Politecnico di Torino

Master thesis for the degree in Physics of Complex Systems

Filippo Zimmaro

Semisupervised classification in the Censored Block Model

External supervisor: Romain Couillet
UGA IDEX DataScience Chair, GIPSA-lab, University Grenoble-Alpes

External co-supervisor: Lorenzo Dall'Amico
PhD student GIPSA lab - Université Grenoble Alpes

Internal supervisor: Alfredo Braunstein
Associate professor DISAT, Politecnico di Torino

October 2021

SCIENCES SORBONNE UNIVERSITÉ

complex systems

gipsa-lab

Contents

1 Introduction

Semisupervised learning finds its place between unsupervised and supervised learning, in the panorama of machine learning algorithms. A semisupervised algorithm works on a partially labelled dataset, i.e. a dataset with a (usually small) fraction of labelled data and a (large) fraction of unlabelled ones. Many real problems fall in the category of semisupervised learning because of the scarcity of labelled data and the aboundance of unlabelled ones, mainly due to the difficulty or cost of the labelling process carried by humans or by expensive experiments [35, 28], opposed to the breakthrough of data collections that characterizes these last few years [17].
The study of semisupervised learning is motivated by its practical value in building better computer algorithms, as well as by its theoretical value in understanding the learning process in machines and humans [14]. Since it merges features of supervised and unsupervised techniques, semisupervised learning is theoretically difficult and not always helpful [4, 10]. Indeed, blindly selecting a semisupervised learning method for a specific task, using the wrong model assumption, can lead to a worse performance with respect to other types of learning [2]. These difficulties are perhaps the causes of its still modest popularity.

A classification task consists in an inference process where we have to cathegorize unlabelled data, on the basis of some assumptions and partial or noisy observations.
Among several proposed approaches to perform classification, graph-based methods involve three steps: graph creation, graph weighting and inference [16, 27]. The creation of a graph $G = (V, E)$ consists in assigning to each vertex a data point and representing the observed similarities through the edges. If the similarities are quantified, to each edge is assigned a weight J_{ij}. The assumption behind graph-based methods is that labels are smooth with respect to the graph, in the sense that if two nodes are connected by a large positive edge, they tend to have the same label, if else the edge is small, or negative, they tend to have different labels [35]. In other words, if two data points are found to be very similar, with high probability they will belong to the same class. After graph construction and weighting, the next step to perform the inference is to identify a cost function to minimize, that penalizes edges with large positive weight to be assigned different labels and viceversa.
In this sense, many different cost functions have been proposed for unsupervised learning, that are expressed in the form of $\hat{\mathbf{y}} A \hat{\mathbf{y}}$, where $\hat{\mathbf{y}}$ is the vector of predicted labels and A is a generic matrix, depending on the cost function, whose spectral properties are informative for the detection of classes (e.g. adjacency matrix, Laplacian [21],...). The algorithms based on the spectrum of such matrices are then called spectral algorithms.

The minimization of the cost function is however an arbitrary method and gives a hard binary classification. Moreover, standard spectral algorithms turn to be suboptimal or even fail to predict classes for sparse graphs (graphs with a low average connectivity), for a multiplicity of reasons [32]. We can refine the inference by applying statistical physics tools, through a Bayesian approach (see [31] for an exhaustive review). In this way, many features can be understood for simple generative models, such as the Stochastic Block Model (SBM, see section A) or the Censored Block Model (CBM). For the Stochastic Block Model, for example, it has been conjectured that no polynomial algorithm can detetect classes below a certain threshold [8]. Furthermore, efficient spectral algorithms that manage to detect communities down to this threshold have been recently developed by the statistical physics community and proved to be optimal for the SBM, namely Non-Backtracking [18] and Bethe-Hessian [24, 7]. These have been then adapted to the unsupervised Censored Block Model in [26].

In a semisupervised framework, the cost function must include also the information surplus given by the revelation of some true labels, or any other kind of surplus information. Efficiently embedding this information is part of the graph construction (and weighting) and is a hard and important problem.
The standard way to proceed is to modify the algorithms of unsupervised learning in order to encode this surplus of information [35]. For example, to the minimization of the cost function with the Laplacian matrix is added the constraint of labelled data to be assigned their revealed true label (Laplacian regularization, see [1, 19]). As the other more standard algorithms, also the recent spectral algorithms developed by statistical physicists have been adapted to semisupervised learning, although not much has been done in this sense yet. In [25], the authors try to adapt the Non-Backtracking matrix to a semisupervised problem by constantly fixing the labelled nodes to their true value in the iterative process to find the Non-Backtracking eigenvector. Morevover, in [33] and [12] is conjectured that the detectability threshold disappears for a SBM of two classes, as soon as we know a fraction of labelled nodes.

The scope of this work is to study a simple generative model such as the Censored Block Model in a semisupervised setting, considering for simplicity the case of binary classification. We tackle the problem from a different perspective with respect to [25], i.e. we focus on efficiently embedding the semisupervised information on the graph and then we apply and analyze standard unsupervised algorithms on the derived configurations, namely naive mean field (section 5) and adjacency matrix (section 6). Interestingly, we find that the inference with the latter, with the proposed graph modification, is substantially unbalanced towards the semisupervised information (section 6.1). In other words, the algorithm does not learn from unlabelled data. A similar problem, for high dimensional gaussian generated data, has been noticed for the standard regularized Laplacian by [20] and solved in [19] with the introduction of a centered adjacency matrix as well as a constraint on the norm of the prediction vector, controlled by a hyperparameter. Similarly, we fix the problem of unbalancing by adjusting the adjacency matrix of the modified graph (section 6.2). Last, we try to give an estimation of the modification factor, initially left as a hyperparameter (section 6.4). Algorithms, simulations and possible extensions are reported in the last two chapters.

2 The Censored Block Model

Starting from a set of n nodes assigned to two classes, the Censored Block Model consists in an Erdős–Rényi random graph with probablity of drawing an edge between two nodes equal to c/n, c determining the level of sparsity of the graph. Moreover, weights J_{ij} are attached to the edges. They are binary random variables (± 1) sorted in the following way: considering that σ_i^* and σ_j^* are respectively the true classes ($\sigma_i^* = \pm 1$) of nodes i, j and $p \in [0.5, 1]$,

$$J_{ij} = \begin{cases} \sigma_i^* \sigma_j^* & \text{w.p.} \quad p \\ -\sigma_i^* \sigma_j^* & \text{w.p.} \quad 1-p \end{cases} \tag{1}$$

In other words, for each edge, with probability p the informative coupling is taken (true information), with probability $1-p$ the uninformative one (misleading information). The task is to construct an inference vector $\hat{\sigma}$ that maximizes the similarity with the vector of the true planted assignment σ^*.

2.1 An example: horse or donkey?

In order to better understand the Censored Block Model, let's consider the following example.

On a table there are n pictures, $n/2$ representing a horse and the rest $n/2$ a donkey. A child, who has never seen horses and donkeys before in his life, wants to learn how to distinguish these two, similar, species. He has no other information but a tool that, choosing randomly a couple of pictures, tells him if the depicted animals are of the same species or not. Unfortunately, this tool does not work perfectly: with some probability $p > 0.5$ it tells the truth, else it tells a false information. In addition, the tool has a limited battery: for each possible couple of pictures, it is applied only with probability c/n (the higher is c the more powerful is the battery).
The child, trying to divide the pictures in the group of horses and the group of donkeys, is exactly attempting to solve the Censored Block Model!

2.2 Mapping to Ising Spin Glass

The Censored Block Model is also called Planted Spin Glass since it is pretty easy to map it to an Ising Spin Glass model with couplings generated according to a planted assignment of spins as in (1). One can write the probability of a coupling J_{ij} to be equal to $J = \pm 1$ as

$$P(J_{ij} = J | \sigma_i^*, \sigma_j^*) = p\mathbf{1}(J_{ij} = \sigma_i^* \sigma_j^*) + (1-p)\mathbf{1}(J_{ij} = -\sigma_i^* \sigma_j^*) \tag{2}$$

Defining a specific inverse temperature that we call β_N

$$\beta_N = \frac{1}{2} \log \frac{p}{1-p} \tag{3}$$

we can rewrite

$$P(J_{ij} = J | \sigma_i^*, \sigma_j^*) = \frac{e^{\beta_N J \sigma_i^* \sigma_j^*}}{2 \cosh \beta_N} \tag{4}$$

that can be intended as the likelihood of observing (J_{ij}) having (σ_i^*, σ_j^*) as planted assignment. With coupling extracted indipendently from the above distribution, we find that the posterior probability of a configuration σ given the observations (J_{ij}), is given by Bayes theorem

$$P(\sigma | (J_{ij})_{ij \in E}) = \frac{P(\sigma) \prod_{ij \in E} P(J_{ij} | \sigma_i, \sigma_j)}{P((J_{ij})_{ij \in E})} \tag{5}$$

Assuming a uniform prior, the posterior probability of the configuration σ given the observations (J_{ij}) is the Boltzmann distribution of an Ising spin-glass model

$$P(\sigma | (J_{ij})_{ij \in E}) = \frac{e^{\beta_N \sum_{ij \in E} J_{ij} \sigma_i \sigma_j}}{Z} \tag{6}$$

Note that a similar calculation can be performed for more than two classes, in which case we would end up with a Potts model [23].

2.3 Gauge invariance and Viana-Bray model

We can go further considering that any Ising Model is invariant under the following transformation (Gauge invariance): for any $\tau_i = \pm 1, \;\; i = 1, ..., n$

$$\sigma_i \to \tilde{\sigma}_i = \sigma_i \tau_i \qquad J_{ij} \to \tilde{J}_{ij} = J_{ij} \tau_i \tau_j \tag{7}$$

It is easy to check indeed that the transformed Boltzmann distribution $\tilde{P}$ verifies $\tilde{P}(\tilde{\sigma}) = P(\sigma)$. Choosing $\tau_i = \sigma_i^* \;\; i = 1, ..., n$, we have that the couplings $\tilde{J}_{ij}$ are positive if informative and negative if uninformative, in other words

$$P(\tilde{J}_{ij} = J) = \frac{e^{\beta_N}}{2 \cosh \beta_N} \mathbf{1}(J = 1) + \frac{e^{-\beta_N}}{2 \cosh \beta_N} \mathbf{1}(J = -1) \tag{8}$$

and the planted assignment of the transformed model is

$$\tilde{\sigma}_i^* = +1 \qquad \forall i \tag{9}$$

So we recovered another mapping, this time to an Ising spin-glass with nodes belonging all to the same class and couplings with a ferromagnetic bias, i.e.

