# Inference and Mutual Information on Random Factor Graphs

**Amin Coja-Oghlan** ✉
Mathematics Institute, Goethe Universität Frankfurt am Main, Germany

**Max Hahn-Klimroth** ✉
Mathematics Institute, Goethe Universität Frankfurt am Main, Germany

**Philipp Loick** ✉
Mathematics Institute, Goethe Universität Frankfurt am Main, Germany

**Noela Müller** ✉
Mathematics Institute, University of Munich, Germany

**Konstantinos Panagiotou** ✉
Mathematics Institute, University of Munich, Germany

**Matija Pasch** ✉
Mathematics Institute, University of Munich, Germany

## Abstract

Random factor graphs provide a powerful framework for the study of inference problems such as decoding problems or the stochastic block model. Information-theoretically the key quantity of interest is the mutual information between the observed factor graph and the underlying ground truth around which the factor graph was created; in the stochastic block model, this would be the planted partition. The mutual information gauges whether and how well the ground truth can be inferred from the observable data. For a very general model of random factor graphs we verify a formula for the mutual information predicted by physics techniques. As an application we prove a conjecture about low-density generator matrix codes from [Montanari: IEEE Transactions on Information Theory 2005]. Further applications include phase transitions of the stochastic block model and the mixed $\boldsymbol{k}$-spin model from physics.

**2012 ACM Subject Classification** Mathematics of computing → Probabilistic inference problems

**Keywords and phrases** Information theory, random factor graphs, inference problems, phase transitions

**Digital Object Identifier** 10.4230/LIPIcs.STACS.2021.24

**Related Version** *Full Version*: `https://arxiv.org/abs/2007.07494`

**Funding** *Amin Coja-Oghlan*: Supported by DFG CO 646/3.
*Max Hahn-Klimroth*: Supported by Stiftung Polytechnische Gesellschaft and DFG FOR 2975.
*Philipp Loick*: Supported by DFG CO 646/3.
*Noela Müller*: Supported by the European Research Council, ERC Grant Agreement 772606–PTRCSP.
*Konstantinos Panagiotou*: Supported by the European Research Council, ERC Grant Agreement 772606–PTRCSP.
*Matija Pasch*: Supported by the European Research Council, ERC Grant Agreement 772606–PTRCSP.

38th International Symposium on Theoretical Aspects of Computer Science (STACS 2021).
Editors: Markus Bläser and Benjamin Monmege; Article No. 24; pp. 24:1–24:15
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

# 1 Introduction

## 1.1 Background and motivation

Since the 1990s there has been an immense interest in inference and learning problems on random graphs. One motivation has been to seize upon random graphs as benchmarks for inference algorithms of all creeds and denominations. An excellent example of this is the stochastic block model; the impressive literature on this model alone is surveyed in [1]. A second, no less salient motivation has been the use of random graphs in probabilistic constructions. Concrete examples include powerful error correcting codes such as low density generator matrix or low density parity check codes, which have since found their way into modern communications standards [20, 31]. Further prominent recent applications include compressed sensing and group testing [2, 14, 15]. It appears hardly a stretch to claim that in terms of real world impact these constructions occupy top ranks among applications of the probabilistic method and, indeed, modern combinatorics generally.

Yet many applications of the probabilistic method to inference problems still lack a satisfactory rigorous justification. Some are supported primarily by empirical evidence, i.e., not much more than a bunch of computer experiments. Quite a few others have been inspired by a versatile but non-rigorous approach from physics known as the "cavity method". But while there has been progress in recent years, vast gaps between the physics predictions and their rigorous vindications remain. One important reason for this is that the random graph models used in practical inference tend to be significantly more intricate than, say, a classical binomial random graph. For instance, a highly popular breed of low-density parity check codes use delicately tailored degree distributions for both the variable nodes and the check nodes of the Tanner graph [31].

In this paper we significantly advance the rigorous state of the art by corroborating important cavity method predictions wholesale for a rich class of inference problems that accommodates the very general choices of degree distributions of interest in high-dimensional Bayesian inference problems and coding theory. Generally, the objective in such inference problems is to recover the ground truth from the observable data. Think, for instance, of retrieving the hidden communities in the stochastic block model or of reconstructing the original message from a noisy codeword. For this broad class of models we rigorously establish the formulas that the cavity method predicts for the mutual information, which is the key information-theoretic potential that gauges precisely how much it is possible in principle to learn about the ground truth. Technically we build upon and extend the methods developed in [11] for random graph models of Erdős-Rényi type. While we follow a similar general proof strategy, the greater generality of the present results necessitates significant upgrades to virtually all of the moving parts. For example, due to the more rigid combinatorial structure of graphs with given degrees many of the manoeuvres that are straightforward for binomial random graphs now require delicate coupling arguments.

We proceed to highlight applications of our main results to three specific problems that have each received a great deal of attention in their own right: low-density generator matrix codes, the stochastic block model and the mixed $\boldsymbol{k}$-spin model, which hails from mathematical physics. Then in Section 2 we state the main results concerning the general class of random factor graph models. Section 3 contains an overview of the proof strategy and a detailed comparison with prior work.

## 1.2 Low-density generator matrix codes

A powerful and instructive class of error-correcting codes, low-density generator matrix ("ldgm") codes are based on random bipartite graphs with given degree distributions. Specifically, let $\boldsymbol{d}, \boldsymbol{k} \geq 0$ be bounded integer-valued random variables, let $n$ be an integer and let $\boldsymbol{m} \sim \mathrm{Po}(n\mathbb{E}[\boldsymbol{d}]/\mathbb{E}[\boldsymbol{k}])$ be a Poisson variable. One vertex class $V = \{x_1, \ldots, x_n\}$ of the graph represents the bits of the original message. The other class $F = \{a_1, \ldots, a_{\boldsymbol{m}}\}$ represents the rows of the code's generator matrix. To obtain the random graph $\boldsymbol{G}$ create for each variable node $x_i$ an independent copy $\boldsymbol{d}_i$ of $\boldsymbol{d}$. Similarly, create an independent copy $\boldsymbol{k}_i$ of $\boldsymbol{k}$ for each check node $a_i$. Then given the event $\left\{\sum_{i=1}^{n} \boldsymbol{d}_i = \sum_{i=1}^{\boldsymbol{m}} \boldsymbol{k}_i\right\}$ that the total degrees on both sides match let $\boldsymbol{G}$ be a random bipartite graph where every $x_i$ has degree $\boldsymbol{d}_i$ and every $a_i$ has degree $\boldsymbol{k}_i$. We tacitly restrict to $n$ such that this event has positive probability.

The generator matrix of the ldgm code is now precisely the $\boldsymbol{m} \times n$ biadjacency matrix $A(\boldsymbol{G})$ of $\boldsymbol{G}$, viewed as a matrix over $\mathbb{F}_2$. Thus, the rows of $A(\boldsymbol{G})$ correspond to the check nodes $a_1, \ldots, a_{\boldsymbol{m}}$, the columns correspond to $x_1, \ldots, x_n$ and the $(i,j)$-entry equals one iff $a_i$ and $x_j$ are adjacent. For a given message $\boldsymbol{x} \in \mathbb{F}_2^n$ the corresponding codeword reads $\boldsymbol{y} = A(\boldsymbol{G})\boldsymbol{x} \in \mathbb{F}_2^{\boldsymbol{m}}$. The receiver on the other end of a noisy channel observes a scrambled version $\boldsymbol{y}^*$ of $\boldsymbol{y}$. Specifically, $\boldsymbol{y}^*$ is obtained from $\boldsymbol{y}$ by flipping every bit with probability $\eta \in (0, 1/2)$ independently. To gauge the potential of the code, the key question is how much information about the original $\boldsymbol{x}$ the receiver can possibly extract from $\boldsymbol{y}^*$. Naturally, the receiver also knows $\boldsymbol{G}$. Hence, we aim to work out the conditional mutual information

$$I(\boldsymbol{x}, \boldsymbol{y}^* \mid \boldsymbol{G}) = \sum_{x \in \mathbb{F}_2^n, y \in \mathbb{F}_2^{\boldsymbol{m}}} \mathbb{P}\left[\boldsymbol{x} = x, \boldsymbol{y}^* = y \mid \boldsymbol{G}\right] \log \frac{\mathbb{P}\left[\boldsymbol{x} = x, \boldsymbol{y}^* = y \mid \boldsymbol{G}\right]}{2^n \mathbb{P}\left[\boldsymbol{y}^* = y \mid \boldsymbol{G}\right]}.$$

