Journal of Machine Learning Research 18 (2018) 1-52 Submitted 6/17; Revised 2/18; Published 4/18

Submatrix localization via message passing

Bruce Hajek B-HAJEK@ILLINOIS.EDU
Department of ECE and Coordinated Science Lab
University of Illinois at Urbana-Champaign
Urbana, IL 61801, USA

Yihong Wu YIHONG.WU@YALE.EDU
Department of Statistics and Data Science
Yale University
New Haven, CT 06520, USA

Jiaming Xu XU972@PURDUE.EDU
Krannert School of Management
Purdue University
West Lafayette, IN 47907, USA

Editor: Qiang Liu

Abstract

The principal submatrix localization problem deals with recovering a $K \times K$ principal submatrix of elevated mean μ in a large $n \times n$ symmetric matrix subject to additive standard Gaussian noise, or more generally, mean zero, variance one, subgaussian noise. This problem serves as a prototypical example for community detection, in which the community corresponds to the support of the submatrix. The main result of this paper is that in the regime $\Omega(\sqrt{n}) \le K \le o(n)$, the support of the submatrix can be weakly recovered (with $o(K)$ misclassification errors on average) by an optimized message passing algorithm if $\lambda = \mu^2 K^2/n$, the signal-to-noise ratio, exceeds $1/e$. This extends a result by Deshpande and Montanari previously obtained for $K = \Theta(\sqrt{n})$ and $\mu = \Theta(1)$. In addition, the algorithm can be combined with a voting procedure to achieve the information-theoretic limit of exact recovery with sharp constants for all $K \ge \frac{n}{\log n}(\frac{1}{8e} + o(1))$. The total running time of the algorithm is $O(n^2 \log n)$.

Another version of the submatrix localization problem, known as noisy biclustering, aims to recover a $K_1 \times K_2$ submatrix of elevated mean μ in a large $n_1 \times n_2$ Gaussian matrix. The optimized message passing algorithm and its analysis are adapted to the bicluster problem assuming $\Omega(\sqrt{n_i}) \le K_i \le o(n_i)$ and $K_1 \asymp K_2$. A sharp information-theoretic condition for the weak recovery of both clusters is also identified.

Keywords: Submatrix localization, biclustering, message passing, spectral algorithms computational complexity, high-dimensional statistics

1. Introduction

The problem of *submatrix detection* and *localization*, also known as *noisy biclustering* (Hartigan, 1972; Shabalin et al., 2009; Kolar et al., 2011; Butucea and Ingster, 2013; Butucea et al., 2015; Ma and Wu, 2015; Chen and Xu, 2014; Cai et al., 2017), deals with finding a submatrix with an elevated mean in a large noisy matrix, which arises in many applica-

tions such as social network analysis and gene expression data analysis. A widely studied statistical model is the following:

$$W = \mu \mathbf{1}_{C_1^*} \mathbf{1}_{C_2^*}^\top + Z, \tag{1}$$

where $\mu > 0$, $\mathbf{1}_{C_1^*}$ and $\mathbf{1}_{C_2^*}$ are indicator vectors of the row and column support sets $C_1^* \subset [n_1]$ and $C_2^* \subset [n_2]$ of cardinality K_1 and K_2, respectively, and Z is an $n_1 \times n_2$ matrix consisting of independent standard normal entries. The objective is to accurately locate the submatrix by estimating the row and column support based on the large matrix W.

For simplicity we start by considering the symmetric version of this problem, namely, locating a principal submatrix, and later extend our theoretic and algorithmic findings to the asymmetric case. To this end, consider

$$W = \mu \mathbf{1}_{C^*} \mathbf{1}_{C^*}^\top + Z, \tag{2}$$

where $C^* \subset [n]$ has cardinality K and Z is an $n \times n$ symmetric matrix with $\{Z_{ij}\}_{1 \le i \le j \le n}$ being mutually independent standard normal. Given the data matrix W, the problem of interest is to recover C^*. This problem has been investigated in (Deshpande and Montanari, 2015; Montanari et al., 2015; Hajek et al., 2017) as a prototypical example of the *hidden community problem*,[1] because the distribution of the entries exhibits a community structure, namely, $W_{i,j} \sim \mathcal{N}(\mu, 1)$ if both i and j belong to C^* and $W_{i,j} \sim \mathcal{N}(0, 1)$ if otherwise.

Assuming that C^* is drawn from all subsets of $[n]$ of cardinality K uniformly at random, we focus on the following two types of recovery guarantees.[2] Let $\xi = \mathbf{1}_{C^*} \in \{0, 1\}^n$ denote the indicator of the community. Let $\widehat{\xi} = \widehat{\xi}(A) \in \{0, 1\}^n$ be an estimator.

- We say that $\widehat{\xi}$ *exactly recovers* ξ if, as $n \to \infty$, $\mathbb{P}[\xi \neq \widehat{\xi}] \to 0$.
- We say that $\widehat{\xi}$ *weakly recovers* ξ if, as $n \to \infty$, $d(\xi, \widehat{\xi})/K \to 0$ in probability, where d denotes the Hamming distance.

The weak recovery guarantee is phrased in terms of convergence in probability, which turns out to be equivalent to convergence in mean. Indeed, the existence of an estimator satisfying $d(\xi, \widehat{\xi})/K \to 0$ is equivalent to the existence of an estimator such that $\mathbb{E}[d(\xi, \widehat{\xi})] = o(K)$ (see (Hajek et al., 2017, Appendix A) for a proof). Clearly, any estimator achieving exact recovery also achieves weak recovery; for bounded K, these two criteria are equivalent.

Intuitively, for a fixed matrix size n, as either the submatrix size K or the signal strength μ decreases, it becomes more difficult to locate the submatrix. A key role is played by the parameter

$$\lambda = \frac{\mu^2 K^2}{n},$$

which is the signal-to-noise ratio for classifying an index i according to the statistic $\sum_j W_{i,j}$, which is distributed according to $\mathcal{N}(\mu K, n)$ if $i \in C^*$ and $\mathcal{N}(0, n)$ if $i \notin C^*$. As shown in

1. A slight variation of the model in (Deshpande and Montanari, 2015; Hajek et al., 2017) is that the data matrix therein is assumed to have zero diagonal. As shown in (Hajek et al., 2017), the absence of the diagonal has no impact on the statistical limit of the problem as long as $K \to \infty$, which is the case considered in the present paper.
2. Exact and weak recovery are called strong consistency and weak consistency in (Amini et al., 2013; Mossel et al., 2015), respectively.

Appendix A, it turns out that if the submatrix size K grows linearly with n, the information-theoretic limits[3] of both weak and exact recovery are easily attainable via thresholding. To see this, note that in the case of $K \asymp n$ simply thresholding the row sums can provide weak recovery in $O(n^2)$ time provided that $\lambda \to \infty$, which coincides with the information-theoretic conditions of weak recovery as proved in (Hajek et al., 2017). Moreover, in this case, one can show that this thresholding algorithm followed by a linear-time voting procedure achieves exact recovery whenever information-theoretically possible. Thus, this paper concentrates on weak and exact recovery in the sublinear regime of

$$\Omega(\sqrt{n}) \leq K \leq o(n). \tag{3}$$

We show that an optimized message passing algorithm provides weak recovery in nearly linear – $O(n^2 \log n)$ – time if $\lambda > 1/e$. This extends the sufficient conditions obtained in (Deshpande and Montanari, 2015) for the regime $K = \Theta(\sqrt{n})$ and $\mu = \Theta(1)$.[4] Our algorithm is the same as the message passing algorithm proposed in (Deshpande and Montanari, 2015), except that we find the polynomial that maximizes the signal-to-noise ratio via Hermite polynomials instead of using the truncated Taylor series as in (Deshpande and Montanari, 2015). The proofs follow closely those in (Deshpande and Montanari, 2015), with the most essential differences described at the end of Section 2.

We observe that $\lambda > 1/e$ is much more stringent than $\lambda > \frac{4K}{n} \log \frac{n}{K}$, the information-theoretic weak recovery threshold established in (Hajek et al., 2017). It is an open problem whether any polynomial-time algorithm can provide weak recovery for $\lambda \leq 1/e$. In addition, we show that if $\lambda > 1/e$, the message passing algorithm followed by a linear-time voting procedure can provide exact recovery whenever information-theoretically possible. This procedure achieves the optimal exact recovery threshold determined in (Hajek et al., 2017) with sharp constants if $K \geq (\frac{1}{8e} + o(1)) \frac{n}{\log n}$. See Section 3.1 for a detailed comparison with information-theoretic limits.

The message passing algorithm is simpler to formulate and analyze for the principal submatrix recovery problem; nevertheless, we show in Section 5 how to adapt the message passing algorithm and its analysis to the biclustering problem. Sharp conditions for exact recovery for the biclustering problem was obtained in (Butucea et al., 2015). We show that calculations in (Butucea et al., 2015) with minor adjustments provide information-theoretic conditions for weak recovery as well. The connection between weak and exact recovery via the voting procedure described in (Hajek et al., 2017) carries over to the biclustering problem.