$$E[\tilde{J}_{ij}] = \tanh \beta_N > 0 \tag{10}$$

as soon as $p > 0.5$. This setting corresponds to the Viana-Bray model, one case of Ising spin-glass with couplings following a Bernoulli distribution like (8). The Viana-Bray model has been analytically studied by the statistical physics community [29] and has the phase diagram reported in figure 1. Following [6], the phase is determined by the temperature at which we decide to solve the problem $T = \beta^{-1}$ and by the ferromagnetic bias of the couplings $E[\tilde{J}_{ij}]$. The latter, as we have seen in (10), depends on β_N, called Nishimori temperature. For large β and β_N (corresponding to a large ferromagnetic bias $E[\tilde{J}_{ij}]$), the system is in the *ferromagnetic phase*, where the typical configuration tends to have all the spins aligned. For large β and small β_N, instead, the system is in the *spin-glass phase*, where the global magnetization is null ($\frac{1}{n}\sum_i E[\sigma_i] = 0$) but local order of the spins can be observed ($\frac{1}{n}\sum_i E[\sigma_i]^2 \neq 0$). For small values of β (large T) and sufficiently small ferromagnetic bias, the system is in the *paramagnetic phase*, where the average magnetization of each spin is null.
Since the planted assignment of the Viana-Bray model to which we mapped the original problem corresponds to all the spins in the same class, the inference is possible if and only if the average magnetization of the Viana-Bray model is different from zero, i.e. if and only if it is in the ferromagnetic phase.
The Nishimori temperature β_N, that can be computed exactly with the knowledge of the generative model through (3) (i.e. knowing the reliability of the tool used by the child), comes from the application of Bayes theorem and corresponds to the optimal temperature at which solving the model to get the best inference. This can be seen also in the phase diagram, as it can be proved that the system at β_N is either in the paramagnetic (undetectable) or in the ferromagnetic (detectable) phase (see the red line in fig. 1). The value of β_N at the paramagnetic-ferromagnetic transitions thus determines the detectability threshold. After applying conjectures on the location of transition lines (and thus tricritical points)[1], the detectability threshold turns to be exactly the one predicted in (68) for the Censored Block Model with two classes,

$$c > \frac{1}{(2p-1)^2} \tag{11}$$

Replacing, or erroneously estimating, β_N with another β will lead to a decrease in performance. For example, looking for the ground state configuration corresponds to set $\beta \to \infty$, which within this picture is manifestly suboptimal. This being said, many algorithms (included the adjacency matrix proposed later) solve the inference problem looking for the ground state of the constructed Hamiltonian[2], so neglecting any entropic term (the typical configuration of extracted J_{ij} is generally not the most probable one) and setting far from the Bayes optimal (Nishimori temperature) regime. This may be justified when we do not dispose of any information on the generative model (i.e. when the child does not know at all how well his tool performs) or in the trivial case of no uncertainty on similarities, $p = 1$; indeed there $\beta_N \to \infty$.

[1] The calculus of the localizations of the transition lines is not reported, see [23] for a slightly more specific treatment.

[2] Looking for the ground state of a cost function, e.g. the coupling Hamiltonian, corresponds to the so called MinCut problem [3] and it is also NP-Complete.

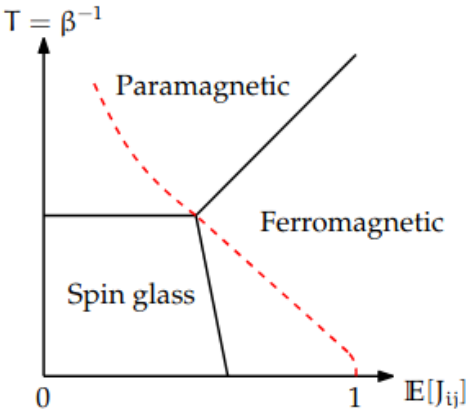

Figure 1: Phase diagram of Viana-Bray model. In red the Nishimori line that passes through the tricritical point and never enters in the spin glass phase. Image taken from [23].

3 The Semisupervised Censored Block Model

The paragraphs before refer all to unsupervised models, i.e. where no further information than the one carried by the edges is present. Now we introduce the Semisupervised Censored Block Model, in which the true label of some nodes is revealed.

In other words, a teacher, who knowns if the pictures represent a donkey or a horse, comes and helps the child by revealing him the correct animal (true label) depicted on some of the pictures.
By exploiting the further information given by the teacher, the child can implement a better prediction on still unknown pictures: he is solving the Semisupervised Censored Block Model!

Intuitively, as soon as the child can apply once his tool, if he perform his prediction on an unlabelled data only considering the similarities, if there are, with the labelled ones, then on average he would infer better than random guess. This is essentially why, as soon as $c > 0$, the detectability threshold disappears. Once again, we could know the generative model (so the probability p, and thus β_N) and being automatically in the Bayes Optimal setting, or not. In the second case, the child can estimate the tool's efficiency (p) by looking at how well it performed on the labelled dataset, as it is better explained at the end of this section.

3.1 Formulation

The graph is the same Erdős–Rényi with average degree c and couplings whose weights are given by (1). Then the true labels of some nodes are revealed, in number $n_l = \nu n$, with ν the fraction of labelled nodes. We refer to n_u as the number of unlabelled nodes, U, L respectively to the set of unlabelled and labelled nodes, σ_u, σ_l as the vectors of respectively unlabelled and labelled, $\hat{\sigma}_u$ as the vector of inference of the unlabelled belonging classes. By σ_i or σ_j we refer instead to single spins.
The goal is to optimize the inference only on unlabelled nodes, which corresponds to maximize

$$O(\hat{\sigma}_u) = \frac{1}{n_u} \hat{\sigma}_u^T \sigma_u^* \tag{12}$$

Since this quantity (12) which we refer as "overlap" is $O \in [-1, 1]$ we define alternatively the following quantity, which we call "score"

$$S(\hat{\sigma}_u) = 2 \left| \frac{1}{n_u} \sum_{i \in U} 1(\hat{\sigma}_i = \sigma_i^*) - 0.5 \right| \tag{13}$$

which has the property to be $S = 0$ for random guess and $S = 1$ for perfect recovery. Moreover, since in the configurations where we can apply spectral algorithms (where only couplings are present) the two true classes are defined unless a global spin flip, (13) takes into account this fact producing the same score for $\sigma_u \to -\sigma_u$. This is why in the simulations that will follow we estimate the algorithm performance through the score defined in (13) instead of the overlap (12).

3.2 Semisupervised Ising Spin Glass

There are various possible ways to embed the information surplus deriving from the knowledge of the true label of some nodes, which we call *semisupervised information*. For example, the semisupervised information can be seen as prior in the Bayes equation (5), or else embedded through modification of the couplings between labelled nodes that become always informative and with infinite modulus. This is the way we procceed in the next section. For now, it is sufficient to see that the semisupervised information, if encoded in the Bayes derivation (5), brings in any case to a different posterior and thus an Ising model that cannot be mapped anymore to a Viana-Bray model.
Let's pretend that we decided to encode the information of the true label of some nodes in the form of a prior, (5) becomes

$$P(\sigma | (J_{ij})_{ij \in E}) \ \propto \ e^{\sum_{i \in L} h_i \sigma_i} e^{\beta_N \sum_{ij \in E} J_{ij} \sigma_i \sigma_j} \tag{14}$$

where $h_i = \pm\infty$ are the infinite field related to the labelled nodes and pointing towards their revealed true polarizations.
Looking at the new Ising model of (14), one can ask if the temperature β_N calculated as (3) is still the optimal temperature at which solving the Ising model (i.e. if considerations made in the section before are still valid), or not: on one side β_N comes from a Bayesian inference and it should be exact, on the other side the semisupervised

information is encoded in a totally arbitrary way. However, if we have the certainty on the true label of some nodes, we set the fields to infinity and no problem arises: the Ising model can be reduced to only unlabelled nodes and β_N will still be optimal[3]. As soon as we do not have the certainty on the label, but only a higher probability for some nodes to be in a class rather than in another, fixing the correct field h_i is tricky as it would influence the whole model (temperature included). The second is an interesting case and very likely in real-life (even the labelling process is not 100% correct!), but for the sake of simplicity is not treated here.[4]
Considering this, we will continue to use β_N to solve the Ising model associated to the Semisupervised CBM. Moreover, we will exploit the fact that β_N, even if not known *a priori*, can be empirically estimated in a semisupervised setting from labelled-labelled connections as

$$\hat{p} = \frac{\sum_{ij \in \vec{E}, i,j \in L} \mathbf{1}(J_{ij} = \sigma_i^* \sigma_j^*)}{\sum_{ij \in \vec{E}, i,j \in L} 1} \tag{15}$$

and thus

$$\hat{\beta}_N = \frac{1}{2} \log \frac{\hat{p}}{1 - \hat{p}} \tag{16}$$

[3]Actually β_N will still be Bayes optimal, but in principle nothing guarantees that it is also the best temperature to solve the new "semisupervised" Ising model. However, arguments for the analogy of the Bayes optimal temperature and the Nishimori temperature for a general model can be found in [31].

[4]One further point argued by [13] is that the introduction of labelled nodes (further constraints) increases frustration, enlarging the Spin-Glass phase and making the free energy landscape more ragged: if we solve numerically the problem with Montecarlo methods at a generic temperature β, with high probability they get stuck in local suboptimal minima.

4 How to include the information on labelled nodes?

In graph-oriented semisupervised clustering we have the advantage of possessing some further information on some nodes, i.e. knowing their true label. The game is then to find a good strategy that reduces the initial problem to an efficiently solvable one. The considerations in this section are not only linked to the Censored Block Model, but can be easily generalized to any model that can be mapped to an Ising Spin Glass.

In the unsupervised setting, the probability of a configuration in the Ising model correspondent to the CBM is, as we have seen,

$$P(\sigma) \equiv P(\sigma_u, \sigma_l) = \frac{e^{\beta \sum_{<ij>\in E} J_{ij}\sigma_i\sigma_j}}{Z} \tag{17}$$

In a semi supervised problem we have information about the labels of some nodes, i.e. σ_l are no longer random variables but are fixed to a value $\sigma_l^* = \pm 1$ according to their true class. We could represent this gain of information as a difference in entropies

$$\Delta S = S[P(\sigma_u, \sigma_l)] - S[P(\sigma_u, \sigma_l|\sigma_l = \sigma_l^*)] \tag{18}$$

where the probability of the first term of the r.h.s. has been defined before and the one of the second is actually restricted on the unlabelled nodes and reads

$$P(\sigma_u|\sigma_l = \sigma_l^*) = \frac{e^{\beta \sum_{i,j\in U,} J_{ij}\sigma_i\sigma_j} e^{\beta \sum_{i\in U, j\in L} J_{ij}\sigma_i\sigma_j^*} e^{\beta \sum_{i,j\in L} J_{ij}\sigma_i^*\sigma_j^*}}{Z} \tag{19}$$

For simplicity of notations, $< ij > \in E$ or $i, j : i < j$ have been omitted under every sum (notice however that when the edge is not present it just contributes with a factor 1). We are interested in finding equivalent forms of (19), in order to reshape the semisupervised problem in a convenient way. Indeed, being approximations, algorithms may work better (or being better understood) on a configuration rather than on another.
The section is organized as follows: first we see two ways to polarize labelled nodes, the second one using only edges and thus permitting the application of spectral algorithms. Then we derive another configuration for the semisupervised problem that reduces the graph to the unlabelled subgraph, paying the price of introducing finite fields. From this perspective, it will be straightforward to identify the two kinds (semisupervised and unsupervised) of information and to notice how in a semisupervised setting the threshold disappears. Last, we introduce another configuration, this time consisting of an approximation of (19), but whose highest probability vector (i.e. its ground state) is shown to be equivalent to the one of (19).