A precise prediction as to its asymptotical value was put forward on the basis of the physicists' cavity method. As most such predictions, the formula comes as a variational problem that asks to optimise a functional called the Bethe free entropy over a space of probability measures. Specifically, let $\mathfrak{P}_*([-1,1])$ be the space of all probability measures $\rho$ on the interval $[-1,1]$ with mean zero. Let $(\boldsymbol{\theta}_{i,j,\rho})_{i,j \geq 1} \subseteq [-1,1]$ be a family of samples from $\rho$. Further, let $(\boldsymbol{J}_i)_{i \geq 1}$ be Rademacher variables, i.e., $\mathbb{P}\left[\boldsymbol{J}_i = 1\right] = \mathbb{P}\left[\boldsymbol{J}_i = -1\right] = 1/2$. In addition, let $(\hat{\boldsymbol{k}}_i)_{i \geq 1}$ be random variables with distribution

$$\mathbb{P}\left[\hat{\boldsymbol{k}}_i = \ell\right] = \frac{\ell \mathbb{P}\left[\boldsymbol{k} = \ell\right]}{\mathbb{E}[\boldsymbol{k}]} \qquad (\ell \geq 0). \tag{1}$$

All of these are independent. Finally, let $\Lambda(z) = z \log(z)$. Then the Bethe free entropy reads

$$\begin{aligned} \mathcal{B}_{\mathrm{ldgm}}(\rho, \eta) = \mathbb{E}\Bigg[\frac{1}{2}\Lambda\Bigg(\sum_{\sigma \in \{0,1\}} \prod_{i=1}^{\boldsymbol{d}} 1 + (-1)^{\sigma} \boldsymbol{J}_i (1 - 2\eta) \prod_{j=1}^{\hat{\boldsymbol{k}}_i - 1} \boldsymbol{\theta}_{i,j,\rho}\Bigg)\Bigg] \\ - \frac{\mathbb{E}[\boldsymbol{d}]}{\mathbb{E}[\boldsymbol{k}]} \mathbb{E}\left[(\boldsymbol{k} - 1)\Lambda\left(1 + \boldsymbol{J}_1 (1 - 2\eta) \prod_{j=1}^{\boldsymbol{k}} \boldsymbol{\theta}_{1,j,\rho}\right)\right]. \end{aligned}$$

▶ **Theorem 1.** *For any $\boldsymbol{d}, \boldsymbol{k}$ and for all $\eta \in (0,1)$ we have*

$$\begin{aligned} \lim_{n \to \infty} \frac{1}{n} I(\boldsymbol{x}, \boldsymbol{y}^* \mid \boldsymbol{G}) = \left(1 + \frac{\mathbb{E}[\boldsymbol{d}]}{\mathbb{E}[\boldsymbol{k}]}\right) \log(2) + \frac{\mathbb{E}[\boldsymbol{d}]}{\mathbb{E}[\boldsymbol{k}]} (\eta \log(\eta) + (1 - \eta) \log(1 - \eta)) \\ - \sup_{\rho \in \mathfrak{P}_*[-1,1]} \mathcal{B}_{\mathrm{ldgm}}(\rho, \eta) \qquad \textit{in probability.} \end{aligned}$$

Theorem 1 completely solves a well known conjecture [25, Conjecture 1] and significantly extends the results from [32, 11], which required the restrictive assumption that the check degree $\boldsymbol{k}$ be constant.

A possible objection to a result such as Theorem 1 might be that the resulting formula appears exceedingly complicated as it leaves us with a potentially difficult variational problem. Yet two points are to be made in defense. First, by vindicating the precise formula predicted by the cavity method, the theorem and its proof show that this technique and the ideas behind it do indeed get to the bottom of the problem. Second, since the formula involves a supremum, any $\rho \in \mathfrak{P}_*[-1,1]$ yields an upper bound on the mutual information. Hence, the heuristic population dynamics algorithm deemed to produce good candidate maximisers and beloved of physicists, can be harnessed to get rigorous bounds in one direction. Finally, in some cases it is possible to precisely identify the maximiser analytically [6, 9].

## 1.3 The stochastic block model

An instructive model of graph clustering, the stochastic block model presumes that a random graph is created in two steps. First each of the $n$ vertices $\{x_1, \ldots, x_n\}$ receives one of $q \geq 2$ possible colours $\boldsymbol{\sigma}^*_{x_i} \in [q]$ uniformly and independently. Then a sparse random graph is created where vertices with the same colour are either more likely to be connected by an edge (assortative case), or less likely (disassortative). Different versions of this model have been proposed. While in the simplest one edges are inserted independently, here we consider a model from [27] that produces a $d$-regular graph. Hence, let $d \geq 3$ be an integer and let $\boldsymbol{G} = \boldsymbol{G}(n, d)$ be a random $d$-regular graph. Further, given a parameter $\beta > 0$ let $\boldsymbol{G}^* = \boldsymbol{G}^*(n, d, \boldsymbol{\sigma}^*)$ be a random graph drawn from the distribution

$$\mathbb{P}\left[\boldsymbol{G}^* = G \mid \boldsymbol{\sigma}^*\right] \propto \exp\left[-\beta \sum_{vw \in E(G)} \mathbf{1}\left\{\boldsymbol{\sigma}^*_v = \boldsymbol{\sigma}^*_w\right\}\right], \tag{2}$$

with the $\propto$-symbol hiding the normalisation required to obtain a probability distribution. Thus, the parameter $\beta$ tunes the penalty that we impose on monochromatic edges by comparison to the null model $\boldsymbol{G}$. At $\beta = 0$ there is no such penalty and $\boldsymbol{G}^*$ and $\boldsymbol{G}$ are identical. But even for positive $\beta$ the random graphs $\boldsymbol{G}, \boldsymbol{G}^*$ may still be indistinguishable and in effect recovering $\boldsymbol{\sigma}^*$ may be impossible. Hence, a fundamental question is for what $q, d, \beta$ it is possible to discriminate between $\boldsymbol{G}, \boldsymbol{G}^*$. Formally, we recall that the *Kullback-Leibler divergence* of $\boldsymbol{G}^*, \boldsymbol{G}$ is defined as $D_{\mathrm{KL}}\left(\boldsymbol{G}^* \| \boldsymbol{G}\right) = \sum_G \mathbb{P}\left[\boldsymbol{G}^* = G\right] \log \frac{\mathbb{P}[\boldsymbol{G}^* = G]}{\mathbb{P}[\boldsymbol{G} = G]}$. The Kullback-Leibler divergence is an information-theoretic potential that gauges the similarity of two random graph models. In particular, if $D_{\mathrm{KL}}\left(\boldsymbol{G}^* \| \boldsymbol{G}\right) = \Omega(n)$, then $\boldsymbol{G}, \boldsymbol{G}^*$ can be told apart because natural observables will take vastly different values on the two models. Whether $D_{\mathrm{KL}}\left(\boldsymbol{G}^* \| \boldsymbol{G}\right) = \Omega(n)$ depends on the value of the Bethe free entropy for the stochastic block model. To be precise, let $\mathcal{P}([q])$ be the set of all probability distributions $(\mu(1), \ldots, \mu(q))$ on $[q]$. We identify $\mathcal{P}([q])$ with the standard simplex in $\mathbb{R}^q$. Further, let $\mathfrak{P}_*([q])$ be the set of all probability measures $\pi$ on $\mathcal{P}([q])$ such that $\int \mu(\sigma) \mathrm{d}\pi(\mu) = 1/q$ for every $\sigma \in [q]$. In other words, the mean of $\pi$ is the barycenter of the simplex. Let $(\boldsymbol{\mu}_{i,\pi})_{i \geq 1}$ be a family of independent samples from $\pi$ and let

$$\begin{aligned}\mathcal{B}_{\mathtt{sbm}}(\pi, \beta) = \mathbb{E}&\left[\frac{\Lambda\left(\sum_{\sigma=1}^{q} \prod_{i=1}^{d} 1 - (1 - \mathrm{e}^{-\beta})\boldsymbol{\mu}_{i,\pi}(\sigma)\right)}{q\left(1 - (1 - \mathrm{e}^{-\beta})/q\right)^d}\right] \\ &- \mathbb{E}\left[\frac{d\Lambda\left(1 - (1 - \mathrm{e}^{-\beta}) \sum_{\sigma=1}^{q} \boldsymbol{\mu}_{1,\pi}(\sigma)\boldsymbol{\mu}_{2,\pi}(\sigma)\right)}{2\left(1 - (1 - \mathrm{e}^{-\beta})/q\right)}\right].\end{aligned}$$

▶ **Theorem 2.** *Let*

$$\beta^* = \inf\left\{\beta > 0 : \sup_{\pi \in \mathfrak{P}_*([q])} \mathcal{B}_{\mathtt{sbm}}(\pi, \beta) > \log(q) + \frac{d}{2}\log\left(1 - (1 - \mathrm{e}^{-\beta})/q\right)\right\}.$$

**(i)** *If* $\beta < \beta^*$*, then* $\lim_{n\to\infty} \frac{1}{n} D_{\mathrm{KL}}\left(\boldsymbol{G}^* \| \boldsymbol{G}\right) = 0$.