The analysis of the message passing algorithm is based on the moment method adopted in (Deshpande and Montanari, 2015). When the noise matrix Z is Gaussian, an alternative technique to analyze message passing algorithms is introduced in (Bayati and Montanari, 2011) and generalized by (Javanmard and Montanari, 2013). A distinct advantage of the

3. In this paper, by information-theoretic limits, we mean the sufficient and necessary conditions for attaining weak or exact recovery by any estimator, regardless of its computational cost.
4. The main results (Theorems 1 and 3) of (Deshpande and Montanari, 2015) assume $\mu = \Theta(1)$ but not $K = \Omega(\sqrt{n})$). This is because, as pointed out at the end of the proof of (Deshpande and Montanari, 2015, Theorem 3), if $K = \omega(\sqrt{N})$, then the spectral method and its proof in (Alon et al., 1998) already work. However, the state evolution analysis of message passing algorithm still assumes $K = \Theta(\sqrt{n})$ as stated in (Deshpande and Montanari, 2015, Lemma 2.2).

moment method in our context is that the Gaussian assumption can be relaxed to a subgaussian assumption. Accordingly, we introduce the following assumption.

Assumption 1 *Given $C^* \in [n]$ and $\mu > 0$, the following holds. W is an $n \times n$ symmetric matrix with $\{W_{ij}\}_{1 \le i \le j \le n}$ being mutually independent random variables. Let $Z_{ij} = W_{ij} - \mu \mathbf{I}_{\{i,j \in C^*\}}$. Then $\mathbb{E}[Z_{ij}] = 0$ for all i, j, and $\mathsf{var}(Z_{ij}) = 1$ for $(i,j) \notin C^* \times C^*$. Finally, there is a constant $\gamma > 0$ that does not depend on n such that $\mathbb{E}\left[e^{sZ_{ij}}\right] \le e^{\gamma s^2/2}$ for $s \in \mathbb{R}$, i.e. the W's and Z's are subgaussian with proxy variance γ.*

The variance of a subgaussian random variable is less than or equal to its proxy variance, so Assumption 1 implies $\gamma \ge 1$, and $\mathsf{var}(Z_{ij}) \le \gamma$ for all $i, j \in [n]$ and all $n \ge 1$. Of course, Assumption 1 holds in the Gaussian case such that the Z_{ij} are all $\mathcal{N}(0,1)$ random variables.

Notation For any positive integer n, let $[n] = \{1, \dots, n\}$. For any set $T \subset [n]$, let $|T|$ denote its cardinality and T^c denote its complement. For two sets S, T, let $S \Delta T = (S \setminus T) \cup (T \setminus S)$ denote the set difference. For an $m \times n$ matrix M, let $\|M\|$ and $\|M\|_{\mathrm{F}}$ denote its spectral and Frobenius norm, respectively. Let $\sigma_i(M)$ denote its singular values ordered decreasingly. For any $S \subset [m], T \subset [n]$, let $M_{ST} \in \mathbb{R}^{|S| \times |T|}$ denote $(M_{ij})_{i \in S, j \in T}$ and for $m = n$ abbreviate $M_S = M_{SS}$. For a vector x, let $\|x\|$ denote its Euclidean norm. We use standard big O notations, e.g., for any sequences $\{a_n\}$ and $\{b_n\}$, $a_n = \Theta(b_n)$ or $a_n \asymp b_n$ if there is an absolute constant $c > 0$ such that $1/c \le a_n/b_n \le c$. All logarithms are natural and we use the convention $0 \log 0 = 0$. Let Φ and Q denote the cumulative distribution function (CDF) and complementary CDF of the standard normal distribution, respectively. For $\epsilon \in [0,1]$, define the binary entropy function $h(\epsilon) \triangleq \epsilon \log \frac{1}{\epsilon} + (1-\epsilon) \log \frac{1}{1-\epsilon}$. We say a sequence of events $\mathcal{E}_n$ holds with high probability, if $\mathbb{P}\{\mathcal{E}_n\} \to 1$ as $n \to \infty$. Denote the Kolmogorov-Smirnov (KS) distance between distributions μ and ν by $d_{\mathrm{KS}}(\mu, \nu) \triangleq \sup_{x \in \mathbb{R}} |\mu((-\infty, x]) - \nu((-\infty, x])|$.

2. Algorithms and main results

To avoid a plethora of factors $\frac{1}{\sqrt{n}}$ in the notation, we describe the message-passing algorithm using the scaled version

$$A = \frac{1}{\sqrt{n}} W. \tag{4}$$

Under Assumption 1, the entries A_{ij} are subgaussian with proxy variance $\frac{\gamma}{n}$, mean 0 or $\frac{\mu}{\sqrt{n}}$, and variance $\frac{1}{n}$ for $(i,j) \notin C^* \times C^*$. This section presents algorithms and theoretical guarantees for the symmetric model (2). Extensions to the asymmetric case for the biclustering problem (1) are given in Section 5.2.

Let $f(\cdot, t) \colon \mathbb{R} \to \mathbb{R}$ be a scalar function for each iteration t. Let $\theta_{i \to j}^{t+1}$ denote the message transmitted from index i to index j at iteration $t+1$, which is given by

$$\theta_{i \to j}^{t+1} = \sum_{\ell \in [n] \setminus \{i,j\}} A_{\ell i} f(\theta_{\ell \to i}^t, t), \quad \forall j \ne i \in [n]. \tag{5}$$

with the initial conditions $\theta^0_{i\to j} \equiv 0$. Moreover, let θ^{t+1}_i denote index i's belief at iteration $t+1$, which is given by

$$\theta^{t+1}_i = \sum_{\ell\in[n]\setminus\{i\}} A_{\ell i} f(\theta^t_{\ell\to i}, t). \tag{6}$$

The form of (5) is inspired by belief propagation algorithms, which have the natural non-backtracking property: the message sent from i to j at time $t+1$ does not depend on the message sent from j to i at time t, thereby reducing the effect of echoes of messages sent by j.

To present an informal derivation of the *state evolution equations*, which track the asymptotic distributions of the messages, let us postulate the following assumptions: Suppose that for each fixed t, as $n\to\infty$: (a) the empirical distribution of $(\theta^t_i : i\in C^*)$ converges to $\mathcal{N}(\mu_t,\tau_t^2)$ and the empirical distribution of $(\theta^t_i : i\in[n]\setminus C^*)$ converges to $\mathcal{N}(0,\tau_t^2)$; (b) $\{\theta^t_{i\to j}\}$ are independent of A; (c) $\theta^t_{i\to j}\approx\theta^t_i$. Then it follows from (6) and $K=o(n)$ that for any $i\in C^*$,

$$\begin{aligned}
\mathbb{E}\left[\theta^{t+1}_i \mid \{\theta^t_{\ell\to i} : \ell\neq i\}\right] &\overset{(b)}{=} \sum_{\ell\in[n]\setminus\{i\}} \mathbb{E}\left[A_{\ell i}\right] f(\theta^t_{\ell\to i}, t) \\
&= \frac{\mu}{\sqrt{n}} \sum_{\ell\in C^*\setminus\{i\}} f(\theta^t_{\ell\to i}, t) \\
&\xrightarrow[n\to\infty]{(a),(c)} \sqrt{\lambda}\,\mathbb{E}\left[f(\mu_t+\tau_t Z, t)\right],
\end{aligned}$$

and for any $i\in[n]$,

$$\begin{aligned}
\mathsf{var}\left(\theta^{t+1}_i \mid \{\theta^t_{\ell\to i} : \ell\neq i\}\right) &\overset{(b)}{=} \sum_{\ell\in[n]\setminus\{i\}} \mathsf{var}\left(A_{\ell i}\right) f(\theta^t_{\ell\to i}, t)^2 \\
&= \frac{1}{n} \sum_{\ell\in[n]\setminus(C^*\cup\{i\})} f(\theta^t_{\ell\to i}, t)^2 + o(1) \\
&\xrightarrow[n\to\infty]{(a),(c)} \mathbb{E}\left[f(\tau_t Z, t)^2\right],
\end{aligned}$$

where Z represents a generic standard normal random variable. Since the conditional means and variances have deterministic limits, those are also the limits of the unconditional means and variances. Therefore, we get the following recursive equations for $t\geq 0$:

$$\mu_{t+1} = \sqrt{\lambda}\mathbb{E}\left[f(\mu_t+\tau_t Z, t)\right], \tag{7}$$

$$\tau_{t+1} = \mathbb{E}\left[f(\tau_t Z, t)^2\right], \tag{8}$$

where the initial conditions are $\mu_0=\tau_0=0$. Following (Deshpande and Montanari, 2015), we call (7) and (8) the *state evolution equations*. The heuristic derivation of state evolution equations given above is certainly not rigorous mainly due to the dependency between $\theta^t_{i\to j}$'s and A. In Section 6, we present a rigorous justification of state evolution equations via the moment method following (Deshpande and Montanari, 2015). A crucial fact that we exploit

is the non-backtracking property of the message passing rule (5), which has the effect of reducing the dependency between $\theta^t_{i\to j}$'s and A.

