Contextual Stochastic Block Models

Yash Deshpande* Andrea Montanari † Elchanan Mossel‡ Subhabrata Sen§

July 23, 2018

Abstract

We provide the first information theoretic tight analysis for inference of latent community structure given a sparse graph along with high dimensional node covariates, correlated with the same latent communities. Our work bridges recent theoretical breakthroughs in the detection of latent community structure without nodes covariates and a large body of empirical work using diverse heuristics for combining node covariates with graphs for inference. The tightness of our analysis implies in particular, the information theoretical necessity of combining the different sources of information. Our analysis holds for networks of large degrees as well as for a Gaussian version of the model.

1 Introduction

Data clustering is a widely used primitive in exploratory data analysis and summarization. These methods discover clusters or partitions that are assumed to reflect a latent partitioning of the data with semantic significance. In a machine learning pipeline, results of such a clustering may then be used for downstream supervised tasks, such as feature engineering, privacy-preserving classification or fair allocation [CMS11, KGB$^+$12, CDPF$^+$17].

At risk of over-simplification, there are two settings that are popular in literature. In *graph clustering*, the dataset of n objects is represented as a symmetric *similarity matrix* $A = (A_{ij})_{1\le i,j\le n}$. For instance, A can be binary, where $A_{ij} = 1$ (or 0) denotes that the two objects i, j are similar (or not). It is, then, natural to interpret A as the adjacency matrix of a graph. This can be carried over to non-binary settings by considering weighted graphs. On the other hand, in more traditional (binary) *classification* problems, the n objects are represented as p-dimensional feature or covariate vectors $b_1, b_2, \cdots, b_n$. This feature representation can be the input for a clustering method such as k-means, or instead used to construct a similarity matrix A, which in turn is used for clustering or partitioning. These two representations are often taken to be mutually exclusive and, in fact, interchangeable. Indeed, just as feature representations can be used to construct similarity matrices, popular spectral methods [NJW02, VL07] implicitly construct a low-dimensional feature representation from the similarity matrices.

This paper is motivated by scenarios where the graph, or similarity, representation $A \in \mathbb{R}^{n\times n}$, and the feature representation $B = [b_1, b_2, \ldots, b_n] \in \mathbb{R}^{p\times n}$ provide *independent*, or *complementary*, information on the latent clustering of the n objects. (Technically, we will assume that A and B are conditionally independent given the node labels.) We argue that in fact in almost all practical graph clustering problems, feature representations provide complementary information of the latent clustering. This is indeed the case in many social and biological networks, see e.g. [NC16] and references within.

As an example, consider the 'political blogs' dataset [AG05]. This is a directed network of political blogs during the 2004 US presidential election, with a link between two blogs if one referred to the other. It is possible to just use the graph structure in order to identify political communities (as was done in

*Department of Mathematics, Massachusetts Institute of Technology
†Departments of Electrical Engineering and Statistics, Stanford University
‡Department of Mathematics, Massachusetts Institute of Technology
§Department of Mathematics, Massachusetts Institute of Technology

[AG05]). Note however that much more data is available. For example we may consider an alternative feature representation of the blogs, wherein each blog is converted to a 'bag-of words' vector of its content. This gives a quite different, and complementary representation of blogs that plausibly reflects their political leaning. A number of approaches can be used for the simple task of predicting leaning from the graph information (or feature information) individually. However, given access to both sources, it is challenging to combine them in a principled fashion.

In this context, we introduce a simple statistical model of complementary graph and high-dimensional covariate data that share latent cluster structure. This model is an intuitive combination of two well-studied models in machine learning and statistics: the *stochastic block model* and the *spiked covariance model* [Abb17, HLL83, JL04]. We focus on the task of uncovering this latent structure and make the following contributions:

Sharp thresholds: We establish a sharp information-theoretic threshold for detecting the latent structure in this model. This threshold is based on non-rigorous, but powerful, techniques from statistical physics.

Rigorous validation: We consider a certain 'Gaussian' limit of the statistical model, which is of independent interest. In this limit, we rigorously establish the correct information-theoretic threshold using novel Gaussian comparison inequalities. We further show convergence to the Gaussian limit predictions as the density of the graph diverges.

Algorithm: We provide a simple, iterative algorithm for inference based on the belief propagation heuristic. For data generated from the model, we empirically demonstrate that the the algorithm achieves the conjectured information-theoretic threshold.

The rest of the paper is organized as follows. The model and results are presented in Section 2. Further related work is discussed in Section 3. The prediction of the threshold from statistical physics techniques is presented in 4, along with the algorithm. While all proofs are presented in the appendix, we provide an overview of the proofs of our rigorous results in Section 5. Finally, we numerically validate the prediction in Section 6.

2 Model and main results

We will focus on the simple case where the n objects form two latent clusters of approximately equal size, labeled $+$ and $-$. Let $v \in \{\pm 1\}^n$ be the vector encoding this partitioning. Then, the observed data is a pair of matrices (A^G, B), where A^G is the adjacency matrix of the graph G and $B \in \mathbb{R}^{p\times n}$ is the matrix of covariate information. Each column b_i, $i \le n$ of matrix B contains the covariate information about vertex i. We use the following probabilistic model: conditional on v, and a latent vector $u \sim \mathsf{N}(0, I_p/p)$:

$$\mathbb{P}(A^G_{ij} = 1) = \begin{cases} c_{\text{in}}/n & \text{with probability } , \\ c_{\text{out}}/n & \text{otherwise.} \end{cases} \tag{1}$$

$$b_i = \sqrt{\frac{\mu}{n}} v_i u + \frac{Z_i}{\sqrt{p}}, \tag{2}$$

where $Z_i \in \mathbb{R}^p$ has independent standard normal entries. It is convenient to parametrize the edge probabilities by the average degree d and the normalized degree separation λ:

$$c_{\text{in}} = d + \lambda\sqrt{d}\,, \qquad c_{\text{out}} = d - \lambda\sqrt{d}\,. \tag{3}$$

Here d, λ, μ are parameters of the model which, for the sake of simplicity, we assume to be fixed and known. In other words, two objects i, j in the same cluster or community are *slightly more likely* to be connected than for objects i, j' in different clusters. Similarly, according to (2), they have *slightly positively* correlated feature vectors b_i, b_j, while objects i, j' in different clusters have negatively correlated covariates $b_i, b_{j'}$.

Note that this model is a combination of two observation models that have been extensively studied: the stochastic block model and the spiked covariance model. The stochastic block model has its roots in sociology literature [HLL83] and has witnessed a resurgence of interest from the computer science and statistics community since the work of Decelle et al. [DKMZ11]. This work focused on the sparse setting where the graph as $O(n)$ edges and conjectured, using the non-rigorous cavity method, the following *phase transition* phenomenon. This was later established rigorously in a series of papers [MNS15, MNS13, Mas14].

Theorem 1 ([MNS15, MNS13, Mas14])**.** *Suppose $d > 1$ is fixed. The graph G is distinguishable with high probability from an Erdös-Renyi random graph with average degree d if and only if $\lambda \geq 1$. Moreover, if $\lambda > 1$, there exists a polynomial-time computable estimate $\widehat{v} = \widehat{v}(A^G) \in \{\pm 1\}^n$ of the cluster assignment satisfying, almost surely:*

$$\liminf_{n\to\infty} \frac{|\langle \widehat{v}, v\rangle|}{n} \geq \varepsilon(\lambda) > 0. \tag{4}$$

In other words, given the graph G, it is possible to non-trivially estimate the latent clustering v if, and only if, $\lambda > 1$.

The covariate model (2) was proposed by Johnstone and Lu [JL04] and has been extensively studied in statistics and random matrix theory. The weak recovery threshold was characterized by a number of authors, including Baik et al [BBAP05], Paul [Pau07] and Onatski et al [OMH$^+$13].

Theorem 2 ([BBAP05, Pau07, OMH$^+$13])**.** *Let $\widehat{v}_1$ be the principal eigenvector of $B^\mathsf{T}B$, where $\widehat{v}_1$ is normalized so that $\|\widehat{v}_1\|^2 = n$. Suppose that $p, n \to \infty$ with $p/n \to 1/\gamma \in (0, \infty)$. Then $\liminf_{n\to\infty} |\langle \widehat{v}_1, v\rangle|/n > 0$ if and only if $\mu > \sqrt{\gamma}$. Moreover, if $\mu < \sqrt{\gamma}$, no such estimator exists.*

In other words, this theorem shows that it is possible to estimate v nontrivally solely from the covariates using, in fact, a spectral method if, and only if $\mu > \sqrt{\gamma}$.

Our first result is the following prediction that establishes the analogous threshold prediction that smoothly interpolates between Theorems 1 and 2.

Claim 3 (Cavity prediction)**.** *Given A^G, B as in Eqs.*(1), (2)*, and assume that $n, p \to \infty$ with $p/n \to 1/\gamma \in (0, \infty)$. Then there exists an estimator $\widehat{v} = \widehat{v}(A^G, B) \in \{\pm 1\}^n$ so that $\liminf |\langle \widehat{v}, v\rangle|/n$ is bounded away from 0 if and only if*

$$\lambda^2 + \frac{\mu^2}{\gamma} > 1\,. \tag{5}$$

We obtain this prediction via the cavity method, a powerful technique from the statistical physics of mean field models [MM09]. This derivation is outlined in Section 4. Theorems 1 and 2 confirm this prediction rigorously in the corner cases, in which either λ or μ vanishes, using sophisticated tools from random matrix theory and sparse random graphs.

Our main result confirms rigorously this claim in the limit of large degrees.

Theorem 4. *Suppose v is uniformly distributed in $\{\pm 1\}^n$ and we observe A^G, B as in* (1), (2)*. Consider the limit $p, n \to \infty$ with $p/n \to 1/\gamma$. Then we have, for some $\varepsilon(\lambda, \mu) > 0$ independent of d,*

$$\liminf_{n\to\infty} \sup_{\widehat{v}(\,\cdot\,)} \frac{|\langle \widehat{v}(A^G, B), v\rangle|}{n} \geq \varepsilon(\lambda, \mu) - o_d(1) \qquad \text{if } \lambda^2 + \mu^2/\gamma > 1, \tag{6}$$

$$\limsup_{n\to\infty} \sup_{\widehat{v}(\,\cdot\,)} \frac{|\langle \widehat{v}(A^G, B), v\rangle|}{n} = o_d(1) \qquad \text{if } \lambda^2 + \mu^2/\gamma < 1. \tag{7}$$

Here the limits hold in probability, the supremum is over estimators $\widehat{v} : (A^G, B) \mapsto \widehat{v}(A^G, B) \in \mathbb{R}^n$, with $\|\widehat{v}(A^G, B)\|_2 = \sqrt{n}$. Here $o_d(1)$ indicates a term independent of n which tends to zero as $d \to \infty$.

In order to establish this result, we consider a modification of the original model in (1), (2), which is of independent interest. Suppose, conditional on $v \in \{\pm 1\}$ and the latent vector u we observe (A, B) as follows:

$$A_{ij} \sim \begin{cases} \mathsf{N}(\lambda v_i v_j/n, 1/n) & \text{if } i < j \\ \mathsf{N}(\lambda v_i v_j/n, 2/n) & \text{if } i = j, \end{cases} \tag{8}$$

$$B_{ai} \sim \mathsf{N}(\sqrt{\mu} v_i u_a/\sqrt{n}, 1/p). \tag{9}$$

This model differs from (1), in that the graph observation A^G is replaced by the observation A which is equal to $\lambda vv^{\mathsf{T}}/n$, corrupted by Gaussian noise. This model generalizes so called 'rank-one deformations' of random matrices [Péc06, KY13, BGN11], as well as the $\mathbb{Z}_2$ synchronization model [ABBS14, Cuc15].

Our main motivation for introducing the Gaussian observation model is that it captures the large-degree behavior of the original graph model. The next result formalizes this intuition: its proof is an immediate generalization of the Lindeberg interpolation method of [DAM16].

Theorem 5. *Suppose $v \in \{\pm 1\}^n$ is uniformly random, and u is independent. We denote by $I(v; A^G, B)$ the* mutual information *of the latent random variables v and the observable data A^G, B. For all λ, μ: we have that:*

$$\lim_{d\to\infty} \limsup_{n\to\infty} \frac{1}{n} |I(v; A^G, B) - I(v; A, B)| = 0, \tag{10}$$

$$\lim_{d\to\infty} \limsup_{n\to\infty} \Big| \frac{1}{n} \frac{\mathrm{d} I(v; A^G, B)}{\mathrm{d}(\lambda^2)} - \frac{1}{4} \mathsf{MMSE}(v; A^G, B) \Big| = 0, \tag{11}$$

where $\mathsf{MMSE}(v; A^G, B) = n^{-2} \mathbb{E}\{\|vv^{\mathsf{T}} - \mathbb{E}\{vv^{\mathsf{T}} | A^G, B\}\|_F^2\}$.

For the Gaussian observation model (8), (9) we can establish a precise weak recovery threshold, which is the main technical novelty of this paper.

Theorem 6. *Suppose v is uniformly distributed in $\{\pm 1\}^n$ and we observe A, B as in* (8), (9). *Consider the limit $p, n \to \infty$ with $p/n \to 1/\gamma$.*

1. *If $\lambda^2 + \mu^2/\gamma < 1$, then for any estimator $\widehat{v} : (A, B) \mapsto \widehat{v}(A, B)$, with $\|\widehat{v}(A, B)\|_2 = \sqrt{n}$, we have $\limsup_{n\to\infty} |\langle \widehat{v}, v\rangle|/n = 0$.*
2. *If $\lambda^2 + \mu^2/\gamma > 1$, let $\widehat{v}(A, B)$ be normalized so that $\|\widehat{v}(A, B)\|_2 = \sqrt{n}$, and proportional the maximum eigenvector of the matrix $M(\xi_*)$, where*

$$M(\xi) = A + \frac{2\mu^2}{\lambda^2 \gamma^2 \xi} B^{\mathsf{T}} B + \frac{\xi}{2} \mathrm{I}_n \,, \tag{12}$$

and $\xi_* = \arg\min_{\xi > 0} \lambda_{\max}(M(\xi))$. Then, $\liminf_{n\to\infty} |\langle \widehat{v}, v\rangle|/n > 0$ in probability.

Theorem 4 is proved by using this threshold result, in conjunction with the universality Theorem 5.

3 Related work

The need to incorporate node information in graph clustering has been long recognized. To address the problem, diverse clustering methods have been introduced— e.g. those based on generative models [NC16, Hof03, ZVA10, YJCZ09, KL12, LM12, XKW+12, HL14, YML13], heuristic model free approaches [BVR17, ZLZ+16, GVB12, ZCY09, NAJ03, GFRS13, DV12, CZY11, SMJZ12, SZLP16], Bayesian methods [CB10, BC11] etc. [BCMM15] surveys other clustering methods for graphs with node and edge attributes. Semisupervised graph clustering [Pee12, EM12, ZMZ14], where labels are available for a few vertices are also somewhat related to our line of enquiry. The literature in this domain is quite vast and extremely diffuse, and thus we do not attempt to provide an exhaustive survey of all related attempts in this direction.