**(ii)** *If* $\beta > \beta^*$*, then* $\lim_{n\to\infty} \frac{1}{n} D_{\mathrm{KL}}\left(\boldsymbol{G}^* \| \boldsymbol{G}\right) > 0$.

Theorem 2 easily implies that for $\beta > \beta^*$ it is information-theoretically possible to recover a non-trivial approximation to $\boldsymbol{\sigma}^*$ from $\boldsymbol{G}^*$. In other words, there exists an exponential time algorithm that likely outputs a colouring $\tau$ of the vertices that has a significantly greater overlap with the ground truth $\boldsymbol{\sigma}^*$ than a random guess. An open question is whether for $\beta > \beta^*$ this problem can even be solved by a polynomial time algorithm. The going conjecture is that in general the answer is "no" and that efficient recoverability kicks in only at a second threshold $\beta^{**} > \beta^*$ for many interesting choices of $q, d$ [12].

## 1.4 The mixed k-spin model

Not only do the main results of this paper facilitate rigorous proofs of physics predictions for problems in computer science, but also, conversely, do we obtain new theorems on problems of keen interest in statistical physics. For example, the mixed $\boldsymbol{k}$-spin model is an important spin glass model [28]; its purpose is to describe the magnetic interactions in metallic alloys. To define the model let $\boldsymbol{k} \geq 2$ be an integer-valued random variable such that $\mathbb{E}[\boldsymbol{k}^{2+\varepsilon}] < \infty$ for some $\varepsilon > 0$ and $\mathbb{P}[\boldsymbol{k} = 2] > 0$. Let $(\boldsymbol{k}_i)_{i\geq 1}$ be a sequence of independent copies of $\boldsymbol{k}$. Moreover, let $d > 0$ and let $\boldsymbol{H} = \boldsymbol{H}_{\boldsymbol{k}}(n, \boldsymbol{m})$ be a (non-uniform) random hypergraph on $V_n = \{x_1, \ldots, x_n\}$ with $\boldsymbol{m} = \mathrm{Po}(dn/\mathbb{E}[\boldsymbol{k}])$ independent hyperedges $a_1, \ldots, a_{\boldsymbol{m}}$ such that $a_i$ comprises $\boldsymbol{k}_i$ vertices, drawn uniformly without replacement. Thus, in the special case that $\boldsymbol{k}$ is constant we obtain the classical binomial random hypergraph. To turn this random hypergraph into a spin glass model we draw for each of its edges $a_i$ an independent standard Gaussian $\boldsymbol{J}_i$. Additionally, let $\beta > 0$ be a parameter, commonly coined the inverse temperature. Then the *Boltzmann distribution* of the model is the probability distribution on $\{\pm 1\}^{V_n}$ defined by

$$\mu_{\boldsymbol{H},\boldsymbol{J},\beta}(\sigma) = \frac{\exp\left(\beta \sum_{i=1}^{\boldsymbol{m}} \boldsymbol{J}_i \prod_{x \in a_i} \sigma_x\right)}{Z(\boldsymbol{H}, \boldsymbol{J}, \beta)} \qquad (\sigma \in \{\pm 1\}^{V_n}),$$

where $Z(\boldsymbol{H}, \boldsymbol{J}, \beta) = \sum_{\tau \in \{\pm 1\}^{V_n}} \exp\left(\beta \sum_{i=1}^{\boldsymbol{m}} \boldsymbol{J}_i \prod_{x \in a_i} \tau_x\right)$. The normalising term $Z(\boldsymbol{H}, \boldsymbol{J}, \beta)$ is known as the *partition function.*

A key question is whether for given $d, \beta, \boldsymbol{k}$ there occur long-range correlations between the magnetic "spins" observed at $x_1, \ldots, x_n$. Formally, let $\boldsymbol{\sigma} \in \{\pm 1\}^{V_n}$ signify a sample from the Boltzmann distribution. Then we say that *long-range correlations are absent* if

$$\lim_{n\to\infty} \frac{1}{n^2} \sum_{x,y \in V_n} \mathbb{E}\left|\mu_{\boldsymbol{H},\boldsymbol{J},\beta}(\{\boldsymbol{\sigma}_x = \boldsymbol{\sigma}_y = 1\}) - \mu_{\boldsymbol{H},\boldsymbol{J},\beta}(\{\boldsymbol{\sigma}_x = 1\})\mu_{\boldsymbol{H},\boldsymbol{J},\beta}(\{\boldsymbol{\sigma}_y = 1\})\right| = 0.$$

In words, the equation expresses that for most pairs $x, y$ of vertices the spins $\boldsymbol{\sigma}_x, \boldsymbol{\sigma}_y$ are essentially independent. If this is violated, we say that long-range correlations are present.

According to physics predictions for a given $\beta > 0$ long-range correlations emerge at a critical value $d_{\beta,\boldsymbol{k}}$ that can be determined in terms of the Bethe free entropy [19, 24]. The methods developed in this paper enable us to corroborate this formula rigorously. Specifically, let $\mathfrak{P}_*([-1,1])$ be the space of all probability measures on $[-1,1]$ with mean zero. Given $\pi \in \mathfrak{P}_*([-1,1])$ let $(\boldsymbol{\mu}_{i,j,\pi})_{i,j\geq 1}$ be a family of independent samples from $\pi$. Additionally, let $(\hat{\boldsymbol{k}}_i)_{i\geq 1}$ be a family of independent random variables with point masses (1) and let $\boldsymbol{d} = \mathrm{Po}(d)$. Then the Bethe free entropy $\mathcal{B}_{\boldsymbol{k}-\mathrm{spin}}(\pi)$ of the $\boldsymbol{k}$-spin model is given by the expression

$$\frac{1}{2}\mathbb{E}\left[\Lambda\left(\sum_{\sigma_1\in\{\pm1\}}\prod_{i=1}^{\boldsymbol{d}}\left(1+\sum_{\sigma_2,\ldots,\sigma_{\hat{\boldsymbol{k}}_i}\in\{\pm1\}}\tanh\left(\beta\boldsymbol{J}_j\prod_{j\in[\hat{\boldsymbol{k}}_i]}\sigma_j\right)\prod_{j=2}^{\hat{\boldsymbol{k}}_i}\frac{1+\sigma_j\boldsymbol{\mu}_{i,j,\pi}}{2}\right)\right)\right]$$
$$-\frac{d}{\mathbb{E}[\boldsymbol{k}]}\mathbb{E}\left[(\boldsymbol{k}-1)\Lambda\left(1+\sum_{\sigma\in\{\pm1\}^{\boldsymbol{k}}}\tanh\left(\beta\boldsymbol{J}_1\prod_{i=1}^{\boldsymbol{k}}\sigma_j\right)\prod_{i=1}^{\boldsymbol{k}}\frac{1+\sigma_i\boldsymbol{\mu}_{1,i,\pi}}{2}\right)\right].$$

▶ **Theorem 3.** *Let* $d_{\beta,\boldsymbol{k}} = \inf\left\{d > 0 : \sup_{\pi\in\mathfrak{P}_*([-1,1])}\mathcal{B}_{\boldsymbol{k}-\mathrm{spin}}(\pi) > \log 2\right\}$.

**(i)** *Long-range correlations are absent for* $d < d_{\beta,\boldsymbol{k}}$.

**(ii)** *For any* $\varepsilon > 0$ *there exists* $d_{\beta,\boldsymbol{k}} < d < d_{\beta,\boldsymbol{k}} + \varepsilon$ *where long-range correlations are present.*

Thus, the point $d_{\beta,\boldsymbol{k}}$, characterised by the Bethe variational principle, marks the onset of complex magnetic interactions in the mixed $\boldsymbol{k}$-spin model. This critical value is known as the *replica symmetry breaking* phase transition in physics jargon. As a further application of the main results we can pinpoint the so-called condensation phase transition of the Potts antiferromagnet on random $d$-regular graphs, another problem of interest in mathematical physics. More details can be found in Section 16 of the full version.

## 2 The mutual information of random factor graphs

The theorems quoted in Section 1 are easy consequences of results on general random factor graph models. These more general theorems, one of which we present next, constitute the main results of the paper.