4.1 Polarizing labelled nodes

The most intuitive way to artificially make a labelled node polarize is by adding in the Hamiltonian an infinite local field that points towards the direction of its assigned class. In other words, for each labelled node we insert a prior in the Bayes theorem equation (5) and we end up with

$$P_{\mathbf{IF}}(\sigma_u, \sigma_l; C|\sigma_l^*) = \frac{e^{\beta(\sum_{i,j\in U} J_{ij}\sigma_i\sigma_j + \sum_{i\in U}(C\sigma_i^*)\sigma_i)}}{Z'} \tag{20}$$

with $C \to \infty$. We call this configuration "**Infinite Fields**". This is similar to what has been done in one of the first papers on the topic [30], although there the fields to polarize labelled nodes are set finite. Obviously, the marginal distribution of the unlabelled $P_{IF}(\sigma_u) = \sum_{\sigma_l=\{\pm 1\}} P_{IF}(\sigma_u, \sigma_l; C|\sigma_l^*)$ is the same as (19), as the probability of a configuration with some labelled polarized oppositely w.r.t. their fields is null.
However, the presence of fields does not allow us to implement spectral algorithms, that can take as inputs only pairwise couplings. Thus we propose another method to artificially polarize labelled nodes using only edges, consisting of setting infinite labelled-labelled couplings, ferromagnetic $(+\infty)$ if the two labelled nodes belong to the same class, antiferromagnetic $(-\infty)$ otherwise. This corresponding to an Ising model here called "**Infinite Couplings**"

$$P_{\mathbf{IC}}(\sigma_u, \sigma_l; C|\sigma_l^*) = \frac{e^{\beta \sum_{i,j\notin L} J_{ij}\sigma_i\sigma_j} e^{\beta \sum_{i,j\in L}(\sigma_i^*\sigma_j^* C)\sigma_i\sigma_j}}{Z} \tag{21}$$

with $C \to \infty$. In order to ensure a correct polarization, one may have to draw a Minimum Spanning Tree (MST) on the subgraph of labelled nodes, otherwise there could be clusters with uncoherent polarizations. Anyways, this turns not to be the case for sufficiently high average connection c of the original ER graph and sufficiently high percentage of labelled nodes ν. Similar ideas have been used in the context of the Stochastic Block Model

by [9], where the authors modify the corresponding Potts Hamiltonian enclosing in further edges the information given by the revelation of some labels. The method is well described and inserted in a wider context in [34]. Although inserting in the graph potentially infinite cuplings (anyways, in other more general problems C could be set not necessarily to ∞ but tuned in order to reflect the reliability of the semisupervised information), this configuration preserves the local tree-like structure of a the original graph, if it is sparse. Last, one can check again that, unless an uninfluential global spin flip symmetry ($P_{\mathbf{IC}}(\sigma_u, \sigma_l) = P_{\mathbf{IC}}(-\sigma_u, -\sigma_l)$), the marginal probability $P_{\mathbf{IC}}(\sigma_u) = \sum_{\sigma_l=\{\pm 1\}} P_{\mathbf{IC}}(\sigma_u, \sigma_l; C|\sigma_l^*)$ with $C \to \infty$ is the same as (19).

4.2 From nodes to fields

The last factor in the numerator of the r.h.s. of (19) is just a constant, so (19) can be rewritten as

$$P(\sigma_u|\sigma_l = \sigma_l^*) = \frac{e^{\beta \sum_{i,j\in U} J_{ij}\sigma_i\sigma_j} e^{\beta \sum_{i\in U, j\in L} J_{ij}\sigma_i\sigma_j^*}}{Z'} \tag{22}$$

Defining for each unlabelled node $i \in U$ a field that encodes its connection with labelled nodes as

$$h_i = \sum_{j\in L: <ij>\in \vec{E}} J_{ij}\sigma_j^* \tag{23}$$

we find that the probability of a configuration after revealing the labels of a set of nodes L becomes the Boltzmann weight of an Ising model restricted to the unlabeled nodes $i \in U$ with fields given by (23), i.e.

$$P_{\mathbf{FF}}(\sigma_u|\sigma_l^*) = \frac{e^{\beta(\sum_{i,j} J_{ij}\sigma_i\sigma_j + \sum_i h_i\sigma_i)}}{Z'} \tag{24}$$

which goes under the name of "**Finite Fields**" and corresponds exactly to (19) restricted to unlabelled nodes.
It is easy to see that independently from the couplings J_{ij} as long as the vector of fields $h \neq 0_{n_u}$ (so for $c > 0$ and $\nu \neq 0$) and $\beta = \beta_N \neq 0$, the expected value of the local field points towards the true magnetization of the node, i.e. $\sigma_i^* \mathrm{E}[h_i] > 0$ and thus solving the Ising model would give predicted magnetizations averagely aligned with fields, thus on average a performance better than random guess. This argument supports once again the disappearance of the threshold in semisupervised problems, showed for the Stochastic Block Model for two classes by [33], which is claimed more rigorously in the upcoming sections. The proposed configuration (24) represents an improvement with respect to [30], since the system to solve has $n - n_l = n_u$ nodes. Moreover, it is clear from the Hamiltonian of (24) that we isolated two different kinds of information:

- the term $\sum_{i,j} J_{ij}\sigma_i\sigma_j$ encodes the "unsupervised" information through the couplings $\{J_{ij}\}$
- the term $\sum_i h_i\sigma_i$ encodes the "semisupervised" information through the fields $\{h_i\}$

Artificially enlarging one term rather than another would correspond to rely more on semisupervised or unsupervised information.

4.3 From fields to edges

Now let's try, starting from (24), to shift the information exclusively on edges, in order to be able to devise spectral algorithms. Basically, we add a further node to the unlabelled graph, namely the new spin σ_+, and rewrite each field associated to a node i as a coupling between node i and the added node. Doing so, we end up with the configuration

$$P_{\mathbf{FC}}(\sigma_u, \sigma_+|\sigma_l^*) = \frac{e^{\beta(\sum_{i,j} J_{ij}\sigma_i\sigma_j + \sum_i h_i\sigma_i\sigma_+)}}{Z''} \tag{25}$$

which we call "**Finite Couplings**".
Computing the marginal on unlabelled of (25)

$$P_{\mathbf{FC}}(\sigma_u|\sigma_l^*) = \sum_{\sigma_+=\pm 1} P_{\mathbf{FC}}(\sigma_u, \sigma_+|\sigma_l^*) = \frac{e^{\beta(\sum_{i,j} J_{ij}\sigma_i\sigma_j + \sum_i h_i\sigma_i)}}{Z''} + \frac{e^{\beta(\sum_{i,j} J_{ij}\sigma_i\sigma_j - \sum_i h_i\sigma_i)}}{Z''} \tag{26}$$

we notice that it does not correspond anymore to (19). However, the most probable configuration (i.e. the ground state) of (25) restricted to unlabelled and (24) are the same:

$$\arg\max_{\sigma_u} P_{\mathbf{FC}}(\sigma_u|\sigma_l^*) = \arg\max_{\sigma_u} P_{\mathbf{FF}}(\sigma_u|\sigma_l^*) \tag{27}$$

We can understand this by noticing that depending on the polarization ± 1 of the spin σ_+ in (25), the unlabelled spins σ_u will arrange to minimize the Hamiltonian in one of the two symmetric configurations $+\sigma_u$ or $-\sigma_u$. Being the coupling term invariant under this global spin flip, whatever polarization σ_+ there will be one of the two symmetric configurations $+\sigma_u$ or $-\sigma_u$ that equivalently minimizes the Hamiltonian. Moreover, these two configurations correspond to the same inference. Thus, the ground state configuration does not really depend on the polarization of the added node σ_+.
Another way to see this is rewriting (26) as

$$P_{\mathbf{FC}}(\sigma_u|\sigma_l^*) = \frac{e^{\beta(\sum_{i,j} J_{ij}\sigma_i\sigma_j)}\cosh(\beta(\sum_i h_i\sigma_i))}{Z''} \tag{28}$$

Arguing that configurations with spins (partially) aligned with fields have a much higher probability to happen and noticing that for this kind of configurations the hyperbolic cosine will be dominated by one of its two terms,

$$\cosh(\beta(\sum_i h_i\sigma_i)) \approx e^{\beta(\sum_i h_i\sigma_i)} \tag{29}$$

unless constants, and the approximation is more valid the greater is β, then (26) will resemble at (19). Specifically, setting $\beta \to \infty$, the configurations turn to be exactly the same (which means that their ground state is the same, too). Having similar free energy profile and same ground state, for not too hard problems and sparse graphs we take (25) as a good approximation of (19) and use it to perform the inference through the adjacency matrix. Indeed in the next section we interpret the adjacency matrix as a ground state search with relaxed constraints as well as the hessian of the mean field free energy. We preferred the Finite Couplings configuration rather than the Infinite Couplings, since the perturbation on the unsupervised adjacency matrix is more tractable (as well as the matrix has reduced size).

5 Inference with Naive Mean Field

5.1 Theoretical background

Before applying the algorithms on the derived configurations, we briefly revisit the derivations and interpretations of the mean field approximation and the adjacency matrix spectral algorithm. The complete derivations are reported in the "Theoretical Background" section of the appendix and will be used in the successive sections.
The Naive Mean Field approximation is derived by the minimization of the mean field variational free energy (section B.1), to underline the analogy with the derivation of Bethe-Hessian (not reported), where instead the variational free energy has to be minimized under Bethe distribution. If the mean field variational free energy brings to a fixed point equation (Naive Mean Field Equation), the Bethe free energy minimization is made through its (Bethe-)Hessian calculated at the paramagnetic $m = 0$ point. The eigenvector correspondent to the smallest eigenvalue of the Hessian represents the steepest directions for which the paramagnetic point is unstable, thus it points towards a minimum of the related free energy and corresponds to an assignment of spin magnetizations correlated to their true classes. We can still apply an analogous idea to the simpler mean field free energy: doing so, we get that the Hessian is bijectively linked with the adjacency matrix J (section B.2). Moreover, we find that in this minimization the inverse temperature ceases to play a role (the eigenvectors of the Adjacency Matrix stay the same under a rescale of the couplings). Thus the adjacency matrix, not including temperature, does not operate in the Bayes optimal regime. This is confirmed by the ideas in section B.3, where it is shown that the adjacency matrix identifies the ground state of an Ising Hamiltonian with no fields (so a problem independent on the inverse temperature β) relaxing the constraint of discreteness of spins. The latter represents a consistent approximation of the original problem. In the last paragraph of the relative section in the appendix, we deepen the connection between Mean Field approximation and the adjacency matrix, that turn to be equivalent if we set β for the mean field as $\beta = \frac{1}{\lambda_{max}}$, where λ_{max} is the largest eigenvalue of the correspondent adjacency matrix, once $\frac{1}{\lambda_{max}}$ is small enough.

