{ij} + \beta\left(A_{\text{pers}}^t\right)_{ij}$. Since μ and ν are stationary distributions,

$$\mathbb{E}\left(A_{\text{new}}^t\right)_{ij} \;=\; \begin{cases}\theta_i\theta_j\mu_1(1-P_{11}) & \text{if } \sigma_i=\sigma_j,\\ \theta_i\theta_j\nu_1(1-Q_{11}) & \text{otherwise,}\end{cases}$$

$$\mathbb{E}\left(A_{\text{pers}}^t\right)_{ij} \;=\; \begin{cases}\theta_i\theta_j\mu_1 P_{11} & \text{if } \sigma_i=\sigma_j,\\ \theta_i\theta_j\nu_1 Q_{11} & \text{otherwise.}\end{cases}$$

Therefore, using $W_{ij} = \sum_{t=2}^{T}\tilde{a}_{ij}$ we have

$$\mathbb{E}W_{ij} \;=\; \begin{cases}(T-1)\theta_i\theta_j\mu_1\left(\alpha(1-P_{11})+\beta P_{11}\right) & \text{if } \sigma_i=\sigma_j\\ (T-1)\theta_i\theta_j\nu_1\left(\alpha(1-Q_{11})+\beta Q_{11}\right) & \text{otherwise.}\end{cases}$$

Since the community labelling are sampled uniformly at random and using the normalization for the θ_i's, we have

$$\bar{d}_i \;=\; (T-1)\theta_i N\,\frac{\mu_1\left(\alpha(1-P_{11})+\beta P_{11}\right)+(K-1)\nu_1\left(\alpha(1-Q_{11})+\beta Q_{11}\right)}{K},$$

together with $\bar{m} = \frac{N^2}{2}(T-1)\frac{\mu_1(\alpha(1-P_{11})+\beta P_{11})+(K-1)\nu_1(\alpha(1-Q_{11})+\beta Q_{11})}{K}$. Hence, we observe that $\theta_i\theta_j(P_{01}-Q_{01}) = \gamma\frac{\bar{d}_i\bar{d}_j}{2\bar{m}}$ where $\gamma = (P_{01}-Q_{01})(T-1)\frac{\mu_1(\alpha(1-P_{11})+\beta P_{11})+(K-1)\nu_1(\alpha(1-Q_{11})+\beta Q_{11})}{K}$. We end the proof using equation (6.3.8). □

Modularity and normalized spectral clustering

The regularized modularity of a partition $\sigma \in [K]^N$ of the graph A is defined as

$$\mathcal{M}(A, \sigma, \gamma) = \sum_{i,j} \delta(\sigma_i, \sigma_j) \left(A_{ij} - \gamma \frac{d_i d_j}{2m} \right)$$

where $d = A1_n$ and γ is a resolution parameter. This can be rewritten as

$$\mathcal{M}(A, \sigma, \gamma) = \operatorname{Tr} \tilde{Z}^T \left(A - \gamma \frac{dd^T}{2m} \right) \tilde{Z}$$

where $\tilde{Z} \in \{0,1\}^{N \times K}$ is the membership matrix associated to the vector σ, that is $\tilde{Z}_{ik} = 1$ for $k = \sigma_i$, and $\tilde{Z}_{ik} = 0$ otherwise. As maximising the modularity over $\sigma \in [K]^N$ is in general NP-complete [Bra+07], it is convenient to perform a continuous relaxation. Following [New13], we transform the problem into

$$\hat{X} = \underset{\substack{X \in \mathbb{R}^{N \times K} \\ X^T D X = I_K}}{\arg\max} \operatorname{Tr} X^T \left(A - \gamma \frac{dd^T}{2m} \right) X. \tag{6.3.9}$$

The predicted membership matrix $\hat{Z}$ is then recovered by performing an approximated solution to the following k-means problem (see [KK10])

$$\left(\hat{Z}, \hat{Y} \right) = \underset{Z \in \mathcal{Z}_{N,K}, Y \in \mathbb{R}^{K \times K}}{\arg\min} \left\| ZY - \hat{X} \right\|_F. \tag{6.3.10}$$

The Lagrangian associated to the optimization problem (6.3.9) is

$$\operatorname{Tr} X^T \left(A - \gamma \frac{dd^T}{2m} \right) X - \operatorname{Tr} \left(\Lambda^T \left(X^T X - I_K \right) \right)$$

where $\Lambda \in \mathbb{R}^{K \times K}$ is a symmetric matrix of Lagrangian multipliers. Up to a change of basis, we can assume that Λ is diagonal. The solution of (6.3.9) verifies

$$\left(A - \gamma \frac{dd^T}{2m} \right) X = DX\Lambda \quad \text{and} \quad X^T D X = I_K,$$

which is a generalized eigenvalue problem: the columns of X are the generalized eigenvectors, and the diagonal elements of Λ are the eigenvalues. In particular, since the constant vector 1_n verifies $(A - \gamma \frac{dd^T}{2m}) 1_n = (1-\gamma) D 1_n$, we conclude that the eigenvalues should be larger than $1 - \gamma$ for the partition to be meaningful.

Multiplying the first equation by 1_n^T leads to $(1-\gamma) d^T X = d^T X \Lambda$, and therefore $d^T X = 0$ (using the previous remark on Λ). The system then simplifies in

$$AX = DX\Lambda \quad \text{and} \quad X^T D X = I_K.$$

Defining a re-scaled vector $U = D^{-1/2} X$ shows that U verifies $D^{-1/2} A D^{-1/2} U = U\Lambda$ and $U^T U = I_K$. Thus, the columns of U are eigenvectors of $D^{-1/2} A D^{-1/2}$ associated to the K largest eigenvalue (or equivalently, the eigenvectors of $\mathcal{L} = I_N - D^{-1/2} A D^{-1/2}$ associated to the K smallest eigenvalues).

6.4 Temporal networks with time-varying communities

In this section, we consider a population of N nodes partitioned into K time-evolving communities. At time t, we denote by $\sigma_i^t \in [K]$ the community membership of node i and by $A_{ij}^t \in \{0,1\}$ the observed interaction between nodes i and j. We investigate methods of recovering the community structure, denoted by $\left(\sigma^1, \cdots, \sigma^T\right)$ where $\sigma^t \in [K]^N$, from an observed adjacency tensor $A = \left(A_{ij}^t\right)$.

6.4.1 Model description

Similarly to several papers on dynamic SBM [Gha+16; MM17; Bar+18], we firstly assume that each node community labels $\sigma_i^{1:T} \in [K]^T$ is a Markov chain of length T with initial probability α and transition probability matrix π. Hence,

$$\mathbb{P}\left(\sigma^{1:T}\right) = \prod_{i=1}^{N} \alpha\left(\sigma_i^1\right) \prod_{t=2}^{T} \pi\left(\sigma_i^{t-1}, \sigma_i^t\right). \tag{6.4.1}$$

For simplicity, we will assume that the initial labels and the transitions are uniform, that is

$$\alpha = \frac{1}{K} 1_K \quad \text{and} \quad \pi = \eta I_K + \frac{1-\eta}{K} 1_K 1_K^T \tag{6.4.2}$$

In other words, a node keeps its label with probability $\eta \in [0,1]$, and choose a label uniformly at random with probability $1-\eta$.

We then assume that the pair interactions between two nodes i and j are Markov processes depending only on the community labelling and on some degree correction parameters $\theta = (\theta_1, \cdots, \theta_N)$. In particular,

$$\mathbb{P}(A \,|\, \sigma, \theta) = \prod_{1 \le i < j \le N} \mathbb{P}\left(A_{ij}^1 \,|\, \sigma_i^1, \sigma_j^1, \theta_i, \theta_j\right) \prod_{t=2}^{T} \mathbb{P}\left(A_{ij}^t \,|\, A_{ij}^{t-1}, \sigma_i^t, \sigma_j^t, \theta_i, \theta_j\right). \tag{6.4.3}$$

We consider a homogeneous model in which the initial distribution is given by

$$\mathbb{P}\left(A_{ij}^1 \,|\, Z_{i1}, Z_{j1}, \theta_i, \theta_j\right) = \begin{cases} \mu^{\theta_i\theta_j}\left(A_{ij}^1\right) & \text{if } \sigma_i^1 = \sigma_j^1 \\ \nu^{\theta_i\theta_j}\left(A_{ij}^1\right) & \text{otherwise,} \end{cases} \tag{6.4.4}$$

and the transition probabilities are

$$\mathbb{P}\left(A_{ij}^t = b \,|\, A_{ij}^{t-1} = a, \sigma_i^t, \sigma_j^t, \theta_i, \theta_j\right) = \begin{cases} P_{ab}^{\theta_i\theta_j} & \text{if } \sigma_i^t = \sigma_j^t \\ Q_{ab}^{\theta_i\theta_j} & \text{otherwise.} \end{cases} \tag{6.4.5}$$

Similarly to Section 6.3.1, the degree-corrected initial distributions are defined by

$$\mu^{\theta_i\theta_j} = \begin{pmatrix} 1 - \theta_i\theta_j\mu_1 \\ \theta_i\theta_j\mu_1 \end{pmatrix} \quad \text{and} \quad \nu^{\theta_i\theta_j} = \begin{pmatrix} 1 - \theta_i\theta_j\nu_1 \\ \theta_i\theta_j\nu_1 \end{pmatrix}, \tag{6.4.6}$$

and transition probability matrices

$$P^{\theta_i\theta_j} = \begin{pmatrix} 1-\theta_i\theta_j P_{01} & \theta_i\theta_j P_{01} \\ 1-P_{11} & P_{11} \end{pmatrix} \quad \text{and} \quad Q^{\theta_i\theta_j} = \begin{pmatrix} 1-\theta_i\theta_j Q_{01} & \theta_i\theta_j Q_{01} \\ 1-Q_{11} & Q_{11} \end{pmatrix}, \tag{6.4.7}$$

with the assumption that $\min_{i,j}\{\theta_i\theta_j\delta\} \leq 1$, where $\delta = \max\{\mu_1, \nu_1, P_{01}, Q_{01}\}$. We normalise the degree correction parameters so that $\sum_i 1(\sigma_i^1 = k)\theta_i = \sum_i 1(\sigma_i^1 = k)$ for all k. Finally, we suppose that the transition probabilities and degree-corrected parameters do not vary with time, to avoid any parameter identifiability issues [MM17].

6.4.2 Online inference as a semi-supervised problem

Time-varying community memberships leads to a contamination of the information given by the past interactions. Indeed, if node i changes its community assignment at time t_1, then one shall not use the interactions of node i during the first t_1 to find its community membership at time $t > t_1$. This *lagging problem* occurs when the layers are temporally correlated and renders the clustering harder. To avoid this issue, we propose an online recovery of the node labels. More specifically:

- at time $t = 1$, we use a static community detection algorithm to output $\hat{\sigma}^1$, a prediction of the initial node labels σ^1 from the observation of the first snapshot A^1;
- at time $t > 1$, we will use the observation of the first t snapshots $A^1, \ldots, A^t$ as well as the previous predictions $\hat{\sigma}^1, \cdots \hat{\sigma}^{t-1}$. This will be treated as a semi-supervised learning problem, where the prediction $\hat{\sigma}^{t-1}$ is seen as a noisy oracle for the true node labelling σ^t.

From the Markov structure, the prediction at time $t > 1$ reduces to predicting σ^t using only the network at time $t-1$ and t and the previous prediction $\hat{\sigma}^{t-1}$. This can be interpreted as a noisy semi-supervised problem, where the previous prediction $\hat{\sigma}^{t-1}$ plays the role of an oracle for the node labels at time t. This oracle is noisy, as it bears two kinds of potential mistakes. Firstly, $\hat{\sigma}^{t-1}$ is not necessarily exactly equal to the perfect community labelling σ^{t-1}. Furthermore, since the node labels vary through time, σ^{t-1} does not precisely correspond to σ^t. Assume that the network data A and community labels σ come from the model describe in Section 6.4.1. We define the rate of mistake of the oracle $\hat{\sigma}^{t-1}$ as

$$\rho_t \;=\; \mathbb{P}\left(\sigma_i^t \neq \hat{\sigma}_i^{t-1}\right).$$

The following Proposition gives the expression of the MAP estimator for this online learning problem.

Proposition 22. *Let $s \in [K]^N$ be a noisy oracle on the node labels at time t, which is supposed to be independent of the observed interactions A. Define the rate of mistake of s as $\rho = \mathbb{P}(s_i \neq \sigma_i)$ and assume this rate is the same for all nodes. A Maximum A Posteriori estimator for the online learning problem described previously is defined by*

$$\hat{\sigma}^t \;=\; \underset{\sigma\in[K]^N}{\arg\max}\, \mathbb{P}\left(\sigma \mid A^t, A^{t-1}, s\right)$$

and is any labelling that maximises

$$\sum_{i,j} \delta(\sigma_i, \sigma_j) \left\{ \ell_{01}^{\theta_i\theta_j} \left(A_{ij}^t - A_{ij}^{t-1} A_{ij}^t \right) + \ell_{10}^{\theta_i\theta_j} \left(A_{ij}^{t-1} - A_{ij}^{t-1} A_{ij}^t \right) + \ell_{11}^{\theta_i\theta_j} A_{ij}^{t-1} A_{ij}^t - \log \frac{Q_{00}^{\theta_i\theta_j}}{P_{00}^{\theta_i\theta_j}} \right\}$$
$$+2\lambda \sum_{i=1}^{N} 1(\sigma_i = s_i)$$

where $\ell_{ab}^{\theta_i\theta_j} = \log \frac{P_{ab}^{\theta_i\theta_j}}{P_{ab}^{\theta_i\theta_j}} - \log \frac{P_{00}^{\theta_i\theta_j}}{P_{00}^{\theta_i\theta_j}}$ *and* $\lambda = \log \frac{1-\rho}{\rho}$.

Proof. By Bayes' rule, $\mathbb{P}(\sigma \mid A^t, A^{t-1}, s, \theta) \propto \mathbb{P}(A^t \mid A^{t-1}, \sigma, s, \theta)\, \mathbb{P}(\sigma \mid A^{t-1}, s, \theta)$ where the proportionality symbol hides a term $\mathbb{P}(A^t \mid A^{t-1}, s, \theta)$ independent of σ.

Since $\mathbb{P}(A^t \mid A^{t-1}, \sigma, s, \theta) = \mathbb{P}(A^t \mid A^{t-1}, \sigma, \theta)$, then by proceeding similarly to the proof of Proposition 21, the log-likelihood term $\log \mathbb{P}(A^t \mid A^{t-1}, \sigma, \theta)$ can be rewritten as

$$\frac{1}{2} \sum_{i,j} \delta(\sigma_i, \sigma_j) \left\{ \ell_{01}^{\theta_i\theta_j} \left(A_{ij}^t - A_{ij}^{t-1} A_{ij}^t \right) + \ell_{10}^{\theta_i\theta_j} \left(A_{ij}^{t-1} - A_{ij}^{t-1} A_{ij}^t \right) + \ell_{11}^{\theta_i\theta_j} A_{ij}^{t-1} A_{ij}^t - \log \frac{Q_{00}^{\theta_i\theta_j}}{P_{00}^{\theta_i\theta_j}} \right\}.$$

The oracle information is equal to

$$\begin{aligned}
\mathbb{P}(\sigma \mid s) &= \prod_{i=1}^{N} \frac{\mathbb{P}(s_i \mid \sigma_i)}{\mathbb{P}(s_i)} \mathbb{P}(\sigma_i) \\
&= (1-\rho)^{|\{i \in [N] \colon \sigma_i = s_i\}|} \rho^{|\{i \in [N] \colon \sigma_i \neq s_i\}|} \left(\frac{1}{K} \right)^N \\
&= \left(\frac{\rho}{1-\rho} \right)^{|\{i \in [N] \colon \sigma_i \neq s_i\}|} (1-\rho)^N \left(\frac{1}{K} \right)^N
\end{aligned}$$

where we used the uniformity of the node labels. □

6.4.3 Continuous relaxation of the MAP

For simplicity of the derivations to come, in this section we restrict the study to $K = 2$.