### 2.1 Random factor graph models

Remarkably many classical problems from combinatorics, statistics and physics can be expressed conveniently in the language of factor graph models [24, 29, 34]. A factor graph $G$ is a bipartite graph whose vertex classes are variable nodes $V(G)$ and factor nodes $F(G)$. The former represent the variables of the combinatorial problem in question, such as the individual bits of a codeword. Generally we assume that these variables range over a domain $\Omega$ of size $q = |\Omega| \geq 2$. Moreover, the factor nodes encode the interactions between the variables, such as the linear relations imposed by the check matrix of a code. Each factor node $a \in F(G)$ comes with a function $\psi_a : \Omega^{\partial a} \to (0,\infty)$ that assigns a positive weight to value combinations of the adjacent variables $\partial a$. The factor graph gives rise to a probability distribution

$$\mu_G(\sigma) = \frac{\psi_G(\sigma)}{Z_G}, \text{ where } \psi_G(\sigma) = \prod_{a\in F(G)}\psi_a(\sigma_{\partial a}) \text{ and } Z_G = \sum_{\tau\in\Omega^{V(G)}}\psi_G(\tau) \quad (\sigma \in \Omega^{V(G)}). \tag{3}$$

To describe problems such as the ones from Section 1 we introduce models where the factor graph itself is random. Specifically, let $\boldsymbol{d}, \boldsymbol{k} \geq 0$ be integer-valued random variables and let $(\boldsymbol{d}_i)_{i\geq 1}$, $(\boldsymbol{k}_i)_{i\geq 1}$ be independent copies of $\boldsymbol{d}, \boldsymbol{k}$. Further, for each $k$ in the support of $\boldsymbol{k}$ let $\Psi_k$ be a finite set of $k$-ary functions $\psi : \Omega^k \to (0, \infty)$. Let $P_k$ be a probability distribution on $\Psi_k$ and let us write $\boldsymbol{\psi}_k$ for a sample from $P_k$. Further, let $\boldsymbol{\psi}$ be a random variable distributed as $\boldsymbol{\psi}_{\boldsymbol{k}}$, let $P$ be the distribution of $\boldsymbol{\psi}_{\boldsymbol{k}}$ and let $k_\psi$ denote the arity of $\psi$.

Now, to construct a factor graph let $V_n = \{x_1, \ldots, x_n\}$ be a set of variable nodes and let $F_{\boldsymbol{m}} = \{a_1, \ldots, a_{\boldsymbol{m}}\}$ be a set of $\boldsymbol{m} \sim \mathrm{Po}(n\mathbb{E}[\boldsymbol{d}]/\mathbb{E}[\boldsymbol{k}])$ factor nodes. We obtain the random factor graph $\boldsymbol{G}$ as follows.

**G1** given the event $\sum_{i=1}^n \boldsymbol{d}_i = \sum_{i=1}^{\boldsymbol{m}} \boldsymbol{k}_i$, choose a bipartite graph on variable and factor nodes such that every $x_i$ has degree $\boldsymbol{d}_i$ and every $a_j$ has degree $\boldsymbol{k}_j$ uniformly at random.

**G2** choose for every factor node $a_i$ a weight function $\psi_{a_i}$ from the distribution $\boldsymbol{\psi}_{\boldsymbol{k}_i}$.

In the language of inference problems the random factor graph $\boldsymbol{G}$ is going to provide a null model because the weight functions in **G2** are independent of the graph structure from **G1**. For instance, in the context of the stochastic block model from Section 1.3, this model plays the role of the purely random graph without a particular underlying colouring.

## 2.2 The teacher-student scheme

The teacher-student scheme organically turns the null model into an inference problem. A helpful metaphor might be to imagine a teacher who attempts to convey a ground truth $\boldsymbol{\sigma}^*$ to a student by presenting examples. The ground truth itself is a random vector chosen uniformly from the space $\Omega^{V_n}$. The set of examples corresponds to a factor graph $\boldsymbol{G}^*$.

To be precise, let $\mathfrak{D}$ be the $\sigma$-algebra generated by the degrees and the total number of factor nodes of the null model $\boldsymbol{G}$. Then the factor graph $\boldsymbol{G}^*$ is chosen from the distribution

$$\mathbb{P}\left[\boldsymbol{G}^* = G \mid \mathfrak{D}, \boldsymbol{\sigma}^*\right] = \frac{\mathbb{P}\left[\boldsymbol{G} = G \mid \mathfrak{D}\right] \psi_G(\boldsymbol{\sigma}^*)}{\mathbb{E}[\psi_{\boldsymbol{G}}(\boldsymbol{\sigma}^*) \mid \mathfrak{D}, \boldsymbol{\sigma}^*]}. \tag{4}$$

Hence, we reweigh the null model **G1**–**G2** according to the ground truth $\boldsymbol{\sigma}^*$, rewarding graphs under which $\boldsymbol{\sigma}^*$ receives a higher weight. In the case of the stochastic block model, $\boldsymbol{G}^*$ matches the reweighing (2) that prefers bichromatic edges. The obvious question is how much of an imprint $\boldsymbol{\sigma}^*$ leaves on the resulting factor graph $\boldsymbol{G}^*$? Before we answer this question in general let us illustrate how the examples from Section 1 fit into the general framework.

▶ **Example 4** (ldgm codes). Let $\Omega = \{+1, -1\}$ with $+1 = (-1)^0$ representing $0 \in \mathbb{F}_2$ and $-1$ representing $1 \in \mathbb{F}_2$. For every degree $k \geq 1$ there are two $k$-ary weight functions $\psi_{\eta,k,\pm 1}$ defined by $\psi_{\eta,k,J}(\sigma) = 1 - (1-2\eta)J\prod_{i=1}^k \sigma_i$ for $\sigma \in \Omega^k$.

The probability distribution $P_k$ is defined by $P(\psi_{\eta,k,J}) = 1/2$. With this setup the bipartite graph structure of the null model $\boldsymbol{G}$ coincides with the bipartite graph introduced in Section 1.2. Moreover, the $\pm 1$-labels of the weight functions (i.e., value of $J$ such that $\psi_{a_i} = \psi_{\eta,\boldsymbol{k}_i,J}$) represent the entries of the vector $\boldsymbol{y}^*$. Thus, while in the null model $\boldsymbol{G}$ these vector entries are purely random, in the reweighted model $\boldsymbol{G}^*$ the labels are distributed precisely as the entries of the vector $\boldsymbol{y}^*$ from the ldgm model.

▶ **Example 5** (stochastic block model). Let $\Omega = [q]$ be a set of $q$ colours. We introduce a single binary weight function $\psi_{\beta,q}(\sigma_1, \sigma_2) = \exp(-\beta\mathbf{1}\{\sigma_1 = \sigma_2\})$ and we let $\boldsymbol{d}$ be the constant random variable $d$. With this weight function the construction (4) coincides with the definition (2) of the stochastic block model.

The main theorem is going to provide a formula for the mutual information of $\boldsymbol{G}^*$ and the ground truth $\boldsymbol{\sigma}^*$, provided that the distribution $P$ on weight functions satisfies a number of easy-to-check conditions. To state these conditions let us denote by $\mathcal{P}(\Omega)$ the set of all probability distributions on $\Omega$, endowed with the topology inherited from Euclidean space. Moreover, let $\mathfrak{P}_*(\Omega)$ signify the space of all probability measures $\pi$ on $\mathcal{P}(\Omega)$ such that $\int_{\mathcal{P}(\Omega)} \mu(\omega)\mathrm{d}\pi(\mu) = 1/q$ for all $\omega \in \Omega$. Finally, for a given $\pi \in \mathfrak{P}_*(\Omega)$ let $(\boldsymbol{\mu}_{i,\pi})_{i\geq 1}$ be independent samples from $\pi$ and recall $\Lambda(x) = x \log x$. The assumptions read as follows.

**DEG** there exists $\varepsilon > 0$ such that $\mathbb{E}[\boldsymbol{d}^{2+\varepsilon}], \mathbb{E}[\boldsymbol{k}^{2+\varepsilon}] < \infty$.

**SYM** there exist reals $\varepsilon, \xi > 0$ such that for all $k \in \operatorname{supp} \boldsymbol{k}$, $\psi \in \Psi_k$, $j \in [k]$, $\omega \in \Omega$ we have

$$\sum_{\sigma \in \Omega^k} \mathbf{1}\{\sigma_j = \omega\} \psi(\sigma) = q^{k-1}\xi, \qquad \varepsilon < \psi(\sigma) < 1/\varepsilon \quad (\sigma \in \Omega^k).$$

**BAL** for every $k \in \operatorname{supp} \boldsymbol{k}$ the function $\mu \in \mathcal{P}(\Omega) \mapsto \sum_{\sigma\in\Omega^k} \mathbb{E}[\boldsymbol{\psi}_k(\sigma)] \prod_{i=1}^k \mu(\sigma_i)$ is concave and attains its maximum at the uniform distribution on $\Omega$.