5.2 Naive Mean Field on the derived semisupervised configurations

Calling m the magnetization vector and β the external parameter interpreted as the inverse temperature, the Naive Mean Field algorithms gives an approximation of the average magnetization of each spin by looking for a fixed point of the celebrated mean field equation, so gives a $\hat{m}$ such that

$$\hat{m} = \tanh\left(\beta(J\hat{m} + h)\right) \tag{30}$$

where J is the Adjacency matrix of the graph and h is the vector of fields. If we apply the Naive Mean field algorithm to the different exactly derived semisupervised configurations listed in section 4, we find that they correspond to the same mean field equation and we get exactly the same performance. Indeed, for example for the configuration with IC whose Hamiltonian is

$$-H(\sigma_u, \sigma_l) = \sum_{i,j \notin L} J_{ij}\sigma_i\sigma_j + \sum_{i,j \in L} (\sigma_i^*\sigma_j^* C)\sigma_i\sigma_j \tag{31}$$

the naive mean field equations are for labelled L and unlabelled U nodes respectively

$$\begin{aligned} &\text{i}\in L \qquad m_i = \tanh\left(\beta(\textstyle\sum_{j\in\partial i, j\in U} J_{ij}m_j + \sum_{j\in\partial i, j\in L}(\sigma_i^*\sigma_j^* C)m_j)\right) \\ &i \in U \qquad m_i = \tanh\left(\beta(\textstyle\sum_{j\in\partial i} J_{ij}m_j\right) \end{aligned} \tag{32}$$

For the labelled nodes $i \in L$, the infinite term, always present by construction, dominates and they always correctly polarize to $m_i = \sigma_i^*$. Once labelled polarized, the system is reduced to unlabelled with equations

$$i \in U \qquad m_i = \tanh\left(\beta(\sum_{j\in\partial i, j\in U} J_{ij}m_j + \sum_{j\in\partial i, j\in L} J_{ij}\sigma_j^*)\right) \tag{33}$$

This is exactly the mean field equation that we would have obtained with finite fields, indeed

$$h_i = \sum_{j \in \partial i, j \in L} J_{ij} \sigma_j^* \tag{34}$$

so of the configuration with Hamiltonian

$$-H(\sigma_u, \sigma_l) = \sum_{i,j \in U} J_{ij} \sigma_i \sigma_j + \sum_{i \in U} h_i \sigma_i \tag{35}$$

and so on for other configurations.

5.3 Disappearance of detectability threshold

More interestingly, we look at results in section B.4 where we showed that the paramagnetic-ferromagnetic transition for the mean field happens at temperature

$$\beta_{PF} = \frac{1}{\lambda_{max}} \tag{36}$$

Specifically, for $\beta > \beta_{PF}$ the naive mean field has a ferromagnetic solution while for $\beta < \beta_{PF}$ it is in the paramagnetic phase. Note that the ferromagnetic solution is always correlated with the planted assignment since exploiting the Gauge invariance we can map the problem to another one with all spins belonging to the same class. Now if we take the configuration (25) with finite couplings, its adjacency matrix reads

$$J = \begin{bmatrix} J_U & h \\ h^T & 0 \end{bmatrix} \tag{37}$$

In the worst case in which the matrix of the unlabelled graph J_U is very sparse, for a non null percentage of labelled nodes $\nu > 0$, the largest eigenvalue λ_{max} of J will be dominated by the dense "semisupervised" perturbation (i.e. the field vector h) as soon as $p > 0.5$ and in the appendix C it is shown that

$$\lambda_{max} \sim \sqrt{n_u} \tag{38}$$

So

$$\beta_{PF} = \frac{1}{\lambda_{max}} \to 0 \tag{39}$$

in the thermodynamic limit $n_u \to \infty$, showing that the mean field has always a ferromagnetic solution, as long as $p > 0.5$. Since every ferromagnetic solution has a non null score in the Gauge transformed model, this proves the absence of the detectability threshold.
A straightforward alternative way to see this, is to neglect at all the connections among unlabelled nodes and apply the mean field equation to the finite field configurations modified in this way. We would obtain the trivial expression of the average magnetization of each node

$$m_i = \tanh \beta h_i \tag{40}$$

and build the inference vector

$$\sigma_i = sign(m_i) = sign(\tanh \beta h_i) = sign(h_i) \tag{41}$$

As soon as we are provided with the information of the true label of some nodes, i.e. semisupervised $\nu > 0$, and the problem is solvable $p > 0.5$, the expected value of h_i is greater than zero and points towards the direction of the true assignment, then $\frac{1}{n_u} \sigma^T \sigma^* > \frac{1}{2}$ and the performance is better than random guess. This trivial algorithm that simply takes the sign of h_i, when it is different from zero, will be called "Simple Comparison" and corresponds to a trivial supervised classification.

6 Inference with Adjacency Matrix

The inference with the adjacency matrix consists in taking as prediction the sign of the entries of the eigenvector associated to the largest eigenvalue λ_{max} of the adjacency matrix J, built from the Hamiltonian of a configuration without fields. Although it might be far from being the optimal polynomial algorithm to infer the planted structure of a semisupervised Censored Block Model, the inference method based on the adjacency matrix still reveals interesting features of semisupervised learning. Indeed, we will notice how if we blindly apply it to a semisupervised configuration (so where in theory we dispose of the same information of the unsupervised case, and a surplus) sometimes we end up with a poorer performance with respect to the unsupervised case. This is like if the child, euphoric for having being helped, is not able to ponder his intuitions by balancing the sources of informations, and surprisingly gets worse than before. This is not only specific of the model we have been working with, but poorer performance of semisupervised models with respect to unsupervised or fully supervised with just the labelled dataset have been identified *partout* in literature [2, 20]. Being suboptimal (and also more complicated!) than unsupervised or fully supervised represents the defeat of a semisupervised learning algorithm. Deeper studies to understand its complexity and exploit the surplus of information are needed. Of course, this is just an innocent and specific perspective on the problem.

Once derived the semisupervised Finite Couplings configuration (25), we look for the eigenvector associated to the largest eigenvalue of the adjacency matrix built from the Hamiltonian (which has only couplings) of (25), i.e. the vector $\hat{\mathbf{f}} = (\hat{f}_u, \hat{f}_+)$, $\hat{f}_u \in R^{n_u}, \hat{f}_+ \in R$, such that

$$J\hat{\mathbf{f}} \equiv \begin{bmatrix} J_U & h \\ h^T & 0 \end{bmatrix} \begin{bmatrix} \hat{f}_u \\ \hat{f}_+ \end{bmatrix} = \lambda_{max} \begin{bmatrix} \hat{f}_u \\ \hat{f}_+ \end{bmatrix} \tag{42}$$

and then take as predictions the sign of the entries of $\hat{f}_u$

$$\hat{\sigma}_u = sign(\hat{f}_u) \tag{43}$$

We showed in the section before that the normalized vector $(\hat{f}_u, \hat{f}_+)$ is the one that solves the unconstrained Hamiltonian minimization

$$\max_{\mathbf{f}} \frac{1}{2}\mathbf{f}^T J\mathbf{f} = \max_{f_u, f_+} \sum_{i,j \in U} J_{ij} f_i f_j + \sum_{i \in U} h_i f_i f_+ \tag{44}$$

"unconstrained" in the sense that the entries of the vector $\mathbf{f}$ can be real numbers (with the constraint of unitary norm, not reported in the equation above), differently from discrete spins $\sigma_i = \pm 1$ in the original ground-state search

$$\max_{\sigma} \frac{1}{2}\sigma^T J\sigma = \max_{\sigma_u, \sigma_+} \sum_{i,j \in U} J_{ij} \sigma_i \sigma_j + \sum_{i \in U} h_i \sigma_i \sigma_+ \tag{45}$$

6.1 The inference is unbalanced: two ways to show

We show that the proposed inference through (43) is deeply unbalanced towards the semisupervised information, i.e. the information coming from the connections with labelled nodes.
From the eigenvalue equation (42) we get a system of $n_u + 1$ equations

$$\begin{cases} J_U \hat{f}_u + h\hat{f}_+ = \lambda_{max} \hat{f}_u \\ h^T \hat{f}_u = \lambda_{max} f_+ \end{cases} . \tag{46}$$

which substituting

$$J_U \hat{f}_u + \frac{hh^T}{\lambda_{max}} \hat{f}_u = \lambda_{max} \hat{f}_u \tag{47}$$

thus we have for the inference eigenvector of unlabelled $\hat{f}_u$

$$(J_U + \frac{hh^T}{\lambda_{max}}) \hat{f}_u = \lambda_{max} \hat{f}_u \tag{48}$$

The largest eigenvector of the adjacency matrix in (42) with J_U not dense goes as

$$\lambda_{max} \sim \sqrt{n_u} \tag{49}$$

and the calculus is reported in the appendix.
Once we know this we can get an intuition on how the inference is performed by expliciting the equation (48): for every $\hat{f}_i$, $\hat{f}_i \in \hat{f}_u$, $i \in U$

$$\hat{f}_i = \frac{1}{\lambda_{max}} \Big(\sum_{j \in \partial i} J_{ij} \hat{f}_j + \sum_{j=1}^{n} \frac{h_i h_j}{\lambda_{max}} \hat{f}_j \Big) \tag{50}$$

The first term of the r.h.s. represents the unsupervised information and it is a sum, on average, of c elements. The second, which represents the semisupervised information, is instead a sum of n_u terms, divided by λ_{max}. It is easy to understand that the second term is $\sqrt{n_u}$ times larger than the first: in the thermodynamic limit, the inference will be based only on the second term

$$\hat{f}_i \approx \frac{1}{\lambda_{max}} \Big(\sum_{j=1}^{n} \frac{h_i h_j}{\lambda_{max}} \hat{f}_j \Big) \tag{51}$$

that corresponds to

$$h h^T \hat{f}_u = \lambda_{max}^2 \hat{f}_u \tag{52}$$

which is solved by $\lambda_{max}^2 = h^T h$ and

$$\hat{f}_u = h \tag{53}$$

the prediction vector would be equal to the vector of fields, and so the sign. Thus the inference would be very similar to the trivial algorithm called "Simple Comparison" explained in the section 5. Not equal because, in the case in which h has some entries h_i equal to zero, the term that encloses the semisupervised information vanishes and the inference would be guided by the unlabelled connections, i.e. the first term in the r.h.s. of (50).

Now we try to develop another way to look at the same problem, i.e. the overrating of the semisupervised information in our proposed setting.

It is easy to see that the solution of the eigenvalue equation (42) for $\hat{f}_u$, assuming without loss of generality that the returned eigenvector is normalized, reads

$$\hat{f}_u = \sqrt{1 - ||\hat{f}_u||^2} \cdot \Big[(\lambda_{max} I_{n_u} - J_U)^{-1} h \Big] \tag{54}$$

where since we infer through taking the sign, the factor in front turns to be irrelevant

$$sign\Big(\sqrt{1 - ||\hat{f}_u||^2} \cdot (\lambda_{max} I_{n_u} - J_U)^{-1} h) \Big) = sign\Big((\lambda_{max} I_{n_u} - J_U)^{-1} h \Big) \tag{55}$$

We then manipulate the expression (54)

$$\Big(\lambda_{max} I_{n_u} - J_U \Big)^{-1} h = \Big(\lambda_{max} (I_{n_u} - \frac{J_U}{\lambda_{max}}) \Big)^{-1} h = \frac{1}{\lambda_{max}} \Big(I_{n_u} - \frac{J_U}{\lambda_{max}} \Big)^{-1} h \tag{56}$$

and discard from now on the $\frac{1}{\lambda_{max}}$ factor. Since $||\frac{J_U}{\lambda_{max}}|| \equiv \max_{ij} |\frac{J_{ij}}{\lambda_{max}}| < 1$, then[5] the expression before becomes proportional to

$$\Big(I_{n_u} + \frac{J_U}{\lambda_{max}} + \frac{J_U^2}{\lambda_{max}^2} + \frac{J_U^3}{\lambda_{max}^3} + ... \Big) h \tag{57}$$

Thus if

$$\Big|\Big| \Big(\sum_{\gamma=1}^{\infty} \frac{J_U^\gamma}{\lambda_{max}^\gamma} \Big) h \Big|\Big| << ||h|| \tag{58}$$

then $sign(\hat{f}_u) \approx sign(h)$ and the information coming from the unlabelled connections is systematically ignored in the inference process. Intuitively, the condition is fulfilled for a sufficiently sparse matrix J_U, and a sufficiently high labelled percentage ν. However, the quantity (58) is difficult to estimate *a priori* and also numerically. In any case, if λ_{max} stayed finite, with high probability the condition is not fulfilled, since the number of non-vanishing elements of the powers of J_U is proportional to n_u. Indeed, in the next paragraph we propose a modified adjacency matrix $\tilde{J}$ whose largest eigenvalue $\tilde{\lambda}_{max}$ stays finite.