Denote by $A_{\text{pers}}^t = A^{t-1} \odot A^t$ the *persisting edges*, by $A_{\text{new}} = A^t - A_{\text{pers}}^t$ the *freshly formed* edges, and by $A_{\text{old}} = A^{t-1} - A_{\text{pers}}$ the *disappearing edges* between time $t-1$ and t. Using a Taylor expansion as in Section 6.3.1, we can approximate the MAP estimator as the maximisation over $x \in \{-1, 1\}^N$ of

$$x^T \left(W - \tau \frac{dd^T}{2m} \right) x - \lambda (s - x)^T (s - x) \tag{6.4.8}$$

where $W^t = \alpha_{01} A_{\text{new}}^t + \alpha_{10} A_{\text{old}}^t + \alpha_{11} A_{\text{pers}}^t$ with $\alpha_{ab} = \log \frac{P_{ab}}{Q_{ab}}$, τ is a resolution parameter $d_i = \sum_{j=1}^N W_{ij}^t$, and $m = \frac{1}{2} \sum_{i=1}^N d_i$. The maximisation of Expression (6.4.8) becomes

equivalent to $\hat{\sigma}^t = \mathrm{sign}(\hat{z})$ where $\hat{z}$ verifies

$$\hat{z} \;=\; \underset{z\in\{-1,1\}^N}{\arg\min}\; -z^T M z + \lambda\,(s-z)^T\,(s-z)$$

with $M = W - \tau\frac{dd^T}{2m}$. This minimisation problem is analogous to the one studied in Chapter 5 on noisy semi-supervised clustering in the SBM. Indeed, we propose the following continuous relaxation

$$\hat{x} \;=\; \underset{\substack{x\in\mathbb{R}^N\\ x^T D x = 2m}}{\arg\min}\; -x^T M x + \lambda\,(s-x)^T\,(s-x)\,,$$

where $D = \mathrm{diag}(d_1,\cdots,d_N)$ and $m = \frac{1}{2}\sum_{i=1}^N d_i$. The solution of this relaxation is computed by mimicking the reasoning of Section 5.2.1. In particular, by denoting the eigen-decomposition of $D^{-1/2}\,(-M+\lambda I_N)\,D^{-1/2}$ by

$$D^{-1/2}\,(-M+\lambda I_N)\,D^{-1/2} \;=\; Q\Delta Q^T$$

with $\Delta = \mathrm{diag}(\delta_1,\ldots,\delta_N)$ and $QQ^T = I_N$ and letting $b = \lambda Q^t s$, we obtain that $\hat{x}$ verifies

$$(-M + \lambda I_N - \gamma_* D)\,\hat{x} \;=\; \lambda s, \tag{6.4.9}$$

where γ_* is the smallest solution of the *explicit secular equation* [GGV89]

$$\sum_{i=1}^n \left(\frac{b_i}{\delta_i - \gamma}\right)^2 - 2m \;=\; 0. \tag{6.4.10}$$

This leads to Algorithm 10.

Algorithm 10: Online clustering of time-varying communities.

Input: Observed graph sequence $A^{1:T} = \left(A^1,\ldots,A^T\right)$; number of communities K; static graph clustering algorithm `algo`; parameters $\alpha_{01},\alpha_{10},\alpha_{11}$ and $\lambda_1,\ldots,\lambda_T$.

Output: Node labelling $\hat{\sigma}^{1:T}$.

Initialize: Compute $\hat{\sigma}^1 \leftarrow \mathtt{algo}\,(A^1)$.

1 **for** $t = 2,\ldots,T$ **do**

2 Compute $W = \alpha_{01}A^t_{\text{new}} + \alpha_{10}A^t_{\text{old}} + \alpha_{11}A^t_{\text{pers}}$.

3 Compute $M = W - \frac{dd^T}{2m}$ where $d_i = \sum_{j=1}^N W_{ij}$ and $m = \frac{1}{2}\sum_{i=1}^N d_i$.

4 Let γ^* be the smallest solution of Equation (6.4.10).

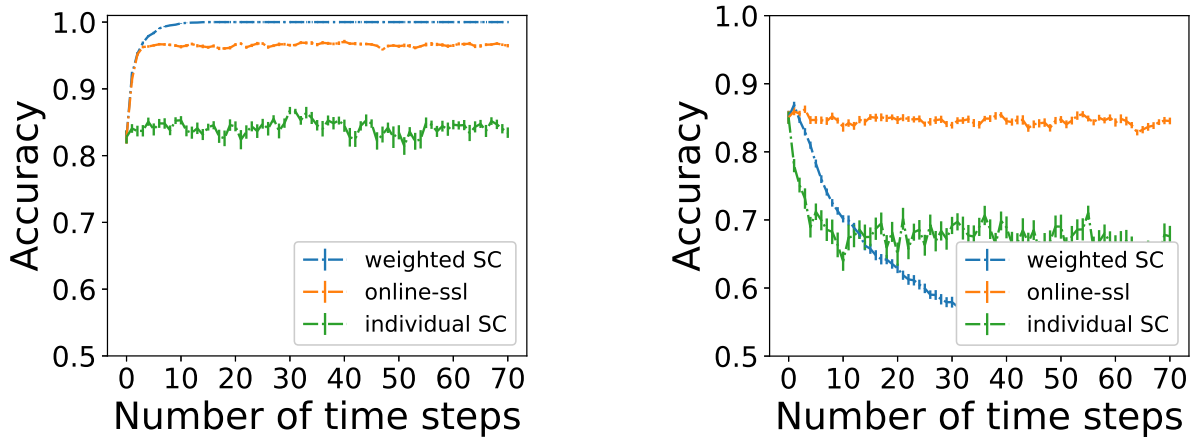
5 Compute $\hat{x}$ as the solution of Equation (6.4.9).

6 Let $\hat{\sigma}^t = \mathrm{sign}(\hat{x})$.

6.4.4 Numerical experiments

We compare in Figure 6.8 the averaged accuracy obtained by Algorithm 10 with Algorithm 9 (spectral clustering with persistent edges) and an algorithm performing spectral clustering on each snapshot individually. In particular, we observe that when $\eta = 1$ (*i.e.,* static community structure), Algorithm 9 is extremely efficient, as expected. Since it takes into account all previous snapshots, it in particular outperforms Algorithm 10. On the contrary, when $\eta \neq 1$, the lagging problem arises, and Algorithm 9 ends up with a very poor accuracy after a few snapshots. On the contrary, Algorithm 10 keeps a very high accuracy over all snapshots.

In Figure 6.8, we choose λ_t to be constant and equal to 0.5, while Figure 6.9 explores other possible values. We observe that when λ_t is equal to a constant in the interval $[0.1, 1]$, Algorithm 10 outputs similar performances. On the other hand, when λ becomes too large, Algorithm 10 gives too much importance to the oracle, and the accuracy becomes worse. In practice, the choice of the parameters λ_t could be optimized from the data, e.g. based on η or on the transition matrices P and Q. Moreover, it would be intuitive to increase λ_t with t, as the confidence in the oracle is higher when more temporal data is available. We leave this as future work.



(a) $\eta = 1$ (b) $\eta = 0.85$

Figure 6.8: Accuracy of Algorithm 10 (*online-ssl*) with $\alpha_{01} = 1, \alpha_{10} = 0$ and $\alpha_{11} = 2$, on time-varying Markov Block Models with 300 nodes and $K = 2$ blocks (with uniform prior), and a stationary Markov edge evolution $\mu_1 = 0.05$, $\nu_1 = 0.02$, $P_{11} = 0.7$ and $Q_{11} = 0.3$. The results are averaged over 25 synthetic graphs, and error bars show the standard error. We compare with Algorithm 5 (*weighted SC* with $\alpha = 1, \beta = 2$) and an algorithm performing Spectral Clustering on each snapshot individually.

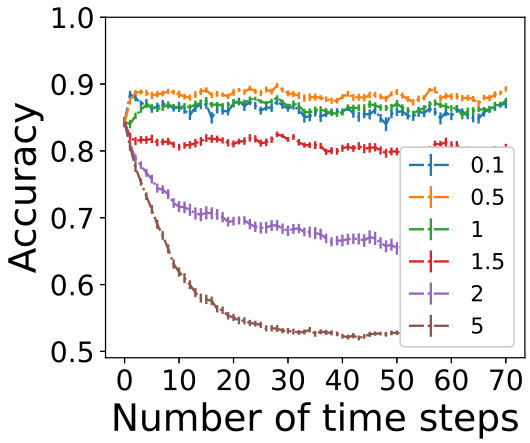


Figure 6.9: Accuracy of Algorithm 10 with $\alpha_{01} = 1$, $\alpha_{10} = 0$ and $\alpha_{22} = 2$ and various values of λ. Simulations are performed on time-varying Markov Block Models with $n = 300$, $K = 2$, $\mu_1 = 0.05$, $\mu_2 = 0.02$, $P_{11} = 0.7$, $Q_{11} = 0.3$ and $\eta = 0.9$. The results are averaged over 25 synthetic graphs, and error bars show the standard error.

Chapter 7

Higher-Order Spectral Clustering for Geometric Graphs

The present Chapter is devoted to clustering geometric graphs. While the standard spectral clustering is often not effective for geometric graphs, we present here an effective generalization, called *higher-order spectral clustering*. It resembles in concept the classical spectral

clustering method but uses for partitioning the eigenvector associated with a higher-order eigenvalue. We establish the consistency of this algorithm for a wide class of geometric graphs which we call Soft Geometric Block Model. A small adjustment of the algorithm provides strong consistency. We also show that our method is effective in numerical experiments even for graphs of modest size.

7.1 Model definition and notations

7.1.1 Notations

Let $\mathbb{T}^d = \mathbb{R}^d/\mathbb{Z}^d$ be the flat unit torus in dimension d represented by $\left[-\frac{1}{2}, \frac{1}{2}\right]^d$. The norm ℓ_∞ in $\mathbb{R}^d$ naturally induces a norm on $\mathbb{T}^d$ such that for any vector $x = (x_1, \dots, x_d) \in \mathbb{T}^d$ we have $\|x\| = \max\limits_{1\le i\le d} |x_i|$.

For a measurable function $F : \mathbb{T}^d \to \mathbb{R}$ and $k \in \mathbb{Z}^d$, we denote $\hat{F}(k) = \int_{\mathbb{T}^d} F(x) e^{-2i\pi\langle k,x\rangle}\, dx$ the Fourier transform of F. The Fourier series of F is given by

$$\sum_{k\in\mathbb{Z}^d} \hat{F}(k) e^{2i\pi\langle k,x\rangle}.$$

For two integrable functions F, $G : \mathbb{T}^d \to \mathbb{R}$, we define the convolution operation $F * G(y) = \int_{\mathbb{T}^d} F(y-x)G(x)\, dx$ and $F^{*m} = F*F*\cdots*F$ (m times). We recall that $\widehat{F * G}(k) = \hat{F}(k)\hat{G}(k)$.

7.1.2 Soft Geometric Block Model

A Soft Geometric Block Model (SGBM) is defined by a dimension d, a number of nodes n, a set of blocks K and a connectivity probability function $F : \mathbb{T}^d \times K \times K \to [0, 1]$. The node set is taken as $V = [n]$. The model is parametrized by a *node labelling* $\sigma : V \to K$ and *nodes' positions* $X = (X_1, \dots, X_n) \in \left(\mathbb{T}^d\right)^n$. We suppose that $F(\cdot, \sigma, \sigma') = F(\cdot, \sigma', \sigma)$ and for any $X \in \mathbb{T}^d$, $F(X)$ depends only on the norm $\|X\|$. The probability of appearance of an edge between nodes i and j is defined by $F(X_i - X_j, \sigma_i, \sigma_j)$. Note that this probability depends only on the distance between X_i and X_j and the labels σ_i, σ_j. Consequently, the model parameters specify the distribution

$$\mathbb{P}_{\sigma,X}(A) = \prod_{1\le i<j\le n} \left(F\left(X_i - X_j, \sigma_i, \sigma_j\right)\right)^{A_{ij}} \left(1 - F\left(X_i - X_j, \sigma_i, \sigma_j\right)\right)^{1-A_{ij}} \tag{7.1.1}$$

of the adjacency matrix $A = (A_{ij})$ of a random graph.

Furthermore, in this Chapter we assume that the model has only two equal size blocks, i. e., $K = \{1, 2\}$, and $\sum_{i=1}^n 1(\sigma_i = 1) = \sum_{i=1}^n 1(\sigma_i = 2) = \frac{n}{2}$. The labels are assigned randomly, that is, the set $\{i \in [n] \colon \sigma_i = 1\}$ is chosen randomly over all the $\frac{n}{2}$-subsets of $[n]$. We assume

that the entries of X and σ are independent and $\forall i \in V$, X_i is uniformly distributed over $\mathbb{T}^d$. Finally, suppose that for any $x \in \mathbb{T}^d$

$$F(x, \sigma, \sigma') \;=\; \begin{cases} F_{\text{in}}(x), & \text{if} \quad \sigma = \sigma', \\ F_{\text{out}}(x), & \text{otherwise,} \end{cases} \tag{7.1.2}$$

where $F_{\text{in}}, F_{\text{out}} : \mathbb{T}^d \to [0,1]$ are two measurable functions. We call these functions *connectivity probability functions*.

The average intra- and inter-community *edge densities* are denoted by μ_{in} and μ_{out}, respectively. Their expressions are given by the first Fourier modes of F_{in} and F_{out}:

$$\mu_{\text{in}} \;=\; \int_{\mathbb{T}^d} F_{\text{in}}(x)dx \quad \text{and} \quad \mu_{\text{out}} \;=\; \int_{\mathbb{T}^d} F_{\text{out}}(x)dx.$$

These quantities will play an important role in the following, as they represent the intensities of interactions between nodes in the same community and nodes in different communities. In particular, the average inside community degree is $\left(\frac{n}{2} - 1\right)\mu_{\text{in}}$, and the average outside community degree is $\frac{n}{2}\mu_{\text{out}}$.

Example 13. An SGBM where $F_{\text{in}}(x) = p_{\text{in}}$ and $F_{\text{out}}(x) = p_{\text{out}}$ with $p_{\text{in}}, p_{\text{out}}$ being constants is an instance of the Stochastic Block Model.

Example 14. An SGBM where $F_{\text{in}}(x) = 1(\|x\| \leq r_{\text{in}})$, $F_{\text{out}}(x) = 1(\|x\| \leq r_{\text{out}})$ with $r_{\text{in}}, r_{\text{out}} \in \mathbb{R}_+$ is an instance of the Geometric Block Model introduced in [Gal+18].