**POS** for any two probability distributions $\pi, \pi' \in \mathfrak{P}_*(\Omega)$ and any $k \in \operatorname{supp} \boldsymbol{k}$ we have

$$\mathbb{E}\left[\Lambda\left(\sum_{\tau\in\Omega^k} \boldsymbol{\psi}_k(\tau) \prod_{i=1}^k \boldsymbol{\mu}_{i,\pi}(\tau_i)\right)\right] + (k-1)\mathbb{E}\left[\Lambda\left(\sum_{\tau\in\Omega^k} \boldsymbol{\psi}_k(\tau) \prod_{i=1}^k \boldsymbol{\mu}_{i,\pi'}(\tau_i)\right)\right]$$
$$\geq \sum_{j=1}^k \mathbb{E}\left[\Lambda\left(\sum_{\tau\in\Omega^k} \boldsymbol{\psi}_k(\tau) \boldsymbol{\mu}_{j,\pi}(\tau_j) \prod_{i\neq j} \boldsymbol{\mu}_{i,\pi'}(\tau_i)\right)\right].$$

The first assumption **DEG** ensures that the factor graphs are “sparse” or, formally, locally finite. Yet **DEG** allows for very general degree distributions, including Poisson and power law distributions. Moreover, conditions **SYM** and **BAL** are symmetry conditions. Roughly speaking, they provide that all the values $\omega \in \Omega$ are on the same footing, i.e., there is no semantic preference for any value. Finally condition **POS** can be viewed as a convexity requirement. This assumption is needed for the technical reason of facilitating the interpolation method, a proof technique that we borrow from mathematical physics. The conditions are easily seen to be satisfied in many models of interest including, of course, the stochastic block model and ldgm codes. Crucially, the assumptions can be checked solely in terms of the weight functions; no random graphs considerations are required. [1]

## 2.3 The mutual information

The main result of the paper vindicates the physicists’ hunch that the mutual information between the teacher’s ground truth $\boldsymbol{\sigma}^*$ and the data $\boldsymbol{G}^*$ presented to the student is determined by the Bethe free entropy. To state the result we introduce the following generic version of the Bethe functional. Let $(\boldsymbol{\psi}_{k,i})_{k,i}$ be a family of independent random weight functions such that $\boldsymbol{\psi}_{k,i}$ is distributed as $\boldsymbol{\psi}_k$. Further, let $(\boldsymbol{h}_{k,i})_i$ with $\boldsymbol{h}_{k,i} \in [k]$ be a family of independent uniformly distributed indices. Given $\pi \in \mathfrak{P}_*(\Omega)$ let $(\mu_{i,j,\pi})_{i,j\geq 1}$ be a family of independent samples from $\pi$. Recalling that $(\hat{\boldsymbol{k}}_i)_{i\geq 1}$ are independent random variables with point masses (1), we define

[1] We point out that **POS** fails to hold in the case of the *assortative* stochastic block model.

$$\mathcal{B}(\pi) = \frac{1}{q}\mathbb{E}\left[\xi^{-\boldsymbol{d}}\Lambda\left(\sum_{\sigma\in\Omega}\prod_{i=1}^{\boldsymbol{d}}\sum_{\tau\in\Omega^{\hat{\boldsymbol{k}}_i}}\mathbf{1}\left\{\tau_{\boldsymbol{h}_{\hat{\boldsymbol{k}}_i,i}}=\sigma\right\}\boldsymbol{\psi}_{\hat{\boldsymbol{k}}_i,i}(\tau)\prod_{j\in[\hat{\boldsymbol{k}}_i]\setminus\{\boldsymbol{h}_{\hat{\boldsymbol{k}}_i,i}\}}\boldsymbol{\mu}_{i,j,\pi}(\tau_j)\right)\right] \qquad (5)$$
$$-\frac{\mathbb{E}[\boldsymbol{d}]}{\xi\mathbb{E}[\boldsymbol{k}]}\mathbb{E}\left[(\boldsymbol{k}-1)\Lambda\left(\sum_{\tau\in\Omega^{\boldsymbol{k}}}\boldsymbol{\psi}_{\boldsymbol{k}}(\tau)\prod_{j=1}^{\boldsymbol{k}}\boldsymbol{\mu}_{1,j,\pi}(\tau_j)\right)\right].$$

The following theorem expresses the mutual information of $\boldsymbol{G}^*$ and $\boldsymbol{\sigma}^*$ given the degrees and the total number of factor nodes as the variational problem of maximising the Bethe functional.

▶ **Theorem 6.** *For any random factor graph model that satisfies the conditions* **DEG**, **SYM**, **BAL** *and* **POS***,*

$$\lim_{n\to\infty}\frac{1}{n}I(\boldsymbol{\sigma}^*,\boldsymbol{G}^*\mid\mathfrak{D}) = \log q + \frac{\mathbb{E}[\boldsymbol{d}]}{\xi\mathbb{E}[\boldsymbol{k}]}\mathbb{E}\left[q^{-k_{\boldsymbol{\psi}}}\sum_{\tau\in\Omega^{k_{\boldsymbol{\psi}}}}\Lambda(\boldsymbol{\psi}(\tau))\right] - \sup_{\pi\in\mathfrak{P}_*(\Omega)}\mathcal{B}(\pi) \qquad (6)$$

*in probability.*

The formula (6) is in line with predictions from [33]. Moreover, the results quoted in Section 1 are immediate consequences of Theorem 6.

## 3 Proof strategy

In this section we survey the proof of Theorem 6. Subsequently we discuss how the strategy compares to prior work, particularly [11]. Throughout we tacitly assume that **DEG**, **SYM**, **BAL** and **POS** are satisfied.

### 3.1 The partition function

The starting point for computing the mutual information is to observe that this quantity is closely connected to the partition function of $\boldsymbol{G}^*$.

▶ **Proposition 7.** *W.h.p. we have*

$$I(\boldsymbol{\sigma}^*,\boldsymbol{G}^*\mid\mathfrak{D})/n = \log q + \frac{\mathbb{E}[\boldsymbol{d}]}{\xi\mathbb{E}[\boldsymbol{k}]}\mathbb{E}\left[q^{-k_{\boldsymbol{\psi}}}\sum_{\tau\in\Omega^{k_{\boldsymbol{\psi}}}}\Lambda(\boldsymbol{\psi}(\tau))\right] - \mathbb{E}[\log Z(\boldsymbol{G}^*)]/n + o(1).$$

Hence, Proposition 7 reduces our task to computing $\mathbb{E}[\log Z(\boldsymbol{G}^*)]$. This is still a formidable challenge because the logarithm sits inside the expectation; hence, routine techniques such as moment calculations do not bite. Instead we will combine two separate techniques. The first is a coupling argument known as the Aizenman-Sims-Starr scheme. This argument will show that $\mathbb{E}[\log Z(\boldsymbol{G}^*)]$ is upper bounded by $\sup_\pi \mathcal{B}(\pi)$. The second component, the interpolation method, will supply the matching lower bound.

What these techniques have in common is that they both boil down to "local" calculations. That is, we need to assess the impact on the partition function $Z(\boldsymbol{G}^*)$ of a small number of local changes such as addition of a few factor or variable nodes to $\boldsymbol{G}^*$. We will perform these computations by way of a probabilistic argument, namely by tracing how they affect the average weight of a sample from the Boltzmann distribution of $\boldsymbol{G}^*$. The key is a simple but powerful fact that trades as the Nishimori identity.

## 3.2 The Nishimori identity

To formulate this identity we need to introduce a slightly modified version of the random factor graph model $\boldsymbol{G}^*$. Recall from (4) that $\boldsymbol{G}^*$ was obtained by first drawing $\boldsymbol{\sigma}^*$ uniformly at random and then reweighting the null model $\boldsymbol{G}$ according to the weight of $\boldsymbol{\sigma}^*$. If we combine these two steps the net effect should be, at least roughly, that a specific $G$ comes up with probability proportional to $Z(G)$, as every $\sigma \in \Omega^{V_n}$ provides $G$ with a $\psi_G(\sigma)$ chance of being sampled. Thus, $\boldsymbol{G}^*$ should be roughly equivalent to the random factor graph model $\hat{\boldsymbol{G}}$ defined by $\mathbb{P}\left[\hat{\boldsymbol{G}} = G \mid \mathfrak{D}\right] \propto Z_G \mathbb{P}\left[\boldsymbol{G} = G \mid \mathfrak{D}\right]$. Indeed, this equivalence turns out to be exact if we make one minimal change. Namely, instead of drawing the ground truth $\boldsymbol{\sigma}^*$ uniformly at random, we draw a sample from the distribution $\mathbb{P}\left[\hat{\boldsymbol{\sigma}} = \sigma \mid \mathfrak{D}\right] \propto \mathbb{E}\left[\psi_{\boldsymbol{G}}(\sigma) \mid \mathfrak{D}\right]$ for $\sigma \in \Omega^{V_n}$. The following is an extension of [11, Proposition 3.10] to the present, more general class of factor graph models with given degrees.