[5] If $||A|| < 1$, then $(I - A)^{-1} \approx I + A + A^2 + A^3 + ...$

6.2 (Un)Intuitively rescaling the labelled-unlabelled connections

From the discussion of the orders of the terms in equation (50), it appears intuitive to rescale by a factor of order $O(\frac{1}{\sqrt{n_u}})$ the overweighted term encoding the semisupervised information, in order to make the two terms comparable. This can be done by rescaling the vector of fields h, representing the connections between labelled and unlabelled in the way

$$h \longrightarrow \frac{\alpha}{\sqrt{n_u}} h \tag{59}$$

where $\alpha \in R$ is a factor of order 1, left free for the moment. Thus the modified adjacency matrix $\tilde{J}$ reads

$$\tilde{J} = \begin{bmatrix} J_U & \frac{\alpha}{\sqrt{n_u}} h \\ \frac{\alpha}{\sqrt{n_u}} h & 0 \end{bmatrix} \tag{60}$$

Through the same calculus of before, we get that now the maximum eigenvalue of $\tilde{J}$ stays practically constant in n_u, $\tilde{\lambda}_{max} \sim O(1)$.

The direct expression, solution of the eigenvalue equation, of the inference vector $\hat{\tilde{f}}_u$ now reads

$$\hat{\tilde{f}}_u = \sqrt{1 - ||\hat{\tilde{f}}_u||^2} \cdot \left[(\tilde{\lambda}_{max}(\alpha)\, I_{n_u} - J_U)^{-1} \frac{\alpha}{\sqrt{n_u}} h \right] \tag{61}$$

where the dependency of the largest eigenvalue by α is explicit, $\tilde{\lambda}_{max}(\alpha)$, and it is practically the only change with respect to the correspondent equation of the unmodified case (54), since the factor $\frac{\alpha}{\sqrt{n_u}}$ is irrelevant for the inference (same for $\sqrt{1 - ||\hat{\tilde{f}}_u||^2}$).

Intuitively, the factor α controls the balance between the two sources of information:

- $\alpha \to 0$ cancels the contribution of the semisupervised information, thus recovering the performance of unsupervised learning
- $\alpha \to +\infty$ gives infinite weight to semisupervised information, recovering the performance of the unmodified algorithm

There will exist then a finite α^* for which the inference is optimally balanced.

6.3 Analogies with Regularization with centered similarities

The derived expression (61) is very similar to the one that the authors of [19] derived from their proposed algorithm, which reads

$$\hat{f}_u^{(Mai)} = (\lambda I_{n_u} - W)^{-1} h' \tag{62}$$

that comes from the constrained maximization problem

$$\begin{cases} \max_{f_u} f^T W f \\ \\ ||f_u||^2 = n_u e^2 \end{cases} . \tag{63}$$

where the only differences are that W is the centered adjacency matrix[6], $h' = W_{ul}\sigma_l^*$ so as before but using the labelled-unlabelled connections of the centered adjacency matrix, and $\lambda(e)$ is a specific function of the parameter e, left free. They proposed the maximization above after having identified that the standard semisupervised Laplacian's effective learning from unlabelled data was negligible, at least for high-dimensional gaussian generated data, in order to try to fix this problem. Through a rigorous theoretical analysis, in [19] is claimed that

- $e \to 0$ recovers the starting unmodified algorithm, in this case Laplacian regularization
- $e \to +\infty$ recovers the performance of unsupervised classification

[6]The original adjacency matrix J is centered through the projection matrix P as $W = PJP$, where $P = (I_n - \frac{1}{n} 1_n 1_n^T)$. Our matrix J, instead, is centered by construction.

- there always exist a e^*, between the two extremes, that leads to a performance gain over Laplacian regularization

Last, note that the function to maximize reads

$$f^T W f = \sum_{i,j=1}^{n} w_{ij} f_i f_j \equiv Q(f) \tag{64}$$

So, the approximation in [19] resides in the choice of a somewhat arbitrary cost function to minimize, i.e. $Q(f)$. The inference is regularized by an "estensivity" condition on the norm of the predicted unlabelled vector and controlled by the hypermarameter e.
In our proposed algorithm, we start by an exactly derived posterior distribution and arrive through an approximation to identify a cost function to minimize very similar to (64), for a different generative model. The approximation resides in passing from this true posterior to the configuration with Finite Couplings (25), whose marginal for the unlabelled (26) is slightly different from (19).
Then we use the adjacency matrix to predict the average magnetization of each spin in the Ising model, which we have seen (appendix B.3) that it can be interpreted also as an unconstrained ground state search. To improve this approximation, after noticing that the inference is unbalanced, we modify the adjacency matrix by scaling the labelled-unlabelled connection of a factor $\alpha/\sqrt{n_u}$. This, with α appropriately set, forces the algorithm to learn also from the unlabelled connections, exactly like the constraint on the norm of $\hat{f}_u$ does in (63). Indeed, α plays the same (inverse) role of e in the method of [19].
Even though we lack of a rigorous mathematical analysis like the one of [19], the cost function that we proposed comes from a somewhat clear derivation. Exploiting this fact, we try to push further and devise a procedure for setting α, at least for some problems, while the correspondent value e in [19] was left as a hyperparameter.

6.4 Estimating the optimal α for easy problems

The exact procedure, once we know the true posterior, would consist in calculating the marginal probability of each node i to be in each class and then take as inferred class the one correspondent to the highest probability. This is of course impossible to obtain in polynomial time, since in order to get the marginals one would have to calculate the probability of each possible configuration: the number of possible configurations scales exponentially with the number of nodes n. Thus there exist algorithms that looks for approximate solutions. One consists in approximating the partition function sum with its highest term, corresponding to the ground state of the Hamiltonian in the posterior distribution. This is still an NP-complete problem and it is justified if the entropy does not play a significant role, i.e. when the temperature of the derived Ising model is low (β_N high, easy problems), and the partition function sum is dominated by its highest term.
In our derivation, however, we performed another approximation: we passed from the exact posterior (19) to the Finite Couplings configuration, which has marginals slightly different from (19), as $2\cosh\left(\beta_N \sum_{i\in U} h_i \sigma_i\right) \neq e^{\beta_N \sum_{i\in U} h_i \sigma_i}$. Once again, for high β_N one of the two terms of the hyperbolic cosine is negligible with respect to the other, at least for the configurations whose probability really contributes (the one with spins at least partially aligned with the fields). Indeed, the ground state (correspondent to $\beta \to \infty$) of (25) to (19) for σ_u are the same.
These considerations motivate us to claim that taking as inference the ground state of the Hamiltonian of the Finite Coupling configuration is a valid approximation for the inference, at least for large β_N (easy problems). Thus, we set α to be the one whose related inference maximizes the value of the opposite Hamiltonian of (25).
In other words, we

- Choose an initial α to start
- Perform the adjacency matrix inference with fields rescaled of $\frac{\alpha}{\sqrt{n_u}}$, obtaining a prediction vector $\hat{\mathbf{f}} = (\hat{f}_u(\alpha), \hat{f}_+(\alpha))$
- Take as inference the sign of this vector, i.e. $\hat{\sigma}_u(\alpha) = sign(\hat{f}_u(\alpha))$ and $\hat{\sigma}_+(\alpha) = sign(\hat{f}_+(\alpha))$
- Calculate the value of the opposite Hamiltonian at that configuration, $-H(\hat{\sigma}_u(\alpha), \hat{\sigma}_+(\alpha))$
- Repeat for different α and choose the one for which $-H(\hat{\sigma}_u(\alpha), \hat{\sigma}_+(\alpha))$ is maximal

A good point is that we do not really know when this approximation is valid. Indicatively, from simulations with two equal classes and low percentage of labelled nodes ν, we observe that the approximation is consistent approximately after the unsupervised detectability threshold, i.e. for c and p such that $c(2p-1)^2 \approx 1$, so once we know c for $p > \frac{1}{2}(1 + \frac{1}{\sqrt{c}})$.

7 Proposed algorithms and Simulations

From the considerations of the previous paragraphs, the following algorithms are proposed:

Algorithm 1: Semisupervised CBM with Naive Mean Field (2 classes)

Input: An undirected graph $G = (V, E)$ with couplings $(J_{ij})_{ij \in E} = \pm 1$, a set of labelled nodes $L \subset V$ with label vector $\sigma_l^* \in R^{|L|}$

Output: Classification vector of unlabelled data, $\hat{\sigma}_u \in R^{|U|}$, $U = V \setminus L$

1. Compute the entries of the field vector h as (23)
2. Estimate the coupling reliability p through (15), if not known *a priori*, then the optimal temperature β_N through (16)
3. Iterate the constructed mean field equation (30), where $\beta = \beta_N$ and J is the adjacency matrix restricted to unlabelled-unlabelled connections, called J_U in (42), until convergence to a fixed point $\hat{m} \in R^{|U|}$
4. Build the classification vector $\hat{\sigma}_u = sign(\hat{m})$

Algorithm 2: Semisupervised CBM with adjacency matrix (2 classes)

Input: An undirected graph $G = (V, E)$ with couplings $(J_{ij})_{ij \in E} = \pm 1$, a set of labelled nodes $L \subset V$ with label vector $\sigma_l^* \in R^{|L|}$

Output: Classification vector of unlabelled data, $\hat{\sigma}_u^* \in R^{|U|}$, $U = V \setminus L$

1. Compute the entries of the field vector h as in (23)
2. Define the adjacency matrix J as in (37). Compute the eigenvector $\hat{\mathbf{f}} = (\hat{f}_u, \hat{f}_+)$ associated to the largest eigenvalue of J
3. Build the classification vector $\hat{\sigma}_u^* = sign(\hat{f}_u(\alpha^*))$

Algorithm 3: Semisupervised CBM with adjusted adjacency matrix (2 classes) for easy problems (sufficiently large c and p, see section 6.4)

Input: An undirected graph $G = (V, E)$ with couplings $(J_{ij})_{ij \in E} = \pm 1$, a set of labelled nodes $L \subset V$ with label vector $\sigma_l^* \in R^{|L|}$

Output: Classification vector of unlabelled data, $\hat{\sigma}_u^* \in R^{|U|}$, $U = V \setminus L$

1. Compute the entries of the field vector h as in (23)
2. Define the Hamiltonian of the configuration with Finite Couplings as $H(\sigma_u, \sigma_+) = \sum_{i,j} J_{ij}\sigma_i\sigma_j + \sum_i h_i\sigma_i\sigma_+$
3. Consider the adjusted adjacency matrix for a general α, $\tilde{J}(\alpha)$ as in (60)
4. Choose a set of possible factors α and repeat the procedure described in section 6.4 using the Hamiltonian above, in order to calculate the best factor α^*
5. Build the classification vector as $\hat{\sigma}_u^* = sign(\hat{f}_u(\alpha^*))$

7.1 Simulations

The algorithm used in simulations are the three above plus the Unsupervised Adjacency Matrix (that operates like the unmodified adjacency matrix but simply neglecting the semisupervised information, so correspondent to set $\alpha = 0$) and the one with adjusted matrix but setting the optimal α^* by optimizing the correspondent scores[7].
Figure 2 shows the defeat of the semisupervised unmodified adjacency matrix (Algorithm 2): for some p, the performance of semisupervised adjacency matrix is poorer than unsupervised adjacency matrix on the same data! In other terms, the inference is strongly biased towards the semisupervised information, as discussed in section

[7] Fixing α^* using the feedback of the performance, which is an unrealistic case: if I already know the planted assignment why should I infer it? Indeed, such algorithm is used just to investigate the reliability of Algorithm 3.