Example 15. We call Waxman Block Model (WBM) an SGBM with $F_{\text{in}}(x) = \min(1, q_{\text{in}}\mathrm{e}^{-s_{\text{in}}\|x\|})$, $F_{\text{out}}(x) = \min(1, q_{\text{out}}\mathrm{e}^{-s_{\text{out}}\|x\|})$. This is a clustered version of the Waxman model [Wax88], which is a particular case of soft geometric random graphs [Pen16].

Formally, *clustering* or *community recovery problem* is the following problem: given the observation of the adjacency matrix A and the knowledge of $F_{\text{in}}, F_{\text{out}}$, we want to recover the latent community labelling σ. Given an estimator $\hat{\sigma}$ of σ, we define the loss ℓ as the ratio of misclassified nodes, up to a global permutation of the labels:

$$\ell\left(\sigma, \hat{\sigma}\right) \;=\; \frac{1}{n} \min_{\pi \in \mathcal{S}_2} \sum_i 1\left(\sigma_i \neq \pi \circ \hat{\sigma}_i\right).$$

Then, $\hat{\sigma}$ is said to be consistent (or equivalently, achieves almost exact recovery) if

$$\forall \epsilon > 0 : \quad \lim_{n \to \infty} \mathbb{P}\left(\ell\left(\sigma, \hat{\sigma}\right) > \epsilon\right) \;=\; 0,$$

and strongly consistent (equivalently, achieves exact recovery) if

$$\lim_{n \to \infty} \mathbb{P}\left(\ell\left(\sigma, \hat{\sigma}\right) > 0\right) \;=\; 0.$$

7.2 The analysis of limiting spectrum

7.2.1 Limit of the spectral measure

Theorem 11. *Consider an SGBM defined by (7.1.1)-(7.1.2). Assume that $F_{\text{in}}(0)$ and $F_{\text{out}}(0)$ are equal to the Fourier series of $F_{\text{in}}(\cdot)$ and $F_{\text{out}}(\cdot)$ evaluated at 0. Let $\lambda_1, \dots, \lambda_n$ be the eigenvalues of A, and*

$$\mu_n \;=\; \sum_{i=1}^{n} \delta_{\lambda_i/n}$$

the spectral measure of the matrix $\frac{1}{n}A$. Then, for all Borel sets B with $\mu(\partial B) = 0$ and $0 \notin \bar{B}$, a.s.,

$$\lim_{n\to\infty} \mu_n(B) = \mu(B),$$

where μ is the following measure:

$$\mu \;=\; \sum_{k\in\mathbb{Z}^d} \delta_{\frac{\widehat{F}_{\text{in}}(k)+\widehat{F}_{\text{out}}(k)}{2}} \;+\; \delta_{\frac{\widehat{F}_{\text{in}}(k)-\widehat{F}_{\text{out}}(k)}{2}}.$$

Remark 17. The limiting measure μ is composed of two terms. The first term, $\sum_{k\in\mathbb{Z}^d} \delta_{\frac{\widehat{F}_{\text{in}}(k)+\widehat{F}_{\text{out}}(k)}{2}}$ corresponds to the spectrum of a random graph with no community structure, and where edges between two nodes at distance x are drawn with probability $\frac{F_{\text{in}}(x)+F_{\text{out}}(x)}{2}$. In other words, it is the null-model of the considered SGBM. Hence, the eigenvectors associated with those eigenvalues bear no community information, but only geometric features.

On the contrary, the second term $\sum_{k\in\mathbb{Z}^d} \delta_{\frac{\widehat{F}_{\text{in}}(k)-\widehat{F}_{\text{out}}(k)}{2}}$ corresponds to the difference between intra- and inter-community edges. In particular, as we shall see later, the ideal eigenvector for clustering is associated with the eigenvalue $\widetilde{\lambda}$ closest to $\lambda_* = n\frac{\widehat{F}_{\text{in}}(0)-\widehat{F}_{\text{out}}(0)}{2}$. Other eigenvectors might mix some geometric and community features and hence are harder to analyze.

Last, the eigenvalue $\widetilde{\lambda}$ is not necessarily the second largest eigenvalue, as the ordering of eigenvalues here depends on the Fourier coefficients $\widehat{F}_{\text{in}}(k)$ and $\widehat{F}_{\text{out}}(k)$, and is in general non trivial.

Remark 18. The assumptions on $F_{\text{in}}(0)$ and $F_{\text{out}}(0)$ are validated for a wide range of reasonable connectivity functions. For instance, by Dini's criterion, all the functions that are differentiable at 0 satisfy these conditions. Another appropriate class consists of piecewise C^1 functions that are continuous at 0 (this follows from the Dirichlet conditions).

Proof. The outline of the proof of Theorem 11 follows closely [Bor08]. First, we show that $\forall m \in \mathbb{N}$, $\lim_{n\to\infty} \mathbb{E}\,\mu_n(P_m) = \mu(P_m)$ where $P_m(t) = t^m$. Second, we use Talagrand's concentration inequality to prove that $\mu_n(P_m)$ is not far from its mean, and conclude with Borel–Cantelli lemma.

(i) By Lemma 30 in Section 7.6, in order to establish the desired convergence it is enough to show that $\lim_{n\to\infty} \mathbb{E}\mu_n(P_m) = \mu(P_m)$ for any $m \in \mathbb{N}$. First,

$$\mathbb{E}\mu_n(P_m) = \frac{1}{n^m}\sum_{i=1}^{n}\mathbb{E}\lambda_i^m = \frac{1}{n^m}\mathbb{E}\operatorname{Tr}A^m. \tag{7.2.1}$$

By definition,

$$\operatorname{Tr}A^m = \sum_{\alpha\in[n]^m}\prod_{j=1}^{m}A_{i_j,i_{j+1}},$$

with $\alpha = (i_1,\dots,i_m) \in [n]^m$ and $i_{m+1} = i_1$. We denote $\mathcal{A}_n^m$ the set of m-permutations of $[n]$, that is $\alpha \in \mathcal{A}_n^m$ iff α is an m-tuple without repetition. We have,

$$\operatorname{Tr}A^m = \sum_{\alpha\in\mathcal{A}_n^m}\prod_{j=1}^{m}A_{i_j,i_{j+1}} + R_m, \tag{7.2.2}$$

where $R_m = \sum_{\alpha\in[n]^m\setminus\mathcal{A}_n^m}\prod_{j=1}^{m}A_{i_j,i_{j+1}}$.

We first bound the quantity R_m. Since $|A_{ij}| \le 1$, we have

$$|R_m| \le \left|[n]^m\setminus\mathcal{A}_n^m\right| = n^m - \frac{n!}{(n-m)!} = \frac{m(m-1)}{2}n^{m-1} + o(n^{m-1}),$$

where we used $\dfrac{n!}{(n-m)!} = n^m - n^{m-1}\sum_{i=0}^{m-1} i + o(n^{m-1})$. Hence

$$\lim_{n\to\infty}\frac{1}{n^m}R_m = 0. \tag{7.2.3}$$

Moreover,

$$\begin{aligned}\mathbb{E}\sum_{\alpha\in\mathcal{A}_n^m}\prod_{j=1}^{m}A_{i_j,i_{j+1}} &= \sum_{\alpha\in\mathcal{A}_n^m}\int_{(\mathbb{T}^d)^m}\prod_{j=1}^{m}F(x_{i_j}-x_{i_{j+1}},\sigma_{i_j},\sigma_{i_{j+1}})dx_{i_1}\dots dx_{i_m}\\ &= \sum_{\alpha\in\mathcal{A}_n^m}G(\alpha)\end{aligned}$$

where $G(\alpha) = \int_{(\mathbb{T}^d)^m}\prod_{j=1}^{m}F(x_{i_j}-x_{i_{j+1}},\sigma_{i_j},\sigma_{i_{j+1}})dx_{i_1}\dots dx_{i_m}$ for $\alpha\in\mathcal{A}_n^m$.

Let us first show that the value of $G(\alpha)$ depends only on the number of consecutive indices corresponding to the nodes from the same community. More precisely, let us define the set $\mathcal{S}(\alpha) = \{j \in [m] : \sigma_{i_j} = \sigma_{i_{j+1}}\}$. Using Lemma 31 in Section 7.6 and the fact that the convolution is commutative, we have

$$G(\alpha) = F_{\text{in}}^{*|\mathcal{S}(\alpha)|} * F_{\text{out}}^{*(m-|\mathcal{S}(\alpha)|)}(0).$$

We introduce the following equivalence relationship in $\mathcal{A}_n^m$: $\alpha \sim \alpha'$ if $|\mathcal{S}(\alpha)| = |\mathcal{S}(\alpha')|$. We notice than $G(\cdot)$ is constant on each equivalence class, and equals to $F_{\text{in}}^{*p} * F_{\text{out}}^{*(m-p)}(0)$ for any $\alpha \in \mathcal{A}_n^m$ such that $|\mathcal{S}(\alpha)| = p$.

Then, let us calculate the cardinal of each equivalence class with $|\mathcal{S}(\alpha)| = p$. First of all, we choose the set $\mathcal{S}(\alpha)$ which can be done in $\binom{m}{p}$ ways if $m-p$ is even and cannot be done if $m-p$ is odd. The latter follows from the fact that p (the number of 'non-changes' in the consecutive community labels) has the same parity as m (the total number of indices) since $i_{m+1} = i_1$. The set $\mathcal{S}(\alpha)$ defines the community labels up to the flip of communities since $\sigma_{i_j} = \sigma_{i_{j+1}}$ for any $j \in \mathcal{S}(\alpha)$ and $\sigma_{i_j} \neq \sigma_{i_{j+1}}$ for $j \in [m] \backslash \mathcal{S}(\alpha)$.

Let $N_1(\alpha)$ be the number of indices i_j with $\sigma_{i_j} = 1$. Consider first the case $\sigma_{i_1} = 1$ and note that $N_1(\alpha)$ is totally defined by the set $\mathcal{S}(\alpha)$. There are $\frac{n}{2}$ possible choices for i_1. Now we have two possibilities. If $\sigma_{i_1} = \sigma_{i_2}$ then we have $\frac{n}{2} - 1$ possible choices for the index i_2 (since $\alpha \in \mathcal{A}_n^m$). Otherwise, if $\sigma_{i_1} \neq \sigma_{i_2}$ then the index i_2 can be chosen in $\frac{n}{2}$ ways. Resuming the above operation, we choose $N_1(\alpha)$ indices from the first community, and it can be done in $n/2(n/2-1)\dots(n/2-N_1(\alpha))$ ways. The indices from the second community can be chosen in $n/2(n/2-1)\dots(n/2-(m-N_1(\alpha)))$ ways. Thus in total the number of possible choices of indices is

$$\begin{aligned}
&\frac{n}{2}\left(\frac{n}{2}-1\right)\dots\left(\frac{n}{2}-N_1(\alpha)\right)\cdot\frac{n}{2}\left(\frac{n}{2}-1\right)\dots\left(\frac{n}{2}-(m-N_1(\alpha))\right)\\
&= \frac{n^m}{2^m} + O(n^{m-1}), \quad n \to \infty.
\end{aligned}$$

The same reasoning applies if $\sigma_{i_1} = 2$. Hence, when n goes to infinity, the cardinal of each equivalence class is

$$|\{\alpha \in \mathcal{A}_n^m \colon |\mathcal{S}(\alpha)| = p\}| \;=\; \begin{cases} 0 & \text{if } m-p \text{ is odd,} \\ 2\binom{m}{p}\frac{n^m}{2^m} + O(n^{m-1}) & \text{otherwise.} \end{cases}$$

This can be rewritten as

$$|\{\alpha \in \mathcal{A}_n^m \colon |\mathcal{S}(\alpha)| = p\}| = \binom{m}{p}\left(1+(-1)^{m-p}\right)\frac{n^m}{2^m} + O(n^{m-1}), \quad n \to \infty.$$

Hence,

$$\begin{aligned}
\mathbb{E}\sum_{\alpha\in\mathcal{A}_n^m}\prod_{j=1}^{m} A_{i_j,i_{j+1}} &= \sum_{p=0}^{m} |\{\alpha \in \mathcal{A}_n^m \colon |\mathcal{S}(\alpha)| = p\}|\, F_{\text{in}}^{*p} * F_{\text{out}}^{*(m-p)}(0)\\
&= \frac{n^m}{2^m}\sum_{p=0}^{m}\binom{m}{p}\left(1+(-1)^{m-p}\right)F_{\text{in}}^{*p} * F_{\text{out}}^{*(m-p)}(0) + O(n^{m-1})\\
&= n^m\left(\left(\frac{F_{\text{in}}+F_{\text{out}}}{2}\right)^{*m}(0) + \left(\frac{F_{\text{in}}-F_{\text{out}}}{2}\right)^{*m}(0)\right) + O(n^{m-1}).
\end{aligned}$$

Therefore, equations (7.2.1), (7.2.2) and (7.2.3) give us:

$$\lim_{n\to\infty}\mathbb{E}\mu_n(P_m) \;=\; \left(\frac{F_{\text{in}}+F_{\text{out}}}{2}\right)^{*m}(0) + \left(\frac{F_{\text{in}}-F_{\text{out}}}{2}\right)^{*m}(0).$$

Finally, since $F_{\text{in}}, F_{\text{out}}$ are equal to their Fourier series at 0, and using $\widehat{F * G}(k) = \hat{F}(k)\hat{G}(k)$, we have

$$\lim_{n\to\infty} \mathbb{E}\mu_n(P_m) \;=\; \sum_{k\in\mathbb{Z}^d} \left(\frac{\widehat{F}_{\text{in}}(k) + \widehat{F}_{\text{out}}(k)}{2}\right)^m + \left(\frac{\widehat{F}_{\text{in}}(k) - \widehat{F}_{\text{out}}(k)}{2}\right)^m = \mu\left(P_m\right). \quad (7.2.4)$$

(ii) For each $m \geq 1$, and n fixed, we define

$$\begin{array}{rccc} Q_m: & SGBM(F_{\text{in}}, F_{\text{out}}) & \longrightarrow & \mathbb{R} \\ & A & \longmapsto & \frac{1}{n^{m-1}} \operatorname{Tr} A^m \end{array}$$

where $SGBM(F_{\text{in}}, F_{\text{out}})$ denotes the set of adjacency matrices of an SGBM random graph with connectivity functions $F_{\text{in}}, F_{\text{out}}$. Note that $Q_m(A) = n\mu_n(P_m)$.

Let $A, \widetilde{A}$ be two adjacency matrices. We denote the Hamming distance by $\mathrm{d}_{\text{Ham}}\left(A, \widetilde{A}\right) = \sum_{i=1}^n \sum_{j=1}^n 1(A_{ij} \neq \widetilde{A}_{ij})$. Using Lemma 34 in Section 7.6, we show that the function Q_m is (m/n)–Lipschitz for the Hamming distance:

$$\left|Q_m(A) - Q_m(\widetilde{A})\right| \;\leq\; \frac{m}{n}\, \mathrm{d}_{\text{Ham}}\left(A, \widetilde{A}\right). \quad (7.2.5)$$

Let M_m be the median of Q_m. Talagrand's concentration inequality [Tal96, Proposition 2.1] states that

$$\mathbb{P}\left(|Q_m - M_m| > t\right) \;\leq\; 4\exp\left(-\frac{n^2t^2}{4m^2}\right), \quad (7.2.6)$$

which after integrating over all t gives

$$|n\mathbb{E}\mu_n\left(P_m\right) - M_m| \;\leq\; \mathbb{E}\,|Q_m(A) - M_m| \;\leq\; \frac{C_m}{n},$$

since $\mathbb{E}X = \int_0^\infty \mathbb{P}(X > t)dt$ for any positive random variable X. The constant C_m is equal to $8m\int_0^\infty e^{-u^2}du$.