▶ **Proposition 8.** *We have*

$$\mathbb{P}\left[\hat{\boldsymbol{G}} = G \mid \mathfrak{D}\right] \mu_G(\sigma) = \mathbb{P}\left[\hat{\boldsymbol{\sigma}} = \sigma \mid \mathfrak{D}\right] \mathbb{P}\left[\boldsymbol{G}^* = G \mid \mathfrak{D}, \boldsymbol{\sigma}^* = \sigma\right]. \tag{7}$$

*Furthermore, $\hat{\boldsymbol{\sigma}}$ and $\boldsymbol{\sigma}^*$ as well as $\boldsymbol{G}^*, \hat{\boldsymbol{G}}$ are mutually contiguous and* $\mathbb{E}[\log Z_{\boldsymbol{G}^*}] = \mathbb{E}[\log Z_{\hat{\boldsymbol{G}}}] + o(n)$.

The proof of Proposition 8 relies on Bayes' formula combined with a somewhat subtle application of local limit theorems and other probabilistic tools. The details can be found in Section 4 of the full version.

## 3.3 Degree pruning

A further preparation is degree pruning. Specifically, while in the random factor graph models $\boldsymbol{G}^*$ and $\hat{\boldsymbol{G}}$ may possess degrees as large as $n^{1/2-\varepsilon}$, the following proposition shows that it suffices to prove the main result (6) for bounded degree sequences.

▶ **Proposition 9.** *Assume that for any integer $L > 0$ and for any $\boldsymbol{d}, \boldsymbol{k}$ such that $\boldsymbol{d}, \boldsymbol{k} \leq L$ the statement* (6) *is true. Then* (6) *holds for all $\boldsymbol{d}, \boldsymbol{k}$ that satisfy* **DEG** *and for which* $\mathbb{E}\left[\boldsymbol{d}\right], \mathbb{E}\left[\boldsymbol{k}\right] > 0$.

The proof of Proposition 9 is based on concentration inequalities and coupling arguments for bipartite graphs with given degree sequences. Hence, we may assume from here on that $\boldsymbol{d}, \boldsymbol{k}$ are bounded.

## 3.4 Cavities and couplings

Two of the main steps towards the proof of Theorem 6, the Aizenman-Sims-Starr scheme and the interpolation method, hinge on comparing random factor graphs with slightly different parameters. For example, we will need to compare a random factor graph $\boldsymbol{G}^*$ with $n$ variable and $\mathrm{Po}(\mathbb{E}[\boldsymbol{d}]n/\mathbb{E}[\boldsymbol{k}])$ factor nodes and a factor graph with $n+1$ variable and the commensurate number of $\mathrm{Po}(\mathbb{E}[\boldsymbol{d}](n+1)/\mathbb{E}[\boldsymbol{k}])$ factor nodes. In the classical case of binomial factor graphs as treated in [11] where factor nodes are drawn independently this coupling would be relatively straightforward. Indeed, we could just add a variable node and a few extra factor nodes to the graph with $n$ variables. However, in the present setting of given degrees matters are much more delicate. For instance, how would you set up such a coupling for the $d$-regular stochastic block model from Section 1.3? Due to the given degrees the graph structure is too rigid to accommodate the necessary local changes.

To cope with this issue we first create a bit of wiggling room for ourselves by slightly reducing the number of factor nodes. This idea has been used in prior work on factor graphs with rigid degree distributions such as [9]. However, matters turn out to be rather more delicate here because we do not just work with purely random factor graphs, but with graphs drawn from the teacher-student model. Thus, we need to take care to meticulously implement the weight shifts in accordance with (4). Hence, for a small but fixed $\varepsilon > 0$ let $\boldsymbol{m}_\varepsilon = \mathrm{Po}((1-\varepsilon)\mathbb{E}[\boldsymbol{d}]n/\mathbb{E}[\boldsymbol{k}])$ be a Poisson variable with a slightly smaller mean than $\boldsymbol{m}$. Because we assume that all degrees are bounded, with probability $1-\exp(-\Omega(n))$ we have $\sum_{i=1}^n \boldsymbol{d}_i \geq \sum_{i=1}^{\boldsymbol{m}_\varepsilon} \boldsymbol{k}_i$. In fact, w.h.p. the total variable degree exceeds the total degree of the first $\boldsymbol{m}_\varepsilon$ factor nodes by $\Omega(n)$. Let $\boldsymbol{G}(n, \boldsymbol{m}_\varepsilon)$ be a random factor graph with variable nodes $x_1, \ldots, x_n$ and factor nodes $a_1, \ldots, a_{\boldsymbol{m}_\varepsilon}$ of degrees $\boldsymbol{k}_1, \ldots, \boldsymbol{k}_{\boldsymbol{m}_\varepsilon}$ drawn uniformly at random subject to the condition that the degree of each $x_i$ remains bounded by $\boldsymbol{d}_i$. Thus, some of the variable nodes will likely have a degree strictly smaller than their "target degree" $\boldsymbol{d}_i$. We refer to these variable degrees as *cavities*. Further, given $\sigma \in \Omega^{V_n}$ let $\boldsymbol{G}^*(n, \boldsymbol{m}_\varepsilon, \sigma)$ be the random factor graph obtained as in (4), i.e., with $\mathfrak{D}_\varepsilon$ denoting the $\sigma$-algebra generated by the degrees and the total number of factors nodes of $\boldsymbol{G}(n, \boldsymbol{m}_\varepsilon)$ we let

$$\mathbb{P}\left[\boldsymbol{G}^*(n, \boldsymbol{m}_\varepsilon, \sigma) = G \mid \mathfrak{D}_\varepsilon\right] \propto \mathbb{P}\left[\boldsymbol{G}(n, \boldsymbol{m}_\varepsilon) = G \mid \mathfrak{D}_\varepsilon\right] \psi_G(\sigma).$$

The following proposition establishes that we can indeed think of $\boldsymbol{G}^*(n, \boldsymbol{m}_\varepsilon + 1, \sigma)$ as being obtained from $\boldsymbol{G}^*(n, \boldsymbol{m}_\varepsilon, \sigma)$ by adding one extra factor node $a_{\boldsymbol{m}_\varepsilon+1}$. Further, for two factor graphs $G, G'$ on the same set of nodes let $G \triangle G'$ be the symmetric difference of their edge sets.

▶ **Proposition 10.** *Assume that $|\sigma^{-1}(\omega)| = n/q + O(\sqrt{n}\log n)$ for all $\omega \in \Omega$. Then there exists a coupling of $\boldsymbol{G}^*(n, \boldsymbol{m}_\varepsilon, \sigma)$ and $\boldsymbol{G}^*(n, \boldsymbol{m}_\varepsilon + 1, \sigma)$ such that*

$$\mathbb{P}\left[\boldsymbol{G}^*(n, \boldsymbol{m}_\varepsilon, \sigma) = \boldsymbol{G}^*(n, \boldsymbol{m}_\varepsilon + 1, \sigma) - a_{\boldsymbol{m}_\varepsilon+1} \mid \mathfrak{D}_\varepsilon\right] = 1 - \tilde{O}(1/n),$$
$$\mathbb{P}\left[|\boldsymbol{G}^*(n, \boldsymbol{m}_\varepsilon, \sigma) \triangle \boldsymbol{G}^*(n, \boldsymbol{m}_\varepsilon + 1, \sigma) - a_{\boldsymbol{m}_\varepsilon+1}| > n^{2/3} \mid \mathfrak{D}_\varepsilon\right] = 1 - \tilde{O}(1/n^2).$$

There is a similar coupling that accommodates the addition of an extra variable node.