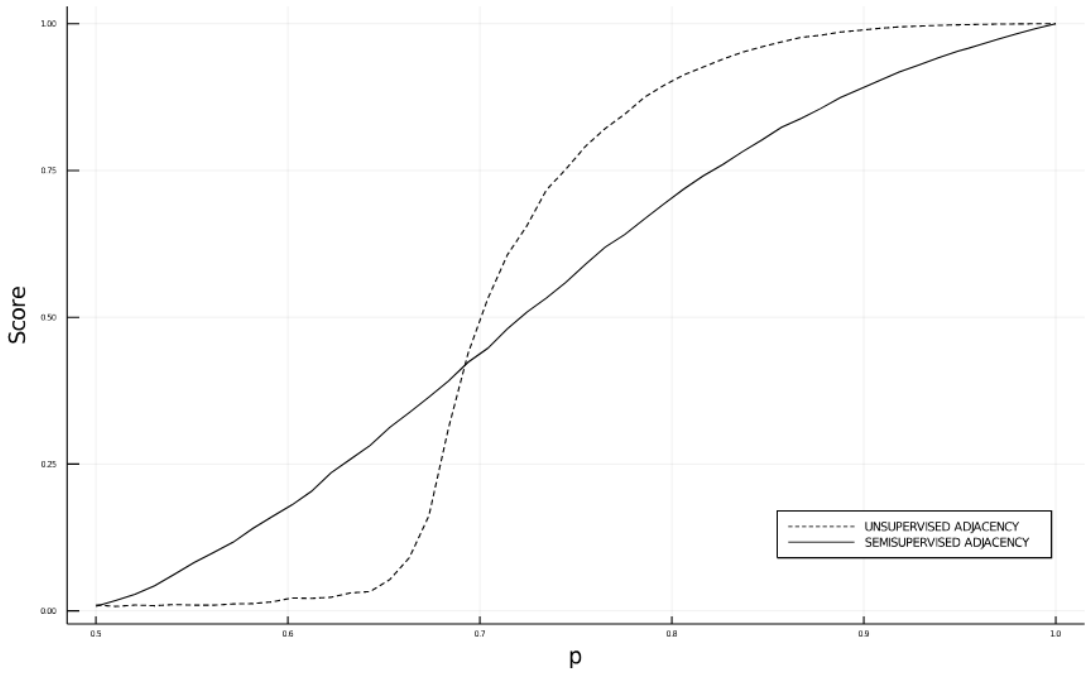

Figure 2: **Performances of unsupervised and semisupervised inference with unmodified adjacency matrix (Algorithm 2).** Performances (the score as in (13)) of the unsupervised ($\nu = 0$) and semisupervised unmodified algorithm for a labelled percentage $\nu = 0.1$ (10% labelled), as a function of the easyness of the problem p. Other parameters: $n = n_u + n_l = 10000$, $c = 10$, planted assignment of equally sized classes. Performances are averaged over 30 trials.

6.1, which gives a manifestly suboptimal performance for easy problems. However, the community detection with semisupervised adjacency, as predicted, does not show any threshold, contrairly to its unsupervised analogous.
Figure 3 shows how this problem can be solved adjusting the adjacency matrix by rescaling fields, whose performance for optimal α^* is always beyond the ones of Unsupervised and unmodified Semisupervised, as discussed in sections 6.2 and 6.3.
Figure 4, instead, shows how reliable is setting α^* through the Hamiltonian minimization, i.e. applying Algorithm 3. It is evident, indeed, that for difficult problem (small p) the approximations mentioned in section 6.4 are not valid and the algorithm does not perform well. For easy problems, instead, the α predicted by the procedure of section 6.4 seems to be approximately the optimal one. In the graph is showed how an estimated threshold for this two regimes can be given by the unsupervised detectability threshold, although this remains an heuristic deduction.

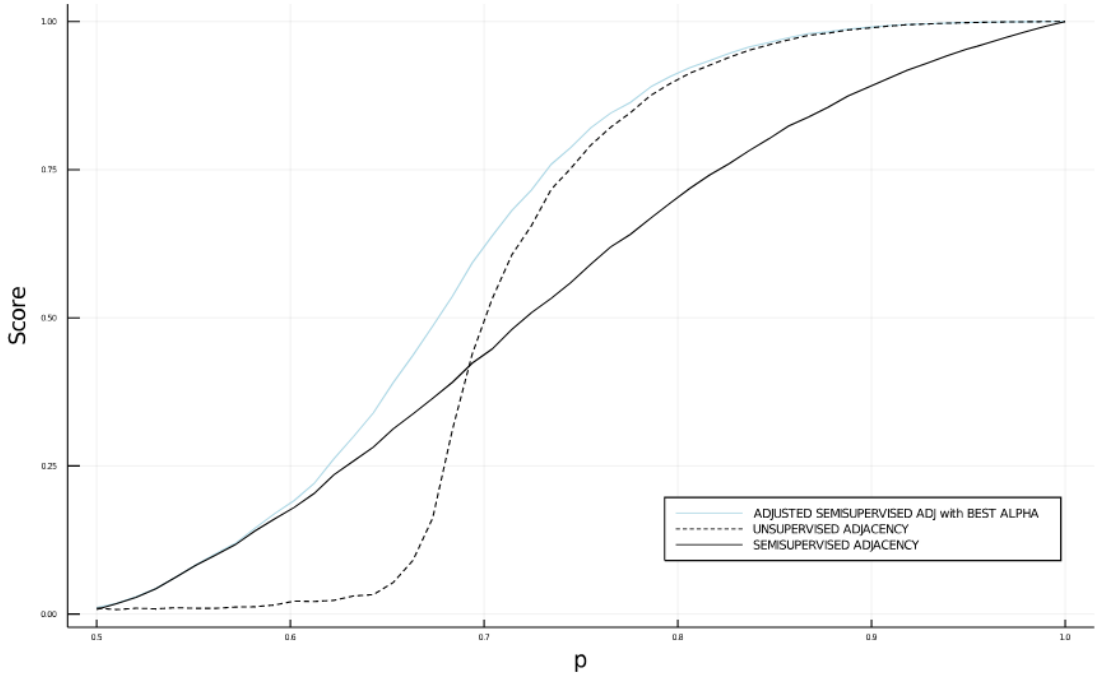

Figure 3: **Performances of modified adjacency matrix with best possible parameter α, compared with unsupervised and semisupervised inference with unmodified adjacency matrix (Algorithm 2).** The graph is the same as fig. 2, and so are the parameters. Here the only addition is the performance of the adjusted adjacency matrix algorithm, with parameter α artificially fixed in order to maximize the score. Remarkably, the performance of such algorithm stays always above the one of both unsupervised and unmodified semisupervised.

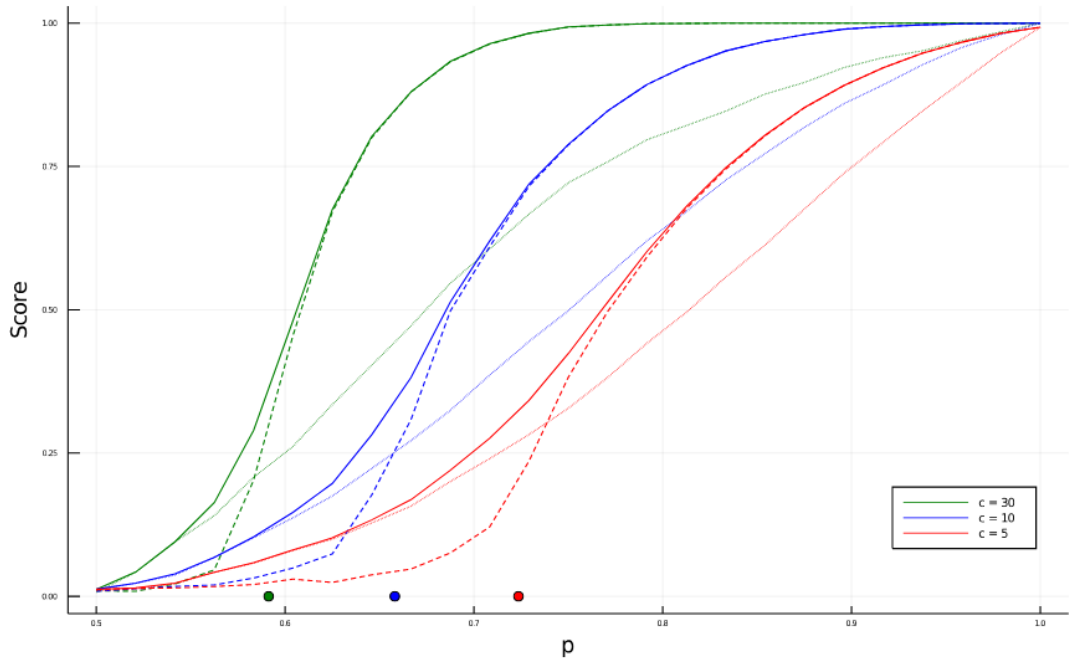

Figure 4: **Comparing the performance of the adjusted adjacency matrix with α set through the ground state with best possible alpha and with unmodified adjacency matrix. Three cases, correspondent to three different average connectivities c, are reported.** Different colors correspond to different average connectivities (levels of sparsity) c. For each color, the different lines represent the performances (the score as in (13)) of different algorithms as function of the easyness of the problem p. The *dense line* represents the adjusted adjacency matrix algorithm with α set in order to produce the best performance, the *light dotted line* represents the unmodified semisupervised adjacency matrix algorithm, i.e. Algorithm 2. The *dashed line* represents the adjusted adjacency matrix algorithm with α set as to maximize the ground state of the correspondent Hamiltonian, i.e. Algorithm 3. For each color (i.e. each c), the dots on the Score=0 axis represent the correspondent unsupervised detectability thresholds. They give a rough estimation on when the Algorithm 3 starts to have good performances. Other parameters: $n = n_u + n_l = 5000$, 5% labelled ($\nu = 0.05$), planted assignment of equally sized classes. Performances are averaged over 30 trials.