Moreover,

$$\begin{aligned} n\left|\mu_n(P_m) - \mathbb{E}\mu_n(P_m)\right| \;&\leq\; |n\mu_n(P_m) - M_m| + |M_m - n\mathbb{E}\mu_n(P_m)| \\ &\leq\; |Q_m - M_m| + \frac{C_m}{n}. \end{aligned}$$

Let $s > 0$. Since C_m/n^2 goes to 0 when n goes to infinity, we can pick n large enough such that $s > \frac{C_m}{n^2}$. Thus, using again inequality (7.2.6), we have

$$\begin{aligned} \mathbb{P}\left(|\mu_n(P_m) - \mathbb{E}\mu_n(P_m)| > s\right) \;&\leq\; \mathbb{P}\left(\frac{1}{n}\,|Q_m - M_m| > s - \frac{C_m}{n^2}\right) \\ &\leq\; 4\exp\left(-\frac{n^4}{4m^2}\left(s - \frac{C_m}{n^2}\right)^2\right). \end{aligned}$$

However, by (7.2.4), $\lim_{n\to\infty} \mathbb{E}\mu_n(P_m) = \mu(P_m)$. Hence $\mu_n(P_m)$ converges in probability to $\mu(P_m)$. Let $s_n = \frac{1}{n^\kappa}$ with $\kappa > 0$, and

$$\epsilon_n = 4\exp\left(-\frac{n^4}{4m^2}\left(s_n - \frac{C_m}{n^2}\right)^2\right).$$

Since $\sum_{n\in\mathbb{N}} \epsilon_n < \infty$ when $\kappa < 2$, an application of Borel–Cantelli lemma shows that the convergence holds in fact almost surely. This concludes the proof. □

7.2.2 Conditions for the isolation of the ideal eigenvalue

As noticed in Remark 17, the ideal eigenvector for clustering is associated with the eigenvalue of the adjacency matrix A closest to $n\frac{\mu_{\text{in}}-\mu_{\text{out}}}{2}$ (recall that $\mu_{\text{in}} = \widehat{F}_{\text{in}}(0)$ and $\mu_{\text{out}} = \widehat{F}_{\text{out}}(0)$). The following proposition shows that this ideal eigenvalue is isolated from other eigenvalues under certain conditions.

Proposition 23. *Consider the adjacency matrix A of an SGBM defined by (7.1.1)-(7.1.2), and assume that:*

$$\widehat{F}_{\text{in}}(k) + \widehat{F}_{\text{out}}(k) \neq \mu_{\text{in}} - \mu_{\text{out}}, \qquad \forall k \in \mathbb{Z}^d, \tag{7.2.7}$$

$$\widehat{F}_{\text{in}}(k) - \widehat{F}_{\text{out}}(k) \neq \mu_{\text{in}} - \mu_{\text{out}}, \qquad \forall k \in \mathbb{Z}^d\backslash\{0\}, \tag{7.2.8}$$

with $\mu_{\text{in}} \neq \mu_{\text{out}}$. Then, the eigenvalue of A closest to $n\frac{\mu_{\text{in}}-\mu_{\text{out}}}{2}$ is of multiplicity one. Moreover, there exists $\epsilon > 0$ such that for large enough n every other eigenvalue is at a distance at least ϵn.

Proof. Let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of A. Let $i^* \in \arg\min_{i\in[n]} \left|\frac{\lambda_i}{n} - \frac{\mu_{\text{in}}-\mu_{\text{out}}}{2}\right|$. We will show that there exists $\epsilon > 0$ such that for large enough n, we have for all $i \neq i^*$:

$$\left|\frac{\lambda_i}{n} - \frac{\mu_{\text{in}} - \mu_{\text{out}}}{2}\right| > \epsilon.$$

Due to condition (7.2.7), and the fact that

$$\lim_{|k|\to\infty} \left(\widehat{F}_{\text{in}}(k) + \widehat{F}_{\text{out}}(k)\right) = 0,$$

there is some fixed $\epsilon_1 > 0$ such that

$$\min_{k\in\mathbb{Z}^d} \left(\left|\frac{\widehat{F}_{\text{in}}(k) + \widehat{F}_{\text{out}}(k)}{2} - \frac{\mu_{\text{in}} - \mu_{\text{out}}}{2}\right|\right) > \epsilon_1.$$

Similarly, condition (7.2.8) ensures the existence of $\epsilon_2 > 0$ such that

$$\min_{k\in\mathbb{Z}^d\backslash\{0\}} \left(\left|\frac{\widehat{F}_{\text{in}}(k) - \widehat{F}_{\text{out}}(k)}{2} - \frac{\mu_{\text{in}} - \mu_{\text{out}}}{2}\right|\right) > \epsilon_2.$$

Denote $\epsilon_3 = \frac{|\mu_{\text{in}} - \mu_{\text{out}}|}{4}$. Let $\epsilon = \min(\epsilon_1, \epsilon_2, \epsilon_3)$ and consider $B = \left[\frac{\mu_{\text{in}} - \mu_{\text{out}}}{2} - \epsilon, \frac{\mu_{\text{in}} - \mu_{\text{out}}}{2} + \epsilon\right]$. By Theorem 11, a.s.,

$$\lim_{n\to\infty} \mu_n(B) = \mu(B) = 1.$$

Therefore, for n large enough the only eigenvalue of A in the interval B is λ_{i^*}. □

The following proposition shows that conditions (7.2.7) and (7.2.8) of Proposition 23 are almost always verified for a GBM.

Proposition 24. *Consider the d-dimensional GBM model, where $F_{\text{in}}, F_{\text{out}}$ are 1-periodic, and defined on the flat torus $\mathbb{T}^d$ by $F_{\text{in}}(x) = 1(\|x\| \le r_{\text{in}})$ and $F_{\text{out}}(x) = 1(\|x\| \le r_{\text{out}})$, with $r_{\text{in}} > r_{\text{out}} > 0$. Denote by $\mathcal{B}_+$ and $\mathcal{B}_-$ the sets of parameters r_{in} and r_{out} defined by negation of conditions (7.2.7) and (7.2.8):*

$$\mathcal{B}^+ = \left\{(r_{\text{in}}, r_{\text{out}}) \in \mathbb{R}_+^2 : \widehat{F}_{\text{in}}(k) + \widehat{F}_{\text{out}}(k) = \mu_{\text{in}} - \mu_{\text{out}} \text{ for some } k \in \mathbb{Z}^d\right\}$$
$$\mathcal{B}^- = \left\{(r_{\text{in}}, r_{\text{out}}) \in \mathbb{R}_+^2 : \widehat{F}_{\text{in}}(k) - \widehat{F}_{\text{out}}(k) = \mu_{\text{in}} - \mu_{\text{out}} \text{ for some } k \in \mathbb{Z}^d \backslash \{0\}\right\}.$$

Then these sets of 'bad' parameters are of zero Lebesgue measure:

$$\text{Leb}(\mathcal{B}^+) = 0; \qquad \text{and} \qquad \text{Leb}(\mathcal{B}^-) = 0.$$

Hence for $\mathcal{B} = \mathcal{B}^+ \cup \mathcal{B}^-$

$$\text{Leb}(\mathcal{B}) = 0.$$

Proof. It is clear that

$$\text{Leb}(\mathcal{B}) \le \text{Leb}(\mathcal{B}^+) + \text{Leb}(\mathcal{B}^-).$$

Thus, it is enough to show that $\text{Leb}(\mathcal{B}^+) = 0$ and $\text{Leb}(\mathcal{B}^-) = 0$. We shall establish the first equality, and the second equality can be proved similarly.

By Lemma 32 in Section 7.6, the negation of condition (7.2.7) for given functions F_{in} and F_{out} is as follows:

$$\exists k = (k_1, \cdots, k_d) \in \mathbb{Z}^d : \; r_{\text{in}}^d \prod_{j=1}^d \text{sinc}(2\pi r_{\text{in}} k_j) + r_{\text{out}}^d \prod_{j=1}^d \text{sinc}(2\pi r_{\text{out}} k_j) = r_{\text{in}}^d - r_{\text{out}}^d,$$

where $\text{sinc}(x) = \begin{cases} \frac{\sin x}{x} & \text{if } x \neq 0 \\ 0 & \text{otherwise} \end{cases}$ is the sinus cardinal function.

Notice that $\lim_{k_j \to \infty} \text{sinc}(2\pi r_{\text{in}} k_j) = 0$ and $\lim_{k_j \to \infty} \text{sinc}(2\pi r_{\text{out}} k_j) = 0$ while the right-hand side of the above equation is fixed. Therefore, this equation can hold only for k from a finite set $\mathcal{K}$. Let us fix some $k = (k_1, \ldots, k_d) \in \mathcal{K}$ and denote

$$\mathcal{B}_k^+ = \left\{(r_{\text{in}}, r_{\text{out}}) \in \mathbb{R}_+^2 : r_{\text{in}}^d \prod_{j=1}^d \text{sinc}(2\pi r_{\text{in}} k_j) + r_{\text{out}}^d \prod_{j=1}^d \text{sinc}(2\pi r_{\text{out}} k_j) = r_{\text{in}}^d - r_{\text{out}}^d\right\}.$$

Let us now fix r_{in}, and consider the condition defining $\mathcal{B}_k^+$ as an equation on r_{out}. Define the functions

$$f_k(x) \;=\; x^d\left(1+\prod_{j=1}^{d}\operatorname{sinc}(2\pi x k_j)\right),$$
$$g_k(x) \;=\; x^d\left(1-\prod_{j=1}^{d}\operatorname{sinc}(2\pi x k_j)\right).$$

Then the mentioned equation takes the form

$$f_k(r_{\text{out}}) \;=\; g_k(r_{\text{in}}). \tag{7.2.9}$$

Consider the function $h_k : \mathbb{C} \to \mathbb{R}$:

$$h_k(z) \;=\; z^d\left(1+\prod_{j=1}^{d}\operatorname{sinc}(2\pi z k_j)\right).$$

Clearly, this function coincides with f_k on $\mathbb{R}$. Moreover, it is holomorphic in $\mathbb{C}$, as $\operatorname{sinc}(z)$ is holomorphic (it can be represented by the series $\sum_{n=0}^{\infty}\frac{(-1)^n}{(2n+1)!}z^{2n}$), and the product of holomorphic functions is again holomorphic. But then the derivative $h_k'(z)$ is also holomorphic, therefore, it has a countable number of zeros in $\mathbb{C}$. Clearly, $h_k' \equiv f_k'$ on $\mathbb{R}$, which yields that f_k' has a countable number of zeros in $\mathbb{R}$.

Hence, $\mathbb{R}_+$ is divided into a countable number of intervals on which the function $f_k(x)$ is strictly monotone. That is, $\mathbb{R}_+ = \sqcup_{i=0}^{\infty}[a_i(k), b_i(k)]$ where $f_{k,i} = f_k\big|_{[a_i(k),b_i(k)]}$ is strictly monotone. Then the function $f_{k,i}^{-1}(x)$ is correctly defined and, since $f_{k,i}$ is measurable and injective, $f_{k,i}^{-1}$ is measurable as well. Consequently, there is a unique solution $r_{\text{out}} = f_{k,i}^{-1}(g_k(r_{\text{in}}))$ of equation (7.2.9) for $r_{\text{in}} \in [\min f_{k,i}; \max f_{k,i}]$. If $r_{\text{in}} \notin [\min f_{k,i}; \max f_{k,i}]$, there is no solution at all.

Therefore, $\mathcal{B}_{k,i}^+ = \left\{\left(r_{\text{in}}, f_{k,i}^{-1}(g_k(r_{\text{in}}))\right) : r_{\text{in}} \in [\min f_{k,i}; \max f_{k,i}]\right\}$ is the graph of some measurable function in $\mathbb{R}_+^2$. Since such a graph has a zero Lebesgue measure (see *e.g.*, [WZ77, Lemma 5.3]), we have:

$$\operatorname{Leb}(\mathcal{B}_k^+) \;=\; \operatorname{Leb}\left(\cup_{i=1}^{\infty}\mathcal{B}_{k,i}^+\right) \;=\; 0.$$

Hence, we can conclude that

$$\operatorname{Leb}(\mathcal{B}^+) \;=\; \operatorname{Leb}\left(\bigcup_{k\in\mathcal{K}}\mathcal{B}_k^+\right) \;\le\; \sum_{k\in\mathcal{K}}\operatorname{Leb}(\mathcal{B}_k^+) \;=\; 0.$$

Carrying out a similar argumentation for $\mathcal{B}^-$ completes the proof. □

7.3 Consistency of higher-order spectral clustering

In this section we show that spectral clustering based on the ideal eigenvector (see Algorithm 11) is consistent for SGBM (Theorem 12). We then show that a simple extra step can in fact achieve strong consistency.

Algorithm 11: Higher-Order Spectral Clustering (HOSC).

Input: Adjacency matrix A, average intra- and inter-cluster edge densities $\mu_{\text{in}}, \mu_{\text{out}}$.
Output: Node labelling $\widetilde{\sigma} \in \{1,2\}^n$.

Global step:
Let $\widetilde{\lambda}$ be the eigenvalue of A closest to $\lambda_* = \frac{(\mu_{\text{in}}-\mu_{\text{out}})}{2}n$, and $\widetilde{v}$ be the associated eigenvector.
for $i = 1, \dots, n$ **do**
    If $\widetilde{v}_i > 0$, let $\widetilde{\sigma}_i = 1$; otherwise, let $\widetilde{\sigma}_i = 2$.

Remark 19. The worst case complexity of the eigenvalue factorization is $O(n^3)$ [Dem+08]. However, when the matrix is sufficiently sparse and the eigenvalues are well separated, the empirical complexity can be close to $O(kn)$, where k is the number of required eigenvalues [Dem+08]. Moreover, since Algorithm 11 uses only the sign of eigenvector elements, a quite rough accuracy can be sufficient for classification purposes.

7.3.1 Weak consistency of higher-order spectral clustering

Theorem 12. *Let us consider the d-dimensional SGBM with connectivity probability functions F_{in} and F_{out} satisfying conditions (7.2.7)-(7.2.8). Then Algorithm 11 is consistent. More precisely, Algorithm 11 misclassifies at most $O(\log n)$ nodes.*

Proof. Let us introduce some notations. Recall that $\mu_{\text{in}} = \widehat{F}_{\text{in}}(0)$ and $\mu_{\text{out}} = \widehat{F}_{\text{out}}(0)$. In the limiting spectrum, the ideal eigenvalue for clustering is

$$\lambda_* = \frac{\mu_{\text{in}} - \mu_{\text{out}}}{2} n.$$

We consider the closest eigenvalue of A to λ_*:

$$\widetilde{\lambda} = \arg\min_{\lambda} (|\lambda - \lambda_*|).$$

Also, let $\widetilde{v}$ be the normalized eigenvector corresponding to $\widetilde{\lambda}$. In this proof, the Euclidean norm $\|\cdot\|_2$ is used.