In terms of rigorous results, [AJC14, LMX15] introduced and analyzed a model with informative edges, but they make the strong and unrealistic requirement that the label of individual edges and each of their endpoints are uncorrelated and are only able to prove one side of their conjectured threshold. The papers [BVR17, ZLZ+16] –among others– rigorously analyze specific heuristics for clustering and provide some guarantees that ensure consistency. However, these results are not optimal. Moreover, it is possible that they only hold in the regime where using either the node covariates or the graph suffices for inference.

Several theoretical works [KMS16, MX16] analyze the performance of local algorithms in the semi-supervised setting, i.e., where the true labels are given for a small fraction of nodes. In particular [KMS16] establishes that for the two community sparse stochastic block model, correlated recovery is impossible given any vanishing proportion of nodes. Note that this is in stark contrast to Theorem 4 (and the Claim for the sparse graph model) above, which posits that given high dimensional covariate information actually shifts the information theoretic threshold for detection and weak recovery. The analysis in [KMS16, MX16] is also local in nature, while our algorithms and their analysis go well beyond the diameter of the graph.

4 Belief propagation: algorithm and cavity prediction

Recall the model (1), (2), where we are given the data (A^G, B) and our task is to infer the latent community labels v. From a Bayesian perspective, a principled approach computes posterior expectation with respect to the conditional distribution $\mathbb{P}(v, u|A^G, B) = \mathbb{P}(v, u, A^G, B)/\mathbb{P}(A^G, B)$. This is, however, not computationally tractable because it requires to marginalize over $v \in \{+1, -1\}^n$ and $u \in \mathbb{R}^p$. At this point, it becomes necessary to choose an approximate inference procedure, such as variational inference or mean field approximations [WJ+08]. In Bayes inference problem on locally-tree like graphs, belief propagation is optimal among local algorithms (see for instance [DM15] for an explanation of why this is the case).

The algorithm proceeds by computing, in an iterative fashion *vertex messages* η_i^t, m_a^t for $i \in [n]$, $a \in [p]$ and *edge messages* $\eta_{i\to j}^t$ for all pairs (i, j) that are connected in the graph G. For a vertex i of G, we denote its neighborhood in G by ∂i. Starting from an initialization $(\eta^{t_0}, m^{t_0})_{t_0=-1,0}$, we update the messages in the following *linear* fashion:

$$\eta_{i\to j}^{t+1} = \sqrt{\frac{\mu}{\gamma}}(B^{\mathsf{T}} m^t)_i - \frac{\mu}{\gamma}\eta_i^{t-1} + \frac{\lambda}{\sqrt{d}} \sum_{k\in\partial i\setminus j} \eta_{k\to i}^t - \frac{\lambda\sqrt{d}}{n} \sum_{k\in[n]} \eta_k^t, \tag{13}$$

$$\eta_i^{t+1} = \sqrt{\frac{\mu}{\gamma}}(B^{\mathsf{T}} m^t)_i - \frac{\mu}{\gamma}\eta_i^{t-1} + \frac{\lambda}{\sqrt{d}} \sum_{k\in\partial i} \eta_{k\to i}^t - \frac{\lambda\sqrt{d}}{n} \sum_{k\in[n]} \eta_k^t, \tag{14}$$

$$m^{t+1} = \sqrt{\frac{\mu}{\gamma}} B\eta^t - \mu m^{t-1}. \tag{15}$$

Here, and below, we will use $\eta^t = (\eta_i^t)_{i\in[n]}$, $m^t = (m_a^t)_{a\in[p]}$ to denote the vectors of vertex messages. After running the algorithm for some number of iterations $t_{\max}$, we return, as an estimate, the sign of the vertex messages $\eta_i^{t_{\max}}$, i.e.

$$\widehat{v}_i(A^G, B) = \operatorname{sgn}(\eta_i^{t_{\max}}). \tag{16}$$

These update equations have a number of intuitive features. First, in the case that $\mu = 0$, i.e. we have no covariate information, the edge messages become:

$$\eta_{i\to j}^{t+1} = \frac{\lambda}{\sqrt{d}} \sum_{k\in\partial i\setminus j} \eta_{k\to i}^t - \frac{\lambda\sqrt{d}}{n} \sum_{k\in[n]} \eta_k^t, \tag{17}$$

which corresponds closely to the spectral power method on the *nonbacktracking walk* matrix of G [KMM+13]. Conversely, when $\lambda = 0$, the updates equations on m^t, η^t correspond closely to the usual power iteration to compute singular vectors of B.

We obtain this algorithm from belief propagation using two approximations. First, we linearize the belief propagation update equations around a certain 'zero information' fixed point. Second, we use an 'approximate message passing' version of the belief propagation updates which results in the addition of the memory terms in Eqs. (13), (14), (15). The details of these approximations are quite standard and deferred to Appendix D. For a heuristic discussion, we refer the interested reader to the tutorials [Mon12, TKGM14] (for the Gaussian approximation) and the papers [DKMZ11, KMM$^+$13] (for the linearization procedure).

As with belief propagation, the behavior of this iterative algorithm, in the limit $p, n \to \infty$ can be tracked using a distributional recursion called *density evolution*.

Definition 1 (Density evolution)**.** *Let $(\bar{m}, U)$ and $(\bar{\eta}, V)$ be independent random vectors such that $U \sim \mathsf{N}(0,1)$, $V \sim \text{Uniform}(\{\pm 1\})$, $\bar{m}, \bar{\eta}$ have finite variance. Further assume that $(\bar{\eta}, V) \stackrel{\text{d}}{=} (-\bar{\eta}, -V)$ and $(\bar{m}, U) \stackrel{\text{d}}{=} (-\bar{m}, -U)$ (where $\stackrel{\text{d}}{=}$ denotes equality in distribution).*

We then define new random pairs $(\bar{m}', U')$ and $(\bar{\eta}', V')$, where $U' \sim \mathsf{N}(0,1)$, $V' \sim \text{Uniform}(\{\pm 1\})$, and $(\bar{\eta}, V) \stackrel{\text{d}}{=} (-\bar{\eta}, -V)$, $(\bar{m}, U) \stackrel{\text{d}}{=} (-\bar{m}, -U)$, via the following distributional equation

$$\bar{m}'\big|_{U'} \stackrel{\text{d}}{=} \mu \mathbb{E}\{V\bar{\eta}\} U' + \big(\mu \mathbb{E}\{\bar{\eta}^2\}\big)^{1/2} \zeta_1, \tag{18}$$

$$\begin{aligned} \bar{\eta}'\big|_{V'=+1} \stackrel{\text{d}}{=} \frac{\lambda}{\sqrt{d}} \Big[\sum_{k=1}^{k_+} \bar{\eta}_k\big|_+ + \sum_{k=1}^{k_-} \bar{\eta}_k\big|_- \Big] - \lambda\sqrt{d}\mathbb{E}\{\bar{\eta}\} \\ + \frac{\mu}{\gamma}\mathbb{E}\{U\bar{m}\} + \Big(\frac{\mu}{\gamma}\mathbb{E}\{\bar{m}^2\}\Big)^{1/2} \zeta_2. \end{aligned} \tag{19}$$

Here we use the notation $X|_Y \stackrel{\text{d}}{=} Z$ to mean that the conditional distribution of X given Y is the same as the (unconditional) distribution of Z. Notice that the distribution of $\bar{\eta}'\big|_{V'=-}$ is determined by the last equation using the symmetry property. Further $\bar{\eta}_k|_+$ and $\bar{\eta}_k|_-$ denote independent random variables distributed (respectively) as $\bar{\eta}|_{V=+}$ and $\bar{\eta}|_{V=-}$. Finally $k_+ \sim \mathsf{Poiss}(d/2 + \lambda\sqrt{d}/2)$, $k_- \sim \mathsf{Poiss}(d/2 - \lambda\sqrt{d}/2)$, $\zeta_1 \sim \mathsf{N}(0,1)$ and $\zeta_2 \sim \mathsf{N}(0,1)$ are mutually independent and independent from the previous random variables.

The density evolution *map, denoted by DE, is defined as the mapping from the law of $(\bar{\eta}, V, \bar{m}, U)$ to the law of $(\bar{\eta}', V', \bar{m}', U')$. With a slight abuse of notation, we will omit V, U, V', U', whose distribution is left unchanged and write*

$$(\bar{\eta}', \bar{m}') = \mathsf{DE}(\bar{\eta}, \bar{m})\,. \tag{20}$$

The following claim is the core of the cavity prediction. It states that the density evolution recursion faithfully describes the distribution of the iterates η^t, m^t.

Claim 7. *Let $(\bar{\eta}^0, V)$, $(\bar{m}^0, U)$ be random vectors satisfying the conditions of definition 1. Define the density evolution sequence $(\bar{\eta}^t, \bar{m}^t) = \mathsf{DE}^t(\bar{\eta}^0, \bar{m}^0)$, i.e. the result of iteratively applying the mapping DE t times.*

Consider the linear message passing algorithm of Eqs. (13) to (15), with the following initialization. We set $(m_r^0)_{r\in[p]}$ conditionally independent given u, with conditional distribution $m_r^0|_u \stackrel{\text{d}}{=} \bar{m}^0|_{U=\sqrt{p}u_r}$. Analogously, $\eta_i^0, \eta_{i\to j}^0$ are conditionally independent given v with $\eta_i^0|_v \stackrel{\text{d}}{=} \bar{\eta}^0|_{V=v_i}$, $\eta_{i\to j}^0|_v \stackrel{\text{d}}{=} \bar{\eta}^0|_{V=v_i}$. Finally $\eta_i^{-1} = \eta_{i\to j}^{-1} = m_r^{-1} = 0$ for all i, j, r.

Then, as $n, p \to \infty$ with $p/n \to 1/\gamma$, the following holds for uniformly random indices $i \in [n]$ and $a \in [p]$:

$$(m_a^t, u_a\sqrt{p}) \stackrel{\text{d}}{\Rightarrow} (\bar{m}^t, U) \tag{21}$$

$$(\eta_i^t, v_i) \stackrel{\text{d}}{\Rightarrow} (\bar{\eta}^t, V). \tag{22}$$

The following simple lemma shows the instability of the density evolution recursion.

Lemma 8. *Under the density evolution mapping, we obtain the random variables* $(\bar{\eta}', \bar{m}') = \mathsf{DE}(\bar{\eta}, \bar{m}'$ *Let* m *and* m' *denote the vector of the first two moments of* $(\bar{\eta}, V, \bar{m}, U)$ *and* $(\bar{\eta}', V', \bar{m}', U')$ *defined as follows:*

$$\mathsf{m} = (\mathbb{E}\{V\bar{\eta}\}, \mathbb{E}\{U\bar{m}\}, \mathbb{E}\{\bar{\eta}^2\}, \mathbb{E}\{\bar{m}^2\})\,, \tag{23}$$

and similarly for m'*. Then, for* $\|\mathsf{m}\|_2 \to 0$*, we have*

$$\mathsf{m}' = \begin{bmatrix} \lambda^2 & \mu/\gamma & 0 & 0 \\ \mu & 0 & 0 & 0 \\ 0 & 0 & \lambda^2 & \mu/\gamma \\ 0 & 0 & \mu & 0 \end{bmatrix} \mathsf{m} + O(\|\mathsf{m}\|^2) \tag{24}$$

In particular, the linearized map $\mathsf{m} \mapsto \mathsf{m}'$ *at* $\mathsf{m} = 0$ *has spectral radius larger than one if and only if* $\lambda^2 + \mu^2/\gamma > 1$*.*

The interpretation of the lemma is as follows. If we choose an initialization $(\bar{\eta}^0, V)$, $(\bar{m}^0, U)$ with $\bar{\eta}^0, \bar{m}^0$ *positively correlated* with V and U, then this correlation increases exponentially over time if and only if $\lambda^2 + \mu^2/\gamma > 1$[1]. In other words, a small initial correlation is amplified.

While we do not have an initialization that is positively correlated with the true labels, a random initialization η^0, m^0 has a random correlation with v, u of order $1/\sqrt{n}$. If $\lambda^2 + \mu^2/\gamma > 1$, this correlation is amplified over iterations, yielding a nontrivial reconstruction of v. On the other hand, if $\lambda^2 + \mu^2/\gamma < 1$ then this correlation is expected to remain small, indicating that the algorithm does not yield a useful estimate.

5 Proof overview

As mentioned above, a key step of our analysis is provided by Theorem 6, which establishes a weak recovery threshold for the Gaussian observation model of Eqs. (8), (9).

The proof proceeds in two steps: first, we prove that, for $\lambda^2 + \mu^2/\gamma < 1$ it is impossible to distinguish between data A, B generated according to this model, and data generated according to the null model $\mu = \lambda = 0$. Denoting by $\mathbb{P}_{\lambda,\mu}$ the law of data A, B, this is proved via a standard second moment argument. Namely, we bound the chi square distance uniformly in n, p

$$\chi^2(\mathbb{P}_{\lambda,\mu}, \mathbb{P}_{0,0}) \equiv \mathbb{E}_{0,0}\left\{\left(\frac{\mathrm{d}\mathbb{P}_{\lambda,\mu}}{\mathrm{d}\mathbb{P}_{0,0}}\right)^2\right\} - 1 \le C\,, \tag{25}$$

and then bound the total variation distance by the chi-squared distance $\|\mathbb{P}_{\lambda,\mu} - \mathbb{P}_{0,0}\|_{TV} \le 1 - (\chi^2(\mathbb{P}_{\lambda,\mu}, \mathbb{P}_{0,0}) + 1)^{-1}$. This in turn implies that no test can distinguish between the two hypotheses with probability approaching one as $n, p \to \infty$. The chi-squared bound also allows to show that weak recovery is impossible in the same regime.

In order to prove that weak recovery is possible for $\lambda^2 + \mu^2/\gamma > 1$, we consider the following optimization problem over $x \in \mathbb{R}^n$, $y \in \mathbb{R}^p$:

$$\text{maximize} \quad \langle x, Ax\rangle + b_*\langle x, By\rangle, \tag{26}$$

$$\text{subject to} \quad \|x\|_2 = \|y\|_2 = 1\,. \tag{27}$$

where $b_* = \frac{2\mu}{\lambda\gamma}$. Denoting solution of this problem by $(\widehat{x}, \widehat{y})$, we output the (soft) label estimates $\widehat{v} = \sqrt{n}\widehat{x}$. This definition turns out to be equivalent to the spectral algorithm in the statement of Theorem 6, and is therefore efficiently computable.

This optimization problem undergoes a phase transition exactly at the weak recovery threshold $\lambda^2 + \mu^2/\gamma = 1$, as stated below.

[1] Notice that both the messages variance $\mathbb{E}(\eta^2)$ and covariance with the ground truth $\mathbb{E}(\eta V)$ increase, but the normalized correlation (correlation divided by standard deviation) increases.