Suppose, for the time being, that message distributions are Gaussian with parameters accurately tracked by the state evolution equations. Then it is reasonable to estimate C^* by selecting those indices i such that θ_i^{t+1} exceeds a given threshold. More specifically, classifying an index i based on θ_i^{t+1} boils down to testing two Gaussian hypotheses with signal-to-noise ratio $\frac{\mu_{t+1}}{\tau_{t+1}}$. This gives guidance for selecting the functions $f(\cdot, t)$ based on μ_t and τ_t to maximize $\frac{\mu_{t+1}}{\tau_{t+1}}$. For $t = 0$ any choice of f is equivalent, so long as $f(0, 0) > 0$. Without loss of generality, for $t \geq 1$, we can assume that the variances are normalized, namely, $\tau_t = 1$ (e.g., we take $f(0, 0) = 1$ to make $\tau_1 = 1$) and choose $f(\cdot, t)$ to be the maximizer of

$$\max_g \{\mathbb{E}[g(\mu_t + Z)] \colon \mathbb{E}[g(Z)^2] = 1\} \tag{9}$$

where $Z \sim \mathcal{N}(0, 1)$. By change of measure, $\mathbb{E}[g(\mu_t + Z)] = \mathbb{E}[g(Z)\rho(Z)]$, where

$$\rho(x) = \frac{\mathrm{d}\mathcal{N}(\mu_t, 1)}{\mathrm{d}\mathcal{N}(0, 1)}(x) = e^{x\mu_t - \mu_t^2/2}. \tag{10}$$

Clearly, the best g aligns with ρ and we obtain

$$f(x, t) = \frac{\rho(x)}{\sqrt{\mathbb{E}[\rho^2(Z)]}} = e^{x\mu_t - \mu_t^2}. \tag{11}$$

With this optimized f, we have $\tau_t \equiv 1$ and the state evolution (7) reduces to

$$\mu_{t+1} = \sqrt{\lambda}\mathbb{E}\left[f(\mu_t + Z, t)\right] = \sqrt{\lambda} e^{\frac{\mu_t^2}{2}},$$

or, equivalently,

$$\mu_{t+1}^2 = \lambda e^{\mu_t^2}. \tag{12}$$

Therefore if $\lambda > 1/e$, then (12) has no fixed point and hence $\mu_t \to \infty$ as $t \to \infty$.

Directly carrying out the above heuristic program, however, seems challenging. To rigorously justify the state evolution equations in Section 6, we rely on the the method of moments, requiring f to be a polynomial, which prompts us to look for the best polynomial of a given degree that maximizes the signal-to-noise ratio. Denoting the corresponding state evolution by $(\widehat{\mu}_t, \widehat{\tau}_t)$, we aim to solve the following finite-degree version of (9):

$$\max\{\mathbb{E}[g(\widehat{\mu}_t + Z)] \colon \mathbb{E}[g(Z)^2] = 1, \deg(g) \leq d\}. \tag{13}$$

As shown in Lemma 7, this problem can be easily solved via Hermite polynomials, which form an orthogonal basis with respect to the Gaussian measure, and the optimal choice, denoted by $f_d(\cdot, t)$, is the best degree-d L_2-approximation of the the likelihood ratio (10), which can be obtained by normalizing the first $d + 1$ terms in the orthogonal expansion of (10). Compared to (Deshpande and Montanari, 2015, Lemma 2.3) which shows the existence of a good choice of polynomial that approximates the ideal state evolution (12) based on Taylor expansions, our approach is to find the best message-passing rule of a given

degree which results in the following state evolution that is optimal among all polynomial f of degree d:

$$\widehat{\mu}_{t+1}^2 = \lambda \sum_{k=0}^{d} \frac{\widehat{\mu}_t^{2k}}{k!}. \tag{14}$$

For any $\lambda > 1/e$, there is an explicit choice of the degree d depending only on λ denoted by $d^*(\lambda)$,[5] so that $\widehat{\mu}_t \to \infty$ as $t \to \infty$ and the state evolution (14) for fixed t correctly predicts the asymptotic behavior of the messages when $n \to \infty$. Therefore, as discussed above, $\widetilde{C}$ produced by thresholding messages θ_i^t, is likely to contain a large portion of C^*, but since $K = o(n)$, it may (and most likely will) also contain a large number of indices not in C^*. Following (Deshpande and Montanari, 2015, Lemma 2.4), we show that the power iteration[6] (a standard spectral method) in Algorithm 1 can remove a large portion of the outlier vertices in $\widetilde{C}$.

Combining message passing plus spectral cleanup yields Algorithm 1 for estimating C^* based on the messages θ_i^t, with theoretical guarantees given in Theorem 1.

Algorithm 1 Message passing

1: Input: $n, K \in \mathbb{N}$, $\mu > 0$, $A \in \mathbb{R}^{n \times n}$, $d^*, t^* \in \mathbb{N}$, and $s^* > 0$.
2: Initialize: $\theta_{i \to j}^0 = 0$ for all $i, j \in [n]$ with $i \neq j$ and $\theta_i^0 = 0$. For $t \geq 0$, define the sequence of degree-d^* polynomials $f_{d^*}(\cdot, t)$ as per Lemma 7 and $\widehat{\mu}_t$ in (14).
3: Run $t^* - 1$ iterations of message passing as in (5) with $f = f_{d^*}$ and compute $\theta_i^{t^*}$ for all $i \in [n]$ as per (6).
4: Find the set $\widetilde{C} = \{i \in [n] : \theta_i^{t^*} \geq \widehat{\mu}_{t^*}/2\}$.
5: (Cleanup via power method) Recall that $A_{\widetilde{C}}$ denotes the restriction of A to the rows and columns with index in $\widetilde{C}$. Sample u^0 uniformly from the unit sphere in $\mathbb{R}^{|\widetilde{C}|}$ and compute $u^{t+1} = A_{\widetilde{C}} u^t / \|A_{\widetilde{C}} u^t\|$ for $0 \leq t \leq \lceil s^* \log n \rceil - 1$. Let $\widehat{u} = u^{\lceil s^* \log n \rceil}$. Return $\widehat{C}$, the set of K indices i in $\widetilde{C}$ with the largest values of $|\widehat{u}_i|$.

Theorem 1 *Fix $\lambda > 1/e$. Let K and μ depend on n in such a way that $\mu^2 K^2/n \to \lambda$ and $\Omega(\sqrt{n}) \leq K \leq o(n)$ as $n \to \infty$. Suppose either C^* is deterministic with $|C^*| \equiv K$, or C^* is random such that $|C^*|/K \to 1$ in probability as $n \to \infty$. Suppose Assumption 1 holds for some $\gamma > 0$. Let $d = d^*(\lambda)$ as in (28). For every $\eta \in (0, 1)$, there exist explicit positive constants t^*, s^*, c depending on λ, η and γ such that Algorithm 1 returns $\widehat{C}$ satisfying $|\widehat{C} \Delta C^*| \leq \eta K$, with probability converging to one as $n \to \infty$, and the total time complexity is bounded by $c(\eta, \lambda, \gamma) n^2 \log n$, where $c(\eta, \lambda, \gamma) \to \infty$ as either $\eta \to 0$ or $\lambda \to 1/e$.*

Remark 2 *Algorithm 1 requires the knowledge of the parameter λ to define the sequence of polynomials $f_{d^*}(\cdot, t)$ and $\widehat{\mu}_t$, and the knowledge of the parameter K in the spectral cleanup*

5. See (28) and Remark 9 for the expression.
6. As far as statistical utility is concerned, we could replace $\widehat{u}$ produced by the power iteration by the leading singular vector of $A_{\widetilde{C}}$, but that would incur a higher time complexity because singular value decomposition in general takes $O(n^3)$ time to compute.

step. To avoid the need to know K, we can simply replace the last step of the spectral clean-up (involving choosing the K coordinates of the largest magnitude of $\widehat{u}$) by applying k-means with $k = 2$ on the set $\{|\widehat{u}_i| : i \in \widetilde{C}\}$. See Appendix C for details. With this modification, Theorem 1 continues to hold as long as λ (or a lower bound thereof) is known in order to set the degree d^* and the iteration number t^*.

Remark 3 *As pointed out in (Deshpande and Montanari, 2015, Remeark 2.5), the effective signal-to-noise ratio λ can be potentially improved by a suitable entrywise pre-processing of the observed matrix W. In particular, in (4) we let $A_{ij} = g(W_{ij})$ for some transformation function $g : \mathbb{R} \to \mathbb{R}$. The optimal transformation g in the Gaussian case for which $\gamma = 1$ is given by the maximizer of*

$$\max_g \left\{ \mathbb{E}\left[g(\mu + Z)\right] - \mathbb{E}\left[g(Z)\right] : \mathbb{E}\left[g(Z)^2\right] = \frac{1}{n} \right\}.$$

In view of (9) and (10), the optimal transformation is the scaled likelihood ratio:

$$A_{ij} = \frac{1}{\sqrt{n(e^{\mu^2} - 1)}} \left(\frac{\mathrm{d}\mathcal{N}(\mu, 1)}{\mathrm{d}\mathcal{N}(0, 1)}(W_{ij}) - 1 \right) = \frac{1}{\sqrt{n(e^{\mu^2} - 1)}} \left(e^{W_{ij}\mu - \mu^2/2} - 1 \right)$$

and the signal-to-noise ratio λ is increased to

$$\widetilde{\lambda} = \frac{K^2}{n} \left(e^{\mu^2} - 1 \right).$$

If the resulting A_{ij} is subgaussian with scale $O(1/n)$, then Theorem 1 still applies. However, even if the results extend, in the regime of $K \gg \sqrt{n}$ which we are mostly interested in, we have $\mu \to 0$ and $\widetilde{\lambda} = \lambda(1 + o(1))$, and thus pre-processing cannot boost the signal-to-noise ratio asymptotically.