The plan of the proof is as follows. We consider the vector

$$v_* = (\underbrace{1/\sqrt{n}, \dots, 1/\sqrt{n}}_{n/2}, \underbrace{-1/\sqrt{n}, \dots, -1/\sqrt{n}}_{n/2})^{\mathrm{T}},$$

where we supposed without loss of generality that the $n/2$ first nodes are in cluster 1, and the $n/2$ last nodes are in cluster 2. The vector v_* gives the perfect recovery by the sign of its coordinates. We shall show that with high probability for some constant $C > 0$

$$\|\widetilde{v} - v_*\|_2 \leq C\sqrt{\frac{\log n}{n}}. \tag{7.3.1}$$

We say that an event occurs *with high probability* (*w. h. p.*) if its probability goes to 1 as $n \to \infty$. With the bounding (7.3.1), we shall then show that at most $o(n)$ of entries of $\widetilde{v}$ have a sign that differ from the sign of the respective entry in v_*; hence $\widetilde{v}$ retrieves almost exact recovery.

In order to establish inequality (7.3.1) we will use the following theorem from [KPJ82].

Theorem 13. *Let A be a real symmetric matrix. If $\widetilde{\lambda}$ is the eigenvalue of A closest to $\rho(v) = \frac{v^T A v}{v^T v}$, δ is the separation of ρ from the next closest eigenvalue and $\widetilde{v}$ is the eigenvector corresponding to $\widetilde{\lambda}$, then*

$$|\sin \angle(v, \widetilde{v})| \leq \frac{\|Av - \rho v\|_2}{\|v\|_2 \delta}.$$

First we deal with $\rho(v_*)$. Since v_* is normalized and real-valued (by the symmetry of A), we have

$$\rho(v_*) = v_*^T A v_*.$$

Denote $u = Av_*$. Then, obviously,

$$u_i = \sum_{j=1}^{n} A_{ij}(v_*)_i = \frac{1}{\sqrt{n}} \sum_{j=1}^{n/2} A_{ij} - \frac{1}{\sqrt{n}} \sum_{j=n/2+1}^{n} A_{ij}. \tag{7.3.2}$$

It is clear that each entry A_{ij} with $i \neq j$ is a Bernoulli random variable with the probability of success either μ_{in} or μ_{out}. This can be illustrated by the element-wise expectation of the adjacency matrix:

$$\mathbb{E}A = \left(\begin{array}{ccc|ccc} \mu_{\text{in}} & \dots & \mu_{\text{in}} & \mu_{\text{out}} & \dots & \mu_{\text{out}} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \mu_{\text{in}} & \dots & \mu_{\text{in}} & \mu_{\text{out}} & \dots & \mu_{\text{out}} \\ \hline \mu_{\text{out}} & \dots & \mu_{\text{out}} & \mu_{\text{in}} & \dots & \mu_{\text{in}} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \mu_{\text{out}} & \dots & \mu_{\text{out}} & \mu_{\text{in}} & \dots & \mu_{\text{in}} \end{array}\right).$$

Let us consider the first term in the right-hand side of (7.3.2) for $i \leq n/2$. Since A_{ij} are independent for fixed i, it is easy to see that $Y_i := \sum_{j=1}^{n/2} A_{ij} \sim \text{Bin}(n/2 - 1, \mu_{\text{in}})$ with the

expectation $\mathbb{E}Y_i = (n/2-1)\mu_{\text{in}}$. Then we can use the Chernoff bound to estimate a possible deviation from the mean. For any $0 < t < 1$

$$\mathbb{P}(|Y_i - \mathbb{E}Y_i| > t\mathbb{E}Y_i) \le e^{-\mathbb{E}Y_i t^2/2}. \tag{7.3.3}$$

Let us take $t = \frac{2\sqrt{\log n}}{\sqrt{(n/2-1)\mu_{\text{in}}}}$. Then for large enough n,

$$\mathbb{P}\left(\left|\sum_{j=1}^{n/2} A_{ij} - \mu_{\text{in}}\frac{n}{2}\right| > \sqrt{2\mu_{\text{in}} n\log n}\right) \;=\; \mathbb{P}\left(|Y_i - \mathbb{E}Y_i| > \sqrt{2\mu_{\text{in}} n\log n}\right) \;\le\; \frac{1}{n^2}.$$

Similarly,

$$\mathbb{P}\left(\left|\sum_{j=n/2+1}^{n} A_{ij} - \mu_{\text{out}}\frac{n}{2}\right| > \sqrt{2\mu_{\text{out}} n\log n}\right) \;\le\; \frac{1}{n^2}$$

and

$$\mathbb{P}\left(\left|u_i - (\mu_{\text{in}} - \mu_{\text{out}})\frac{\sqrt{n}}{2}\right| > \sqrt{2(\mu_{\text{in}} + \mu_{\text{out}})\log n}\right) \;\le\; \frac{2}{n^2}. \tag{7.3.4}$$

Denote $\gamma_n = \sqrt{2(\mu_{\text{in}} + \mu_{\text{out}})\log n}$ and notice that $\gamma_n = \Theta(\sqrt{\log n})$. By the union bound, we have for large enough n

$$\mathbb{P}\left(\exists i \le \frac{n}{2}\colon \left|u_i - (\mu_{\text{in}} - \mu_{\text{out}})\frac{\sqrt{n}}{2}\right| > \gamma_n\right) \;\le\; \frac{n}{2}\cdot\frac{2}{n^2} \;=\; \frac{1}{n}. \tag{7.3.5}$$

By the same argumentation,

$$\mathbb{P}\left(\exists i > \frac{n}{2}\colon \left|u_i - (\mu_{\text{out}} - \mu_{\text{in}})\frac{\sqrt{n}}{2}\right| > \gamma_n\right) \le \frac{1}{n}. \tag{7.3.6}$$

Now let us calculate $\rho(v_*)$:

$$\rho(v_*) \;=\; \sum_{i=1}^{n}(v_*)_i u_i \;=\; \frac{1}{\sqrt{n}}\sum_{i=1}^{n/2} u_i - \frac{1}{\sqrt{n}}\sum_{i=n/2+1}^{n} u_i.$$

We already established that $u_i \sim (\mu_{\text{in}} - \mu_{\text{out}})\frac{\sqrt{n}}{2}$ for $i \le \frac{n}{2}$ (which means that $\lim \frac{2u_i}{(\mu_{in}-\mu_{out})\sqrt{n}} = 1$ w.h.p.) and, therefore, that $\frac{1}{\sqrt{n}}\sum_{i=1}^{n/2} u_i \sim (\mu_{\text{in}} - \mu_{\text{out}})\frac{n}{4}$. More precisely, by (7.3.5),

$$\mathbb{P}\left(\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n/2} u_i - \frac{(\mu_{\text{in}} - \mu_{\text{out}})n}{4}\right| > \frac{\gamma_n\sqrt{n}}{2}\right) \;\le\; \frac{1}{n}.$$

In the same way, by (7.3.6),

$$\mathbb{P}\left(\left|\frac{1}{\sqrt{n}}\sum_{i=\frac{n}{2}+1}^{n} u_i - \frac{(\mu_{\text{out}} - \mu_{\text{in}})n}{4}\right| > \frac{\gamma_n\sqrt{n}}{2}\right) \;\le\; \frac{1}{n}.$$

Finally,

$$\mathbb{P}\left(\left|\rho(v_*) - \frac{(\mu_{\text{in}} - \mu_{\text{out}})n}{2}\right| > \gamma_n\sqrt{n}\right) \leq \frac{2}{n}. \tag{7.3.7}$$

Now let us denote $w = Av_* - \rho(v_*)v_* = u - \rho(v_*)v_*$. As we already know, $u_i \sim (\mu_{\text{in}} - \mu_{\text{out}})\frac{\sqrt{n}}{2}$ and $(\rho(v_*)v_*)_i \sim (\mu_{\text{in}} - \mu_{\text{out}})\frac{\sqrt{n}}{2}$ for $i \leq \frac{n}{2}$. Clearly, for $i \leq \frac{n}{2}$

$$|w_i| \leq \left|u_i - \frac{(\mu_{\text{in}} - \mu_{\text{out}})\sqrt{n}}{2}\right| + \left|\frac{(\mu_{\text{in}} - \mu_{\text{out}})\sqrt{n}}{2} - \frac{1}{\sqrt{n}}\rho(v_*)\right|.$$

Then

$$\begin{aligned}\mathbb{P}\left(|w_i| > \gamma_n\right) \leq\ & \mathbb{P}\left(\left|u_i - \frac{(\mu_{\text{in}} - \mu_{\text{out}})\sqrt{n}}{2}\right| > \gamma_n\right) + \\ & + \mathbb{P}\left(\left|\frac{(\mu_{\text{in}} - \mu_{\text{out}})\sqrt{n}}{2} - \frac{1}{\sqrt{n}}\rho(v_*)\right| > \gamma_n\right).\end{aligned}$$

A similar bound can be derived for the case $i > n/2$. Taking into account that $\rho(v_*)$ does not depend on i, using the union bound and equations (7.3.4) and (7.3.7), we get that

$$\mathbb{P}\left(\max_i |w_i| > 2\gamma_n\right) \leq n \cdot \frac{2}{n^2} + \frac{2}{n} = \frac{4}{n}.$$

One can readily see that $\|w\|_2 \leq \sqrt{n \cdot \max_i w_i^2} = \sqrt{n}\max_i |w_i|$. Thus, we finally can bound the Euclidean norm of the vector w:

$$\mathbb{P}\left(\|w\|_2 > 2\gamma_n\sqrt{n}\right) \leq \frac{4}{n} \to 0, \quad n \to \infty.$$

Now we can use Theorem 13. According to this result,

$$|\sin\angle(v_*, \widetilde{v})| \leq \frac{\|Av_* - v_*\rho(v_*)\|_2}{\|v_*\|_2\delta} = \frac{\|w\|_2}{\delta} \leq \frac{2\gamma_n\sqrt{n}}{\delta} \quad w.h.p.,$$

where $\delta = \min_i |\lambda_i(A) - \rho(v_*)|$ over all $\lambda_i \neq \widetilde{\lambda}$. Since we have assumed that (7.2.7) and (7.2.8) hold, by Proposition 23, $\delta > \epsilon n$. Then, since v_* is normalized, a simple geometric consideration guarantees that

$$\|v_* - \widetilde{v}\|_2 \leq \sqrt{2}\,|\sin\angle(v_*, \widetilde{v})| \leq \frac{2\sqrt{2}\gamma_n\sqrt{n}}{\epsilon n} = \frac{2\sqrt{2}\gamma_n}{\epsilon\sqrt{n}} \quad w.h.p. \tag{7.3.8}$$

Let us denote *the number of errors* by

$$r = \left|\{i \in [n]\colon \operatorname{sign}\left((v_*)_i\right) \neq \operatorname{sign}\left(\widetilde{v}_i\right)\}\right|.$$

If we now remember that the vector v_* consists of $\pm\frac{1}{\sqrt{n}}$, it is clear that for any i with $\operatorname{sign}((v_*)_i) \neq \operatorname{sign}(\widetilde{v}_i)$

$$|(v_*)_i - \hat{v}_i| > \frac{1}{\sqrt{n}}.$$

The number of such coordinates is r. Therefore,

$$\|v_* - \widetilde{v}\|_2^2 \;\geq\; r\left(\frac{1}{\sqrt{n}}\right)^2 \;=\; \frac{r}{n}.$$

Then, by (7.3.8), the following chain of inequalities holds:

$$\frac{r}{n} \;\leq\; \|v_* - \widetilde{v}\|_2^2 \;\leq\; \frac{8\gamma_n^2}{\epsilon^2 n} \;=\; \frac{16(\mu_{\text{in}} + \mu_{\text{out}})\log n}{\epsilon^2 n} \quad w.h.p.$$

Hence, with high probability

$$r \;\leq\; \frac{16(\mu_{\text{in}} + \mu_{\text{out}})\log n}{\epsilon^2} = O(\log n), \quad n \to \infty.$$

Thus, the vector $\widetilde{v}$ provides almost exact recovery. This completes the proof. □

7.3.2 Strong consistency of higher-order spectral clustering with local improvement

In order to derive a strong consistency result, we shall add an extra step to Algorithm 11. Given $\widetilde{\sigma}$, the prediction of Algorithm 11, we classify each node to be in the community where it has the most neighbors, according to the labelling $\widetilde{\sigma}$. We summarize this procedure in Algorithm 12, and Theorem 14 states the exact recovery result.

Algorithm 12: Higher-Order Spectral Clustering with Local Improvement (HOSC-LI).

Input: Adjacency matrix A, average intra- and inter-cluster edge densities $\mu_{\text{in}}, \mu_{\text{out}}$.
Output: Node labelling $\widehat{\sigma} \in \{1,2\}^n$.

Global step:
Let $\widetilde{\sigma}$ be the output of Algorithm 11.
Local improvement:
for $i = 1, \ldots, n$ **do**
    Set $\hat{\sigma}_i := \arg\max_{k\in\{1,2\}} \sum_{j\neq i} 1\left(\widetilde{\sigma}_j = k\right) a_{ij}$.

Remark 20. The local improvement step runs in $O(nd_{\max})$ operations, where $d_{\max}$ is the maximum degree of the graph. Albeit the local improvement step being convenient for the theoretical proof, we will see in Section 7.4 (Figure 7.3) that in practice Algorithm 11 already works well often giving 100% accuracy even without local improvement.

Theorem 14. *Let us consider the d-dimensional SGBM defined by (7.1.1)-(7.1.2), and connectivity probability functions F_{in} and F_{out} satisfying conditions (7.2.7)-(7.2.8). Then Algorithm 12 provides exact recovery for the given SGBM.*

Proof. We need to prove that the almost exact recovery of Algorithm 11 (established in Theorem 12) can be transformed into exact recovery by the local improvement step. This step consists in counting neighbours in the obtained communities. For each node i we count the number of neighbours in both supposed communities determined by the sign of the vector $\widetilde{v}$ coordinate:

$$\begin{aligned}\widetilde{Z}_1(i) &= \sum_{\mathrm{sign}(\widetilde{v}_j)=1} A_{ij},\\ \widetilde{Z}_2(i) &= \sum_{\mathrm{sign}(\widetilde{v}_j)=-1} A_{ij}.\end{aligned}$$

According to Algorithm 12, if $\widetilde{Z}_1(i) > \widetilde{Z}_2(i)$, we assign the label $\widehat{\sigma}_i = 1$ to node i, otherwise we label it as $\widehat{\sigma}_i = 2$. Suppose that some node i is still misclassified after this procedure and our prediction does not coincide with the true label: $\widehat{\sigma}_i \neq \sigma_i$. Let us assume without loss of generality that $\sigma_i = 1$ and, therefore, $\widehat{\sigma}_i = 2$. Then, clearly, $\widetilde{Z}_2(i) > \widetilde{Z}_1(i)$.