Lemma 9. *Denote by $T = T_{n,p}(A, B)$ the value of the optimization problem (26).*

(i) If $\lambda^2 + \frac{\mu^2}{\gamma} < 1$, then, almost surely

$$\lim_{n,p\to\infty} T_{n,p}(A, B) = 2\sqrt{1 + \frac{b_*^2\gamma}{4}} + b_* \,. \tag{28}$$

(ii) If $\lambda, \mu > 0$, and $\lambda^2 + \frac{\mu^2}{\gamma} > 1$ then there exists $\delta = \delta(\lambda, \mu) > 0$ such that, almost surely

$$\lim_{n,p\to\infty} T_{n,p}(A, B) = 2\sqrt{1 + \frac{b_*^2\gamma}{4}} + b_* + \delta(\lambda, \mu) \,. \tag{29}$$

(iii) Further, define

$$\tilde{T}_{n,p}(\tilde{\delta}; A, B) = \sup_{\|x\|=\|y\|=1, |\langle x,v\rangle|<\tilde{\delta}\sqrt{n}} \Big[\langle x, Ax\rangle + b_*\langle x, By\rangle\Big].$$

Then for each $\delta > 0$, there exists $\tilde{\delta} > 0$ sufficiently small, such that, amlost surely

$$\lim_{n,p\to\infty} \tilde{T}_{n,p}(\tilde{\delta}; A, B) < 2\sqrt{1 + \frac{b_*^2\gamma}{4}} + b_* + \frac{\delta}{2} \,. \tag{30}$$

The first two points imply that $T_{n,p}(A, B)$ provide a statistic to distinguish between $\mathbb{P}_{0,0}$ and $\mathbb{P}_{\lambda,\mu}$ with probability of error that vanishes as $n, p \to \infty$ if $\lambda^2 + \mu^2/\gamma > 1$. The third point (in conjunction with the second one) guarantees that the maximizer $\widehat{x}$ is positively correlated with v, and hence implies weak recovery.

In fact, we prove a stronger result that provides an asymptotic expression for the value $T_{n,p}(A, B)$ for all λ, μ. We obtain the above phase-transition result by specializing the resulting formula in the two regimes $\lambda^2 + \mu^2/\gamma < 1$ and $\lambda^2 + \mu^2/\gamma > 1$. We prove this asymptotic formula by Gaussian process comparison, using Sudakov-Fernique inequality. Namely, we compare the Gaussian process appearing in the optimization problem of Eq. (26) with the following ones:

$$\mathcal{X}_1(x, y) = \frac{\lambda}{n}\langle x, v_0\rangle^2 + \langle x, \widetilde{g}_x\rangle + b_*\sqrt{\frac{\mu}{n}}\langle x, v_0\rangle\langle y, u_0\rangle + \langle y, \widetilde{g}_y\rangle \,, \tag{31}$$

$$\mathcal{X}_2(x, y) = \frac{\lambda}{n}\langle x, v_0\rangle^2 + \frac{1}{2}\langle x, \widetilde{W}_x x\rangle + b_*\sqrt{\frac{\mu}{n}}\langle x, v_0\rangle\langle y, u_0\rangle + \frac{1}{2}\langle y, \widetilde{W}_y y\rangle \,, \tag{32}$$

where $\widetilde{g}_x$, $\widetilde{g}_y$ are isotropic Gaussian vectors, with suitably chosen variances, and $\widetilde{W}_x$, $\widetilde{W}_y$ are GOE matrices, again with properly chosen variances. We prove that $\max_{x,y} \mathcal{X}_1(x, y)$ yields an upper bound on $T_{n,p}(A, B)$, and $\max_{x,y} \mathcal{X}_2(x, y)$ yields a lower bound on the same quantity.

Note that maximizing the first process $\mathcal{X}_1(x, y)$ essentially reduces to solving a separable problem over the coordinates of x and y and hence to an explicit expression. On the other hand, maximizing the second process leads (after decoupling the term $\langle x, v_0\rangle\langle y, u_0\rangle$) to two separate problems, one for the vector x, and the other for y. Each of the two problems reduce to finding the maximum eigenvector of a rank-one deformation of a GOE matrix, a problem for which we can leverage on significant amount of information from random matrix theory. The resulting upper and lower bound coincide asymptotically.

As is often the case with Gaussian comparison arguments, the proof is remarkably compact, and somewhat surprising (it is unclear a priori that the two bounds should coincide asymptotically). While upper bounds by processes of the type of $\mathcal{X}_1(x, y)$ are quite common in random matrix theory, we think that the lower bound by $\mathcal{X}_2(x, y)$ (which is crucial for proving our main theorem) is novel and might have interesting generalizations.

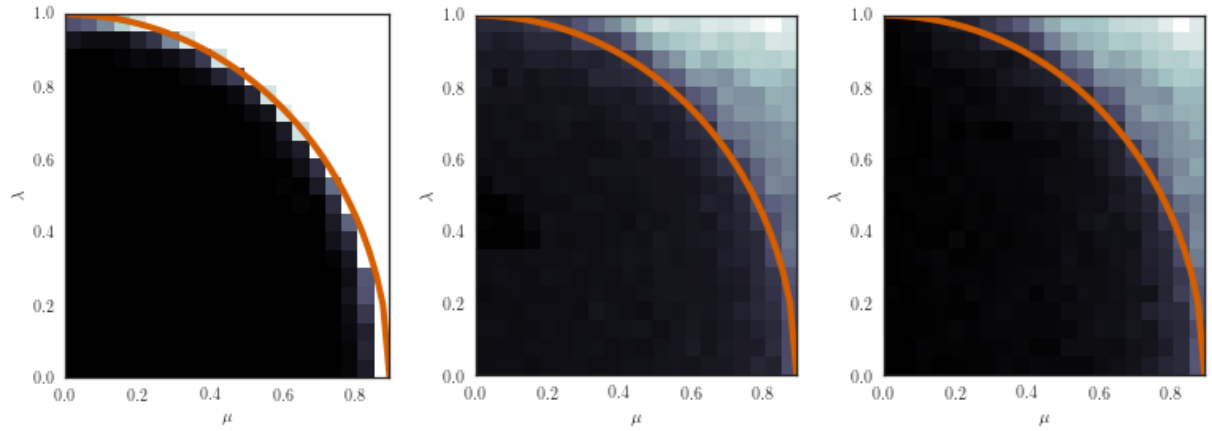

Figure 1: (Left) Empirical probability of rejecting the null (lighter is higher) using BP test. (Middle) Mean overlap $|\langle \widehat{v}^{\mathsf{BP}}, v\rangle/n|$ and (Right) mean covariate overlap $|\langle \widehat{u}^{\mathsf{BP}}, u\rangle|$ attained by BP estimate.

6 Experiments

We demonstrate the efficacy of the full belief propagation algorithm, restated below:

$$\eta_i^{t+1} = \sqrt{\frac{\mu}{\gamma}} \sum_{q\in[p]} B_{qi} m_q^t - \frac{\mu}{\gamma}\Big(\sum_{q\in[p]} \frac{B_{qi}^2}{\tau_q^t}\Big) \tanh(\eta_i^{t-1}) + \sum_{k\in\partial i} f(\eta_{k\to i}^t;\rho) - \sum_{k\in[n]} f(\eta_k^t;\rho_n)\,, \tag{33}$$

$$\eta_{i\to j}^{t+1} = \sqrt{\frac{\mu}{\gamma}} \sum_{q\in[p]} B_{qi} m_q^t - \frac{\mu}{\gamma}\Big(\sum_{q\in[p]} \frac{B_{qi}^2}{\tau_q^t}\Big) \tanh(\eta_i^{t-1}) + \sum_{k\in\partial i\setminus j} f(\eta_{k\to i}^t;\rho) - \sum_{k\in[n]} f(\eta_k^t;\rho_n)\,, \tag{34}$$

$$m_q^{t+1} = \frac{\sqrt{\mu/\gamma}}{\tau_q^{t+1}} \sum_{j\in[n]} B_{qj} \tanh(\eta_j^t) - \frac{\mu}{\gamma\tau_q^{t+1}}\Big(\sum_{j\in[n]} B_{qj}^2 \mathrm{sech}^2(\eta_j^t)\Big) m_q^{t-1} \tag{35}$$

$$\tau_q^{t+1} = \left(1+\mu - \frac{\mu}{\gamma}\sum_{j\in[n]} B_{qj}^2 \mathrm{sech}^2(\eta_j^t)\right)^{-1}. \tag{36}$$

Here the function $f(;\rho)$ and the parameters ρ, ρ_n are defined as:

$$f(z;\rho) \equiv \frac{1}{2}\log\Big(\frac{\cosh(z+\rho)}{\cosh(z-\rho)}\Big)\,, \tag{37}$$

$$\rho \equiv \tanh^{-1}(\lambda/\sqrt{d})\,, \tag{38}$$

$$\rho_n \equiv \tanh^{-1}\Big(\frac{\lambda\sqrt{d}}{n-d}\Big). \tag{39}$$

We refer the reader to Appendix D for a derivation of the algorithm. As demonstrated in Appendix D, the BP algorithm in Section 4 is obtained by linearizing the above in η.

In our experiments, we perform 100 Monte Carlo runs of the following process:

1. Sample A^G, B from $\mathbb{P}_{\lambda,\mu}$ with $n = 800, p = 1000, d = 5$.
2. Run BP algorithm for $T = 50$ iterations with random initialization $\eta_i^0, \eta_i^{-1}, m_a^0, m_a^{-1} \sim_{\text{iid}} \mathsf{N}(0, 0.01)$. yielding vertex and covariate iterates $\eta^T \in \mathbb{R}^n$, $m^T \in \mathbb{R}^p$.
3. Reject the null hypothesis if $\|\eta^T\|_2 > \|\eta^0\|_2$, else accept the null.

4. Return estimates $\widehat{v}_i^{\mathsf{BP}} = \mathrm{sgn}(\eta_i^T)$, $\widehat{u}_a^{\mathsf{BP}} = m_a^T / \left\|m^T\right\|_2$.

Figure 1 (left) shows empirical probabilities of rejecting the null for $(\lambda, \mu) \in [0,1] \times [0, \sqrt{\gamma}]$. The next two plots display the mean overlap $|\langle \widehat{v}^{\mathsf{BP}}, v\rangle/n|$ and $\langle \widehat{u}^{\mathsf{BP}}, u\rangle / \|u\|$ achieved by the BP estimates (lighter is higher overlap). Below the theoretical curve (red) of $\lambda^2 + \mu^2/\gamma = 1$, the null hypothesis is accepted and the estimates show negligible correlation with the truth. These results are in excellent agreement with our theory.

Acknowledgements

A.M. was partially supported by grants NSF DMS-1613091, NSF CCF-1714305 and NSF IIS-1741162. E.M was partially supported by grants NSF DMS-1737944 and ONR N00014-17-1-2598. Y.D would like to acknowledge Nilesh Tripuraneni for discussions about this paper.

References

[Abb17] Emmanuel Abbe, *Community detection and stochastic block models: recent developments*, arXiv preprint arXiv:1703.10146 (2017).

[ABBS14] Emmanuel Abbe, Afonso S Bandeira, Annina Bracher, and Amit Singer, *Decoding binary node labels from censored edge measurements: Phase transition and efficient recovery*, IEEE Transactions on Network Science and Engineering **1** (2014), no. 1, 10–22.

[AG05] Lada A Adamic and Natalie Glance, *The political blogosphere and the 2004 us election: divided they blog*, Proceedings of the 3rd international workshop on Link discovery, ACM, 2005, pp. 36–43.

[AJC14] Christopher Aicher, Abigail Z Jacobs, and Aaron Clauset, *Learning latent block structure in weighted networks*, Journal of Complex Networks **3** (2014), no. 2, 221–248.

[BBAP05] Jinho Baik, Gérard Ben Arous, and Sandrine Péché, *Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices*, Annals of Probability (2005), 1643–1697.

[BC11] Ramnath Balasubramanyan and William W Cohen, *Block-lda: Jointly modeling entity-annotated text and entity-entity links*, Proceedings of the 2011 SIAM International Conference on Data Mining, SIAM, 2011, pp. 450–461.

[BCMM15] Cécile Bothorel, Juan David Cruz, Matteo Magnani, and Barbora Micenkova, *Clustering attributed graphs: models, measures and methods*, Network Science **3** (2015), no. 3, 408–444.

[BGN11] Florent Benaych-Georges and Raj Rao Nadakuditi, *The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices*, Advances in Mathematics **227** (2011), no. 1, 494–521.

[BVR17] Norbert Binkiewicz, Joshua T Vogelstein, and Karl Rohe, *Covariate-assisted spectral clustering*, Biometrika **104** (2017), no. 2, 361–377.

[CB10] Jonathan Chang and David M Blei, *Hierarchical relational models for document networks*, The Annals of Applied Statistics (2010), 124–150.

[CDMF+09] Mireille Capitaine, Catherine Donati-Martin, Delphine Féral, et al., *The largest eigenvalues of finite rank deformation of large wigner matrices: convergence and nonuniversality of the fluctuations*, The Annals of Probability **37** (2009), no. 1, 1–47.

[CDPF+17] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq, *Algorithmic decision making and the cost of fairness*, Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2017, pp. 797–806.

[CMS11] Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate, *Differentially private empirical risk minimization*, Journal of Machine Learning Research **12** (2011), no. Mar, 1069–1109.

[Cuc15] Mihai Cucuringu, *Synchronization over z 2 and community detection in signed multiplex networks with constraints*, Journal of Complex Networks **3** (2015), no. 3, 469–506.

[CZY11] Hong Cheng, Yang Zhou, and Jeffrey Xu Yu, *Clustering large attributed graphs: A balance between structural and attribute similarities*, ACM Transactions on Knowledge Discovery from Data (TKDD) **5** (2011), no. 2, 12.

[DAM16] Yash Deshpande, Emmanuel Abbe, and Andrea Montanari, *Asymptotic mutual information for the balanced binary stochastic block model*, Information and Inference: A Journal of the IMA **6** (2016), no. 2, 125–170.

[DKMZ11] Aurelien Decelle, Florent Krzakala, Cristopher Moore, and Lenka Zdeborová, *Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications*, Physical Review E **84** (2011), no. 6, 066106.

[DM15] Yash Deshpande and Andrea Montanari, *Finding Hidden Cliques of Size N/e in Nearly Linear Time*, Foundations of Computational Mathematics **15** (2015), no. 4, 1069–1128.

[DV12] TA Dang and Emmanuel Viennet, *Community detection based on structural and attribute similarities*, International conference on digital society (icds), 2012, pp. 7–12.

[EM12] Eric Eaton and Rachael Mansbach, *A spin-glass model for semi-supervised community detection.*, AAAI, 2012, pp. 900–906.

[GFRS13] Stephan Gunnemann, Ines Farber, Sebastian Raubach, and Thomas Seidl, *Spectral subspace clustering for graphs with feature vectors*, Data Mining (ICDM), 2013 IEEE 13th International Conference on, IEEE, 2013, pp. 231–240.

[GSV05] D. Guo, S. Shamai, and S. Verdú, *Mutual information and minimum mean-square error in gaussian channels*, IEEE Trans. Inform. Theory **51** (2005), 1261–1282.

[GVB12] Jaume Gibert, Ernest Valveny, and Horst Bunke, *Graph embedding in vector spaces by node attribute statistics*, Pattern Recognition **45** (2012), no. 9, 3072–3083.