After the message passing algorithm and spectral cleanup are applied in Algorithm 1, a final linear-time voting procedure is deployed to obtain weak or exact recovery, leading to Algorithm 2 next. As in (Deshpande and Montanari, 2015), we consider a threshold estimator for each vertex i based on a sum over $\widehat{C}$ given by $r_i = \sum_{j \in \widehat{C}} A_{ij}$. Intuitively, r_i can be viewed as the aggregated "votes" received by the index i in $\widehat{C}$, and the algorithm picks the set of K indices with the most significant "votes". To show that this voting procedure succeeds in weak recovery, a key step is to prove that r_i is close to $\sum_{j \in C^*} A_{ij}$. If $\mu = \Theta(1)$ as in (Deshpande and Montanari, 2015), given that $|\widehat{C} \triangle C^*| = o(K)$, the error incurred by summing over $\widehat{C}$ instead of over C^* could be bounded by truncating A_{ij} to a large magnitude. However, for $\mu \to 0$ that approach fails. Our approach is to introduce the clean-up procedure in Algorithm 2 based on the *successive withholding* method described in (Hajek et al., 2017) (see also (Condon and Karp, 2001; Mossel et al., 2014) for variants of this method). In particular, we randomly partition the set of vertices into $1/\delta$ subsets. One at a time, one subset, say S, is withheld to produce a reduced set of vertices S^c, on which we apply Algorithm 1. The estimate obtained from S^c is then used by the voting procedure to classify the vertices in S. The analysis of the two stages is decoupled because conditioned on C^*, the outcome of Algorithm 1 depends only on A_{S^c}, which is independent of A_{SS^c} used in the voting.

Algorithm 2 Message passing plus voting

1: Input: $n, K \in \mathbb{N}$, $\mu > 0$, $A \in \mathbb{R}^{n\times n}$, $\delta \in (0,1)$ with $1/\delta, n\delta \in \mathbb{N}$, $d^*, t^* \in \mathbb{N}$, and $s^* > 0$.
2: Partition $[n]$ into $1/\delta$ subsets S_k of size $n\delta$ randomly.
3: (Approximate recovery) For each $k = 1, \dots, 1/\delta$, run Algorithm 1 (message passing for approximate recovery) with input $\big(n(1-\delta), \lceil K(1-\delta)\rceil, \mu, A_{S_k^c}, d^*, t^*, s^*\big)$ which outputs $\widehat{C}_k$.
4: (Clean up) For each $k = 1, \dots, 1/\delta$ compute $r_i = \sum_{j\in\widehat{C}_k} A_{ij}$ for all $i \in S_k$ and return C', the set of K indices in $[n]$ with the largest values of r_i.

The following theorem provides a sufficient condition for the message passing plus voting cleanup procedure (Algorithm 2) to achieve weak recovery, and, if an additional sufficient condition is also satisfied, exact recovery.

Theorem 4 *Suppose K and μ depend on n in such a way that $\frac{\mu^2K^2}{n} \to \lambda$ for some fixed $\lambda > 1/e$, and $\Omega(\sqrt{n}) \le K \le o(n)$ as $n \to \infty$. Suppose Assumption 1 holds for some $\gamma > 0$ and $|C^*| = K$. Let $\delta > 0$ be such that $\lambda e(1-\delta) > 1$. Define $d^* = d^*(\lambda(1-\delta))$ as per (28). Then there exist positive constants t^*, s^*, c determined explicitly by δ, λ and γ, such that*

1. *(Weak recovery) Algorithm 2 returns C' with $|C'\Delta C^*|/K \to 0$ in probability as $n \to \infty$.*

2. *(Exact recovery) Furthermore, assume that*

$$\liminf_{n\to\infty} \frac{\sqrt{K}\mu}{\sqrt{2\log K} + \sqrt{2\log n}} > \sqrt{\gamma}. \tag{15}$$

Let $\delta > 0$ be chosen such that for all sufficiently large n,

$$\min\left\{\lambda e(1-\delta), \frac{K\mu(1-2\delta)}{\sqrt{2K\gamma\log K} + \sqrt{2K\gamma\log(n-K)} + \delta\sqrt{K}}\right\} > 1.$$

Then Algorithm 2 returns C' with $\mathbb{P}\{C' \neq C^*\} \to 0$ as $n \to \infty$.

The total time complexity is bounded by $c(\delta,\lambda,\gamma)n^2\log n$, where $c(\delta,\lambda,\gamma) \to \infty$ as $\delta \to 0$ or $\lambda \to 1/e$.

Remark 5 *Theorem 4 ensures Algorithm 2 achieves exact recovery if both (15) and $\lambda > 1/e$ hold; it is of interest to compare these two conditions. Note that*

$$\frac{\sqrt{K}\mu}{\sqrt{2\log K} + \sqrt{2\log n}} = \sqrt{\lambda e} \times \sqrt{\frac{n}{8eK\log n}} \frac{2}{(1 + \sqrt{\log K/\log n})}.$$

Hence, if $\liminf_{n\to\infty} K\log n/n \ge 1/(8e\gamma)$, then $\liminf_{n\to\infty} \log K/\log n = 1$; thus (15) implies $\lambda > 1/e$ and hence (15) alone is sufficient for Algorithm 2 to succeed. If instead $\limsup_{n\to\infty} K\log n/n \le 1/(8e\gamma)$, then $\lambda > 1/e$ implies (15) and thus $\lambda > 1/e$ alone is sufficient for Algorithm 2 to succeed. The asymptotic regime considered in (Deshpande and

Montanari, 2015) entails $K = \Theta(\sqrt{n})$, in which case the condition $\lambda > 1/e$ is sufficient for exact recovery, as shown in (Deshpande and Montanari, 2015). The idea of upgrading weak recovery to exact recovery via a local voting procedure has also appeared in (Abbe et al., 2016; Mossel et al., 2015; Abbe and Sandon, 2015; Yun and Proutiere, 2015) under the context of stochastic block models with community sizes scaling linearly in n. As shown in (Hajek et al., 2017, Corollary 4) in the Gaussian case for which $\gamma = 1$, the condition (15)*, with strict inequality replaced by greater than or equal, is necessary for exact recovery.*

We finish this section by discussing the connections and distinctions to the previous work (Deshpande and Montanari, 2015). Versions of Theorems 1 and 4 are given in (Deshpande and Montanari, 2015) for the case $K = \Theta(\sqrt{n})$ and $\mu = \Theta(1)$. We extend the range of K to $\Omega(\sqrt{n}) \le K \le o(n)$, showing that the message passing plus a cleanup procedure achieves the optimal exact recovery threshold in the Gaussian case with sharp constants if $K \ge (\frac{1}{8e} + o(1))\frac{n}{\log n}$. The algorithms and proofs are nearly the same; we comment here on the main difficulties we encountered when allowing $K/\sqrt{n} \to \infty$ and $\mu \to 0$.

First, a key ingredient in the proof of Theorem 1 is Lemma 6. Its proof is based on the moment method and a larger K requires modification of bounds from (Deshpande and Montanari, 2015) used in calculating the moments of messages, i.e., $\mathbb{E}\left[(\theta^t_{i\to j})^m\right]$ for fixed $m \in \mathbb{N}$, by a more careful counting argument. We refer the interested readers to Remark 31 right after the proof of Lemma 6 for more details.

Secondly, after the message passing algorithm and spectral cleanup are applied in Algorithm 1, a final cleanup procedure is applied to obtain weak recovery or exact recovery (when possible). As in (Deshpande and Montanari, 2015), we consider a threshold estimator for each vertex i based on a sum over $\widehat{C}$. If $K = \Theta(\sqrt{n})$ as assumed in (Deshpande and Montanari, 2015), then λ being a constant implies that the mean μ is bounded away from zero. In this case if $|\widehat{C} \triangle C^*| = o(K)$, the error incurred by summing over $\widehat{C}$ instead of over C^* could be bounded by truncating A_{ij} to a large magnitude $\bar{\rho}$ and bounding the difference of sums by $\bar{\rho}\left|C^* \triangle \widehat{C}\right| = o(K) \ll \mu K$. However, for $K \gg \sqrt{n}$ with vanishing μ this approach fails. Instead, we rely on the cleanup procedure in Algorithm 2 which entails running Algorithm 1 for $1/\delta$ times on subsampled vertices. A related difference we encounter is that if K is large enough then the condition $\lambda > 1/e$ alone is not sufficient for exact recovery, but adding the information-theoretic condition (15) suffices.

Lastly, the method of moments requires $f(\cdot, t)$ to be a polynomial so that the exponential function (11), which results in the ideal state evolution (12), cannot be directly applied. It is shown in (Deshpande and Montanari, 2015, Lemma 2.3) that for any $\lambda > 1/e$ and any threshold M there exists $d^* = d^*(\lambda, M)$ so that taking f to be the truncated Taylor series of (11) up to degree d^* results in the state evolution $\widehat{\mu}_t$ which exceeds M after some finite time $t^*(\lambda, M)$; however, no explicit formula of d^*, which is needed to instantiate Algorithm 1, is provided. Although in principle this does not pose any algorithmic problem as d^* can be found by an exhaustive search in $O(1)$ time independent of n, it is more satisfactory to find the best polynomial message passing rule explicitly which maximizes the signal-to-noise ratio for a given degree (Lemma 7) and provides an explicit formula of d^* as a function of λ only (Remark 9).