Let us denote by $Z_1(i)$ and $Z_2(i)$ degrees of node i in the communities defined by the true labels σ:

$$Z_j(i) = \sum_{\sigma_i=j} A_{ij}, \quad j = 1, 2.$$

Since $\mathrm{sign}(\widetilde{v}_j)$ coincides with the true community partition for all but $C\log n$ nodes (see the end of the proof of Theorem 12),

$$\left|\widetilde{Z}_j(i) - Z_j(i)\right| \le C\log n, \quad j = 1, 2,$$

which implies that

$$\widetilde{Z}_1(i) \ge Z_1(i) - C\log n \quad \text{and} \quad \widetilde{Z}_2(i) \le Z_2(i) + C\log n.$$

Hence, taking into account that $\widetilde{Z}_2(i) > \widetilde{Z}_1(i)$,

$$Z_2(i) + 2C\log n > Z_1(i),$$

which means that the inter-cluster degree of node i is asymptotically not less than its intra-cluster degree (since $Z_j(i) = \Theta(n) \quad w.h.p.$). Intuitively, this should happen very seldom, and Lemma 33 in Section 7.6 gives an upper bound on the probability of this event. Thus, by Lemma 33, for large n,

$$\begin{aligned}&\mathbb{P}(Z_2(i) + 2C\log n > Z_1(i)) = \mathbb{P}(Z_1(i) - Z_2(i) < 2C\log n) \le\\ &\le \mathbb{P}\left(Z_1(i) - Z_2(i) < \sqrt{2(\mu_{\mathrm{in}} + \mu_{\mathrm{out}})n\log n}\right) \le \frac{1}{n} \to 0, \quad n \to \infty.\end{aligned}$$

Then each node is classified correctly with high probability and Theorem 14 is proved. □

7.4 Numerical results

7.4.1 Higher-order spectral clustering on 1-dimensional GBM

Let us consider a 1-dimensional GBM, defined in Example 14. We first emphasize two important points of the theoretical study: the ideal eigenvector for clustering is not necesarily the Fiedler vector, and some eigenvectors, including the Fiedler vector, could correspond to geometric configurations.

Figure 7.1 shows the accuracy (*i.e.,* the ratio of correctly labelled nodes, up to a global permutation of the labels if needed, divided by the total number of nodes) of each eigenvector for a realization of 1-dimensional GBM. We see that, although the Fiedler vector is not suitable for clustering, there is nonetheless one eigevector that stands out of the crowd.

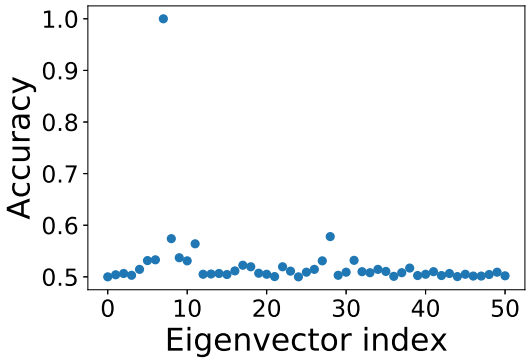
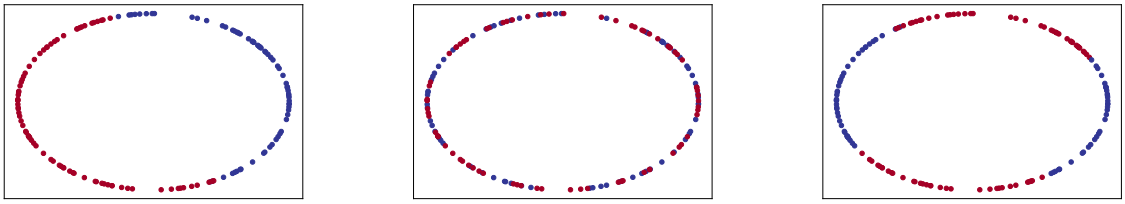


Figure 7.1: Accuracy obtained on a 1-dimensional GBM ($n = 2000$, $r_{\text{in}} = 0.08$, $r_{\text{out}} = 0.02$) using the different eigenvectors of the adjacency matrix. The eigenvector of index k corresponds to the eigenvector associated with the k-th largest eigenvalue of A.

Then, in Figure 7.2 we draw the nodes of a GBM according to their respective position. We then show the clusters predicted by some eigenvectors. We see some geometric configurations (Figures 7.2a and 7.2c), while the eigenvector leading to the perfect accuracy corresponds to index 4 (Figure 7.2b).

Figure 7.3 shows the evolution of the accuracy of Algorithms 11 and 12 when the number of nodes n increases. As expected, the accuracy increases with n. Moreover, we see no significant effect of using the local improvement of Algorithm 12. Thus, we conduct all the rest of numerical experiments with the basic Algorithm 11.

In Figure 7.4, we illustrate the statement of Proposition 24: for some specific values of the pair $(r_{\text{in}}, r_{\text{out}})$, the Conditions (7.2.7) and (7.2.8) do not hold, and Algorithm 11 is not guaranteed to work. We observe in Figure 7.4 that these pairs of bad values exactly correspond to the moments when the index of the ideal eigenvector jumps from one value to another.



(a) $k = 2$ (b) $k = 4$ (c) $k = 8$

Figure 7.2: Example of clustering done using the eigenvector associated to the k-th largest eigenvalue of the adjacency matrix of a 1-dimensional GBM ($n = 150$, $r_{\text{in}} = 0.2$, $r_{\text{out}} = 0.05$). For clarity edges are not shown, and nodes are positioned on a circle according to their true positions. The Fiedler vector ($k = 2$) is associated with a geometric cut, while the 4-th eigenvector corresponds to the true community labelling and leads to the perfect accuracy. The vector $k = 8$ is associated with yet another geometric cut.

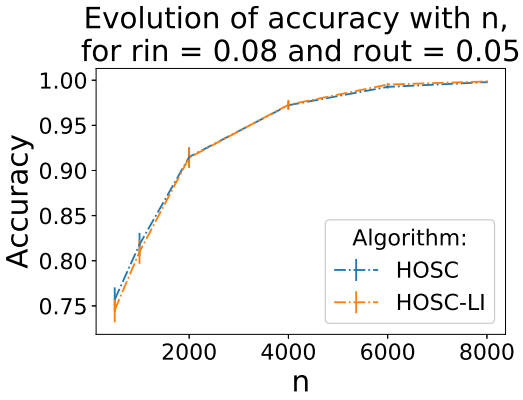


Figure 7.3: Accuracy obtained on 1-dimensional GBM as a function of n, when $r_{\text{in}} = 0.08$ and $r_{\text{out}} = 0.05$, for Algorithm 11 and Algorithm 12. Results are averaged over 100 realisations, and error bars show the standard error.

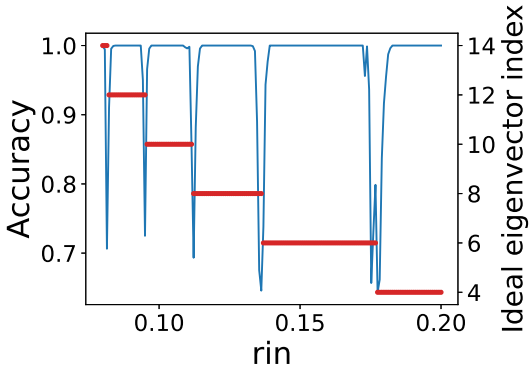


Figure 7.4: Evolution of accuracy (blue curve) with respect to r_{in}, for a GBM with $n = 3000$ and $r_{\text{out}} = 0.06$. Results are averaged over 10 realisations. By the red curve we show the index of the ideal eigenvector, again with respect to r_{in}.

Finally, we compare in Figure 7.5 the accuracy of Algorithm 11 with the motif counting algorithms presented in references [Gal+18] and [Gal+19]. Those algorithms consist in counting the number of common neighbors, and clustering accordingly. We call Motif Counting 1 (resp. Motif Counting 2) the algorithm of reference [Gal+18] (resp. of reference [Gal+19]). We thank the authors for providing us the code used in their papers. We observed that with present realizations the motif counting algorithms take much more time than HOSC takes. For example on a GBM with $n = 3000$, $r_{\text{in}} = 0.08$ and $r_{\text{out}} = 0.04$, HOSC takes 8 seconds, while Motif Counting 1 takes 130 seconds and Motif Counting 2 takes 60 seconds on a laptop with 1.90GHz CPU and 15.5 GB memory.

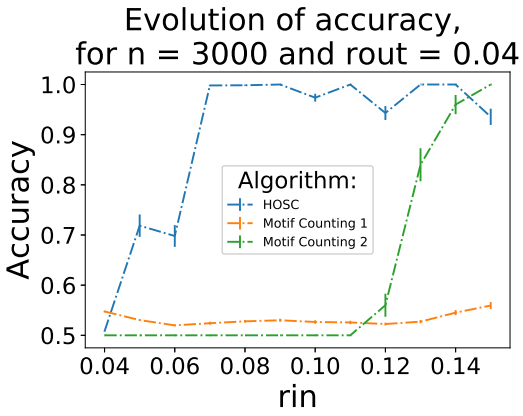


Figure 7.5: Accuracy obtained on 1-dimensional GBM for different clustering methods. Motif Counting 1 corresponds to the algorithm described in [Gal+18] and Motif Counting 2 to the algorithm described in [Gal+19]. Results are averaged over 50 realisations, and error bars show the standard error.

7.4.2 Waxman Block Model

Let us now consider the Waxman Block Model introduced in Example 15. Recall that $F_{\text{in}}(x) = \min(1, q_{\text{in}} e^{-s_{\text{in}} x})$ and $F_{\text{out}}(x) = \min(1, q_{\text{out}} e^{-s_{\text{out}} x})$, where $q_{\text{in}}, q_{\text{out}}, s_{\text{in}}, s_{\text{out}}$ are four positive real numbers. We have the following particular situations:

- if $s_{\text{out}} = 0$, then $F_{\text{out}}(x) = q_{\text{out}}$ and the inter-cluster interactions are independent of the nodes' positions. If $s_{\text{in}} = 0$ as well, we recover the SBM;
- if $q_{\text{in}} = e^{r_{\text{in}} s_{\text{in}}}$ and $q_{\text{out}} = e^{r_{\text{out}} s_{\text{out}}}$, then in the limit $s_{\text{in}}, s_{\text{out}} \gg 1$ we recover the 1-dimensional GBM.

We show in Figure 7.6 the accuracy of Algorithm 11 on a WBM. In particular, we see that we do not need $\mu_{\text{in}} > \mu_{\text{out}}$, and we can recover disassociative communities. However, there are obvious dips when q_{in} is close to q_{out} or s_{in} is close to s_{out}. It is clear that if $q_{\text{in}} = q_{\text{out}}$ on the left-hand side picture or $s_{\text{in}} = s_{\text{out}}$ on the right-hand side picture, one cannot distinguish two communities in the graph. Thus, for small n, we observe some ranges around these 'bad' values where Algorithm 11 fails. As expected, the dips become narrower when n increases.

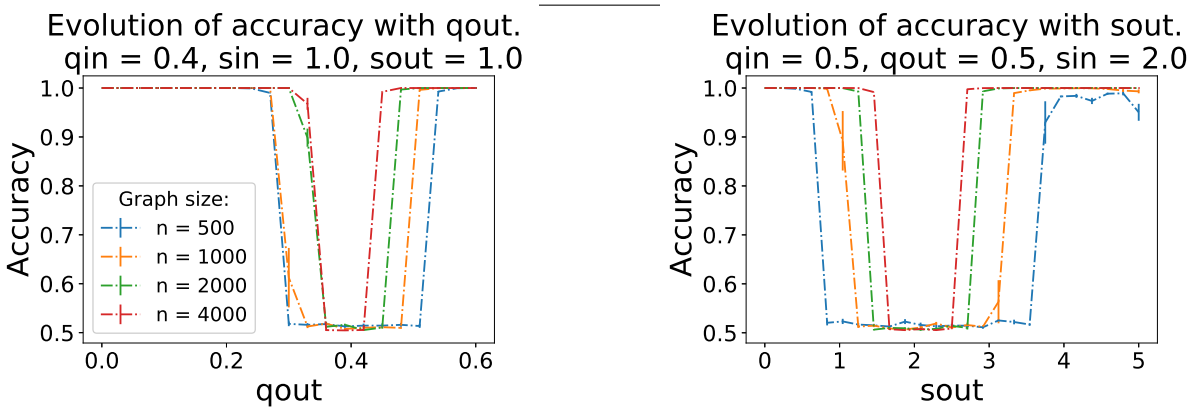


Figure 7.6: Accuracy obtained on a 1-dimensional Waxman Block Model. Results are averaged over 10 realizations. Same colors in the two plots correspond to the same graph size.

7.5 Conclusions and future research

In the present chapter we devised an effective algorithm for clustering geometric graphs. This algorithm is close in concept to the classical spectral clustering method but it makes use of the eigenvector associated with a higher-order eigenvalue. It provides weak consistency for a wide class of graph models which we call the Soft Geometric Block Model, under some mild conditions on the Fourier transform of F_{in} and F_{out}. A small adjustment of the algorithm leads to strong consistency. Moreover, our method was shown to be effective in numerical simulations even for graphs of modest size.

The problem stated in the current paper might be investigated further in several directions. One of them is a possible study on the SGBM with more than two clusters. The situation here is quite different from the SBM where the spectral clustering algorithm with few eigenvectors associated with the smallest non-zero eigenvalues provides good performance. In the SGBM even the choice of such eigenvectors is not trivial, since the 'optimal' eigenvalue for distinguishing two clusters is often not the smallest one.

Another natural direction of research is the investigation of the sparse regime, since all our theoretical results concern the case of degrees linear in n (this assumption is used for the analysis of the adjacency matrix spectrum and for finding the spectral gap around the 'ideal' eigenvalue $\tilde{\lambda}$). In sparser regimes, there are effective algorithms for some particular cases of the SGBM (e. g., for the GBM), but there is no established threshold when exact recovery is theoretically possible. Unfortunately, the method of the current paper without revision is not appropriate for this situation, and the technique will very likely be much more complicated.

Finally, considering weighted geometric graphs could be an interesting task for applications where clustering of weighted graphs is pertinent. For instance, the functions F_{in} and F_{out} can be considered as weights on the edges in a graph. We believe that most of the results of the paper may be easily transferred to this case.

7.6 Auxiliary results

7.6.1 Hamburger moment problem for the limiting measure

Lemma 30. *Assume that $F : \mathbb{T}^d \to \mathbb{R}$ is such that $F(0)$ is equal to the Fourier series of $F(x)$ evaluated at 0 and $0 \leq F(x) \leq 1$. Consider the measure μ defined by the function F:*

$$\mu = \sum_{k \in \mathbb{Z}^d} \delta_{\widehat{F}(k)}.$$

Then μ is defined uniquely by the sequence of its moments $\{M_n\}_{n=1}^{\infty}$.

Proof. It is enough to show that Carleman's condition holds true for μ (see [AK65]):

$$\sum_{n=1}^{\infty} M_{2n}^{-\frac{1}{2n}} = +\infty. \tag{7.6.1}$$

As one can easily see,

$$M_{2n} = \sum_{k \in \mathbb{Z}^d} \widehat{F}^{2n}(k). \tag{7.6.2}$$

From the bounds $0 \leq F(x) \leq 1$ it follows that $\widehat{F}(k) \leq 1$. Then it is clear that $\widehat{F}^n(k) \leq \widehat{F}(k)$ for any $n \in \mathbb{N}$. We can write

$$M_{2n} = \sum_{k \in \mathbb{Z}^d} \widehat{F}^{2n}(k) \leq \sum_{k \in \mathbb{Z}^d} \widehat{F}(k) = F(0) \leq 1.$$

Here we used the assumption that $F(0)$ equals its Fourier series evaluated at 0. Then

$$M_{2n}^{-\frac{1}{2n}} \geq 1.$$

Thus, the series in the right-hand side of (7.6.1) is divergent and Carleman's condition is verified.

□

7.6.2 m-times convolution

Lemma 31. *Let $m \in \mathbb{N}$ and $F_1, \ldots, F_m$ be integrable functions over $\mathbb{T}^d$. Then,*

$$F_1 * \cdots * F_m(0) = \int_{(\mathbb{T}^d)^m} \prod_{j=1}^{m} F_j(z_j - z_{j+1})\, dz_1 \ldots dz_m$$

with the notation $z_{m+1} = z_1$.