[HL14] Tuan-Anh Hoang and Ee-Peng Lim, *On joint modeling of topical communities and personal interest in microblogs*, International Conference on Social Informatics, Springer, 2014, pp. 1–16.

[HLL83] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt, *Stochastic blockmodels: First steps*, Social networks **5** (1983), no. 2, 109–137.

[Hof03] Peter D Hoff, *Random effects models for network data*, na, 2003.

[JL04] Iain M Johnstone and Arthur Yu Lu, *Sparse principal components analysis*, Unpublished manuscript (2004).

[KGB+12] Virendra Kumar, Yuhua Gu, Satrajit Basu, Anders Berglund, Steven A Eschrich, Matthew B Schabath, Kenneth Forster, Hugo JWL Aerts, Andre Dekker, David Fenstermacher, et al., *Radiomics: the process and the challenges*, Magnetic resonance imaging **30** (2012), no. 9, 1234–1248.

[KL12] Myunghwan Kim and Jure Leskovec, *Latent multi-group membership graph model*, arXiv preprint arXiv:1205.4546 (2012).

[KMM+13] Florent Krzakala, Cristopher Moore, Elchanan Mossel, Joe Neeman, Allan Sly, Lenka Zdeborová, and Pan Zhang, *Spectral redemption in clustering sparse networks*, Proceedings of the National Academy of Sciences **110** (2013), no. 52, 20935–20940.

[KMS16] Varun Kanade, Elchanan Mossel, and Tselil Schramm, *Global and local information in clustering labeled block models*, IEEE Transactions on Information Theory **62** (2016), no. 10, 5906–5917.

[KY13] Antti Knowles and Jun Yin, *The isotropic semicircle law and deformation of wigner matrices*, Communications on Pure and Applied Mathematics **66** (2013), no. 11, 1663–1749.

[LM12] Jure Leskovec and Julian J Mcauley, *Learning to discover social circles in ego networks*, Advances in neural information processing systems, 2012, pp. 539–547.

[LMX15] Marc Lelarge, Laurent Massoulié, and Jiaming Xu, *Reconstruction in the labelled stochastic block model*, IEEE Transactions on Network Science and Engineering **2** (2015), no. 4, 152–163.

[Mas14] Laurent Massoulié, *Community detection thresholds and the weak ramanujan property*, Proceedings of the forty-sixth annual ACM symposium on Theory of computing, ACM, 2014, pp. 694–703.

[MM09] M. Mézard and A. Montanari, *Information, Physics and Computation*, Oxford, 2009.

[MNS13] Elchanan Mossel, Joe Neeman, and Allan Sly, *A proof of the block model threshold conjecture*, Combinatorica (2013), 1–44.

[MNS15] ———, *Reconstruction and estimation in the planted partition model*, Probability Theory and Related Fields **162** (2015), no. 3-4, 431–461.

[Mon12] A. Montanari, *Graphical Models Concepts in Compressed Sensing*, Compressed Sensing: Theory and Applications (Y.C. Eldar and G. Kutyniok, eds.), Cambridge University Press, 2012.

[MRZ15] Andrea Montanari, Daniel Reichman, and Ofer Zeitouni, *On the limitation of spectral methods: From the gaussian hidden clique problem to rank-one perturbations of gaussian tensors*, Advances in Neural Information Processing Systems, 2015, pp. 217–225.

[MX16] Elchanan Mossel and Jiaming Xu, *Local algorithms for block models with side information*, Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science, ACM, 2016, pp. 71–80.

[NAJ03] Jennifer Neville, Micah Adler, and David Jensen, *Clustering relational data using attribute and link information*, Proceedings of the text mining and link analysis workshop, 18th international joint conference on artificial intelligence, San Francisco, CA: Morgan Kaufmann Publishers, 2003, pp. 9–15.

[NC16] Mark EJ Newman and Aaron Clauset, *Structure and inference in annotated networks*, Nature Communications **7** (2016), 11863.

[NJW02] Andrew Y Ng, Michael I Jordan, and Yair Weiss, *On spectral clustering: Analysis and an algorithm*, Advances in neural information processing systems, 2002, pp. 849–856.

[OMH+13] Alexei Onatski, Marcelo J Moreira, Marc Hallin, et al., *Asymptotic power of sphericity tests for high-dimensional data*, The Annals of Statistics **41** (2013), no. 3, 1204–1231.

[Pau07] Debashis Paul, *Asymptotics of sample eigenstructure for a large dimensional spiked covariance model*, Statistica Sinica **17** (2007), no. 4, 1617.

[Péc06] Sandrine Péché, *The largest eigenvalue of small rank perturbations of hermitian random matrices*, Probability Theory and Related Fields **134** (2006), no. 1, 127–173.

[Pee12] Leto Peel, *Supervised blockmodelling*, arXiv preprint arXiv:1209.5561 (2012).

[SMJZ12] Arlei Silva, Wagner Meira Jr, and Mohammed J Zaki, *Mining attribute-structure correlated patterns in large attributed graphs*, Proceedings of the VLDB Endowment **5** (2012), no. 5, 466–477.

[SZLP16] Laura M Smith, Linhong Zhu, Kristina Lerman, and Allon G Percus, *Partitioning networks with node attributes by compressing information flow*, ACM Transactions on Knowledge Discovery from Data (TKDD) **11** (2016), no. 2, 15.

[TKGM14] Eric W Tramel, Santhosh Kumar, Andrei Giurgiu, and Andrea Montanari, *Statistical estimation: From denoising to sparse regression and hidden cliques*, arXiv preprint arXiv:1409.5557 (2014).

[VL07] Ulrike Von Luxburg, *A tutorial on spectral clustering*, Statistics and computing **17** (2007), no. 4, 395–416.

[WJ+08] Martin J Wainwright, Michael I Jordan, et al., *Graphical models, exponential families, and variational inference*, Foundations and Trends® in Machine Learning **1** (2008), no. 1–2, 1–305.

[XKW+12] Zhiqiang Xu, Yiping Ke, Yi Wang, Hong Cheng, and James Cheng, *A model-based approach to attributed graph clustering*, Proceedings of the 2012 ACM SIGMOD international conference on management of data, ACM, 2012, pp. 505–516.

[YJCZ09] Tianbao Yang, Rong Jin, Yun Chi, and Shenghuo Zhu, *Combining link and content for community detection: a discriminative approach*, Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2009, pp. 927–936.

[YML13] Jaewon Yang, Julian McAuley, and Jure Leskovec, *Community detection in networks with node attributes*, Data Mining (ICDM), 2013 IEEE 13th international conference on, IEEE, 2013, pp. 1151–1156.

[ZCY09] Yang Zhou, Hong Cheng, and Jeffrey Xu Yu, *Graph clustering based on structural/attribute similarities*, Proceedings of the VLDB Endowment **2** (2009), no. 1, 718–729.

[ZLZ+16] Yuan Zhang, Elizaveta Levina, Ji Zhu, et al., *Community detection in networks with node features*, Electronic Journal of Statistics **10** (2016), no. 2, 3153–3178.

[ZMZ14] Pan Zhang, Cristopher Moore, and Lenka Zdeborová, *Phase transitions in semisupervised clustering of sparse networks*, Physical Review E **90** (2014), no. 5, 052802.

[ZVA10] Hugo Zanghi, Stevenn Volant, and Christophe Ambroise, *Clustering based on random graph model embedding vertex features*, Pattern Recognition Letters **31** (2010), no. 9, 830–836.

A Proof of Theorem 6

We establish Theorem 6 in this section. First, we introduce the notion of contiguity of measures

Definition 2. *Let $\{P_n\}$ and $\{Q_n\}$ be two sequences of probability measures on the measurable space $(\Omega_n, \mathcal{F}_n)$. We say that P_n is contiguous to Q_n if for any sequence of events A_n with $Q_n(A_n) \to 0$, $P_n(A_n) \to 0$.*

It is standard that for two sequences of probability measures P_n and Q_n with P_n contiguous to Q_n, $\limsup_{n\to\infty} d_{\mathrm{TV}}(P_n, Q_n) < 1$. The following lemma provides sufficient conditions for establishing contiguity of two sequence of probability measures.

Lemma 10 (see e.g. [MRZ15]). *Let P_n and Q_n be two sequences of probability measures on $(\Omega_n, \mathcal{F}_n)$. Then P_n is contiguous to Q_n if*

$$\mathbb{E}_{Q_n}\Big[\Big(\frac{\mathrm{d}P_n}{\mathrm{d}Q_n}\Big)^2\Big]$$

exists and remains bounded as $n \to \infty$.

Our next result establishes that asymptotically error-free detection is impossible below the conjectured detection boundary.

Lemma 11. *Let $\lambda, \mu > 0$ with $\lambda^2 + \frac{\mu^2}{\gamma} < 1$. Then $\mathbb{P}_{\lambda,\mu}$ is contiguous to $\mathbb{P}_{0,0}$.*

To establish that consistent detection is possible above this boundary, we need the following lemma. Recall the matrices A, B from the Gaussian model (8), (9).

Lemma 12. *Let $b_* = \frac{2\mu}{\lambda\gamma}$. Define*

$$T = \sup_{\|x\|=\|y\|=1} \Big[\langle x, Ax\rangle + b_*\langle x, By\rangle\Big].$$

(i) Under $\mathbb{P}_{0,0}$, as $n, p \to \infty$, $T \to 2\sqrt{1 + \frac{b_*^2\gamma}{4}} + b_*$ almost surely.

(ii) Let $\lambda, \mu > 0$, $\varepsilon > 0$, with $\lambda^2 + \frac{\mu^2}{\gamma} > 1 + \varepsilon$. Then as $n, p \to \infty$,

$$\mathbb{P}_{\lambda,\mu}\Big(T > 2\sqrt{1 + \frac{b_*^2\gamma}{4}} + b_* + \delta\Big) \to 1,$$

where $\delta := \delta(\varepsilon) > 0$.

(iii) Further, define

$$\tilde{T}(\tilde{\delta}) = \sup_{\|x\|=\|y\|=1, 0<\langle x,v\rangle<\tilde{\delta}\sqrt{n}} \Big[\langle x, Ax\rangle + b_*\langle x, By\rangle\Big].$$

Then for each $\delta > 0$, there exists $\tilde{\delta} > 0$ sufficiently small, such that as $n, p \to \infty$,

$$\mathbb{P}_{\lambda,\mu}\Big(\tilde{T}(\tilde{\delta}) < 2\sqrt{1 + \frac{b_*^2\gamma}{4}} + b_* + \frac{\delta}{2}\Big) \to 1.$$

We defer the proofs of Lemma 11 and Lemma 12 to Sections A.1 and Section A.5 respectively, and complete the proof of Theorem 6, armed with these results.

Proof of Theorem 6. The proof is comparatively straightforward, once we have Lemma 11 and 12. Note that Lemma 11 immediately implies that $\mathbb{P}_{\lambda,\mu}$ is contiguous to $\mathbb{P}_{0,0}$ for $\lambda^2 + \frac{\lambda^2}{\gamma} < 1$.

Next, let $\lambda, \mu > 0$ such that $\lambda^2 + \frac{\mu^2}{\gamma} > 1 + \varepsilon$ for some $\varepsilon > 0$. In this case, consider the test which rejects the null hypothesis H_0 if $T > 2\sqrt{1 + \frac{b_*^2\gamma}{4}} + b_* + \delta$. Lemma 12 immediately implies that the Type I and II errors of this test vanish in this setting.

Finally, we prove that weak recovery is possible whenever $\lambda^2 + \frac{\mu^2}{\gamma} > 1$. To this end, let $(\hat{x}, \hat{y})$ be the maximizer of $\langle x, Ax\rangle + b_*\langle y, Bx\rangle$, with $\|x\| = \|y\| = 1$. Combining parts (ii) and (iii) of Lemma 12, we conclude that $\hat{x}$ achieves weak recovery of the community assignment vector. □

A.1 Proof of Lemma 11

Fix $\lambda, \mu > 0$ satisfying $\lambda^2 + \frac{\mu^2}{\gamma} < 1$. We start with the likelihood,

$$L(u,v) = \frac{\mathrm{d}\mathbb{P}_{\lambda,\mu}}{\mathrm{d}\mathbb{P}_{0,0}} = L_1(u,v)L_2(u,v),$$

$$L_1(u,v) = \exp\Big[\frac{\lambda}{2}\langle A, vv^T\rangle - \frac{\lambda^2 n}{4}\Big]. \tag{40}$$

$$L_2(u,v) = \exp\Big[p\sqrt{\frac{\mu}{n}}\langle B, uv^T\rangle - \frac{\mu p}{2}\|u\|^2\Big]. \tag{41}$$

We denote the prior joint distribution of (u,v) as π, and set

$$L_\pi = \mathbb{E}_{(u,v)\sim\pi}\Big[L(u,v)\Big].$$

To establish contiguity, we bound the second moment of L_π under the null hypothesis, and appeal to Lemma 10. In particular, we denote $\mathbb{E}_0[\cdot]$ to be the expectation operator under the distribution $P_{(0,0)}$ and compute

$$\mathbb{E}_0[L_\pi^2] = \mathbb{E}_0[\mathbb{E}_{(u_1,v_1),(u_2,v_2)}\Big[L(u_1,v_1)L(u_2,v_2)\Big]\Big] = \mathbb{E}_{(u_1,v_1),(u_2,v_2)}\Big[\mathbb{E}_0\Big[L(u_1,v_1)L(u_2,v_2)\Big]\Big],$$

where $(u_1,v_1),(u_2,v_2)$ are i.i.d. draws from the prior π, and the last equality follows by Fubini's theorem. We have, using (40) and (41),

$$\begin{aligned}&L(u_1,v_1)L(u_2,v_2)\\ &= \exp\Big[-\frac{\lambda^2 n}{2} - \frac{\mu p}{2n}\Big(\|u_1\|^2 + \|u_2\|^2\Big) + \frac{\lambda}{2}\Big\langle A, v_1v_1^T + v_2v_2^T\Big\rangle + p\sqrt{\frac{\mu}{n}}\Big\langle B, u_1v_1^T + u_2v_2^T\Big\rangle\Big].\end{aligned}$$

Taking expectation under $\mathbb{E}_0[\cdot]$, upon simplification, we obtain,

$$\mathbb{E}_0[L_\pi^2] = \mathbb{E}_{(u_1,v_1),(u_2,v_2)}\Big[\exp\Big[\frac{\lambda^2}{2n}\langle v_1,v_2\rangle^2 + \frac{\mu p}{n}\langle u_1,u_2\rangle\langle v_1,v_2\rangle\Big]\Big] \tag{42}$$

$$= \mathbb{E}_{(u_1,v_1),(u_2,v_2)}\Big[\exp\Big[n\Big(\frac{\lambda^2}{2}\Big(\frac{\langle v_1,v_2\rangle}{n}\Big)^2 + \frac{\mu}{\gamma}\langle u_1,u_2\rangle\frac{\langle v_1,v_2\rangle}{n}\Big)\Big]\Big] \tag{43}$$

$$= \mathbb{E}\Big[\exp\Big[n\Big(\frac{\lambda^2}{2}X^2 + \frac{\mu}{\gamma}XY\Big)\Big]\Big] \tag{44}$$

Here that $X, Y \in [-1,+1]$ are independent, with X distributed as the normalized sum of n Radamacher random variables, and Y as the first coordinate of a uniform vector on the unit sphere. In particular, defining $h(s) = -((1+s)/2)\log((1+s)) - ((1-s)/2)\log((1-s))$, and denoting by f_Y the density of Y, we have, for $s \in (2/n)\mathbb{Z}$

$$\mathbb{P}\big(X = s\big) = \frac{1}{2^n}\binom{n}{n(1+s/2)} \tag{45}$$

$$\le \frac{C}{n^{1/2}}\, e^{nh(s)} \tag{46}$$

$$f_Y(y) = \frac{\Gamma(p/2)}{\Gamma((p-1)/2)\Gamma(1/2)}(1-y^2)^{(p-3)/2} \tag{47}$$

$$\le C\sqrt{n}(1-y^2)^{p/2}\,. \tag{48}$$

Approximating sums by integrals, and using $h(s) \le -s^2/2$, we get

$$\mathbb{E}_0[L_\pi^2] \le Cn\int_{[-1,1]^2}\exp\Big\{n\Big[\frac{\lambda^2}{2}s^2 + \frac{\mu}{\gamma}sy + h(s) + \frac{1}{2\gamma}\log(1-y^2)\Big\}\mathrm{d}s\mathrm{d}y \tag{49}$$

$$\le Cn\int_{\mathbb{R}^2}\exp\Big\{n\Big[\frac{\lambda^2}{2}s^2 + \frac{\mu}{\gamma}sy - \frac{s^2}{2} - \frac{y^2}{2\gamma}\Big]\Big\}\mathrm{d}s\mathrm{d}y \le C'\,. \tag{50}$$

The last step holds for $\lambda^2 + \mu^2/\gamma < 1$.