3. Statistical optimality and computational considerations

3.1 Comparison with information-theoretic limits in the Gaussian case

As noted in the introduction, in the regime $K = \Theta(n)$, a thresholding algorithm based on row sums provides weak and, if a voting procedure is also used, exact recovery whenever it is informationally possible in the Gaussian case. In this subsection, we compare the performance of the message passing algorithms to the information-theoretic limits on the recovery problem in the regime (3). Throughout this section we restrict attention to the Gaussian case, such that $Z_{ij} \sim \mathcal{N}(0,1)$ and $\gamma = 1$. Also, for converse results, we assume the true community C^* is a subset of $[n]$ of cardinality K selected uniformly at random. Notice that the comparison here takes into account the sharp constant factors. Information-theoretic limits for the biclustering problem are discussed in Section 5.1.

Weak recovery The information-theoretic threshold for weak recovery has been determined in (Hajek et al., 2017, Theorem 2), which, in the regime of (3), boils down to the following: If

$$\liminf_{n\to\infty} \frac{K\mu^2}{4\log\frac{n}{K}} > 1, \tag{16}$$

then weak recovery is possible; conversely, if weak recovery is possible, then

$$\liminf_{n\to\infty} \frac{K\mu^2}{4\log\frac{n}{K}} \geq 1. \tag{17}$$

This implies that the minimal signal-to-noise ratio for weak recovery is

$$\lambda \geq (4+\epsilon)\frac{K}{n}\log\frac{n}{K}$$

for any $\epsilon > 0$, which vanishes in the sublinear regime of $K = o(n)$. In contrast, in the regime of (3), message passing (Algorithm 1) demands a non-vanishing signal-to-noise ratio, namely, $\lambda > 1/e$, to achieve weak recovery. No polynomial-time algorithm is known to succeed if $\lambda \leq 1/e$, suggesting that computational complexity might incur a severe penalty on the statistical optimality when $K = o(n)$.

Exact recovery When the submatrix size satisfies (3), if (15) with $\gamma = 1$ holds, then exact recovery is possible; conversely, if exact recovery is possible, then

$$\liminf_{n\to\infty} \frac{\sqrt{K}\mu}{\sqrt{2\log K} + \sqrt{2\log n}} \geq 1. \tag{18}$$

See (Hajek et al., 2017, Corollary 4). This implies that the minimal signal-to-noise ratio for exact recovery is

$$\lambda \geq (2+\epsilon)\frac{K}{n}\left(\sqrt{\log n} + \sqrt{\log K}\right)^2 \tag{19}$$

for any $\epsilon > 0$. Consequently, we find that the critical submatrix size for which message passing (plus cleanup) can achieve optimal exact recovery is $\frac{n}{\log n}$. Specifically,

- $K = \omega\left(\frac{n}{\log n}\right)$. In this regime, the right hand side of (19) goes to ∞ and hence the minimal signal-to-noise ratio for exact recovery is much higher than that of weak recovery via message passing, namely, $1/e$. Thus, exact recovery can be attainable in polynomial-time by message-passing plus voting clean-up (Algorithm 2).

- $K = \Theta(\frac{n}{\log n})$. In this regime, if we let $K = \frac{\rho n}{\log n}$, the right hand side of (19) is at least $1/e$ if $\rho \geq \frac{1}{8e}$ and strictly less than $1/e$ otherwise. In view of Theorem 4, we conclude that message passing plus voting cleanup (Algorithm 2) achieves the sharp threshold of exact recovery if
$$K \geq \left(\frac{1}{8e} + o(1)\right) \frac{n}{\log n}. \tag{20}$$

- $K = o\left(\frac{n}{\log n}\right)$. In this regime, the right hand side of (19) is $o(1)$. No polynomial-time algorithm (including semidefinite programming relaxation (Hajek et al., 2016)) is known to achieve weak, let alone exact, recovery, when $\lambda = o(1)$.

A counterpart of this conclusion for the biclustering problem is obtained in Remark 20 in terms of the submatrix sizes.

3.2 Comparison with the spectral limit

It is reasonable to conjecture that $\lambda > 1$ is the spectral limit for recoverability by spectral estimation methods. This conjecture is rather vague, because it is difficult to define what constitutes spectral methods. Nevertheless, some evidence for this conjecture is provided by (Deshpande and Montanari, 2015, Proposition 1.1), which, in turn, is based on results on the spectrum of a random matrix perturbed by adding a rank-one deterministic matrix (Knowles and Yin, 2013, Theorem 2.7).

The message passing framework used in this paper itself provides some evidence for the conjecture. Indeed, if $f(x, 0) \equiv 1$ and $f(x, t) = x$ for all $t \geq 1$, the iterates θ^t are close to what is obtained by iterated multiplication by the matrix A, beginning with the all one vector, which is the power method for computation of the eigenvector corresponding to the principal eigenvalue of A.[7] To be more precise, with this linear f the message passing equation (5) can be expressed in terms of powers of the *non-backtracking matrix* $\mathbf{B} \in \mathbb{R}^{\binom{n}{2} \times \binom{n}{2}}$ associated with the matrix A, defined by $B_{ef} = A_{e_1,e_2} \mathbf{1}_{\{e_2 = f_1\}} \mathbf{1}_{\{e_1 \neq f_2\}}$, where $e = (e_1, e_2)$ and $f = (f_1, f_2)$ are directed pairs of indices. Let $\Theta^t \in \mathbb{R}^{n(n-1)}$ denote the messages on directed edges with $\Theta^t_e = \theta^t_{e_1 \to e_2}$. Then, (5) simply becomes $\Theta^t = \mathbf{B}^t \mathbf{1}$. To evaluate the performance of this method, we turn to the state evolution equations (7) and (8), which yield $\mu_t = \lambda^{t/2}$ and $\tau_t = 1$ for all $t \geq 1$. Therefore, by a simple variation of Algorithm 1 and Theorem 1, if $\lambda > 1$, the linear message passing algorithm can provide weak recovery.

For the submatrix *detection* problem, namely, testing $\mu = 0$ (pure noise) versus $\mu > 0$, as opposed to support recovery, if λ is fixed with $\lambda > 1$, a simple thresholding test based

7. If we included i, j in the summation in (5) and (6), then we would have $\theta^t = A^t \mathbf{1}$ exactly. Since the entries of A are $O_P(1/\sqrt{n})$, we expect this only incurs a small difference to the sum for finite number of iterations.

on the largest eigenvalue of the matrix A provides detection error probability converging to zero (Féral and Péché, 2007), while if $\lambda < 1$ no test based solely on the eigenvalues of A can achieve vanishing probability of error (Montanari et al., 2015). It remains, however, to establish a solid connection between the detection and estimation problem for submatrix localization for spectral methods.

3.3 Computational barriers in the Gaussian case

A recent line of work (Kolar et al., 2011; Ma and Wu, 2015; Chen and Xu, 2014; Cai et al., 2017) has uncovered a fascinating interplay between statistical optimality and computational efficiency for the *recovery* problem and the related *detection* and *estimation* problem.[8] Assuming the hardness of the planted clique problem, rigorous computational lower bounds have been obtained in (Ma and Wu, 2015; Cai et al., 2017) through reduction arguments in the Gaussian case. In particular, it is shown in (Ma and Wu, 2015) that when $K = n^\alpha$ for $0 < \alpha < 2/3$, merely achieving the information-theoretic limits of detection within any constant factor (let alone sharp constants) is as hard as detecting the planted clique; the same hardness also carries over to exact recovery in the same regime. Furthermore, it is shown that the hardness of estimating this type of matrix, which is both low-rank and sparse, highly depends on the loss function (Ma and Wu, 2015, Section 5.2). For example, for $K = \Theta(\sqrt{n})$, entry-wise thresholding attains an $O(\log n)$ factor of the minimax mean-square error; however, if the error is gauged in squared operator norm instead of Frobenius norm, attaining an $O(\sqrt{n}/\log n)$ factor of the minimax risk is as hard as solving planted clique. Similar reductions have been shown in (Cai et al., 2017) for exact recovering of the submatrix of size $K = n^\alpha$ and the planted clique recovery problem for any $0 < \alpha < 1$.

The results in (Ma and Wu, 2015; Cai et al., 2017) revealed that the difficulty of submatrix localization crucially depends on the size and planted clique hardness kicks in if $K = n^{1-\Theta(1)}$. In search of the exact phase transition point where statistical and computational limits depart, we further zoom into the regime of $K = n^{1-o(1)}$. We showed in (Hajek et al., 2016) no computational gap exists in the regime $K = \omega(n/\log n)$, since a semidefinite programming relaxation of the maximum likelihood estimator can achieve the information limit for exact recovery with sharp constants. The current paper further pushes the boundary to $K \geq \frac{n}{\log n}(\frac{1}{8e} + o(1))$, in which case the sharp information limits can be attained in nearly linear-time via message passing plus clean-up. However, as soon as $K \leq \frac{n}{\log n}(\frac{1}{8e} - \epsilon)$ for any $\epsilon > 0$, a gap emerges between the statistical limits and the sufficient condition of message passing plus clean-up, given by $\lambda > 1/e$.

4. Proofs of algorithm correctness

We first justify the state evolution equations via the following key lemma, which establishes the asymptotic normality of the empirical distribution of messages with mean and variance given by (7) and (8). A version of this lemma is proved in (Deshpande and Montanari,

8. The papers (Kolar et al., 2011; Ma and Wu, 2015; Chen and Xu, 2014; Cai et al., 2017) considered the biclustering version of the submatrix localization problem (1).