Proof. With the change of variable $u_i = z_i - z_{i+1}$ for $i = 1 \dots m-1$, we have

$$\begin{aligned}
&\int_{(\mathbb{T}^d)^m} \prod_{j=1}^{m} F_j(z_j - z_{j+1})\, dz_1 \dots dz_m \\
&= \int_{\mathbb{T}^d} dz_1 \int_{(\mathbb{T}^d)^{m-1}} \prod_{i=1}^{m-1} F_i(u_i) F_m(-u_1 - \dots - u_{m-1}) du_1 \dots du_{m-1}
\end{aligned}$$

We notice that

$$\begin{aligned}
&\int_{\mathbb{T}^d} du_{m-1} F_{m-1}(u_{m-1}) F_m(-u_1 - \dots - u_{m-1}) \\
&= F_{m-1} * F_m(-u_1 - \dots - u_{m-2}).
\end{aligned}$$

Hence,

$$\begin{aligned}
&\int_{(\mathbb{T}^d)^{m-1}} \prod_{i=1}^{m-1} F_i(u_i) F_m(-u_1 - \dots - u_{m-1}) du_1 \dots du_{m-1} \\
&= F_1 * \dots * F_m(0).
\end{aligned}$$

Therefore,

$$\int_{(\mathbb{T}^d)^m} \prod_{j=1}^{m} F_j(z_j - z_{j+1})\, dz_1 \dots dz_m \;=\; \int_{\mathbb{T}^d} dz_1 F_1 * \dots * F_m(0) \;=\; F_1 * \dots * F_m(0),$$

which ends the proof. □

7.6.3 Fourier transform of the square wave

Lemma 32. *Let $0 < r < \frac{1}{2}$. Let $F : \mathbb{R}^d \to \mathbb{R}$ be 1-periodic such that $F(x) = 1\,(\|x\| \le r)$ for $x \in \mathbb{T}^d$. Then,*

$$\hat{F}(k) \;=\; 2r^d \prod_{j=1}^{d} \operatorname{sinc}(2\pi k_j r),$$

where $k = (k_1, \dots, k_d) \in \mathbb{Z}^d$ and $\operatorname{sinc}(x) = \begin{cases} 1, & \text{if } x = 0, \\ \frac{\sin x}{x}, & \text{otherwise.} \end{cases}$

Proof. We shall use the set $[-1/2, 1/2]^d$ as a representation of $\mathbb{T}^d$. Let us first notice that for $x \in [-1/2, 1/2]^d$

$$F(x) = 1(\|x\| \le r) \;=\; 1\left(\max_{1 \le j \le d} |x_j| \;\le\; r\right) \;=\; \prod_{j=1}^{d} 1\left(|x_j| \;\le\; r\right).$$

Then

$$\begin{aligned}\hat{F}(k) &= \int_{\left[-\frac{1}{2},\frac{1}{2}\right]^d} F(x)e^{-2\pi i\langle k,x\rangle}dx \\ &= \int_{\left[-\frac{1}{2},\frac{1}{2}\right]^d} \prod_{j=1}^{d} 1(|x_j|\le r)e^{-2\pi i k_j x_j}dx_1\dots dx_d \\ &= \prod_{j=1}^{d}\int_{-1/2}^{1/2} 1(|x_j|\le r)e^{-2\pi i k_j x_j}dx_j.\end{aligned}$$

Let us consider some $1\le j\le d$. If $k_j=0$, we have $\int_{-1/2}^{1/2}1(|x_j|\le r)dx_j=\int_{-r}^{r}dx=2r$. Moreover, for $k_j\neq 0$,

$$\hat{F}(k) = \int_{-1/2}^{1/2}1(|x_j|\le r)e^{-2\pi i k_j x_j}dx_j = \int_{-r}^{r}e^{-2\pi i k_j x_j}dx_j = \frac{e^{-2\pi i k_j r}-e^{2\pi i k_j r}}{-2\pi i k_j} =$$
$$= \frac{\sin(2\pi k_j r)}{\pi k_j} = 2r\frac{\sin(2\pi k_j r)}{2\pi k_j r} = 2r\,\mathrm{sinc}(2\pi k_j r).$$

Hence, $\hat{F}(k) = 2r^d\prod_{j=1}^{d}\mathrm{sinc}(2\pi k_j r)$. □

7.6.4 Number of neighbours in different clusters

Lemma 33. *Let us consider the SGBM with connectivity probability functions F_{in} and F_{out} for which $\mu_{\mathrm{in}}=\widehat{F}_{\mathrm{in}}(0)>\widehat{F}_{\mathrm{out}}(0)=\mu_{\mathrm{out}}$. Denote by $Z_{\mathrm{in}}(i)$ (resp., $Z_{\mathrm{out}}(i)$) the 'intra-cluster' (resp., 'inter-cluster') degree of i:*

$$Z_{\mathrm{in}}(i) = \sum_{j:\,\sigma_j=\sigma_i} A_{ij};$$
$$Z_{\mathrm{out}}(i) = \sum_{j:\,\sigma_j\neq\sigma_i} A_{ij}.$$

Denote $\mathcal{B}_i := \left\{Z_{\mathrm{in}}(i)-Z_{\mathrm{out}}(i) < \sqrt{2(\mu_{\mathrm{in}}+\mu_{\mathrm{out}})n\log n}\right\}$. Then

$$\mathbb{P}\left(\cup_{i=1}^{n}\mathcal{B}_i\right)\le\frac{1}{n}.$$

Proof. Let us fix $i\in[n]$. Clearly, $Z_{\mathrm{in}}(i)\sim\mathrm{Bin}(\frac{n}{2}-1,\mu_{\mathrm{in}})$ and $Z_{\mathrm{out}}(i)\sim\mathrm{Bin}(\frac{n}{2},\mu_{\mathrm{out}})$. We again use Chernoff inequality (7.3.3). By this bound, for $t=\frac{2\sqrt{\log n}}{\sqrt{(n/2-1)\mu_{\mathrm{in}}}}$ and large enough n

$$\mathbb{P}\left(\left|Z_{\mathrm{in}}(i)-\mu_{\mathrm{in}}\frac{n}{2}\right| > \sqrt{2\mu_{\mathrm{in}}n\log n}\right) \le \frac{1}{n^2}.$$

By the same reason, Z_{out} is well concentrated around its mean $\mu_{\mathrm{out}}\frac{n}{2}$:

$$\mathbb{P}\left(\left|Z_{\mathrm{out}}(i)-\mu_{\mathrm{out}}\frac{n}{2}\right| > \sqrt{2\mu_{\mathrm{out}}n\log n}\right) \le \frac{1}{n^2}.$$

Therefore, since $\mu_{\text{in}} > \mu_{\text{out}}$,

$$\mathbb{P}(\mathcal{B}_i) = \mathbb{P}\left(Z_{\text{in}}(i) - Z_{\text{out}}(i) \;<\; \sqrt{2(\mu_{\text{in}} + \mu_{\text{out}})n \log n}\right) \;\leq\; \frac{1}{n^2}.$$

By the union bound,

$$\mathbb{P}(\mathcal{B}) \;\leq\; n\mathbb{P}(\mathcal{B}_1) \;\leq\; \frac{1}{n},$$

which proves the lemma. □

7.6.5 Trace operator is Lipschitz

Lemma 34. *Let $A, \widetilde{A} \in \{0,1\}^{n\times n}$ be two adjacency matrices, and $m \geq 1$. Then,*

$$\left|\operatorname{Tr} A^m - \operatorname{Tr} \widetilde{A}^m\right| \;\leq\; m\, n^{m-2}\, \mathrm{d}_{\text{Ham}}\left(A, \widetilde{A}\right).$$

Proof. Since A and $\widetilde{A}$ are adjacency matrices of graphs without self-loops, we have $\operatorname{Tr} A = 0 = \operatorname{Tr} \widetilde{A}$. Hence $\left|\operatorname{Tr} A - \operatorname{Tr} \widetilde{A}\right| = 0 \leq \frac{1}{n}\, \mathrm{d}_{\text{Ham}}\left(A, \widetilde{A}\right)$, and the statement holds for $m = 1$.

Let us now consider $m \geq 2$. We have

$$\begin{aligned}\left|\operatorname{Tr}\left(A^{m-1}\right) - \operatorname{Tr}(\widetilde{A}^{m-1})\right| \;&=\; \left|\sum_{i_1,\ldots,i_m}\left(\prod_{j=1}^{m} A_{i_j i_{j+1}} - \prod_{j=1}^{m} \widetilde{A}_{i_j,i_{j+1}}\right)\right| \\ &\leq\; \sum_{i_1,\ldots,i_m}\left|\prod_{j=1}^{m} A_{i_j i_{j+1}} - \prod_{j=1}^{m} \widetilde{A}_{i_j,i_{j+1}}\right|,\end{aligned}$$

with the notation $i_{m+1} = i_1$. The quantity $\prod_{j=1}^{m} A_{i_j i_{j+1}}$ is equal to one if $A_{i_j i_{j+1}} = 1$ for all $j = 1, \ldots, m$, and equals zero otherwise. Hence, the difference $\left|\prod_{j=1}^{m} A_{i_j i_{j+1}} - \prod_{j=1}^{m} \widetilde{A}_{i_j,i_{j+1}}\right|$ is non-zero and is equal to one in two scenarii:

- $A_{i_j i_{j+1}} = 1$ for all $j = 1, \ldots, m$, while there is a j' such that $\widetilde{A}_{i_{j'} i_{j'+1}} = 0$,
- there is a j' such that $A_{i_{j'} i_{j'+1}} = 0$ and $\widetilde{A}_{i_j i_{j+1}} = 1$ for all $j = 1, \ldots, m$.

Thus,

$$\begin{aligned}\left|\prod_{j=1}^{m} A_{i_j i_{j+1}} - \prod_{j=1}^{m} \widetilde{A}_{i_j,i_{j+1}}\right| \;=\; &1\left(\forall j\; A_{i_j i_{j+1}} = 1\right) 1\left(\exists j' \colon \widetilde{A}_{i_{j'} i_{j'+1}} = 0\right) + \\ &+ 1\left(\exists j' \colon A_{i_{j'} i_{j'+1}} = 0\right) 1\left(\forall j\; \widetilde{A}_{i_j i_{j+1}} = 1\right).\end{aligned}$$

But,

$$1\left(\forall j\ A_{i_j i_{j+1}} = 1\right) 1\left(\exists j' \colon \widetilde{A}_{i_{j'} i_{j'+1}} = 0\right) \ \leq\ \prod_{j=1}^{m} 1(A_{i_j i_{j+1}} = 1) \sum_{j=1}^{m} 1\left(\widetilde{A}_{i_j i_{j+1}} = 0\right)$$
$$\leq\ \sum_{j=1}^{m} 1(A_{i_j i_{j+1}} = 1) 1(\widetilde{A}_{i_j i_{j+1}} = 0).$$

Similarly,

$$1\left(\exists j' \colon A_{i_{j'} i_{j'+1}} = 0\right) 1\left(\forall j\ \widetilde{A}_{i_j i_{j+1}} = 1\right) \ \leq\ \sum_{j=1}^{m} 1(A_{i_j i_{j+1}} = 0) 1(\widetilde{A}_{i_j i_{j+1}} = 1).$$

Therefore,

$$\left| \prod_{j=1}^{m} A_{i_j i_{j+1}} - \prod_{j=1}^{m} \widetilde{A}_{i_j, i_{j+1}} \right| \ \leq\ \sum_{j=1}^{m} 1\left(A_{i_j i_{j+1}} \neq \widetilde{A}_{i_j i_{j+1}}\right).$$

This leads to

$$\sum_{i_1,\ldots,i_m} \left| \prod_{j=1}^{m} A_{i_j i_{j+1}} - \prod_{j=1}^{m} \widetilde{A}_{i_j, i_{j+1}} \right| \ \leq\ \sum_{i_1,\ldots,i_m} \sum_{j=1}^{m} 1\left(A_{i_j i_{j+1}} \neq \widetilde{A}_{i_j i_{j+1}}\right) \ \leq\ m\, n^{m-2}\ \mathrm{d}_{\mathrm{Ham}}(A, \widetilde{A}),$$

where the last line holds since for all $j = 1, \ldots, m$ and $m \geq 2$

$$\sum_{i_1,\ldots,i_m} 1\left(A_{i_j i_{j+1}} \neq \widetilde{A}_{i_j i_{j+1}}\right) \ =\ n^{m-2} \sum_{i_j, i_{j+1}} 1\left(A_{i_j i_{j+1}} \neq \widetilde{A}_{i_j i_{j+1}}\right) \ =\ n^{m-2}\, \mathrm{d}_{\mathrm{Ham}}\left(A, \widetilde{A}\right).$$

□

Bibliography

[Abb18] Emmanuel Abbe. "Community detection and stochastic block models: Recent developments". In: *Journal of Machine Learning Research* 18 (2018), pp. 1–86.

[ABS17] Emmanuel Abbe, Francois Baccelli, and Abishek Sankararaman. "Community detection on Euclidean random graphs". In: *arXiv preprint arXiv:1706.09942v5* (2017).

[ABH16] Emmanuel Abbe, Afonso S. Bandeira, and Georgina Hall. "Exact recovery in the stochastic block model". In: *IEEE Transactions on Information Theory* 62.1 (2016), pp. 471–487.

[AS15] Emmanuel Abbe and Colin Sandon. "Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery". In: *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*. IEEE. 2015, pp. 670–688.

[Abb+20] Emmanuel Abbe et al. "Entrywise eigenvector analysis of random matrices with low expected rank". In: *Annals of statistics* 48.3 (2020), p. 1452.

[AG05] Lada A Adamic and Natalie Glance. "The political blogosphere and the 2004 US election: divided they blog". In: *Proceedings of the 3rd international workshop on Link discovery*. 2005, pp. 36–43.

[AK65] Naum Iljič Aheizer and N Kemmer. *The classical moment problem and some related questions in analysis*. Oliver & Boyd Edinburgh, 1965.

[ABD20] Konstantin Avrachenkov, Andrei Bobu, and Maximilien Dreveton. "Higher-Order Spectral Clustering for Geometric Graphs". In: *arXiv preprint arXiv:2009.11353* (2020).

[AD19] Konstantin Avrachenkov and Maximilien Dreveton. "Almost Exact Recovery in Label Spreading". In: *International Workshop on Algorithms and Models for the Web-Graph*. Springer. 2019, pp. 30–43.

[ADL21] Konstantin Avrachenkov, Maximilien Dreveton, and Lasse Leskelä. "Recovering Communities in Temporal Networks Using Persistent Edges". In: *International Conference on Computational Data and Social Networks*. Springer. 2021, pp. 243–254.

[ADL22] Konstantin Avrachenkov, Maximilien Dreveton, and Lasse Leskelä. "Community recovery in non-binary and temporal stochastic block models". In: *arXiv preprint arXiv:2008.04790* (2022).

[AKL18] Konstantin Avrachenkov, Arun Kadavankandy, and Nelly Litvak. "Mean field analysis of personalized PageRank with implications for local graph clustering". In: *Journal of Statistical Physics* 173.3-4 (2018), pp. 895–916.

[Avr+12] Konstantin Avrachenkov et al. “Generalized optimization framework for graph-based semi-supervised learning”. In: *Proceedings of the 2012 SIAM International Conference on Data Mining*. SIAM. 2012, pp. 966–974.