8 Conclusions

From the results, overall of sections 4 and 6, we can derive the following conclusions and perspectives:

- The semisupervised CBM, as long as $c \geq 1$ and $\nu > 0$, has no detectability threshold.
- The configuration of Finite Couplings (25) does not come exactly from the Bayes derivation, but it is an approximation of the correspondent Ising model, whose limits of validity are pretty unclear. However, their ground state are equivalent and thus the adjacency matrix gives still good results.
- The use of a centered similarity matrix avoids that the Min-Cut problem is solved by cutting the minimum number of edges, i.e. assigning most of the nodes to the same class. In other terms, this is the same problem noticed by Mai in [20] and solved indeed by proposing the centered adjacency matrix. Our generative model, the CBM, produces naturally a matrix with positive and negative weights, thus this problem does not hold and we developed indeed a very similar algorithm as [19].
- What we did is basically to perform a classification only with labelled-unlabelled connections (a sort of easy supervised classification), then we encoded this information on unlabelled under the form of a prior, for each unlabelled node. Such prior is represented in the graph as a positive or negative connection of the unlabelled with one added node, representing the labelled. In the algorithm used for the inference on such graph, the inference might be consistently unbalanced towards one of the two kinds of information, as we showed in the case of the adjacency matrix. The weight (or in other terms, the *reliability*) of the prior is controlled by the introduction of a further parameter α. How to choose this α is still an open problem, but it seems reasonable (and consistent with simulations, at least for easy problems) to choose the one whose correspondent prediction vector minimizes the hamiltonian of the the Finite Couplings configuration. This is a general structure that can be repeated in more realistic problems, when the couplings are not binary random variables but general similarity measures.
- Although already the adjacency matrix gives good performances, the natural extension of this work is to apply to the derived semisupervised configurations the recently proposed Bethe-Hessian (Non-Backtracking) algorithms. We can ask if Bethe-Hessian, that comes from a more fine approximation and that considers also temperature, may work better than the adjacency matrix for difficult problems also in this semisupervised setting. The answer seems to be positive.

References

[1] Mikhail Belkin, Irina Matveeva, and Partha Niyogi. "Regularization and semi-supervised learning on large graphs". In: *International Conference on Computational Learning Theory*. Springer. 2004, pp. 624–638.

[2] 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: *COLT*. 2008, pp. 33–44.

[3] Avrim Blum et al. "Semi-supervised learning using randomized mincuts". In: *Proceedings of the twenty-first international conference on Machine learning*. 2004, p. 13.

[4] Fabio Gagliardi Cozman, Ira Cohen, Marcelo Cesar Cirelo, et al. "Semi-supervised learning of mixture models". In: *ICML*. Vol. 4. 2003, p. 24.

[5] Lorenzo Dall'Amico, Romain Couillet, and Nicolas Tremblay. "A unified framework for spectral clustering in sparse graphs". In: *arXiv preprint arXiv:2003.09198* (2020).

[6] Lorenzo Dall'Amico, Romain Couillet, and Nicolas Tremblay. "Nishimori meets Bethe: a spectral method for node classification in sparse weighted graphs". In: *arXiv preprint arXiv:2103.03561* (2021).

[7] Lorenzo Dall'Amico, Romain Couillet, and Nicolas Tremblay. "Revisiting the bethe-hessian: improved community detection in sparse heterogeneous graphs". In: *arXiv preprint arXiv:1901.09715* (2019).

[8] Aurelien Decelle et al. "Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications". In: *Physical Review E* 84.6 (2011), p. 066106.

[9] Eric Eaton and Rachael Mansbach. "A spin-glass model for semi-supervised community detection". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 26. 1. 2012.

[10] David Elworthy. "Does Baum-Welch re-estimation help taggers?" In: *arXiv preprint cmp-lg/9410012* (1994).

[11] Santo Fortunato. "Community detection in graphs". In: *Physics reports* 486.3-5 (2010), pp. 75–174.

[12] Aram Galstyan, Greg Ver Steeg, and Armen E Allahverdyan. "Statistical Mechanics of Semi–Supervised Clustering in Sparse Graphs". In: *arXiv preprint arXiv:1101.4227* (2011).

[13] Gad Getz, Noam Shental, and Eytan Domany. "Semi-Supervised Learning–A Statistical Physics Approach". In: *arXiv preprint cs/0604011* (2006).

[14] Bryan R Gibson, Timothy T Rogers, and Xiaojin Zhu. "Human semi-supervised learning". In: *Topics in cognitive science* 5.1 (2013), pp. 132–172.

[15] Simon Heimlicher, Marc Lelarge, and Laurent Massoulié. "Community detection in the labelled stochastic block model". In: *arXiv preprint arXiv:1209.2910* (2012).

[16] Tony Jebara, Jun Wang, and Shih-Fu Chang. "Graph construction and b-matching for semi-supervised learning". In: *Proceedings of the 26th annual international conference on machine learning*. 2009, pp. 441–448.

[17] Rob Kitchin. *The data revolution: Big data, open data, data infrastructures and their consequences*. Sage, 2014.

[18] Florent Krzakala et al. "Spectral redemption in clustering sparse networks". In: *Proceedings of the National Academy of Sciences* 110.52 (2013), pp. 20935–20940.

[19] 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.

[20] Xiaoyi Mai and Romain Couillet. "Revisiting and improving semi-supervised learning: a large dimensional approach". In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2019, pp. 3547–3551.

[21] Russell Merris. "Laplacian matrices of graphs: a survey". In: *Linear algebra and its applications* 197 (1994), pp. 143–176.

[22] Cristopher Moore. "The computer science and physics of community detection: Landscapes, phase transitions, and hardness". In: *arXiv preprint arXiv:1702.00467* (2017).

[23] Alaa Saade. "Spectral inference methods on sparse graphs: theory and applications". In: *arXiv preprint arXiv:1610.04337* (2016).

[24] Alaa Saade, Florent Krzakala, and Lenka Zdeborová. "Spectral clustering of graphs with the bethe hessian". In: *arXiv preprint arXiv:1406.1880* (2014).

[25] Alaa Saade et al. "Fast randomized semi-supervised clustering". In: *Journal of Physics: Conference Series*. Vol. 1036. 1. IOP Publishing. 2018, p. 012015.

[26] Alaa Saade et al. "Spectral detection in the censored block model". In: *2015 IEEE International Symposium on Information Theory (ISIT)*. IEEE. 2015, pp. 1184–1188.

[27] Celso André R de Sousa, Solange O Rezende, and Gustavo EAPA Batista. "Influence of graph construction on semi-supervised learning". In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2013, pp. 160–175.

[28] Jesper E Van Engelen and Holger H Hoos. "A survey on semi-supervised learning". In: *Machine Learning* 109.2 (2020), pp. 373–440.

[29] L Viana and Allan J Bray. "Phase diagrams for dilute spin glasses". In: *Journal of Physics C: Solid State Physics* 18.15 (1985), p. 3037.

[30] Fei Wang et al. "Semi-supervised mean fields". In: *Artificial Intelligence and Statistics*. PMLR. 2007, pp. 596–603.

[31] Lenka Zdeborová and Florent Krzakala. "Statistical physics of inference: Thresholds and algorithms". In: *Advances in Physics* 65.5 (2016), pp. 453–552.

[32] Pan Zhang. "Robust spectral detection of global structures in the data by learning a regularization". In: *Advances in Neural Information Processing Systems* 29 (2016), pp. 541–549.

[33] Pan Zhang, Cristopher Moore, and Lenka Zdeborová. "Phase transitions in semisupervised clustering of sparse networks". In: *Physical Review E* 90.5 (2014), p. 052802.

[34] Lei Zhu et al. "A Brief Review of Spin-Glass Applications in Unsupervised and Semi-supervised Learning". In: *International Conference on Neural Information Processing*. Springer. 2016, pp. 579–586.

[35] Xiaojin Zhu and Andrew B Goldberg. "Introduction to semi-supervised learning". In: *Synthesis lectures on artificial intelligence and machine learning* 3.1 (2009), pp. 1–130.

A Where does the Censored Block Model come from?

In this section we give a context to the synthetic generative model used in the dissertation, the Censored Block Model. We show how it can be intended as a specific case of the Labelled Stochastic Block Model, which is in turn a generalization of the Stochastic Block Model.

A.1 The Stochastic Block Model

One of the most used generative model for community detection is the *Stochastic Block Model (SBM)* [5], which basically consists in an Erdős–Rényi random graph in which the probability of drawing an edge between two vertices is not uniform, but depends on latent variables carried by the vertices (i.e. their classes). Roughly speaking, for an assortative graph the probability $p_{\sigma,\sigma'}$ of drawing an edge between the vertices σ, σ' is higher if they belong to the same class ($\sigma = \sigma'$) than to different classes. This disomogeneity, that reflects the class structure, allows an algorithm to infer the latent variables of the nodes, i.e. estimate their true class and perform a classification task. Recently, it has been proved that detection for the SBM is feasible only in a certain regime [8]: setting $p_{\sigma,\sigma'} = \frac{c_{\sigma,\sigma'}}{n}$, where n is the number of vertices, for the case of equally sized communities we have that

$$c_{\sigma,\sigma'} = c_{in}1(\sigma = \sigma') + c_{out}1(\sigma \neq \sigma') \tag{65}$$

where c_{in} is the average connectivity for a node with vertices of the same class, c_{out} as the average connectivity with vertices of different classes. The average connectivity of the whole graph, c is then

$$c = \frac{c_{in} + (q-1)c_{out}}{q} \tag{66}$$

where q is the number of classes. The conjecture of impossibility of detection claims that a polynomial algorithm can detect the hidden communities on a SBM graph better than random guess, for $n \to \infty$ and equal sized binary classes if and only if

$$|c_{in} - c_{out}| > q\sqrt{c}$$

Standard spectral algorithms, such as Laplacian or adjacency matrices, are well adapted for dense graphs (large connectivities) but for a series of reasons, some mentioned in this text, their performance becomes poor or even null close to the detectability threshold, characterizing inference in sparse graphs. It is important to study this regime since real-life graphs are usually sparse [11]. Recent progresses in detecting communities down to the threshold have been made through the introduction of the non-backtracking matrix [18] and Bethe-Hessian [24], which ensure a detection as soon as it is theoretically feasible.

A.2 The Labelled Stochastic Block Model

The Labelled Stochastic Block Model (LSBM) is a generalization of the Stochastic Block Model in the sense that the information carried by the edges is not anymore capture only by their presence/absence, but also by the weight (label) attached to each edge. Note that the label is intended as weight, although in the text we call "label" the latent variables specifying the class assignment of a node, not to the edges. As for the presence/absence of an edge, the probability of having the label l on an edge reflects the latent class assigment and is given by

$$p_{\sigma,\sigma'}(l) = p_{in}(l)1(\sigma = \sigma') + p_{out}(l)1(\sigma \neq \sigma') \tag{67}$$

In [15] it is conjectured that in symmetric LSBM with equal classes size, for two classes $q = 2$, detection is possible if and only if

$$\frac{1}{2}\sum_{l} \frac{(c_{in}p_{in}(l) - c_{out}p_{out}(l))^2}{c_{in}p_{in}(l) + c_{out}p_{out}(l)} > 1 \tag{68}$$

Note that for $p_{in}(l) = p_{out}(l) = p_{in}(l')$ for whatever "labels" (weights) l, l', we recover the Stochastic Block Model.

A.3 The Censored Block Model as a specific case of LSBM

The Censored Block Model is a specific case of the LSBM [23] in which the connectivity does not reflect the latent assignment anymore, exclusively encoded by the labels (weights) attached to the edges. Specifically, $c_{in} = c_{out} = c$, the labels (weights) can be $l = \pm 1$ and $p_{in}(+1) = p_{out}(-1) = p$, for some $p \in [0.5, 1]$ and the other two probabilities as consequence of normalization $p_{out}(+1) = p_{in}(-1) = 1 - p$. Substituting these specific values in the equation of the threshold (68), we note that for sufficiently small p and c the inequality is not satisfied and detection becomes impossible, showing the presence of a phase transition.