2015) by assuming $\mu = \Theta(1)$ and $K = \Theta(\sqrt{n})$. The proof is given in Section 6 using the method of moments, closely following (Deshpande and Montanari, 2015).

Lemma 6 *Let $f(\cdot, t)$ be a finite-degree polynomial for each $t \geq 0$. Let K and μ depend on n such that $\frac{K^2\mu^2}{n} \equiv \lambda$ for some $\lambda > 0$ and $\Omega(\sqrt{n}) \leq K \leq o(n)$. Suppose Assumption 1 holds for some $\gamma > 0$, and suppose either C^* is deterministic with $|C^*| \equiv K$, or C^* is random such that $|C^*|/K \to 1$ in probability as $n \to \infty$. Let $A = W/\sqrt{n}$ and set $\theta^0_{i\to j} = 0$. Consider the message passing algorithm defined by (5) and (6). Then for each fixed t, as $n \to \infty$,*

$$d_{\mathrm{KS}}\left(\frac{1}{|C^*|}\sum_{i\in C^*}\delta_{\theta_i^t},\ \mathcal{N}(\mu_t, \tau_t^2)\right) \xrightarrow{p} 0,$$
$$d_{\mathrm{KS}}\left(\frac{1}{n-|C^*|}\sum_{i\notin C^*}\delta_{\theta_i^t},\ \mathcal{N}(0, \tau_t^2)\right) \xrightarrow{p} 0,$$

where μ_t and τ_t are defined in (7) and (8), respectively; $\frac{1}{|C^*|}\sum_{i\in C^*}\delta_{\theta_i^t}$ and $\frac{1}{n-|C^*|}\sum_{i\notin C^*}\delta_{\theta_i^t}$ are the empirical distributions of θ_i^t for $i \in C^*$ and $i \notin C^*$, respectively.

Next we prove Theorems 1-4. Lemma 6 implies that if $i \in C^*$, then $\theta_i^t \sim \mathcal{N}(\mu_t, \tau_t^2)$; if $i \notin C^*$, then $\theta_i^t \sim \mathcal{N}(0, \tau_t^2)$. Ideally, one would pick the optimal $f(x,t) = e^{\mu_t(x-\mu_t)}$ which result in the optimal state evolution $\mu_{t+1} = \sqrt{\lambda}e^{\mu_t^2/2}$ and $\tau_t = 1$ for all $t \geq 1$. Furthermore, if $\lambda > 1/e$, then $\mu_t \to \infty$ as $t \to \infty$, and thus we can hope to estimate C^* by selecting the indices i such that θ_i^t exceeds a certain threshold. The caveat is that Lemma 6 needs f to be a polynomial of finite degree. Next we proceed to find the best degree-d polynomial for iteration t, denoted by $f_d(\cdot, t)$, which maximizes the signal to noise ratio.

Recall that the Hermite polynomials $\{H_k : k \geq 0\}$ are the orthogonal polynomials with respect to the standard normal distribution (cf. (Szegö, 1975, Section 5.5)), given by

$$H_k(x) = (-1)^k\frac{\varphi^{(k)}(x)}{\varphi(x)} = \sum_{i=0}^{\lfloor k/2\rfloor}(-1)^i(2i-1)!!\binom{k}{2i}x^{k-2i}, \tag{21}$$

where φ denotes the standard normal density and $\varphi^{(k)}(x)$ is its k-th derivative; in particular, $H_0(x) = 1, H_1(x) = x, H_2(x) = x^2 - 1$. Furthermore, $\deg(H_k) = k$ and $\{H_0, \ldots, H_d\}$ span all polynomials of degree at most d. For $Z \sim \mathcal{N}(0,1)$, $\mathbb{E}[H_m(Z)H_n(Z)] = m!\delta_{m=n}$ and $\mathbb{E}[H_k(\mu + Z)] = \mu^k$ for all $\mu \in \mathbb{R}$; hence the relative density $\frac{\mathrm{d}\mathcal{N}(\mu,1)}{\mathrm{d}\mathcal{N}(0,1)}(x) = e^{\mu x - \mu^2/2}$ admits the following expansion:

$$e^{\mu x - \mu^2/2} = \sum_{k=0}^{\infty}H_k(x)\frac{\mu^k}{k!}. \tag{22}$$

Truncating and normalizing the series at the first $d+1$ terms immediately yields the solution to (13) as the best degree-d L_2-approximation to the relative density, described as follows:

Lemma 7 *Fix $d \in \mathbb{N}$ and define $\widehat{\mu}_t$ according to the iteration (14) with $\widehat{\mu}_0 = 0$, namely,*

$$\widehat{\mu}_{t+1}^2 = \lambda G_d(\widehat{\mu}_t^2), \qquad G_d(\mu) = \sum_{k=0}^{d}\frac{\mu^k}{k!}. \tag{23}$$

Define

$$f_d(x,t) = \sum_{k=0}^{d} a_k H_k(x), \tag{24}$$

where $a_k \triangleq \frac{\widehat{\mu}_t^k}{k!}(\sum_{k=0}^{d} \frac{\widehat{\mu}_t^{2k}}{k!})^{-1/2}$. *Then* $f_d(\cdot,t)$ *is the unique maximizer of (13) and the state evolution (7) and (8) with* $f = f_d$ *coincides with* $\tau_t = 1$ *and* $\mu_t = \widehat{\mu}_t$. *Furthermore, for any* $d \geq 2$ *the equation*

$$G_d(a) = aG_{d-1}(a) \tag{25}$$

has a unique positive solution, denoted by a_d^*. *Let*

$$\lambda_d^* = \frac{1}{G_{d-1}(a_d^*)} \tag{26}$$

and define $\lambda_1^* = 1$. *Then*

1. *for any* $d \in \mathbb{N}$ *and any* $\lambda > \lambda_d^*$, $\widehat{\mu}_t \to \infty$ *as* $t \to \infty$ *and hence for any* $M > 0$,

$$t^*(\lambda, M) = \inf\{t : \widehat{\mu}_t > M\} \tag{27}$$

   is finite;

2. $\lambda_d^* \downarrow 1/e$ *monotonically as* $d \to \infty$ *according to* $\lambda_d^* = 1/e + \frac{1/e^2 + o(1)}{(d+1)!}$.

Remark 8 *The best affine update gives* $\lambda_1^* = 1$; *for the best quadratic update,* $a_2^* = \sqrt{2}$ *and hence* $\lambda_2^* = \frac{1}{1+\sqrt{2}} \approx 0.414$. *More values of the threshold are given below, which converges to* $1/e \approx 0.368$ *rapidly.*

d	1	2	3	4	5
λ_d^*	1	0.414	0.376	0.369	0.368

Remark 9 *Let*

$$d^*(\lambda) = \inf\{d \in \mathbb{N} : \lambda_d^* < \lambda\}, \tag{28}$$

which is finite for any $\lambda > 1/e$. *Then for any* $d \geq d^*$, $\widehat{\mu}_t \to \infty$ *as* $t \to \infty$. *As* λ *approaches the critical value* $1/e$, *the degree* $d^*(\lambda)$ *blows up according to* $d^*(\lambda) = \Theta(\log \frac{1}{\lambda e - 1} / \log\log \frac{1}{\lambda e - 1})$, *as a consequence of the last part of Lemma 7.*

Remark 10 (Best affine message passing) *For* $d = 1$, *the best state evolution is given by*

$$\widehat{\mu}_{t+1}^2 = \lambda(1 + \widehat{\mu}_t^2)$$

and the corresponding optimal update rule is

$$f_1(x,t) = \frac{1 + \widehat{\mu}_t x}{\sqrt{1 + \widehat{\mu}_t^2}}.$$

This is strictly better than $f(x,t) = x$ *described in Section 3.2 which gives* $\widehat{\mu}_{t+1}^2 = \lambda \widehat{\mu}_t^2$; *nevertheless, in order to have* $\widehat{\mu}_t \to \infty$ *we still need to assume the spectral limit* $\lambda \geq 1$.

Proof [Proof of Lemma 7] Note that any degree-d polynomial g can be written in terms of the linear combination:

$$f_d(x,t) = \sum_{k=0}^{d} c_k H_k(x),$$

where the coefficients $\{c_k\}$ satisfy $\mathbb{E}[g^2(Z)] = \sum_{k=0}^{d} k! c_k^2 = 1$. By a change of measure, $\mathbb{E}[g(\widehat{\mu}_t + Z)] = \mathbb{E}[g(Z)e^{\widehat{\mu}_t Z - \widehat{\mu}_t^2/2}] = \sum_{k=0}^{d} c_k \widehat{\mu}_t^k$, in view of the orthogonal expansion (22). Thus, to solve the maximization problem (13), it is equivalent to solving

$$\max_{c_k} \left\{ \sum_{k=0}^{d} c_k \widehat{\mu}_t^k : \sum_{k=0}^{d} k! c_k^2 = 1 \right\}.$$

By Cauchy-Schwarz inequality, the optimal coefficients and the optimal polynomial $f_d(\cdot, t)$ are given by (24), resulting in the following state evolution

$$\widehat{\mu}_{t+1} = \sqrt{\lambda} \max\{\mathbb{E}[g(\widehat{\mu}_t + Z)] \colon \mathbb{E}[g(Z)^2] = 1, \deg(g) \le d\} = \left(\lambda \sum_{k=0}^{d} \frac{\widehat{\mu}_t^{2k}}{k!} \right)^{1/2},$$

which is equivalent to (23).