[AFH13] Konstantin E Avrachenkov, Jerzy A Filar, and Phil G Howlett. *Analytic Perturbation Theory and its Applications*. SIAM, 2013.

[Bar+18] Paolo Barucca et al. “Disentangling group and link persistence in dynamic stochastic block models”. In: *Journal of Statistical Mechanics: Theory and Experiment* 2018.123407 (2018), pp. 1–18.

[BLP08] Shai Ben-David, Tyler Lu, and Dávid Pál. “Does Unlabeled Data Provably Help? Worst-case Analysis of the Sample Complexity of Semi-Supervised Learning”. In: *Conference on Learning Theory*. 2008.

[BC09] Peter J Bickel and Aiyou Chen. “A nonparametric view of network models and Newman–Girvan and other modularities”. In: *Proceedings of the National Academy of Sciences* 106.50 (2009), pp. 21068–21073.

[Bil61] Patrick Billingsley. “Statistical Methods in Markov Chains”. In: *The Annals of Mathematical Statistics* 32.1 (1961), pp. 12–40.

[Blo+08] Vincent D Blondel et al. “Fast unfolding of communities in large networks”. In: *Journal of statistical mechanics: theory and experiment* 2008.10 (2008), P10008.

[Bor08] Charles Bordenave. “Eigenvalues of Euclidean random matrices”. In: *Random Structures & Algorithms* 33.4 (2008), pp. 515–532.

[Bra+07] Ulrik Brandes et al. “On finding graph clusterings with maximum modularity”. In: *International Workshop on Graph-Theoretic Concepts in Computer Science*. Springer. 2007, pp. 121–132.

[Cal+20] Jeff Calder et al. “Poisson Learning: Graph Based semi-supervised learning at very low label rates”. In: *International Conference on Machine Learning*. PMLR. 2020, pp. 1306–1316.

[CSZ06] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. *Semi-Supervised Learning*. Adaptive computation and machine learning. MIT Press, 2006.

[Che+21] Yuxin Chen et al. “Spectral Methods for Data Science: A Statistical Perspective”. In: *Foundations and Trends® in Machine Learning* 14.5 (2021), pp. 566–806.

[CCC02] Fabio Gagliardi Cozman, Ira Cohen, and M Cirelo. “Unlabeled Data Can Degrade Classification Performance of Generative Classifiers”. In: *Proceedings of Flairs-02*. 2002, pp. 327–331.

[Dem+08] James W Demmel et al. “Performance and accuracy of LAPACK’s symmetric tridiagonal eigensolvers”. In: *SIAM Journal on Scientific Computing* 30.3 (2008), pp. 1508–1526.

[Dha+22] Souvik Dhara et al. "Spectral recovery of binary censored block models". In: *Proceedings of the 2022 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. SIAM. 2022, pp. 3389–3416.

[FO05] Uriel Feige and Eran Ofek. "Spectral techniques applied to sparse random graphs". In: *Random Structures & Algorithms* 27.2 (2005), pp. 251–275.

[Fie75] Miroslav Fiedler. "A property of eigenvectors of nonnegative symmetric matrices and its application to graph theory". In: *Czechoslovak mathematical journal* 25.4 (1975), pp. 619–633.

[For10] Santo Fortunato. "Community detection in graphs". In: *Physics Reports* 486.3 (2010), pp. 75–174. ISSN: 0370-1573.

[FB14] Julie Fournet and Alain Barrat. "Contact Patterns among High School Students". In: *PLOS ONE* 9.9 (Sept. 2014), pp. 1–17.

[Gal+18] Sainyam Galhotra et al. "The geometric block model". In: *Thirty-Second AAAI Conference on Artificial Intelligence*. 2018.

[Gal+19] Sainyam Galhotra et al. "Connectivity of Random Annulus Graphs and the Geometric Block Model". In: *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2019)* 145 (2019), p. 53.

[GGV89] Walter Gander, Gene H Golub, and Urs Von Matt. "A constrained eigenvalue problem". In: *Linear Algebra and its applications* 114 (1989), pp. 815–839.

[Gao+17] Chao Gao et al. "Achieving optimal misclassification proportion in stochastic block models". In: *The Journal of Machine Learning Research* 18.1 (2017), pp. 1980–2024.

[Gha+16] Amir Ghasemian et al. "Detectability thresholds and optimal algorithms for community structure in dynamic networks". In: *Physical Review X* 6.3 (2016), p. 031005.

[GN02] M. Girvan and M. E. J. Newman. "Community structure in social and biological networks". In: *Proceedings of the National Academy of Sciences* 99.12 (2002), pp. 7821–7826. ISSN: 0027-8424.

[GTP21] Martijn M Gösgens, Alexey Tikhonov, and Liudmila Prokhorenkova. "Systematic analysis of cluster similarity indices: How to validate validation measures". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 3799–3808.

[GP10] Leo J Grady and Jonathan R Polimeni. *Discrete calculus: Applied analysis on graphs for computational science*. Vol. 3. Springer, 2010.

[HW08] Jake M Hofman and Chris H Wiggins. "Bayesian approach to network modularity". In: *Physical review letters* 100.25 (2008), p. 258701.

[HLL83] Paul W. Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. "Stochastic blockmodels: First steps". In: *Social Networks* 5.2 (2 1983), pp. 109–137.

[HJ12] Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 2012.

[Iba99] Yukito Iba. "The Nishimori line and Bayesian statistics". In: *Journal of Physics A: Mathematical and General* 32.21 (Jan. 1999), pp. 3875–3888.

[JŁR00] Svante Janson, Tomasz Łuczak, and Andrzej Ruciński. *Random Graphs*. Wiley, 2000, pp. xii+333. ISBN: 0-471-17541-2.

[JL15] Varun Jog and Po-Ling Loh. "Recovering communities in weighted stochastic block models". In: *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. 2015, pp. 1308–1315.

[KPJ82] W Kahan, B.N. Parlett, and E. Jiang. "Residual bounds on approximate eigensystems of nonnormal matrices". In: *SIAM Journal on Numerical Analysis* 19.3 (1982), pp. 470–484.

[KN11] Brian Karrer and M. E. J. Newman. "Stochastic blockmodels and community structure in networks". In: *Phys. Rev. E* 83 (1 2011), p. 016107.

[KK10] Amit Kumar and Ravindran Kannan. "Clustering with spectral norm and the k-means algorithm". In: *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*. IEEE. 2010, pp. 299–308.

[KSS04] Amit Kumar, Yogish Sabharwal, and Sandeep Sen. "A simple linear time (1+/spl epsiv/)-approximation algorithm for k-means clustering in any dimensions". In: *45th Annual IEEE Symposium on Foundations of Computer Science*. IEEE. 2004, pp. 454–462.

[LLV17] Can M. Le, Elizaveta Levina, and Roman Vershynin. "Concentration and regularization of random graphs". In: *Random Structures & Algorithms* 51.3 (2017), pp. 538–561.

[LCB98] Yann LeCun, Corinna Cortes, and Christopher JC Burges. *The MNIST database of handwritten digits*. 1998. URL: http://yann.lecun.com/exdb/mnist/.

[LL20] Jing Lei and Kevin Z Lin. "Bias-adjusted spectral clustering in multi-layer stochastic block models". In: *arXiv preprint arXiv:2003.08222* (2020).

[LR15] Jing Lei and Alessandro Rinaldo. "Consistency of spectral clustering in stochastic block models". In: *The Annals of Statistics* 43.1 (2015), pp. 215–237.

[Lei+17] Yang Lei et al. "Ground truth bias in external cluster validity indices". In: *Pattern Recognition* 65 (2017), pp. 58–70. ISSN: 0031-3203.

[Les] Lasse Leskelä. *Random graphs and network statistics*.

[LPW08] David A. Levin, Yuval Peres, and Elizabeth L. Wilmer. *Markov Chains and Mixing Times*. http://pages.uoregon.edu/dlevin/MARKOV/: American Mathematical Society, 2008, pp. xviii+371. ISBN: 978-0-8218-4739-8.

[LV06] Friedrich Liese and Igor Vajda. "On divergences and informations in statistics and information theory". In: *IEEE Transactions on Information Theory* 52.10 (2006), pp. 4394–4412.

[MC21] Xiaoyi Mai and Romain Couillet. "Consistent Semi-Supervised Graph Regularization for High Dimensional Data". In: *Journal of Machine Learning Research* 22.94 (2021), pp. 1–48.

[Mas14] Laurent Massoulié. "Community detection thresholds and the weak Ramanujan property". In: *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*. 2014, pp. 694–703.

[MFB15] Rossana Mastrandrea, Julie Fournet, and Alain Barrat. "Contact Patterns in a High School: A Comparison between Data Collected Using Wearable Sensors, Contact Diaries and Friendship Surveys". In: *PLOS ONE* 10.9 (Sept. 2015), pp. 1–26.

[MM17] Catherine Matias and Vincent Miele. "Statistical clustering of temporal networks through a dynamic stochastic block model". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 79.4 (2017), pp. 1119–1141.

[Mei07] Marina Meilă. "Comparing clusterings – An information based distance". In: *Journal of Multivariate Analysis* 98.5 (2007), pp. 873–895.

[MH01] Marina Meilă and David Heckerman. "An experimental comparison of model-based clustering methods". In: *Machine learning* 42.1 (2001), pp. 9–29.

[MM09] Marc Mézard and Andrea Montanari. *Information, physics, and computation*. Oxford University Press, 2009.

[Moo17] Cristopher Moore. "The Computer Science and Physics of Community Detection: Landscapes, Phase Transitions, and Hardness". In: *Bulletin of the EATCS* 121 (2017).

[MNS15] Elchanan Mossel, Joe Neeman, and Allan Sly. "Reconstruction and estimation in the planted partition model". In: *Probability Theory and Related Fields* (2015).

[MNS16] Elchanan Mossel, Joe Neeman, and Allan Sly. "Consistency thresholds for the planted bisection model". In: *Electronic Journal of Probability* (2016).

[MNS18] Elchanan Mossel, Joe Neeman, and Allan Sly. "A proof of the block model threshold conjecture". In: *Combinatorica* 38.3 (2018), pp. 665–708.

[Muc+10] Peter J. Mucha et al. "Community Structure in Time-Dependent, Multiscale, and Multiplex Networks". In: *Science* 328.5980 (2010), pp. 876–878. ISSN: 0036-8075.

[New13] M. E. J. Newman. "Spectral methods for community detection and graph partitioning". In: *Phys. Rev. E* 88 (4 2013), p. 042822.

[New16] Mark EJ Newman. "Equivalence between modularity optimization and maximum likelihood methods for community detection". In: *Physical Review E* 94.5 (2016), p. 052315.

[NG04] Mark EJ Newman and Michelle Girvan. "Finding and evaluating community structure in networks". In: *Physical Review E* 69.2 (2004), p. 026113.

[NJW02] Andrew Y Ng, Michael I Jordan, and Yair Weiss. "On spectral clustering: Analysis and an algorithm". In: *Advances in neural information processing systems*. 2002, pp. 849–856.

[Pam+19] A Roxana Pamfil et al. "Relating modularity maximization and stochastic block models in multilayer networks". In: *SIAM Journal on Mathematics of Data Science* 1.4 (2019), pp. 667–698.

[PC16] Subhadeep Paul and Yuguo Chen. "Consistent community detection in multi-relational data through restricted multi-layer stochastic blockmodel". In: *Electron. J. Statist.* 10.2 (2016), pp. 3807–3870.

[Pei19] Tiago P Peixoto. "Bayesian stochastic blockmodeling". In: *Advances in network clustering and blockmodeling* (2019), pp. 289–332.

[Pen16] Mathew D Penrose. "Connectivity of soft random geometric graphs". In: *Annals of Applied Probability* 26.2 (2016), pp. 986–1028.

[RB06] Jörg Reichardt and Stefan Bornholdt. "Statistical mechanics of community detection". In: *Physical Review E* 74.1 (2006), p. 016110.

[SM00] Jianbo Shi and Jitendra Malik. "Normalized cuts and image segmentation". In: *IEEE Transactions on pattern analysis and machine intelligence* 22.8 (2000), pp. 888–905.

[Sin95] Karen B. Singer-Cohen. "Random intersection graphs". Thesis (Ph.D.)–The Johns Hopkins University. PhD thesis. Johns Hopkins University, 1995, p. 219.

[Sta+20] Ljubiša Stanković et al. "Data Analytics on Graphs Part III: Machine Learning on Graphs, from Graph Topology to Applications". In: *Foundations and Trends® in Machine Learning* 13 (2020), pp. 332–530.

[Tal96] Michel Talagrand. "A new look at independence". In: *Annals of Probability* 24.1 (Jan. 1996), pp. 1–34.

[Tro15] Joel A. Tropp. "An Introduction to Matrix Concentration Inequalities". In: *Foundations and Trends® in Machine Learning* 8.1-2 (2015), pp. 1–230. ISSN: 1935-8237.

[VH14] Tim Van Erven and Peter Harremos. "Rényi divergence and Kullback-Leibler divergence". In: *IEEE Transactions on Information Theory* 60.7 (2014), pp. 3797–3820.

[Von07] Ulrike Von Luxburg. "A tutorial on spectral clustering". In: *Statistics and computing* 17.4 (2007), pp. 395–416.

[WW93] Dorothea Wagner and Frank Wagner. "Between min cut and graph bisection". In: *International Symposium on Mathematical Foundations of Computer Science*. Springer. 1993, pp. 744–750.

[WB17] Y. X. Rachel Wang and Peter J. Bickel. "Likelihood-based model selection for stochastic block models". In: *Ann. Statist.* 45.2 (Apr. 2017), pp. 500–528.

[Wax88] Bernard M Waxman. "Routing of multipoint connections". In: *IEEE Journal on Selected Areas in Communications* 6.9 (1988), pp. 1617–1622.

[WZ77] Richard L Wheeden and Antoni Zygmund. *Measure and integral: An introduction to real analysis*. Chapman & Hall/CRC Pure and Applied Mathematics. Taylor & Francis, 1977.

[XJL20] Min Xu, Varun Jog, and Po-Ling Loh. "Optimal rates for community estimation in the weighted stochastic block model". In: *The Annals of Statistics* 48.1 (2020), pp. 183–204.

[YWS15] Yi Yu, Tengyao Wang, and Richard J Samworth. "A useful variant of the Davis–Kahan theorem for statisticians". In: *Biometrika* 102.2 (2015), pp. 315–323.

[YP16] Se-Young Yun and Alexandre Proutière. "Optimal cluster recovery in the labeled stochastic block model". In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*. NIPS'16. Barcelona, Spain: Curran Associates Inc., 2016, pp. 973–981. ISBN: 978-1-5108-3881-9.

[ZZ16] Anderson Y Zhang and Harrison H Zhou. "Minimax rates of community detection in stochastic block models". In: *The Annals of Statistics* 44.5 (2016), pp. 2252–2280.

[ZR18] Yilin Zhang and Karl Rohe. "Understanding regularized spectral clustering via graph conductance". In: *Advances in Neural Information Processing Systems*. 2018, pp. 10631–10640.

[Zho+04] Dengyong Zhou et al. "Learning with local and global consistency". In: *Advances in neural information processing systems*. 2004, pp. 321–328.

[ZGL03] Xiaojin Zhu, Zoubin Ghahramani, and John D. Lafferty. "Semi-supervised learning using Gaussian fields and harmonic functions". In: *Proceedings of the 20th International Conference on Machine Learning*. 2003, pp. 912–919.