Next, we turn to the proof of Lemma 12. This is the main technical contribution of this paper, and uses a novel Gaussian process comparison argument based on Sudakov-Fernique comparison.

A.2 A Gaussian process comparison result

Let $Z \sim \mathbb{R}^{p\times n}$ and $W \sim \mathbb{R}^{n\times n}$ denote random matrices with independent entries as follows.

$$W_{ij} \sim \begin{cases} \mathsf{N}(0,\rho/n) & \text{if } i<j \\ \mathsf{N}(0,2\rho/n) & \text{if } i=j \end{cases} \tag{51}$$

$$\text{where } W_{ij} = W_{ji},$$

$$Z_{ai} \sim \mathsf{N}(0,\tau/p). \tag{52}$$

For an integer $N>0$, we let $\mathbb{S}^N$ denote the sphere of radius $\sqrt{N}$ in N dimensions, i.e. $\mathbb{S}^N = \{x \in \mathbb{R}^N : \|x\|_2^2 = N\}$. Furthermore let $u_0 \in \mathbb{S}^p$ and $v_0 \in \{\pm 1\}^n$ be fixed vectors. We denote the standard inner product between vectors $x, y \in \mathbb{R}^N$ as $\langle x, y\rangle = \sum_i x_i y_i$. The normalized version will be useful as well: we define $\langle x, y\rangle_N \equiv \sum_i x_i y_i/N$.

We are interested in characterizing the behavior of the following optimization problem in the limit high-dimensional limit $p, n \to \infty$ with constant aspect ratio $n/p = \gamma \in (0,\infty)$.

$$\mathsf{OPT}(\lambda,\mu,b) \equiv \frac{1}{n}\mathbb{E}\max_{(x,y)\in\mathbb{S}^n\times\mathbb{S}^p}\Big[\big(\frac{\lambda}{n}\langle x,v_0\rangle^2 + \langle x, Wx\rangle\big) + b\,\big(\sqrt{\frac{\mu}{np}}\langle x,v_0\rangle\,\langle y,u_0\rangle + \langle y, Zx\rangle\big)\Big].$$

We now introduce two different comparison processes which give upper and lower bounds to $\mathsf{OPT}(\lambda,\mu,b)$. Their asymptotic values will coincide in the high dimensional limit $n,p\to\infty$ with $n/p=\gamma$. Let g_x, g_y, W_x and W_y be:

$$g_x \sim \mathsf{N}(0,(4\rho + b^2\tau)\mathrm{I}_n) \tag{53}$$

$$g_y \sim \mathsf{N}(0, b^2\tau n/p\mathrm{I}_p), \tag{54}$$

$$(W_x)_{ij} \sim \begin{cases} \mathsf{N}(0,(4\rho+b^2\tau)/n) & \text{if } i<j \\ \mathsf{N}(0,2(4\rho+b^2\tau)/n) & \text{if } i=j \end{cases} \tag{55}$$

$$(W_y)_{ij} \sim \begin{cases} \mathsf{N}(0,b^2\tau n/p^2) & \text{if } i<j \\ \mathsf{N}(0,2b^2\tau n/p^2) & \text{if } i=j \end{cases} \tag{56}$$

Proposition 13. *We have*

$$\begin{aligned} \mathsf{OPT}(\lambda,\mu,b) &\le \frac{1}{n}\mathbb{E}\max_{(x,y)\in\mathbb{S}^n\times\mathbb{S}^p}\frac{\lambda}{n}\langle x,v_0\rangle^2 + \langle x,g_x\rangle + b\sqrt{\frac{\mu}{np}}\langle x,v_0\rangle\langle y,u_0\rangle + \langle y,g_y\rangle \\ \mathsf{OPT}(\lambda,\mu,b) &\ge \frac{1}{n}\mathbb{E}\max_{(x,y)\in\mathbb{S}^n\times\mathbb{S}^p}\frac{\lambda}{n}\langle x,v_0\rangle^2 + \frac{1}{2}\langle x,W_x x\rangle + b\sqrt{\frac{\mu}{np}}\langle x,v_0\rangle\langle y,u_0\rangle + \frac{1}{2}\langle y,W_y y\rangle \end{aligned} \tag{57}$$

Proof. The proof is via Sudakov-Fernique inequality. First we compute the distances induced by the three processes. For any pair $(x,y),(x',y')$:

$$\begin{aligned} &\frac{1}{4n}\big(\mathbb{E}\{(\langle x,Wx\rangle + b\langle y,Zx\rangle - \langle x',Wx'\rangle - b\langle y',Zx'\rangle)^2\}\big) = \rho(1-\langle x,x'\rangle_n^2) + \frac{b^2\tau}{2}(1-\langle x,x'\rangle_n\langle y,y'\rangle_p) \\ &\frac{1}{n}\big(\mathbb{E}\{(\langle x,g_x\rangle + \langle y,g_y\rangle - \langle x',g_x\rangle - \langle y',g_y\rangle)^2\}\big) = 2(4\rho+b^2\tau)(1-\langle x,x'\rangle_n) + 2b^2\tau(1-\langle y,y'\rangle_p) \\ &\frac{1}{4n}\big(\mathbb{E}\{(\langle x,W_x x\rangle + \langle y,W_y y\rangle - \langle x',W_x x\rangle - \langle y',W_y y'\rangle)^2\}\big) = (\rho+\frac{b^2\tau}{4})(1-\langle x,x'\rangle_n^2) + \frac{b^2\tau}{4}(1-\langle y,y'\rangle_p^2). \end{aligned}$$

This immediately gives:

$$\begin{aligned}
&\frac{1}{n}\big(\mathbb{E}\{(\langle x, Wx\rangle + b\langle y, Zx\rangle - \langle x', Wx'\rangle - b\langle y', Zx'\rangle)^2\}\big) - \\
&\frac{1}{n}\big(\mathbb{E}\{(\langle x, g_x\rangle + \langle y, g_y\rangle - \langle x', g_x\rangle - \langle y, g_y'\rangle)^2\}\big) \\
&= -4\rho(1 - \langle x, x'\rangle_n)^2 - 2b^2\tau(1 - \langle x, x'\rangle_n)(1 - \langle y, y'\rangle_p) \le 0, \\
&\frac{1}{4n}\big(\mathbb{E}\{(\langle x, Wx\rangle + b\langle y, Zx\rangle - \langle x', Wx'\rangle - b\langle y', Zx'\rangle)^2\}\big) - \\
&\frac{1}{4n}\big(\mathbb{E}\{(\langle x, W_x x\rangle + \langle y, W_y y\rangle - \langle x', W_x x\rangle - \langle y', W_y y'\rangle)^2\}\big) \\
&= \frac{b^2\tau}{4}(\langle x, x'\rangle_n - \langle y, y'\rangle_p)^2 \ge 0.
\end{aligned}$$

The claim follows. □

An immediate corollary of this is the following tight characterization for the null value, i.e. the case when $\mu = \lambda = 0$:

Corollary 14. *For any ρ, τ as n, p diverge with $n/p \to \gamma$, we have*

$$\lim_{n\to\infty} \mathsf{OPT}(0, 0) = \sqrt{4\rho + b^2\tau} + b\sqrt{\frac{\tau}{\gamma}} \tag{58}$$

Note that this upper bound generalizes the maximum eigenvalue and singular value bounds of W, Z respectively. In particular, the case $\tau = 0$ corresponds to the maximum eigenvalue of W, which yields $\mathsf{OPT} = 2\sqrt{\rho}$ while the maximum singular value of Z can be recovered by setting ρ to 0 and b to 1, yielding $\mathsf{OPT} = \sqrt{\tau}(1 + \gamma^{-1/2})$. Corollary 14 demonstrates the limit for the case when $\mu = \lambda = 0$. The following theorem gives the limiting value when λ, μ may be nonzero.

Theorem 15. *Suppose $\mathsf{G} : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}$ is as follows:*

$$\mathsf{G}(\kappa, \sigma^2) = \begin{cases} \kappa/2 + \sigma^2/2\kappa & \text{if } \kappa^2 \ge \sigma^2, \\ \sigma & \text{otherwise.} \end{cases} \tag{59}$$

Then the optimal value $\mathsf{OPT}(\lambda, \mu)$ is

$$\lim_{n\to\infty} \mathsf{OPT}(\lambda, \mu) = \min_{t \ge 0} \left\{ \mathsf{G}(2\lambda + b\mu t, 4\rho + b^2\tau) + \gamma^{-1}\mathsf{G}(b/t, b^2\gamma\tau) \right\}. \tag{60}$$

If the minimum above occurs at $t = t_*$ such that $\mathsf{G}'(2\lambda + b\mu t_*, 4\rho + b^2\tau) = \partial_\kappa \mathsf{G}(\kappa, 4\rho + b^2\tau)|_{\kappa = 2\lambda + b\mu t_*} > 0$, then $\lim_{n\to\infty} \mathsf{OPT}(\lambda, \mu) > \sqrt{4\rho + b^2\tau} + \gamma^{-1}\sqrt{\frac{\tau}{\gamma}}$.

A.3 Proof of Theorem 15: the upper bound

The following lemma removes the effect of the projection of g_x (g_y) along v_0 (resp. u_0). Let $F(x, y) = \frac{1}{n}[\lambda x_1^2 + \langle x, g_x\rangle + b\sqrt{\mu}x_1 y_1 + \langle y, g_y\rangle]$. Further, let $\widetilde{g}_x$ ($\widetilde{g}_y$) be the vectors obtained by setting the first coordinate of g_x (resp. g_y) to zero, and $\widetilde{F}(x, y) = \frac{1}{n}[\lambda x_1^2 + \langle x, \widetilde{g}_x\rangle + b\sqrt{\mu}x_1 y_1 + \langle y, \widetilde{g}_y\rangle]$.

Lemma 16. *The optima of F and $\widetilde{F}$ differ by at most $o(1)$. More precisely:*

$$\left| \mathbb{E} \max_{x,y} F(x, y) - \mathbb{E} \max_{x,y} \widetilde{F}(x, y) \right| = O\Big(\frac{1}{\sqrt{n}}\Big).$$

Proof. For any x, y:

$$F(x,y) = \frac{1}{n}\Big(\lambda x_1^2 + \langle x, g_x\rangle + \sqrt{\mu}x_1 y_1 + \langle y, g_y\rangle\Big) = \widetilde{F}(x,y) + \frac{1}{n}(x_1(g_x)_1 + y_1(g_y)_1)$$
$$\Big|F(x,y) - \widetilde{F}(x,y)\Big| \le \frac{1}{n}(\sqrt{n}|(g_x)_1| + \sqrt{p}|(g_y)_1|).$$

Maximizing each side over x, y and taking expectation yields the lemma. □

With this in hand, we can concentrate on computing the maximum of $\widetilde{F}(x,y)$.

Lemma 17. *Let $\widetilde{g}_x$ ($\widetilde{g}_y$) be the projection of g_x (resp. g_y) orthogonal to the first basis vector. Then*

$$\limsup_{n\infty} \mathbb{E} \max_{(x,y)\in\mathbb{S}^n\times\mathbb{S}^p} \widetilde{F}(x,y) \le \min_{t\le 0} \mathsf{G}(2\lambda + b\mu t, 4\rho + b^2\tau) + \frac{1}{\gamma}\mathsf{G}(b/t, b^2\gamma\tau) \tag{61}$$

Proof. Since $\widetilde{F}(x,y)$ increases if we align the signs of x_1 and y_1 to $+1$, we can assume that they are positive. Furthermore, for fixed, positive x_1, y_1, $\widetilde{F}$ is maximized if the other coordinates align with $\widetilde{g}_x$ and $\widetilde{g}_y$ respectively. Therefore:

$$\begin{aligned}
\max_{x,y} \widetilde{F}(x,y) &= \max_{x_1\in[0,\sqrt{n}], y_1\in[0,\sqrt{p}]} \frac{\lambda x_1^2}{n} + \sqrt{1-\frac{x_1^2}{n}}\frac{\|\widetilde{g}_x\|}{\sqrt{n}} + \frac{b\sqrt{\mu x_1 y_1}}{n} + \sqrt{1-\frac{y_1^2}{p}}\frac{\sqrt{p}\,\|\widetilde{g}_y\|}{n} \\
&= \max_{m_1,m_2\in[0,1]} \lambda m_1 + \sqrt{1-m_1}\frac{\|\widetilde{g}_x\|}{\sqrt{n}} + b\sqrt{\frac{\mu m_1 m_2 p}{n}} + \sqrt{1-m_2}\frac{\sqrt{p}\,\|\widetilde{g}_y\|}{n} \\
&\le \max_{m_1,m_2\in[0,1]} \Big(\lambda + \frac{b\mu t}{2}\Big)m_1 + \sqrt{1-m_1}\frac{\|\widetilde{g}_x\|}{\sqrt{n}} + \frac{p}{n}\Big(\frac{bm_2}{2t} + \sqrt{1-m_2}\frac{\|\widetilde{g}_y\|}{\sqrt{p}}\Big) \\
&= \mathsf{G}(2\lambda + b\mu t, \|\widetilde{g}_x\|^2/n) + \frac{1}{\gamma}\mathsf{G}\Big(\frac{b}{t}, \|\widetilde{g}_y\|^2/p\Big),
\end{aligned} \tag{62}$$

where the first equality is change of variables, the second inequality is the fact that $2\sqrt{ab} = \min_{t\ge 0}(at + b/t)$, and the final equality is by direct calculus.