▶ **Proposition 11.** *Assume that $|\sigma^{-1}(\omega)| = n/q + O(\sqrt{n}\log n)$ for all $\omega \in \Omega$. Given the degree $\boldsymbol{\gamma}$ of $x_{n+1}$ in $\boldsymbol{G}^*(n+1, \boldsymbol{m}_\varepsilon + \boldsymbol{\gamma}, \sigma)$ then there exists a coupling of $\boldsymbol{G}^*(n, \boldsymbol{m}_\varepsilon, \sigma)$ and $\boldsymbol{G}^*(n+1, \boldsymbol{m}_\varepsilon + \boldsymbol{\gamma}, \sigma)$ such that*

$$\mathbb{P}\left[\boldsymbol{G}^*(n, \boldsymbol{m}_\varepsilon, \sigma) = \boldsymbol{G}^*(n+1, \boldsymbol{m}_\varepsilon + \boldsymbol{\gamma}, \sigma) - x_{n+1} - \partial x_{n+1} \mid \mathfrak{D}_\varepsilon\right] = 1 - \tilde{O}(1/n),$$
$$\mathbb{P}\left[|\boldsymbol{G}^*(n, \boldsymbol{m}_\varepsilon, \sigma) = \boldsymbol{G}^*(n+1, \boldsymbol{m}_\varepsilon + \boldsymbol{\gamma}, \sigma) - x_{n+1} - \partial x_{n+1}| > n^{2/3} \mid \mathfrak{D}_\varepsilon\right] = 1 - \tilde{O}(1/n^2).$$

The orders $\tilde{O}(1/n), \tilde{O}(1/n^2)$ of the error terms in Propositions 10 and 11 are vital to facilitate the computation of the partition function. On a technical level, the tools that we develop for proving these propositions, and particularly for dealing with the fragile combinatorics of the factor graph models with given degrees, constitute the main novelty of the paper. This is where we most visibly add to and improve over the machinery developed in prior work. The details can be found in Section 4.3 of the full version.

## 3.5 Aizenman-Sims-Starr and interpolation

Propositions 10 and 11 in combination with a trick known as the Aizenman-Sims-Starr scheme yield the desired upper bound on the partition function.

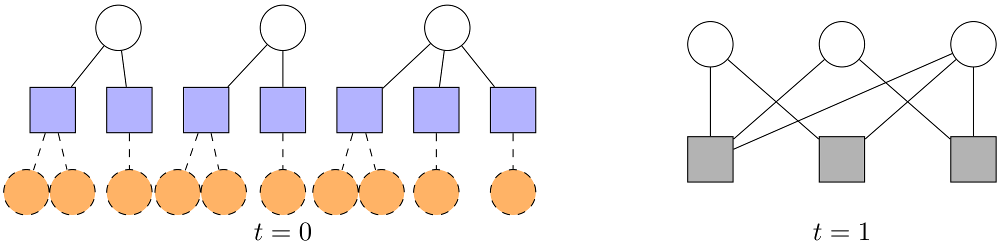

**Figure 1** Illustration of the interpolation method at "times" $t = 0$ and $t = 1$.

▶ **Proposition 12.** *We have* $\mathbb{E}[\log Z(\boldsymbol{G}^*)] \le n \sup_{\pi \in \mathfrak{P}_*(\Omega)} \mathcal{B}(\pi) + o(n)$.

To prove Proposition 12 it suffices to establish the corresponding upper bound for $\boldsymbol{G}^*(n, \boldsymbol{m}_\varepsilon, \boldsymbol{\sigma}^*)$. This is because similar but simpler arguments as in the proof of Proposition 10 show that $\mathbb{E}[\log Z(\boldsymbol{G}^*)] = \mathbb{E}[\log Z(\boldsymbol{G}^*(n, \boldsymbol{m}_\varepsilon, \boldsymbol{\sigma}^*)] + O(\varepsilon n)$. Its proof can be found in Section 13 of the full version. Now, the Aizenman-Sims-Starr scheme for calculating the latter quantity is to write a telescoping sum

$$\begin{aligned}&\mathbb{E}[\log Z(\boldsymbol{G}^*(n, \boldsymbol{m}_\varepsilon, \boldsymbol{\sigma}^*))]\\ &\quad = \sum_{N=0}^{n-1} \mathbb{E}[\log Z(\boldsymbol{G}^*(N+1, \boldsymbol{m}_\varepsilon(N+1), \boldsymbol{\sigma}^*_{N+1}))] - \mathbb{E}[\log Z(\boldsymbol{G}^*(N, \boldsymbol{m}_\varepsilon(N), \boldsymbol{\sigma}^*_N))].\end{aligned}$$

Hence, it suffices to bound the individual summands on the r.h.s., i.e., the differences

$$\mathbb{E}[\log Z(\boldsymbol{G}^*(n+1, \boldsymbol{m}_\varepsilon(n+1), \boldsymbol{\sigma}^*_{n+1}))] - \mathbb{E}[\log Z(\boldsymbol{G}^*(n, \boldsymbol{m}_\varepsilon(n), \boldsymbol{\sigma}^*_n))]. \tag{8}$$

To this end we couple these two random factor graphs. This is where Propositions 10 and 11 enter the fray. Specifically, we think of both these factor graphs as being obtained from a smaller factor graph $\boldsymbol{G}^*_0$ that with variables nodes $x_1, \ldots, x_n$ and slightly fewer factor nodes than either of the two target factor graphs. Then we obtain $\boldsymbol{G}^*(n, \boldsymbol{m}_\varepsilon(n), \boldsymbol{\sigma}^*_n)$ by adding a few random factors to $\boldsymbol{G}^*_0$. Similarly, we obtain $\boldsymbol{G}^*(n+1, \boldsymbol{m}_\varepsilon(n+1), \boldsymbol{\sigma}^*_{n+1})$ from $\boldsymbol{G}^*_0$ by adding a few new random factor nodes as well as a new variable node $x_{n+1}$ along with a number of adjacent factor nodes. Crucially, Propositions 10 and 11 provide the necessary accuracy to trace the impact of these manipulations on the partition function, and the Bethe functional emerges organically as an upper bound on (8).

To obtain the matching lower bound we seize upon the interpolation method. The basic idea is to set up a family of random factor graph models parametrised by time $t \in [0,1]$ such that the model at time $t = 1$ coincides with $\boldsymbol{G}^*(n, \boldsymbol{m}_\varepsilon, \boldsymbol{\sigma}^*)$ while the model at time $t = 0$ is so simple that its partition function can be read off easily. In fact, the partition function of the $t = 0$ model turns out to be the Bethe free entropy. To derive the desired lower bound we prove that the derivative of the log-partition function remains non-negative as we increase $t$. As in the Aizenman-Sims-Starr scheme, the computation of the derivative can be reduced to tracing the impact of local changes. Hence, once more we bring Proposition 10 to bear, this time in combination with the convexity assumption **POS**, to prove the following.

▶ **Proposition 13.** *We have* $\mathbb{E}[\log Z(\boldsymbol{G}^*)] \ge n \sup_{\pi \in \mathfrak{P}_*(\Omega)} \mathcal{B}(\pi) + o(n)$.

Finally, combining Proposition 7–13, we obtain Theorem 6.

## 3.6 Discussion

There has been a great deal of interest in inference problems on random factor graphs recently. The substantial literature on the stochastic block model alone, much of it devoted to corroborating the predictions from [12], is surveyed in [1, 26]. The literature on applications to modern coding theory until about 2008 is surveyed in [31]; important newer contributions include [20, 21]. Further recent applications include compressed sensing [14, 15], group testing [2, 10], code-division multiple access [17, 30] and the patient zero problem [3]. Apart and beyond this rigorous literature, there is a vast body of work based on either physics techniques such as the cavity method or computer experiments.

The great variety of concrete problems studied individually underscores the potential of generic proof techniques or, even better, general theorems that rigorise these predictions wholesale. A first contribution has been made by Coja-Oghlan, Krzalaka, Perkins and Zdeborová [11], who studied the teacher-student model on binomial random factor graph models. While the general proof strategy that we pursue here is guided by that paper, the present factor graph models are more general by allowing prescribed degree sequences for both the variable and factor nodes. From an application viewpoint this generality is highly desirable because, for example, the quality of an error correcting code or a group testing scheme can be boosted by optimising the degree distribution [31]. However, from a technical viewpoint this generality comes at the cost of losing (conditional) independence among the factor nodes. This issue is well known in random graph theory, where random graphs with given degrees require far more intricate proofs than, e.g., the Erdős–Rényi model [18]. Here, these difficulties are exacerbated by the fact that we study not just the plain random graph, which serves as a our null model, but the reweighted random graph distribution induced by the teacher-student scheme. In effect, many of the steps that were straightforwards in [11] become rather delicate due to stochastic dependencies. The key tool that allows us to cope with these dependencies is Proposition 10. Thus, while we follow the strategy from [11] of combining the Aizenman-Sims-Starr scheme with the interpolation method and although we adopt some of the technical ingredients from that work such as the "pinning lemma", the greater generality of the model leads us to crystallise and improve over the previous approach.