B Theoretical background

B.1 Naive Mean Field derivation from Variational Free Energy

It is possible to show that for any distribution over the states of the system $q = \{q_\sigma\}$ the correspondent free energy $\mathcal{F}(q)$ will be always greater or equal that the free energy F of the Boltzmann distribution p. In other words

$$F = \min_q \mathcal{F}(q) \qquad p = \arg\min_q \mathcal{F}(q) \tag{69}$$

where $\mathcal{F}(q)$ is made of an energetic and an entropic term

$$\mathcal{F}(q) = \sum_\sigma q_\sigma E_\sigma + \frac{1}{\beta} \sum_\sigma q_\sigma \log q_\sigma \tag{70}$$

The aim is to use an arbitrary trial distribution, e.g. $p_{MF}(\sigma)$, and minimize the correspondent free energy $\mathcal{F}(p_{MF})$. Choosing

$$p_{MF}(\sigma) = p_1(\sigma_1)p_2(\sigma_2)...p_N(\sigma_N) \tag{71}$$

noticing that the average magnetization for each spin i

$$m_i = \sum_{\sigma_i} \sigma_i p_i(\sigma_i) = p_i(+1) - p_i(-1) \tag{72}$$

can be used to parametrize the single node marginal

$$p_i(\sigma_i) = \frac{1 + \sigma_i m_i}{2} \tag{73}$$

the mean field variational free energy of the Ising model with Hamiltonian

$$H = -\sum_i h_i \sigma_i - \sum_{i<j} J_{ij} \sigma_i \sigma_j \tag{74}$$

can be written only in terms of the local average magnetizations

$$\mathcal{F}(m) = -\sum_i h_i m_i - \sum_{i<j} J_{ij} m_i m_j + \frac{1}{\beta} \sum_i \sum_{\sigma_i} \frac{1 + \sigma_i m_i}{2} \log\big(\frac{1 + \sigma_i m_i}{2}\big) \tag{75}$$

So looking for stationary points of the above variational free energy, i.e. setting the condition

$$0 = \frac{\partial \mathcal{F}}{\partial m_i} \tag{76}$$

we get the Naive Mean Field equation (30) for each node i

$$m_i = \tanh\left(\beta\Big(\sum_{j \in \partial i} J_{ij} m_j + h_i\Big)\right) \tag{77}$$

The case $J_{ij} = J > 0$, $h_i = h = 0$ corresponds to the well-known Ising Ferromagnetic model that presents an average magnetization $m_i = m$, different from zero if $m = 0$ is an unstable extreme of the mean field free energy. The stability of $m = 0$ depends on the coupling constant $\{J_{ij}\}$ and on the inverse temperature β. For non-uniform couplings and fields, things are more complicated: the free energy profile may get more and more rugged and present multiple (suboptimal) minima with different entropies. As discussed before, the phase diagram gets more complicated (see fig. 1). Anyways, as we will see later, we can still use the (in)stability of the paramagnetic point to extract information, as long as it is an extreme, which is always the case in the absence of fields in the Hamiltonian.

B.2 Adjacency Matrix as Hessian of the Mean Field Free Energy at the paramagnetic point

Starting from the mean field free energy expressed in terms of the local average magnetizations (75) in the case of no fields (so when a spectral algorithm can be applied)

$$\mathcal{F}(m) = -\sum_{i<j} J_{ij} m_i m_j + \frac{1}{\beta} \sum_i \sum_{\sigma_i} \frac{1+\sigma_i m_i}{2} \log\big(\frac{1+\sigma_i m_i}{2}\big) \tag{78}$$

If we compute, as before, the gradient of the free energy we find that the paramagnetic point $m = 0$ is an extreme. Following [6], if we compute the Hessian of (78) at the paramagnetic point we get (after multiplying by β)

$$\mathcal{H}^{\mathcal{MF}}(m) = I_n - \beta J \tag{79}$$

where I_n is the identity matrix and J is the adjacency matrix of the graph. The eigenvector associated to the smallest eigenvalue of the Hessian gives the steepest direction for which the paramagnetic point is a maximum. Relying on the fact that this eigenvector points towards the direction of the minimum of the variational free energy, we can assume that this eigenvector is correlated to the true assignment of spins and thus is optimal for the inference. From the equation above we see that the eigenvector associated to the smallest eigenvalue of the Hessian of the variational free energy is equal to the one associated to the largest eigenvalue of the adjacency matrix. Thus one can implement an algorithm that simply calculates the largest isolated eigenvalue of the adjacency matrix J and infer the class of each node by taking the sign of the entries of the associated eigenvector. The problem, for the unsupervised case, is that the spectrum of the Adjacency matrix J often does not have an isolated largest eigenvalue as soon as it is theoretically possible to perform the inference, due a series of reasons. The most important one is that for sparse graphs the eigenvector associated to the largest eigenvalue of J is much influenced the node with highest connection degree and poorly correlated to the structure of the graph. This is why alternative methods, such as Bethe-Hessian (Non-Backtracking), have been proposed by the community for sparse graphs and proved to work down to the detectability threshold [18, 24]. It is useful to remark that in this picture, the inverse temperature β does not play any role. In other terms, the problem is not set in a Bayes optimal setting. Other minimizations of variational free energy, such as Bethe-Hessian, perform the inference taking into account β and thus potentially in a Bayes optimal setting.

B.3 Adjacency Matrix as solver of ground state search with continous spins

In this section we sketch another idea to see the Adjacency Matrix, alternative and orthogonal to the one proposed in the context of minimizing variational mean field free energy. As said, when dealing with an Ising model we have to minimize the free energy, made of an energy and an entropic term. The entropic term is negligible only at $T = 0$ ($\beta \to \infty$), where the only possible state is the ground state. Thus, by ignoring the entropic term we set in a specific regime $\beta \to \infty$ which is far to be Bayes optimal. Anyways, since the generative model is usually unknown and estimating the Nishimori temperature is hard, one can think that performing inference by neglecting completely the temperature and just estimating the ground state can still make sense. Notice that we are kind of overfitting, as we are giving too much importance to the observed coupling realization (strongly influenced by the noise) and nothing to the assumptions on the generative model. Indeed [22] explains that finding the ground state is equivalent to take the Maximum A Posteriori estimator (MAP) while the more correct way to infer the planted assignment is to derive the marginals of each vertex and assign it to the most likely class (for two classes it corresponds to evaluating the sign of the average magnetization m, as we have done so far). Finding the ground state corresponds to minimize the Hamiltonian, i.e. maximize

$$-H(\sigma) = \sum_{i,j \in \vec{E}} J_{ij}\sigma_i\sigma_j + \sum_i h_i\sigma_i \tag{80}$$

that when fields are not present consists in

$$\max_{\sigma=\{\pm 1\}} \sum_{i,j \in \vec{E}} J_{ij}\sigma_i\sigma_j \tag{81}$$

Now if we relax the constraint of spins to be boolean variables and we call them $f_i \in \mathrm{R}$, we have the associated continous maximization problem

$$\max_{f_i \in \mathrm{R}} \sum_{i,j \in \vec{E}} J_{ij} f_i f_j \tag{82}$$

which can be rewritten calling $f \in \mathrm{R}^n$ the vector of continous spins and J the Adjacency matrix as

$$\max_{f \in \mathrm{R}^n} f^T J f \tag{83}$$

unless an irrelevant $1/2$ factor. It is easy to see that the problem is solved by setting f equal to the eigenvector associated to the greatest eigenvalue of the adjacency matrix J. Notice that setting the constraint of f to be of unitary norm $f^T f = 1$ just makes the problem convex and does not affect the inference once we take

$$\hat{\sigma} = sign(f) \tag{84}$$

B.4 Adjacency Matrix and the high T expansion of the Naive Mean Field Equation

We propose another relation between the Adjacency Matrix and the Naive Mean Field method. In the section before the adjacency matrix came out in searching the ground state, i.e. at $T = 0$ so neglecting completely the entropic term. Now we set in the high temperature regime, $\beta \to 0$, when the only contribution for the typical configuration comes from the entropy. The Naive Mean Field Equation without fields reads

$$m = \tanh(\beta J m) \tag{85}$$

becomes for $\beta \to 0$, from the Taylor expansion of $\tanh x$,

$$m \approx \beta J m \tag{86}$$

equal to

$$\frac{1}{\beta} m \approx J m \tag{87}$$

Now recall the eigenvalue equation for the Adjacency Matrix

$$J m = \lambda m \tag{88}$$

(87) and (88) are equivalent for

$$\beta = \frac{1}{\lambda_{max}} \tag{89}$$

Moreover, looking at the linearized mean field equation (87), for $\beta\lambda_{max} < 1$ there is no other fixed point than $m = 0$, that corresponds to the paramagnetic phase, while for $\beta\lambda_{max} > 1$ a different fixed point exists, which corresponds to the ferromagnetic phase. The temperature that satisfies $\beta\lambda_{max} = 1$, i.e.

$$\beta_{PF} = \frac{1}{\lambda_{max}} \tag{90}$$

is thus the temperature of the mean field ferromagnetic-paramagnetic transition.

C Order of the largest eigenvalue of J and $\tilde{J}$

Here we show an approximate calculus to obtain the dependency on n_u of the largest eigenvalue λ_{max} on the symmetric adjacency matrix $J \in R^{n_u \times n_u}$, made by the sparse block J_U and the dense vector h, $h_i \sim O(1)$

$$J = \begin{bmatrix} J_U & h \\ h^T & 0 \end{bmatrix} \tag{91}$$

as well as the largest eigenvalue $\tilde{\lambda}_{max}$ of the adjusted adjacency matrix $\tilde{J}$, whose expression is the same as J but with $h_i \sim O(\frac{1}{\sqrt{n_u}})$, since the fields are rescaled.
The eigenvalue equation

$$\begin{bmatrix} J_U & h \\ h^T & 0 \end{bmatrix} \begin{bmatrix} \hat{f}_u \\ \hat{f}_+ \end{bmatrix} = \lambda_{max} \begin{bmatrix} \hat{f}_u \\ \hat{f}_+ \end{bmatrix} \tag{92}$$

rewritten

$$\begin{cases} J_U \hat{f}_u + h\hat{f}_+ = \lambda_{max}\hat{f}_u \\ h^T \hat{f}_u = \lambda_{max}\hat{f}_+ \end{cases} . \tag{93}$$

$$\begin{cases} (-J_U + \lambda_{max})\hat{f}_u = h\hat{f}_+ \\ \hat{f}_+ = \frac{h^T \hat{f}_u}{\lambda_{max}} \end{cases} . \tag{94}$$

$$(-J_U + \lambda_{max})\hat{f}_u = h\frac{h^T \hat{f}_u}{\lambda_{max}} \tag{95}$$

so for each $\hat{f}_i \in \hat{f}_u$, $i \in U$

$$\lambda_{max}(-(J_U \hat{f}_u)_i + \lambda_{max}\hat{f}_i) = h_i(h^T \hat{f}_u) \tag{96}$$

taking the order of terms, considering that $-(J_U \hat{f}_u)_i \sim O(c)O(\hat{f}_i) \sim O(\hat{f}_i)$ since J_U is sparse, while $h^T \hat{f}_u \sim O(n_u)O(h_i^2)O(\hat{f}_i)$ since h is dense, then $O(\hat{f}_i)$ simplifies and

$$O(\lambda_{max})\big[O(1) + O(\lambda_{max})\big] = O(h_i^2)O(n_u) \tag{97}$$

Thus, for J, $O(h_i) \sim O(1)$

$$\lambda_{max} \sim \sqrt{n_u} \tag{98}$$

while for $\tilde{J}$, $O(h_i) \sim O(\frac{1}{\sqrt{n_u}})$

$$\tilde{\lambda}_{max} \sim O(1) \tag{99}$$