Next we analyze the behavior of the iteration (23). The case of $d = 1$ follows from the obvious fact that $\widehat{\mu}_{t+1}^2 = \lambda(\widehat{\mu}_t^2 + 1)$ diverges if and only if $\lambda \ge 1$. For $d \ge 2$, note that G_d is a strictly convex function with $G_d(0) = 1$ and $G_d' = G_{d-1}$. Also, $(G_d(a) - aG_{d-1}(a))' = -aG_d''(a) < 0$. Thus, $G_d(a) - aG_{d-1}(a)$ is strictly decreasing on $a > 0$ with value $\frac{1}{d!}$ at $a = 1$ and limit $-\infty$ as $a \to \infty$, so (25) has a unique positive solution a_d^* and it satisfies $a_d^* > 1$. Furthermore, $(G_d(a) - aG_{d-1}(a))'\big|_{a=1} = -\sum_{k=0}^{d-2} \frac{1}{k!}$, so by Taylor's theorem,

$$G_d(a) - aG_{d-1}(a) = \frac{1}{d!} - (a-1) \sum_{k=0}^{d-2} \frac{1}{k!} + O((a-1)^2),$$

yielding

$$a_d^* = 1 + \frac{1}{d! \sum_{k=0}^{d-2} \frac{1}{k!}} + O(1/(d!)^2).$$

Consider next the values of λ such that $\widehat{\mu}_t$ diverges. For very large λ, $G_d(a)$ dominates a/λ pointwise and $\widehat{\mu}_t$ diverges. The critical value of λ is when $G_d(a)$ and a/λ meet tangentially, namely,

$$\lambda G_{d-1}(a) = 1, \quad \lambda G_d(a) = a,$$

whose solution is given by $a = a_d^*$ and $\lambda = \lambda_d^*$, where

$$\begin{aligned}
\lambda_d^* &\triangleq \frac{1}{G_{d-1}(a_d^*)} = \frac{1}{G_{d-1}(1) + G_{d-1}'(1)(a_d^* - 1) + O((a_d^* - 1)^2)} \\
&= \frac{1}{\sum_{k=0}^{d} \frac{1}{k!} + O(1/(d!)^2)} = 1/e + \frac{\sum_{k=d+1}^{\infty} 1/k! + O(1/(d!)^2)}{e \sum_{k=0}^{d} 1/k!} \\
&= 1/e + \frac{1/e^2 + o(1)}{(d+1)!}.
\end{aligned}$$

Thus, λ_d^* is the minimum value such that for all $\lambda > \lambda_d^*$, $\lambda G_d(a) > a$ for all $a > 0$, so that starting from any $\widehat{\mu}_t \geq 0$ we have $\widehat{\mu}_t \to \infty$ monotonically. The fact λ_d^* is decreasing in d follows from the fact G_d is pointwise increasing in d. ■

Lemmas 6 and 7 immediately imply the following partial recovery results.

Lemma 11 *Assume that $\lambda > 1/e$ and $\Omega(\sqrt{n}) \leq K \leq o(n)$. Fix any $\epsilon \in (0,1)$. Let $M = \sqrt{8\log(1/\epsilon)}$ and run the message passing algorithm for t iterations with $f = f_{d^*}$, $d^* = d^*(\lambda)$ as in (28), and $t = t^*(\lambda, M)$ as in (27). Let $\widetilde{C} = \{i : \theta_i^{t^*} \geq \widehat{\mu}_{t^*}/2\}$. Then with probability converging to one as $n \to \infty$,*

$$\frac{1}{K}|\widetilde{C} \cap C^*| \geq 1 - \epsilon \tag{29}$$

$$K(1-\epsilon) \leq |\widetilde{C}| \leq n\epsilon. \tag{30}$$

Proof Notice that

$$|\widetilde{C} \cap C^*| = \sum_{i \in C^*} \mathbf{1}_{\left\{\theta_i^{t^*} \geq \widehat{\mu}_{t^*}/2\right\}}.$$

By the choice of $f = f_d$ in (24), we have $\tau_t = 1$ for all $t \geq 1$. It follows from Lemma 6 that

$$\lim_{n\to\infty} \frac{1}{K}|\widetilde{C} \cap C^*| = \mathbb{P}\left\{\widehat{\mu}_{t^*} + Z \geq \widehat{\mu}_{t^*}/2\right\}, \tag{31}$$

where the convergence is in probability. Notice that we have used $d = d^*(\lambda)$ and $t = t^*(\lambda, M)$ defined by (28) and (27) in Lemma 7. Thus $\widehat{\mu}_{t^*} \geq M = \sqrt{8\log(1/\epsilon)}$ and

$$\mathbb{P}\left\{\widehat{\mu}_{t^*} + Z \leq \widehat{\mu}_{t^*}/2\right\} = Q(\widehat{\mu}_{t^*}/2) \leq e^{-\widehat{\mu}_{t^*}^2/8} \leq \epsilon,$$

which, in view of (31), implies (29) with probability converging to one as $n \to \infty$. Similarly, Lemma 6 implies that in probability

$$\lim_{n\to\infty} \frac{1}{n}|\widetilde{C} \backslash C^*| = \mathbb{P}\left\{Z \geq \widehat{\mu}_{t^*}/2\right\} = Q(\widehat{\mu}_{t^*}/2) \leq \epsilon. \tag{32}$$

Since $K = o(n)$, we have $\mathbb{P}\{K(1-\epsilon) \leq |\widetilde{C}| \leq n\epsilon\} \to 1$. ■

Although $\widetilde{C}$ contains a large portion of C^*, since $|\widetilde{C}|$ is linear in n with high probability, i.e., $|\widetilde{C}|/n \to Q(\widehat{\mu}_{t^*}/2)$ by Lemma 6, it is bound to contain a large number of outlier indices. The next lemma, closely following (Deshpande and Montanari, 2015, Lemma 2.4), shows that given the conclusion of Lemma 11, the power iteration in Algorithm 1 can remove most of the outlier indices in $\widetilde{C}$. Its proof is presented in Appendix B.

Lemma 12 *Suppose $\lambda = \frac{\mu^2 K^2}{n} \geq 1/e$, $K \to \infty$, $|C^*|/K \to 1$ in probability, and $\widetilde{C}$ is a set (possibly depending on A) such that (29) – (30) hold for some $0 < \epsilon < \epsilon_0$, where $0 < \epsilon_0 \leq 1/2$ is determined by $1 - \epsilon_0 = 8e\sqrt{4\gamma h(\epsilon_0) + 10\gamma\epsilon_0}$. Let*

$$s^* = \frac{2}{\log(\sqrt{\lambda}(1-\epsilon)/(8\sqrt{4e\gamma h(\epsilon) + 10e\gamma\epsilon}))}, \tag{33}$$

where $h(\epsilon) \triangleq \epsilon \log \frac{1}{\epsilon} + (1-\epsilon)\log \frac{1}{1-\epsilon}$ *is the binary entropy function. Then* $\widehat{C}$ *with* $|\widehat{C}| \equiv K$ *produced by Algorithm 1 returns* $|\widehat{C} \Delta C^*| \le \eta(\epsilon, \lambda, \gamma) K$*, with probability converging to one as* $n \to \infty$*, where*

$$\eta(\epsilon, \lambda, \gamma) = 2\epsilon + e\gamma \frac{4608 h(\epsilon) + 11520\epsilon}{\lambda(1-\epsilon)^2}. \tag{34}$$

Proof [Proof of Theorem 1] Given $\eta \in (0,1)$, choose an arbitrary $\epsilon \in (0, \epsilon_0)$ such that $\eta(\epsilon, \lambda, \gamma)$ defined in (34) is at most η. With t^* specified in Lemma 11 and s^* specified in Lemma 12, the probabilistic performance guarantee in Theorem 1 readily follows by combining Lemmas 11 and 12. The time complexity of Algorithm 1 follows from the fact that for both the BP algorithm and the power method each iteration has complexity $O(n^2)$ and Algorithm 1 entails running BP and the power method for t^* and $\lceil s^* \log n \rceil$ iterations respectively; both t^* and s^* are constants depending only on η, λ, and γ. ■

Proof [Proof of Theorem 4] (Weak recovery) Fix $k \in [1/\delta]$ and let $C_k^* = C^* \cap S_k^c$. Define the $n(1-\delta) \times n(1-\delta)$ matrix $A_k \triangleq A_{S_k^c}$, which corresponds to the submatrix localization problem for a planted community C_k^* whose size has a hypergeometric distribution, resulting from sampling without replacement, with parameters $(n, K, (1-\delta)n)$ and mean $(1-\delta)K$. By a result of (Hoeffding, 1963), the distribution of $|C_k^*|$ is convex order dominated by the distribution that would result from sampling with replacement, namely, the Binom $\big(n(1-\delta), \frac{K}{n}\big)$ distribution. In particular, Chernoff bounds for Binom$(n(1-\delta), \frac{K}{n})$ also hold for $|C_k^*|$, so $|C_k^*|/((1-\delta)K) \to 1$ in probability as $n \to \infty$. Note that $\frac{((1-\delta)K)^2 \mu^2}{n(1-\delta)} \to \lambda(1-\delta)$ and $\lambda(1-\delta)e > 1$ by the choice of δ. Let $d^*(\lambda(1-\delta))$ be given in (28), i.e.,

$$d^*(\lambda(1-\delta)) = \inf\{d \in \mathbb{N} : \lambda_d^* < \lambda(1-\delta)\}.$$