Now let t_* be any minimizer of $\mathsf{G}(2\lambda + b\mu t, 4\rho + b^2\tau) + \gamma^{-1}\mathsf{G}(b/t, b^2\gamma\tau)$. We may assume that $t_* \notin \{0, \infty\}$, otherwise we can use $t_*(\varepsilon)$, an ε-approximate minimizer in $(0,\infty)$ in the argument below. Since the above holds for any t, we have:

$$\max_{x,y} \widetilde{F}(x,y) \le \mathsf{G}(2\lambda + b\mu t_*, \|\widetilde{g}_x\|^2/n) + \gamma^{-1}\mathsf{G}(b/t_*, \|\widetilde{g}_y\|^2/p). \tag{63}$$

By the strong law of large numbers, $\|\widetilde{g}_x\|^2/n \to 4\rho + b^2\tau$ and $\|\widetilde{g}_y\|^2/p \to b^2\gamma\tau$ almost surely. Further, as $\mathsf{G}(\kappa, \sigma^2)$ is continuous in the second argument on $(0,\infty)$, when $\kappa \notin \{0,\infty\}$, almost surely:

$$\limsup \max_{x,y} \widetilde{F}(x,y) \le \mathsf{G}(2\lambda + b\mu t_*, 4\rho + b^2\tau) + \gamma^{-1}\mathsf{G}(b/t_*, b^2\gamma\tau). \tag{64}$$

Taking expectations and using bounded convergence yields the lemma. □

We can now prove the upper bound.

Theorem 15, upper bound. Using Proposition 13, Lemma 16 and Lemma 17 in order:

$$\begin{aligned}
\mathsf{OPT}(\lambda,\mu) &\le \mathbb{E}\{\max_{x,y} F(x,y)\} && (65)\\
&\le \mathbb{E}\{\max_{x,y} \widetilde{F}(x,y)\} + o(n^{-1/3}) && (66)\\
&\le \min_t \mathsf{G}(2\lambda + b\mu t, 4\rho + b^2\tau) + \frac{1}{\gamma}\mathsf{G}(b/t, b^2\gamma\tau) + o(n^{-1/3}). && (67)
\end{aligned}$$

Taking limit $p \to \infty$ yields the result. □

A.4 Proof of Theorem 15: the lower bound

Recall that t_* denotes the optimizer of the upper bound $\mathsf{G}(2\lambda+b\mu t, 4\rho+b^2\tau)+\gamma^{-1}\mathsf{G}(b/t, b^2\gamma\tau)$. By stationarity, we have:

$$b\mu\mathsf{G}'(2\lambda+b\mu t_*, 4\rho+b^2\tau) - \frac{b}{\gamma t_*^2}\mathsf{G}'(\frac{b}{t_*}, b^2\gamma\tau) = 0. \tag{68}$$

Now we proceed in two cases. First, suppose $\mathsf{G}'(2\lambda+b\mu t_*, 4\rho+b^2\tau)=0$. In this case $\mathsf{G}'(b/t_*, b^2\gamma\tau)/t_*^2=0$, whence $\mathsf{G}'(b/t_*, b^2\gamma\tau)=0$. Indeed, the case when $t_*=\infty$ also satisfies this. However, this also implies that $2\lambda+b\mu t_* \le \sqrt{4\rho+b^2\tau}$ and $t_* \ge (\gamma\tau)^{-1/2}$, whereby $\mathsf{G}(2\lambda+b\mu t_*, 4\rho+b^2\tau)=\sqrt{4\rho+b^2\tau}$ and $\mathsf{G}'(b/t_*, b^2\gamma\tau) = b\sqrt{\gamma\tau}$. In this case we consider $\tilde{x}, \tilde{y}$ to be the principal eigenvectors of W_x, W_y rescaled to norms $\sqrt{n}, \sqrt{p}$ respectively and, hence using (57),

$$\mathsf{OPT}(\lambda,\mu,b) \ge \frac{1}{2n}\mathbb{E}\Big[\langle\tilde{x}, W_x\tilde{x}\rangle + \langle\tilde{y}, W_y\tilde{y}\rangle\Big] - o(1). \tag{69}$$

By standard results on GOE matrices the right hand side converges to $\sqrt{4\rho+b^2\tau}+b\sqrt{\frac{\tau}{\gamma}}$ implying the required lower bound.

Now consider the case that $\mathsf{G}'(2\lambda+b\mu t_*, 4\rho+b^2\tau)>0$. Importantly, by stationarity we have

$$t_*^2 = \frac{\mathsf{G}'(bt_*^{-1}, b^2\gamma\tau)}{\mu\gamma\mathsf{G}'(2\lambda+b\mu t_*, 4\rho+b^2\tau)}, \tag{70}$$

and that t_* is finite since the numerator is decreasing in t_*. The key ingredient to prove the lower bound is the following result on the principal eigenvalue/eigenvector of a deformed GOE matrix.

Theorem 18 ([CDMF+09, KY13])**.** *Suppose $W \in \mathbb{R}^{n\times n}$ is a GOE matrix with variance σ^2, i.e. $W_{ij} = W_{ji} \sim \mathsf{N}(0, (1+\delta_{ij}\sigma^2/p)$ and $A = \kappa v_0 v_0^\mathsf{T} + W$ where v_0 is a unit vector. Then the following holds almost surely and in expectation:*

$$\lim_{n\to\infty}\lambda_1(A) = 2\mathsf{G}(\kappa,\sigma^2) = \begin{cases} 2\sigma & \text{if } \kappa<\sigma \\ \kappa+\sigma^2/\kappa & \text{if } \kappa>\sigma. \end{cases} \tag{71}$$

$$\lim_{n\to\infty}\langle v_1(A), v_0\rangle^2 = 2\mathsf{G}'(\kappa,\sigma^2) = \begin{cases} 0 & \text{if } \kappa<\sigma, \\ 1-\sigma^2/\kappa^2 & \text{if } \kappa>\sigma. \end{cases}, \tag{72}$$

where G' denotes the derivative with respect to the first argument.

For the prescribed t_*, define:

$$H(x,y) = \Big(\lambda+\frac{b\mu t_*}{2}\Big)\frac{\langle x, v_0\rangle^2}{n^2} + \frac{\langle x, W_x x\rangle}{2n} + \frac{p}{n}\Big(\frac{b\langle y, u_0\rangle^2}{2t_* p^2} + \frac{\langle y, W_y y\rangle}{2p}\Big) \tag{73}$$

Let $\tilde{x}, \tilde{y}$ be the principal eigenvector of $(2\lambda+b\mu t_*)v_0 v_0^\mathsf{T}/n + W_x$, $bt_*^{-1}u_0u_0^\mathsf{T}/p + W_y$, rescaled to norm $\sqrt{n}$ and $\sqrt{p}$ respectively. Further, we choose the sign of $\tilde{x}$ so that $\langle\tilde{x}, v_0\rangle \ge 0$, and analogously for $\tilde{y}$. Now, fixing an $\varepsilon>0$, we have by Theorem 18, for every p large enough:

$$H(\tilde{x},\tilde{y}) \ge \mathsf{G}(2\lambda+b\mu t_*, 4\rho+b^2\tau) + \gamma^{-1}\mathsf{G}(bt_*^{-1}, b^2\gamma\tau) - \varepsilon \tag{74}$$

$$\frac{\langle\tilde{x}, v_0\rangle}{n} = \sqrt{2\mathsf{G}'(2\lambda+b\mu t_*, 4\rho+b^2\tau)} + O(\varepsilon) \tag{75}$$

$$\frac{\langle\tilde{y}, u_0\rangle}{p} = \sqrt{2\mathsf{G}'(bt_*^{-1}, b^2\gamma\tau)} + O(\varepsilon) \tag{76}$$

We have, therefore:

$$\mathsf{OPT}(\lambda,\mu,b) \geq \mathbb{E}\Big[H(\tilde{x},\tilde{y}) + \Big(\frac{b}{n}\sqrt{\frac{\mu}{np}}\langle\tilde{x},v_0\rangle\langle\tilde{y},u_0\rangle - \frac{b\mu t\langle\tilde{x},v_0\rangle^2}{2n^2} - \frac{b\langle y,u_0\rangle^2}{2tnp}\Big)\Big] \tag{77}$$

$$\geq \mathsf{G}(2\lambda + b\mu t_*, 4\rho + b^2\tau) + \gamma^{-1}\mathsf{G}(bt_*^{-1}, b^2\gamma\tau) + O(\varepsilon(t_* \vee t_*^{-1}))$$
$$+ \Big(2\sqrt{\frac{\mu}{\gamma}\mathsf{G}'(2\lambda + b\mu t_*, 4\rho + b^2\tau)\mathsf{G}'(bt_*^{-1}, b^2\gamma\tau)} - b\mu t_*\mathsf{G}'(2\lambda + b\mu t_*, 4\rho + b^2\tau)$$
$$- \frac{\mathsf{G}'(bt_*^{-1}, b^2\gamma\tau)}{\gamma t_*}\Big)$$
$$\geq \mathsf{G}(2\lambda + b\mu t_*, 4\rho + b^2\tau) + \gamma^{-1}\mathsf{G}(bt_*^{-1}, b^2\gamma\tau) + O(\varepsilon(t_* \vee t_*^{-1})). \tag{78}$$

Here the first inequality since we used a specific guess $\tilde{x}, \tilde{y}$, the second using Theorem 18 and the final inequality follows since the remainder term vanishes due to Eq. (70). Taking expectations and letting ε going to 0 yields the required lower bound.

Given Corollary 14 and Theorem 15, it is not too hard to establish Lemma 12, which we proceed to do next.

A.5 Proof of Lemma 12

Recall $b_* = \frac{2\mu}{\lambda\gamma}$. Part (i) follows directly from Corollary 14, upon setting $\rho = \tau = 1$, and $b = b_*\sqrt{\gamma}$. To establish part (ii), we use Theorem 15. In particular, it suffices to establish that with this specific choice of $b = b_*\sqrt{\gamma}$, for any (λ,μ) with $\lambda^2 + \mu^2/\gamma > 1$, the minimizer t_* of $G(2\lambda + b\mu t, 4 + b^2) + \gamma^{-1}G(b/t, b^2\gamma)$ satisfies $G'(2\lambda + b\mu t_*, 4 + b^2) > 0$. Let us assume, if possible, that $G(2\lambda + b\mu t_*, 4 + b^2) = 0$. Using the stationary point condition (68), in this case $G'(b/t_*, b^2\gamma) = 0$. Next, using the definition of G (59), observe that this implies

$$t_* > \frac{1}{\sqrt{\gamma}}, \quad 2\lambda + \frac{2\mu^2}{\lambda\sqrt{\gamma}}t_* < \sqrt{4 + \frac{4\mu^2}{\lambda^2\gamma}}.$$

These imply:

$$\frac{2}{\lambda}\Big(\lambda^2 + \frac{\mu^2}{\gamma}\Big) < 2\lambda + 2\frac{\mu^2 t_*}{\lambda\mu\sqrt{\gamma}} \tag{79}$$

$$< \sqrt{4 + \frac{4\mu^2}{\lambda^2\gamma}} \tag{80}$$

$$= \frac{2}{\lambda}\sqrt{\lambda^2 + \frac{\mu^2}{\gamma}}. \tag{81}$$

That this is impossible whenever $\lambda^2 + \frac{\mu^2}{\gamma} > 1$. This establishes part (ii). To establish part (iii), we again use the upper bound from Proposition 13, and note that for $0 < \langle x, v\rangle < \tilde{\delta}\sqrt{n}$,

$$\mathbb{E}[\tilde{T}(\tilde{\delta})] \leq \lambda\tilde{\delta}^2 + \sqrt{4 + b_*^2} + \max_{\|y\|=1}\{b_*\sqrt{\mu}\tilde{\delta}\langle u, y\rangle + \frac{1}{\gamma}\langle y, g\rangle\},$$

where $g \sim \mathsf{N}(0, b^2\gamma I_p/p)$. The proof follows using continuity in $\tilde{\delta}$. This completes the proof.

B Proof of Lemma 8

Recall the distributional recursion specified by density evolution (Definition 1).

$$\bar{m}'|_U \stackrel{d}{=} \mu U \mathbb{E}[V\bar{\eta}] + \zeta_1\sqrt{\mu\mathbb{E}[\bar{\eta}^2]},$$
$$\bar{\eta}'|_{V'=+1} \stackrel{d}{=} \frac{\lambda}{\sqrt{d}}\Big[\sum_{k=1}^{k_+}\bar{\eta}_k|_+ + \sum_{k=1}^{k_-}\bar{\eta}_k|_-\Big] - \lambda\sqrt{d}\mathbb{E}[\bar{\eta}] + \frac{\mu}{\gamma}\mathbb{E}[U\bar{m}] + \zeta_2\sqrt{\frac{\mu}{\gamma}\mathbb{E}[\bar{m}^2]},$$

where $V \sim U(\{\pm 1\})$, $U \sim \mathsf{N}(0,1)$, $k_+ \sim \text{Poisson}\Big(\frac{d+\lambda\sqrt{d}}{2}\Big)$, $k_- \sim \text{Poisson}\Big(\frac{d-\lambda\sqrt{d}}{2}\Big)$, $\zeta_1, \zeta_2 \sim \mathsf{N}(0,1)$ are all mutually independent. Further, $\{\bar{\eta}_k|_+\}$ are iid random variables, distributed as $\bar{\eta}|_{V=+1}$. Similarly, $\{\bar{\eta}_k|_-\}$, are iid random variables, distributed as $\bar{\eta}|_{V=-1}$. Finally, we require the collections to be mutually independent, and independent of the other auxiliary variables defined above.