What are alternatives to the present strategy of combining the Aizenman-Sims-Starr scheme with the interpolation method? A classical approach to inference problems on random graphs is the second moment method [5]. Unfortunately, this approach does not generally allow for tight information-theoretic results. The reason is that the precise formula for the mutual information or the information-theoretic threshold in, e.g., the stochastic block model comes in terms of the optimiser of the Bethe free entropy functional. The distribution $\pi$ where the maximum is obtained mirrors the outcome of a complicated message passing process. Intuitively, $\pi$ is an idealised version of the empirical distribution of Belief Propagation messages that whiz around the factor graph upon convergence when launched from either a uniform initialisation or from the completely polarised initialisation corresponding to the ground truth. In some examples this fixed point can be characterised precisely and, unsurprisingly, turns out to be anything but trivial [6]. But we cannot expect the expressiveness required for such a complicated object from a plain second moment computation. A second conceptually elementary approach is to actually compute the message passing fixed point by hand, e.g., via the contraction method. But due to the intricacy of the calculations this method has been pushed through in only a few special cases [27].

Further powerful techniques include spatial coupling [16] and the adaptive interpolation method [7]. Both potentially allow for precise results. The basic idea behind spatial coupling is to convert the given model into a factor graph model with a superimposed geometric structure. A plus of spatial coupling is that it sometimes allows for better inference algorithms. A disadvantage is that the construction has to be carried out case-by-case. By comparison, the adaptive interpolation method has the advantage of being technically relatively clean. However, at least on sparse models its combinatorial nuts and bolts appear to be roughly equivalent to the combination of Aizenman-Sims-Starr and the interpolation argument used here. Furthermore, the latter approach has the merit of being closer in spirit to the physicists' cavity calculation. In addition, at this time the adaptive interpolation method has not been extended to models with given general degree sequences.

Further, there has been quite some work on dense random factor graph models where each variable appears in a constant fraction of factor nodes. Examples are spiked matrix/tensor models [8] or models of neural networks such as the Hopfield model [4, 23]. These methods are closer in nature to the classical Sherrington-Kirkpatrick model [28]. It seems fair to say that more is known about dense models than sparse ones because certain central limit theorem-like simplifications arise. In some cases, the Bethe variational principle reduces to a finite-dimensional or even scalar optimisation problem [13, 22].

To conclude we note that the study of inference problems typically comes in two instalments: an information-theoretic view that asks for thresholds beyond which in principle sufficient information is available to form a non-trivial estimate of the ground truth and an algorithmic view interested in polynomial-time algorithms. While the two perspectives might appear disparate at first glance, information-theoretic results on inference problems like in this paper in combination with tools such as spatial coupling have in the past led to efficient algorithms capable of attaining the information-theoretic thresholds [10, 15]. We view this as an exciting avenue for future research.

## References

1 E. Abbe and A. Montanari. Conditional random fields, planted constraint satisfaction and entropy concentration. *Theory of Computing*, 11:413–443, 2015.

2 M. Aldridge, O. Johnson, and J. Scarlett. Group testing: an information theory perspective. *Foundations and Trends in Communications and Information Theory*, 2019.

3 F. Altarelli, A. Braunstein, L. Dall'Asta, A. Lage-Castellanos, and R. Zecchina. Bayesian inference of epidemics on networks via belief propagation. *Physical review letters*, 112:118701, 2014.

4 D. Amit, H. Gutfreund, and H. Sompolinsky. Storing infinite numbers of patterns in a spin-glass model of neural networks. *Physical Review Letters*, 55:1530, 1985.

5 J. Banks, C. Moore, J. Neeman, and P. Netrapalli. Information-theoretic thresholds for community detection in sparse networks. *Proc. 29th COLT*, pages 383–416, 2016.

6 V. Bapst, A. Coja-Oghlan, S. Hetterich, F. Rassmann, and D. Vilenchik. The condensation phase transition in random graph coloring. *Communications in Mathematical Physics*, 341:543–606, 2016.

7 J. Barbier, C. Chan, and N. Macris. Mutual information for the stochastic block model by the adaptive interpolation method. *Proc. IEEE International Symposium on Information Theory*, pages 405–409, 2019.

8 J. Barbier and N. Macris. The adaptive interpolation method for proving replica formulas. applications to the Curie–Weiss and Wigner spike models. *Journal of Physics A: Mathematical and Theoretical*, 52:294002, 2019.

9 A. Coja-Oghlan, A. Ergür, P. Gao, S. Hetterich, and M. Rolvien. The rank of sparse random matrices. *Proc. 31st SODA*, pages 579–591, 2020.

10 A. Coja-Oghlan, O. Gebhard, M. Hahn-Klimroth, and P. Loick. Optimal group testing. *Proceedings of Machine Learning Research (COLT)*, 2020.

11 A. Coja-Oghlan, F. Krzakala, W. Perkins, and L. Zdeborová. Information-theoretic thresholds from the cavity method. *Advances in Mathematics*, 333:694–795, 2018.

12 A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Phys. Rev. E*, 84:066106, 2011.

13 M. Dia, N. Macris, F. Krzakala, T. Lesieur, and L. Zdeborová. Mutual information for symmetric rank-one matrix estimation: A proof of the replica formula. *Advances in Neural Information Processing Systems*, pages 424–432, 2016.

14 D. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52:1289–1306, 2006.

15 D. Donoho, A. Javanmard, and A. Montanari. Information-theoretically optimal compressed sensing via spatial coupling and approximate message passing. *IEEE Transactions on Information Theory*, 59:7434–7464, 2013.

16 A. Giurgiu, N. Macris, and R. Urbanke. Spatial coupling as a proof technique and three applications. *IEEE Transactions on Information Theory*, 62:5281–5295, 2016.

17 D. Guo and C. Wang. Multiuser detection of sparsely spread cdma. *IEEE journal on selected areas in communications*, 26:421–431, 2008.

18 S. Janson, T. Łuczak, and A. Rucinski. Random graphs. *John Wiley & Sons*, 45, 2011.

19 F. Krzakala, A. Montanari, F. Ricci-Tersenghi, G. Semerjian, and L. Zdeborova. Gibbs states and the set of solutions of random constraint satisfaction problems. *Proc. National Academy of Sciences*, 104:10318–10323, 2007.

20 S. Kudekar, T. Richardson, and R. Urbanke. Spatially coupled ensembles universally achieve capacity under belief propagation. *IEEE Transactions on Information Theory*, 59:7761–7813, 2013.

21 S. Kumar, A. Young, N. Macris, and H. Pfister. Threshold saturation for spatially coupled ldpc and ldgm codes on bms channels. *IEEE Transactions on Information Theory*, 60:7389–7415, 2014.

22 M. Lelarge and L. Miolane. Fundamental limits of symmetric low-rank matrix estimation. *Conference on Learning Theory (COLT)*, pages 1297–1301, 2017.

23 M. Mézard. Mean-field message-passing equations in the Hopfield model and its generalizations. *Physical Review E*, 95:022117, 2017.

24 M. Mézard and A. Montanari. Information, physics and computation. *Oxford University Press*, 2009.

25 A. Montanari. Tight bounds for ldpc and ldgm codes under map decoding. *IEEE Transactions on Information Theory*, 51:3221–3246, 2005.

26 C. Moore. The computer science and physics of community detection: landscapes, phase transitions, and hardness. *Bull. EATCS*, 121, 2017.

27 E. Mossel, J. Neeman, and A. Sly. Reconstruction and estimation in the planted partition model. *Probability Theory and Related Fields*, 162:431–461, 2015.

28 D. Panchenko. The Sherrington-Kirkpatrick model. *Springer*, 2013.

29 J. Pearl. Probabilistic reasoning in intelligent systems: networks of plausible inference. *Elsevier*, 2014.

30 J. Raymond and D. Saad. Sparsely spread cdma – a statistical mechanics-based analysis. *Journal of physics A: mathematical and theoretical*, 40:12315, 2007.

31 T. Richardson and R. Urbanke. Modern coding theory. *Cambridge University Press*, 2012.

32 J. van den Brand and N. Jaafari. The mutual information of ldgm codes. *arXiv*, 2017. `arXiv:1707.04413`.

33 L. Zdeborová and F. Krzakala. Phase transition in the coloring of random graphs. *Phys. Rev. E*, 76:031131, 2007.

34 L. Zdeborová and F. Krzakala. Statistical physics of inference: thresholds and algorithms. *Advances in Physics*, 65:453–552, 2016.