Given these distributional recursions, we compute the vector of moments

$$\mathbb{E}[V'\bar{\eta}'] = \lambda^2\mathbb{E}[V\bar{\eta}] + \frac{\mu}{\gamma}\mathbb{E}[U\bar{m}]$$
$$\mathbb{E}[U'\bar{m}'] = \mu\mathbb{E}[V\bar{\eta}]$$
$$\mathbb{E}[\bar{\eta}'^2] = \lambda^2\mathbb{E}[\bar{\eta}^2] + \frac{\mu^2}{\gamma^2}\mathbb{E}^2[U\bar{m}] + \frac{\mu}{\gamma}\mathbb{E}[\bar{m}^2] + 2\frac{\lambda^2}{\gamma}\mathbb{E}[U\bar{m}]\mathbb{E}[V\bar{\eta}].$$
$$\mathbb{E}[\bar{m}'^2] = \mu^2\mathbb{E}^2[V\bar{\eta}] + \mu\mathbb{E}[\bar{\eta}^2]$$

Thus the induced mapping on moments $\phi^{\mathsf{DE}} : \mathbb{R}^4 \to \mathbb{R}^4$, $\phi^{\mathsf{DE}}(z_1, z_2, z_3, z_4) = (\phi_1, \phi_2, \phi_3, \phi_4)$, with

$$\phi_1 = \lambda^2 z_1 + \frac{\mu}{\gamma}z_2$$
$$\phi_2 = \mu z_1$$
$$\phi_3 = \frac{\mu^2}{\gamma^2}z_2^2 + \frac{2\lambda^2}{\gamma}z_1z_2 + \lambda^2 z_3 + \frac{\mu}{\gamma}z_4,$$
$$\phi_4 = \mu^2 z_1^2 + \mu z_3.$$

The Jacobian of ϕ^{DE} at 0 is, up to identical row/column permutation:

$$J = \begin{bmatrix}\lambda^2 I_2 & \frac{\mu}{\gamma}I_2 \\ \mu I_2 & 0\end{bmatrix}.$$

By direct computation, we see that z is an eigenvalue of J if and only if $z^2 - \lambda^2 z - \frac{\mu^2}{\gamma} = 0$. Consider the quadratic function $f(z) = z^2 - \lambda^2 z - \frac{\mu^2}{\gamma}$ and note that $f(0) < 0$. Thus to check whether f has a root with magnitude greater than 1, it suffices to check its value at $z = 1, -1$. Note that if $\lambda^2 + \frac{\mu^2}{\gamma} > 1$, $f(1) < 0$ and thus J has an eigenvalue greater than 1. Conversely, if $\lambda^2 + \frac{\mu^2}{\gamma} < 1$, $f(1) > 0$ and $f(-1) = 1 + \lambda^2 - \frac{\mu^2}{\gamma} > 1 - \frac{\mu^2}{\gamma} > 0$. This completes the proof.

C Proof of Theorem 4

We prove Theorem 4 in this Section. Recall the matrix mean square errors

$$\mathsf{MMSE}(v; A, B) = \frac{1}{n(n-1)}\mathbb{E}\Big[\|vv^T - \mathbb{E}[vv^T|A, B]\|_F^2\Big],$$
$$\mathsf{MMSE}(v; A^G, B) = \frac{1}{n(n-1)}\mathbb{E}\Big[\|vv^T - \mathbb{E}[vv^T|A^G, B]\|_F^2\Big].$$

The following lemma is immediate from Lemma 4.6 in [DAM16].

Lemma 19. *Let $\widehat{v} = \widehat{v}(A, B)$ be any estimator so that $\|\widehat{v}\|_2 = \sqrt{n}$. Then*

$$\liminf_{n\to\infty} \frac{\langle \widehat{v}, v\rangle}{n} > 0 \text{ in probability } \Rightarrow \limsup_{n\to\infty} \mathsf{MMSE}(v; A, B) < 1. \tag{82}$$

Furthermore, if $\limsup_{n\to\infty} \mathsf{MMSE}(v; A, B) < 1$, there exists an estimator $\widehat{s}(A, B)$ with $\|\widehat{s}(A, B)\|_2 = \sqrt{n}$ so that, in probability:

$$\liminf_{n\to\infty} \frac{\langle \widehat{s}, v\rangle}{n} > 0. \tag{83}$$

Indeed, the same holds for the observation model A^G, B.

Proof of Theorem 4. Consider first the case $\lambda^2 + \frac{\mu^2}{\gamma} < 1$. For any $\theta \in [0, \lambda]$, $\theta^2 + \mu^2/\gamma < 1$ as well. Suppose we have $A(\theta), B$ according to model (8), (9) where λ is replaced with θ. By Theorem 6 (applied at θ) and the second part of Lemma 19, $\liminf_{n\to\infty} \mathsf{MMSE}(v; A(\theta), B) = 1$. Using the I-MMSE identity [GSV05], this implies

$$\lim_{n\to\infty} \frac{1}{n}(I(v; A(\theta), B) - I(v; A(0), B)) = \frac{\theta^2}{4}. \tag{84}$$

By Theorem 5, for all $\theta \in [0, \lambda]$

$$\lim_{d\to\infty} \lim_{n\to\infty} \frac{1}{n}(I(v; A^G(\theta), B) - I(v; A^G(0), B) = \frac{\theta^2}{4}, \tag{85}$$

$$\text{and, therefore } \lim_{n\to\infty} \mathsf{MMSE}(v; A^G, B) = 1 \tag{86}$$

This implies, via the first part of Lemma 19 that for any estimator $\widehat{v}(A^G; B)$, we have $\limsup_{n\to\infty} |\langle \widehat{v}, v\rangle|/n = 0$ in probability, as required.

Conversely, consider the case $\lambda^2 + \frac{\mu^2}{\gamma} > 1$. We may assume that $\mu^2/\gamma < 1$, as otherwise the result follows from Theorem 2. Let $\lambda_0 = (1 - \mu^2/\gamma)^{1/2}$.

Now, by the same argument for Eqs.(84), (85), we obtain for all $\theta_1, \theta_2 \in [\lambda_0, \lambda]$:

$$\limsup_{n\to\infty} \frac{1}{n}(I(v; A(\theta_1), B) - I(v; A(\theta_2), B)) < \frac{\theta_1^2 - \theta_2^2}{4}. \tag{87}$$

Applying Theorem 5, we have for all $\theta_1, \theta_2, \theta \in [\lambda_0, \lambda]$:

$$\lim_{d\to\infty} \limsup_{n\to\infty} \frac{1}{n}(I(v; A^G(\theta_1), B) - I(v; A^G(\theta_2), B)) < \frac{\theta_1^2 - \theta_2^2}{4} \tag{88}$$

$$\text{and therefore, } \limsup \mathsf{MMSE}(v; A^G(\theta), B) < 1. \tag{89}$$

Applying then Lemma 19 implies that we have an estimator $\widehat{s}(A^G, B)$ with non-trivial overlap i.e. in probability:

$$\lim_{d\to\infty} \liminf_{n\to\infty} \frac{\langle \widehat{s}, v\rangle}{n} > 0. \tag{90}$$

This completes the proof.

□

D Belief propagation: derivation

In this section we will derive the belief propagation algorithm. Recall the observation model for $(A^G, B) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{p\times n}$ in Eqs. (1), (2):

$$A^G_{ij} = \begin{cases} 1 & \text{with probability } \frac{d+\lambda\sqrt{d}v_iv_j}{n} \\ 0 & \text{otherwise.} \end{cases} \tag{91}$$

$$B_{qi} = \sqrt{\frac{\mu}{n}}u_qv_i + Z_{qi}, \tag{92}$$

where u_q and Z_{qi} are independent $\mathsf{N}(0, 1/p)$ variables.

We will use the following conventions throughout this section to simplify some of the notation. We will index nodes in the graph, i.e. elements in $[n]$ with $i, j, k \dots$ and covariates, i.e. elements in $[p]$ with $q, r, s, \dots$. We will use '$\simeq$' to denote equality of probability distributions (or densities) up to an omitted proportionality constant, that may change from line to line. We will omit the superscript G in A^G. In the graph G, we will denote neighbors of a node i with ∂i and non-neighbors with ∂i^c.

We start with the posterior distribution of u, v given the data A, B:

$$\mathrm{d}\mathbb{P}\{u, v|A, B\} = \frac{\mathrm{d}\mathbb{P}\{A, B|u, v\}}{\mathrm{d}\mathbb{P}\{A, B\}}\mathrm{d}\mathbb{P}\{u, v\} \tag{93}$$

$$\simeq \prod_{i<j}\Big(\frac{d+\lambda\sqrt{d}v_iv_j}{n}\Big)^{A_{ij}}\Big(1-\frac{d+\lambda\sqrt{d}v_iv_j}{n}\Big)^{1-A_{ij}}$$
$$\cdot\prod_{q,i}\exp\Big(\sqrt{\frac{\mu p^2}{n}}B_{qi}u_qv_i\Big)\prod_q\exp\Big(-\frac{p(1+\mu)}{2}u_q^2\Big). \tag{94}$$

The belief propagation algorithm operates 'messages' $\nu^t_{i\to j}, \nu^t_{q\to i}, \nu^t_{i\to q}$ which are probability distributions. They represent the marginals of the variables v_i, u_q in the absence of variables v_j, u_q, in the posterior distribtuion $\mathrm{d}\mathbb{P}\{u, v|A, B\}$. We denote by $\mathbb{E}^t_{i\to j}, \mathbb{E}^t_{q\to i}, \mathbb{E}^t_{i\to q}$ expectations with respect to these distributions. The messages are are computed using the following update equations:

$$\nu^{t+1}_{i\to j}(v_i) \simeq \prod_{q\in[p]}\mathbb{E}^t_{q\to i}\Big\{\exp\Big(\sqrt{\frac{\mu p^2}{n}}B_{qi}v_iu_q\Big)\Big\}\prod_{k\in\partial i\setminus j}\mathbb{E}^t_{k\to i}\Big(\frac{d+\lambda\sqrt{d}v_iv_k}{n}\Big)\prod_{k\in\partial i^c\setminus j}\mathbb{E}^t_{k\to i}\Big(1-\frac{d+\lambda\sqrt{d}v_iv_k}{n}\Big), \tag{95}$$

$$\nu^{t+1}_{i\to q}(v_i) \simeq \prod_{r\in[p]\setminus q}\mathbb{E}^t_{r\to i}\Big\{\exp\Big(\sqrt{\frac{\mu p^2}{n}}B_{ri}v_iu_r\Big)\Big\}\prod_{k\in\partial i}\mathbb{E}^t_{k\to i}\Big(\frac{d+\lambda\sqrt{d}v_iv_k}{n}\Big)\prod_{k\in\partial i^c}\mathbb{E}^t_{k\to i}\Big(1-\frac{d+\lambda\sqrt{d}v_iv_k}{n}\Big), \tag{96}$$

$$\nu^{t+1}_{q\to i}(u_q) \simeq \exp\Big(-\frac{p(1+\mu)u_q^2}{2}\Big)\prod_{j\neq i}\mathbb{E}^t_{j\to q}\Big\{\exp\Big(\sqrt{\frac{\mu p^2}{n}}B_{qj}v_ju_q\Big)\Big\}. \tag{97}$$

As is standard, we define ν^t_i, ν^t_q in the same fashion as above, except without the removal of the incoming message.

D.1 Reduction using Gaussian ansatz

The update rules (95), (96), (97) are in terms of probability distributions, i.e. measures on the real line or $\{\pm1\}$. We reduce them to update rules on real numbers using the following analytical ansatz. The measure $\nu^t_{i\to j}$ on $\{\pm1\}$ can be summarized using the log-odds ratio:

$$\eta^t_{i\to j} \equiv \frac{1}{2}\log\frac{\nu^t_{i\to j}(+1)}{\nu^t_{i\to j}(-1)}, \tag{98}$$

and we similarly define $\eta^t_{i\to q}$, η^t_i. In order to reduce the densities $\nu^t_{q\to i}$, we use the Gaussian ansatz:

$$\nu^t_{q\to i} = \mathsf{N}\Big(\frac{m^t_{q\to i}}{\sqrt{p}}, \frac{\tau^t_{q\to i}}{p}\Big). \tag{99}$$

With Equations (98) and (99) we can now simplify Equations (95) to (97). The following lemma computes the inner marginalizations in Equations (95) to (97). We omit the proof.

Lemma 20. *With $\nu^t, \mathbb{E}^t$ as defined as per Equations (95) to (97) and η^t, m^t, τ^t as in Equations (98) and (99) we have*

$$\mathbb{E}^t_{q\to i} \exp\Big(\sqrt{\frac{\mu p^2}{n}} B_{qi} v_i u_q\Big) = \exp\Big(\sqrt{\frac{\mu p}{n}} B_{qi} v_i m^t_{q\to i} + \frac{\mu p}{2n} B^2_{qi} \tau^t_{q\to i}\Big), \tag{100}$$

$$\mathbb{E}^t_{i\to j}\Big(\frac{d + \lambda\sqrt{d} v_i v_j}{n}\Big) = \frac{d}{n}\Big(1 + \frac{\lambda v_j}{\sqrt{d}} \tanh(\eta^t_{i\to j})\Big), \tag{101}$$

$$\mathbb{E}^t_{i\to j}\Big(1 - \frac{d + \lambda\sqrt{d} v_i v_j}{n}\Big) = 1 - \frac{d}{n}\Big(1 + \frac{\lambda v_j}{\sqrt{d}} \tanh(\eta^t_{i\to j})\Big), \tag{102}$$

$$\mathbb{E}^t_{i\to q} \exp\Big(p\sqrt{\frac{\mu}{n}} B_{qi} v_i u_q\Big) = \frac{\cosh(\eta^t_{i\to q} + p\sqrt{\mu/n} B_{qi} u_q)}{\cosh \eta^t_{i\to q}}. \tag{103}$$

The update equations take a simple form using the following definitions

$$f(z;\rho) \equiv \frac{1}{2}\log\Big(\frac{\cosh(z+\rho)}{\cosh(z-\rho)}\Big), \tag{104}$$

$$\rho \equiv \tanh^{-1}(\lambda/\sqrt{d}), \tag{105}$$

$$\rho_n \equiv \tanh^{-1}\Big(\frac{\lambda\sqrt{d}}{n-d}\Big). \tag{106}$$

With this, we first compute the update equation for the node messages η^{t+1}. Using Equations (95), (96) and (100) to (103):

$$\eta^{t+1}_{i\to j} = \sqrt{\frac{\mu}{\gamma}} \sum_{q\in[p]} B_{qi} m^t_{q\to i} + \sum_{k\in\partial i\setminus j} f(\eta^t_{k\to i};\rho) - \sum_{k\in\partial i\setminus j} f(\eta^t_{k\to i};\rho_n), \tag{107}$$

$$\eta^{t+1}_{i\to q} = \sqrt{\frac{\mu}{\gamma}} \sum_{r\in[p]\setminus q} B_{ri} m^t_{r\to i} + \sum_{k\in\partial i} f(\eta^t_{k\to i};\rho) - \sum_{k\in\partial i^c} f(\eta^t_{k\to i};\rho_n), \tag{108}$$

$$\eta^{t+1}_{i} = \sqrt{\frac{\mu}{\gamma}} \sum_{q\in[p]} B_{qi} m^t_{q\to i} + \sum_{k\in\partial i} f(\eta^t_{k\to i};\rho) - \sum_{k\in\partial i^c} f(\eta^t_{k\to i};\rho_n). \tag{109}$$

Now we compute the updates for $m^t_{a\to i}, \tau^t_{a\to i}$. We start from Equations (97) and (100), and use Taylor approximation assuming u_q, B_{jq} are both $O(1/\sqrt{p})$, as the ansatz (99) suggests.

$$\log \nu^{t+1}_{q\to i}(u_q) = \text{const.} + \frac{-p(1+\mu)}{2} u_q^2 + \sum_{j\in[n]\setminus i} \log\cosh\Big(\eta^t_{j\to q} + p\sqrt{\frac{\mu}{n}} B_{qj} u_q\Big) \tag{110}$$

$$= \text{const.} + \frac{-p(1+\mu)}{2} u_q^2 + \Big(p\sqrt{\frac{\mu}{n}} \sum_{j\in[n]\setminus i} B_{qj} \tanh(\eta^t_{j\to q})\Big) u_q + \Big(\frac{p^2\mu}{2n} \sum_{j\in[n]} B^2_{qj} \mathrm{sech}^2(\eta^t_{j\to q})\Big) u_q^2 + O\Big(\frac{1}{\sqrt{n}}\Big). \tag{111}$$

Note that here we compute $\log \nu^{t+1}$ only up to constant factors (with slight abuse of the notation '$\simeq$'). It follows from this quadratic approximation that:

$$\tau_{q\to i}^{t+1} = \Big(1+\mu-\frac{\mu}{\gamma}\sum_{j\in[n]\setminus i} B_{qj}^2 \mathrm{sech}^2(\eta_{j\to q}^t)\Big)^{-1}, \tag{112}$$

$$m_{q\to i}^{t+1} = \tau_{q\to i}^{t+1}\sqrt{\frac{\mu}{\gamma}}\sum_{j\in[n]\setminus i} B_{qj}\tanh(\eta_{j\to q}^t) \tag{113}$$

$$= \frac{\sqrt{\mu/\gamma}\sum_{j\in[n]\setminus i} B_{qj}\tanh(\eta_{j\to q}^t)}{1+\mu-\mu\gamma^{-1}\sum_{j\in[n]} B_{qj}^2\mathrm{sech}^2(\eta_{j\to q}^t)}. \tag{114}$$

Updates computing m_q^{t+1}, τ_q^{t+1} are analogous.

D.2 From message passing to approximate message passing

The updates for η^t, m^t derived in the previous section require keeping track of $O(np)$ messages. In this section, we further reduce the number of messages to $O(dn+p)$, i.e. linear in the size of the input graph observation.

The first step is to observe that the dependence of $\eta_{i\to j}^t$ on j is negligible when j is not a neighbor of i in the graph G. This derivation is similar to the presentation in [DKMZ11]. As $\sup_{z\in\mathbb{R}} f(z;\rho) \le \rho$. Therefore, if i, j are not neighbors in G:

$$\eta_{i\to j}^t = \eta_i^t - f(\eta_{j\to i}^{t-1};\rho_n) \tag{115}$$

$$= \eta_i^t + O(\rho_n) = \eta_i^t + O\Big(\frac{1}{n}\Big). \tag{116}$$

Now, for a pair i, j not connected, by Taylor expansion and the fact that $\partial_z f(z;\rho) \le \tanh(\rho)$,

$$f(\eta_{i\to j}^t;\rho_n) - f(\eta_i^t;\rho_n) = O\Big(\frac{\tanh(\rho_n)}{n}\Big) = O\Big(\frac{1}{n^2}\Big). \tag{117}$$

Therefore, the update equation for $\eta_{i\to j}^{t+1}$ satisfies:

$$\eta_{i\to j}^{t+1} = \sqrt{\frac{\mu}{\gamma}}\sum_{q\in[p]} B_{qi} m_{q\to i}^t + \sum_{k\in\partial i\setminus j} f(\eta_{k\to i}^t;\rho) - \sum_{k\in[n]} f(\eta_k^t;\rho_n) + O\Big(\frac{1}{n}\Big), \tag{118}$$

$$\eta_i^{t+1} = \eta_{i\to j}^{t+1} + f(\eta_{j\to i}^t;\rho). \tag{119}$$

Similarly for $\eta_{i\to q}^{t+1}$ we have:

$$\eta_{i\to q}^{t+1} = \sqrt{\frac{\mu}{\gamma}}\sum_{r\in[p]\setminus q} B_{ri} m_{r\to i}^t + \sum_{k\in\partial i} f(\eta_{k\to i}^t;\rho) - \sum_{k\in[n]} f(\eta_k^t;\rho_n) + O\Big(\frac{1}{n}\Big). \tag{120}$$

Ignoring $O(1/n)$ correction term, the update equations reduce to variables $(\eta_{i\to j}^t, \eta_i^t)$ where i, j are neighbors.

We now move to reduce updates for $\eta_{i\to q}^t$ and $m_{q\to i}^t$ to involving $O(n)$ variables. This reduction is more subtle then that of $\eta_{i\to j}^t$, where we are able to simply ignore the dependence of $\eta_{i\to j}^t$ on j if $j \notin \partial i$. We follow a derivation similar to that in [Mon12]. We use the ansatz:

$$\eta_{i\to q}^t = \eta_i^t + \delta\eta_{i\to q}^t \tag{121}$$

$$m_{q\to i}^t = m_q^t + \delta m_{q\to i}^t \tag{122}$$

$$\tau_{q\to i}^t = \tau_q^t + \delta\tau_{q\to i}^t, \tag{123}$$

where the corrections $\delta\eta^t_{i\to q}, \delta m^t_{q\to i}, \delta\tau^t_{q\to i}$ are $O(1/\sqrt{n})$. From Equations (97) and (120) at iteration t:

$$\eta^t_i + \delta\eta^t_{i\to q} = \sqrt{\frac{\mu}{\gamma}} \sum_{r\in[p]\setminus q} B_{ri}(m^{t-1}_r + \delta m^{t-1}_{r\to i}) + \sum_{k\in\partial i} f(\eta^{t-1}_{k\to i};\rho) - \sum_k f(\eta^{t-1}_k;\rho_n) \tag{124}$$

$$= \sqrt{\frac{\mu}{\gamma}} \sum_{r\in[p]} B_{ri}(m^{t-1}_r + \delta m^{t-1}_{r\to i}) + \sum_{k\in\partial i} f(\eta^{t-1}_{k\to i};\rho) - \sum_k f(\eta^{t-1}_k;\rho_n) - \sqrt{\frac{\mu}{\gamma}}\big(B_{qi}m^{t-1}_q + B_{qi}\delta m^{t-1}_{q\to i}\big). \tag{125}$$

Notice that the last term is the only term that depends on q. Further, since $B_{qi}\delta m^{t-1}_{q\to i} = O(1/n)$ by our ansatz, we may safely ignore it to obtain

$$\eta^t_i = \sqrt{\frac{\mu}{\gamma}} \sum_{r\in[p]} B_{ri}(m^{t-1}_r + \delta m^{t-1}_{r\to i}) + \sum_{k\in\partial i} f(\eta^{t-1}_{k\to i};\rho) - \sum_k f(\eta^{t-1}_k;\rho_n) \tag{126}$$

$$\delta\eta^t_{i\to q} = -\sqrt{\frac{\mu}{\gamma}} B_{qi} m^{t-1}_q. \tag{127}$$

We now use the update equation for $\tau^{t+1}_{q\to i}$:

$$\tau^{t+1}_q = \left(1 + \mu - \frac{\mu}{\gamma}\sum_{j\in[n]} B^2_{qj}\mathrm{sech}^2(\eta^t_j + \delta\eta^t_{j\to q})\right)^{-1} + O(1/n) \tag{128}$$

$$= \left(1 + \mu - \frac{\mu}{\gamma}\sum_{j\in[n]} B^2_{qj}\big((\mathrm{sech}^2(\eta^t_j) - 2\mathrm{sech}^2(\eta^t_j)\tanh(\eta^t_j)\delta\eta^t_{i\to q}\big)\right)^{-1} + O(1/n), \tag{129}$$

where we expanded the equation to linear order in $\delta\eta^t_{i\to q}$ and ignored higher order terms. By the identification Equation (127):

$$\tau^{t+1}_q = \left(1 + \mu - \frac{\mu}{\gamma}\sum_{j\in[n]} B^2_{qj}\mathrm{sech}^2(\eta^t_j) + 2\Big(\frac{\mu}{\gamma}\Big)^{3/2}\sum_{j\in[n]} B^3_{qj}\mathrm{sech}^2(\eta^t_j)\tanh(\eta^t_j)m^{t-1}_q\right)^{-1} + O(1/n). \tag{130}$$

Notice here, that there is no term that explicitly depends on i and the final term is $O(1/\sqrt{n})$ since $B_{qj} = O(1/\sqrt{n})$. Therefore, ignoring lower order terms, we have the identification:

$$\tau^{t+1}_q = \left(1 + \mu - \frac{\mu}{\gamma}\sum_{j\in[n]} B^2_{qj}\mathrm{sech}^2(\eta^t_j)\right)^{-1}, \tag{131}$$

$$\delta\tau^{t+1}_{q\to i} = 0. \tag{132}$$

Now we simplify the update for $m_{q\to i}^{t+1}$ using Taylor expansion to first order:

$$m_q^{t+1} + \delta m_{q\to i}^{t+1} = \frac{\sqrt{\mu/\gamma}}{\tau_q^{t+1}} \sum_{j\in[n]\setminus i} B_{qj} \tanh(\eta_j^t + \delta\eta_{j\to q}^t) \tag{133}$$

$$= \frac{\sqrt{\mu/\gamma}}{\tau_q^{t+1}} \sum_{j\in[n]\setminus i} \left(B_{qj} \tanh(\eta_j^t) + B_{qj}\mathrm{sech}^2(\eta_i^t)\delta\eta_{j\to q}^t\right) \tag{134}$$

$$= \frac{\sqrt{\mu/\gamma}}{\tau_q^{t+1}} \sum_{j\in[n]\setminus i} \left(B_{qj} \tanh(\eta_j^t) - \sqrt{\frac{\mu}{\gamma}} B_{qj}^2 \mathrm{sech}^2(\eta_j^t) m_q^{t-1}\right) \tag{135}$$

$$= \frac{\sqrt{\mu/\gamma}}{\tau_q^{t+1}} \sum_{j\in[n]} B_{qj} \tanh(\eta_j^t) - \frac{\mu}{\gamma\tau_q^{t+1}} \left(\sum_{j\in[n]} B_{qj}^2 \mathrm{sech}^2(\eta_j^t)\right) m_q^{t-1}$$

$$- \frac{\sqrt{\mu/\gamma}}{\tau_q^{t+1}} \left(B_{qi} \tanh(\eta_i^t) - \sqrt{\mu/\gamma} B_{qi}^2 \mathrm{sech}^2(\eta_i^t) m_q^{t-1}\right). \tag{136}$$

Only the final term is dependent on i, therefore we can identify:

$$m_q^{t+1} = \frac{\sqrt{\mu/\gamma}}{\tau_q^{t+1}} \sum_{j\in[n]} B_{qj} \tanh(\eta_j^t) - \frac{\mu}{\gamma\tau_q^{t+1}} \left(\sum_{j\in[n]} B_{qj}^2 \mathrm{sech}^2(\eta_j^t)\right) m_q^{t-1}, \tag{137}$$

$$\delta m_{q\to i}^{t+1} = -\frac{\sqrt{\mu/\gamma}}{\tau_q^{t+1}} B_{qi} \tanh(\eta_i^t). \tag{138}$$

Here, as before, we ignore the lower order term in $\delta m_{q\to i}^{t+1}$. Now we can substitute the identification Equation (138) back in Equation (126) at iteration $t+1$:

$$\eta_i^{t+1} = \sqrt{\frac{\mu}{\gamma}} \sum_{r\in[p]} B_{ri} m_r^t - \frac{\mu}{\gamma} \sum_{r\in[p]} \frac{B_{ri}^2}{\tau_r^t} \tanh(\eta_i^{t-1}) + \sum_{k\in\partial i} f(\eta_{k\to i}^t; \rho) - \sum_k f(\eta_k^t; \rho_n). \tag{139}$$

Collecting the updates for $\eta_i^t, \eta_{i\to j}^t, m_q^t$ we obtain the approximate message passing algorithm:

$$\eta_i^{t+1} = \sqrt{\frac{\mu}{\gamma}} \sum_{q\in[p]} B_{qi} m_q^t - \frac{\mu}{\gamma} \left(\sum_{q\in[p]} \frac{B_{qi}^2}{\tau_q^t}\right) \tanh(\eta_i^{t-1}) + \sum_{k\in\partial i} f(\eta_{k\to i}^t; \rho) - \sum_{k\in[n]} f(\eta_k^t; \rho_n)\,, \tag{140}$$

$$\eta_{i\to j}^{t+1} = \sqrt{\frac{\mu}{\gamma}} \sum_{q\in[p]} B_{qi} m_q^t - \frac{\mu}{\gamma} \left(\sum_{q\in[p]} \frac{B_{qi}^2}{\tau_q^t}\right) \tanh(\eta_i^{t-1}) + \sum_{k\in\partial i\setminus j} f(\eta_{k\to i}^t; \rho) - \sum_{k\in[n]} f(\eta_k^t; \rho_n)\,, \tag{141}$$

$$m_q^{t+1} = \frac{\sqrt{\mu/\gamma}}{\tau_q^{t+1}} \sum_{j\in[n]} B_{qj} \tanh(\eta_j^t) - \frac{\mu}{\gamma\tau_q^{t+1}} \left(\sum_{j\in[n]} B_{qj}^2 \mathrm{sech}^2(\eta_j^t)\right) m_q^{t-1} \tag{142}$$

$$\tau_q^{t+1} = \left(1 + \mu - \frac{\mu}{\gamma} \sum_{j\in[n]} B_{qj}^2 \mathrm{sech}^2(\eta_j^t)\right)^{-1}. \tag{143}$$

D.3 Linearized approximate message passing

This algorithm results from expanding the updates Equations (140) to (143) to linear order in the messages $\eta_i^t, \eta_{i\to j}^t$:

$$\eta_i^{t+1} = \sqrt{\frac{\mu}{\gamma}} \sum_{q\in[p]} B_{qi} m_q^t - \frac{\mu}{\gamma}\Big(\sum_{q\in[p]} \frac{B_{qi}^2}{\tau_q^t}\Big)\eta_i^{t-1} + \frac{\lambda}{\sqrt{d}} \sum_{k\in\partial i} \eta_{k\to i}^t - \frac{\lambda\sqrt{d}}{n} \sum_{k\in[n]} \eta_k^t \tag{144}$$

$$\eta_{i\to j}^{t+1} = \sqrt{\frac{\mu}{\gamma}} \sum_{q\in[p]} B_{qi} m_q^t - \frac{\mu}{\gamma}\Big(\sum_{q\in[p]} \frac{B_{qi}^2}{\tau_q^t}\Big)\eta_i^{t-1} + \frac{\lambda}{\sqrt{d}} \sum_{k\in\partial i\setminus j} \eta_{k\to i}^t - \frac{\lambda\sqrt{d}}{n} \sum_{k\in[n]} \eta_k^t \tag{145}$$

$$m_q^{t+1} = \frac{\sqrt{\mu/\gamma}}{\tau_q^{t+1}} \sum_{j\in[n]} B_{qj}\eta_j^t - \frac{\mu}{\gamma\tau_q^{t+1}}\Big(\sum_{j\in[n]} B_{qj}^2\Big) m_q^{t-1} \tag{146}$$

$$\tau_q^{t+1} = \left(1 + \mu - \frac{\mu}{\gamma}\sum_{j\in[n]} B_{qj}^2\right)^{-1}. \tag{147}$$

This follows from the linear approximation $f(z;\rho) = \tanh(\rho)z$ for small z. The algorithm given in the main text follows by using the law of large numbers to approximate $\sum_{j\in[n]} B_{qj}^2 \approx 1/\gamma$, $\sum_{q\in[p]} B_{qj}^2 \approx 1$, and hence $\tau 4_q \approx 1$.