Choose an arbitrary $\epsilon \in (0, \epsilon_0)$ to satisfy $\eta(\epsilon, \lambda(1-\delta), \gamma) \le \delta$, i.e.,

$$2\epsilon + e\gamma \frac{4608 h(\epsilon) + 11520\epsilon}{\lambda(1-\delta)(1-\epsilon)^2} \le \delta.$$

Define $\widehat{\mu}_t$ recursively according to (14) with λ replaced by $\lambda(1-\delta)$ and $\widehat{\mu}_0 = 0$, i.e.,

$$\widehat{\mu}_{t+1}^2 = \lambda(1-\delta) \sum_{k=0}^{d} \frac{\widehat{\mu}_t^{2k}}{k!}.$$

Define $t^*(\delta, \lambda, \gamma)$ according to (27) with $M = 8\log(1/\epsilon)$, and $s^*(\delta, \lambda, \gamma)$ according to (33) with λ replaced by $\lambda(1-\delta)$. Then Theorem 1 with n and K replaced by $n(1-\delta)$ and $\lceil K(1-\delta) \rceil$ implies that as $n \to \infty$,

$$\mathbb{P}\left\{ |\widehat{C}_k \triangle C_k^*| \le \delta K \text{ for } 1 \le k \le 1/\delta \right\} \to 1.$$

Given $(C_k^*, \widehat{C}_k)$, each of the random variables $r_i \sqrt{n}$ for $i \in S_k$ is conditionally subgaussian with proxy variance at most $K\gamma$. Furthermore, on the event, $\mathcal{E}_k = \{|\widehat{C}_k \triangle C_k^*| \le \delta K\}$,

$$|\widehat{C}_k \cap C_k^*| \ge |\widehat{C}_k| - |\widehat{C}_k \triangle C_k^*| = \lceil K(1-\delta) \rceil - |\widehat{C}_k \triangle C_k^*| \ge K(1-2\delta).$$

Therefore, on the event $\mathcal{E}_k$, for $i \in S_k \cap C^*$, $r_i\sqrt{n}$ has mean greater than or equal to $K(1-2\delta)\mu$, and for $i \in S_k \backslash C^*$, r_i has mean zero.

Define the following set by thresholding

$$C_o' = \{i \in [n] : r_i \geq (1-2\delta)\sqrt{\lambda}/2\}$$

The number of indices in S_k incorrectly classified by $C_o' \cap S_k$ satisfies (use $|S_k| = \delta n$):

$$\mathbb{E}\left[\left|(C_o' \cap S_k)\Delta(C^* \cap S_k)\right|\right] \leq \delta n e^{-\Omega(n/K)},$$

where the last inequality follows because r_i is subgaussian with proxy variance at most $\gamma K/n$. Summing over $k \in [1/\delta]$ yields $\mathbb{E}\left[|C_o'\Delta C^*|\right] \leq ne^{-\Omega(n/K)}$. By Markov's inequality,

$$\mathbb{P}\left\{|C_o'\Delta C^*| \geq K^2/n\right\} \leq \frac{n^2}{K^2}e^{-\Omega(n/K)} \overset{K=o(n)}{=} o(1).$$

Instead of C_o', Algorithm 2 outputs C' which selects the K indices in $[n]$ with the largest values of r_i. Applying the same argument as that at the end of the proof of Lemma 12, we get $|C^*\triangle C'| \leq 2|C^*\triangle C_o'| + ||C^*| - K|$, and hence $|C^*\triangle C'|/K \to 0$ in probability.

(Exact recovery) By the union bound and Chernoff's bound for subgaussian random variables, the maximum of m subgaussian random variables X_i with zero mean and proxy variance at most γ satisfies that

$$\begin{aligned}\mathbb{P}\left\{\max_{1\leq i\leq m} X_i \geq \sqrt{\gamma}\left(\sqrt{2\log m} + t\right)\right\} &\leq m\exp\left(-\left(\sqrt{2\log m}+t\right)^2/2\right)\\ &= \exp\left(-t\sqrt{2\log m} - t^2/2\right).\end{aligned}$$

It follows that $\max_{1\leq i\leq m} X_i$ is at most $\sqrt{2\gamma\log m} + o_P(1)$ as $m \to \infty$. Also, for $k \in [1/\delta]$, $|S_k \cap C^*| \leq |C^*| = K$ and $|S_k\backslash C^*| \leq |[n]\backslash C^*| = n - K$. Therefore,

$$\min_{i\in S_k\cap C^*} r_i\sqrt{n} \geq K(1-2\delta)\mu - \sqrt{2K\gamma\log K} + o_P(\sqrt{K}) \tag{35}$$

$$\max_{j\in S_k\backslash C^*} r_i\sqrt{n} \leq \sqrt{2K\gamma\log(n-K)} + o_P(\sqrt{K}). \tag{36}$$

Since k ranges over a finite number of values, namely, $[1/\delta]$, (35) and (36) continue to hold with left-hand sides replaced by $\min_{i\in C^*} r_i\sqrt{n}$ and $\max_{j\notin C^*} r_i\sqrt{n}$, respectively. Therefore, by the choice of δ, $\min_{i\in C^*} r_i\sqrt{n} > \max_{j\in[n]\backslash C^*} r_i\sqrt{n}$ with probability converging to one as $n \to \infty$ and so $C' = C^*$ with probability converging to one as well.

(Time complexity) The running time of Algorithm 2 is dominated by invoking Algorithm 1 for a constant number, $1/\delta$, of times, and the number of iterations within Algorithm 1 is $(t^* + s^*\log n)n^2$, with both t^* and $s^* \to \infty$ as either $\delta \to 0$ or $\lambda \to 1/e$. In particular, the threshold comparisons require $O(n^2)$ computations. Thus, the total complexity of Algorithm 2 is as stated in the theorem. ■

5. The Gaussian biclustering problem

We return to the biclustering problem where the goal is to locate a submatrix whose row and column support need not coincide. Consider the model (1) parameterized by $(n_1, n_2, K_1, K_2, \mu)$ indexed by a common n with $n \to \infty$. In Section 5.1 we present the information limits for weak and exact recovery for the Gaussian bicluster model. The sharp conditions given for exact recovery are from (Butucea et al., 2015), and calculations from (Butucea et al., 2015) with minor adjustment provide conditions for weak recovery as well. Section 5.2 shows how the optimized message passing algorithm and its analysis can be extended from the symmetric case to the asymmetric case for biclustering and compares its performance to the fundamental limits. As originally observed in (Hajek et al., 2017) for recovering the principal submatrix, the connection between weak and exact recovery via the voting procedure extends to the biclustering problem as well. Note that for the sake of simplicity, we focus on Gaussian biclustering where the noise matrix Z is Gaussian; the results can be readily extended to the case where Z has subguassian entries as we did in the symmetric case.

5.1 Information-theoretic limits for Gaussian biclustering

Information-theoretic conditions ensuring exact recovery of both C_1^* and C_2^* by the maximum likelihood estimator (MLE), i.e.,

$$(\widehat{C}_1^{\mathsf{MLE}}, \widehat{C}_2^{\mathsf{MLE}}) = \arg\max_{\substack{|C_1|=K_1 \\ |C_2|=K_2}} \sum_{\substack{i\in C_1 \\ j\in C_2}} W_{ij}$$

are obtained in (Butucea et al., 2015). While (Butucea et al., 2015) does not focus on conditions for weak recovery, the calculations therein combined with the voting procedure for exact recovery described in (Hajek et al., 2017) in fact resolve the information limits for both weak and exact recovery in the bicluster Gaussian model. Throughout this section we assume that $K_i = o(n_i)$ for $i = 1, 2$. For the converse results we assume C_i^* is a subset of $[n_i]$ of cardinality K_i selected uniformly at random for $i = 1, 2$, with C_1^* independent of C_2^*. Let

$$\lambda_i = \frac{K_i^2 \mu^2}{n_i}, \quad \text{for } i = 1, 2.$$

Theorem 13 (Weak recovery thresholds for Gaussian biclustering) *If*

$$\liminf_{n\to\infty} \frac{\mu\sqrt{K_1 K_2}}{\sqrt{2(K_1 \log(n_1/K_1) + K_2 \log(n_2/K_2))}} > 1, \tag{37}$$

then both C_1^* and C_2^* can be weakly recovered by the MLE. Conversely, if both C_1^* and C_2^* can be weakly recovered by some estimator, then

$$\liminf_{n\to\infty} \frac{\mu\sqrt{K_1 K_2}}{\sqrt{2(K_1 \log(n_1/K_1) + K_2 \log(n_2/K_2)}} \geq 1. \tag{38}$$

If C_2^* (C_1^*) can be weakly recovered, one can further obtain exact recovery of C_1^* (C_2^*) via a voting cleanup procedure similar to Algorithm 2; it uses the method of successive withholding. We give the voting procedure in Algorithm 3 for exact recovery of C_1^* based on weak recovery of C_2^*; exact recovery of C_2^* based on weak recovery of C_1^* is analogous.

Algorithm 3 Weak recovery of C_2^* plus cleanup for exact recovery of C_1^*

1: Input: $n_1, n_2, K_1, K_2 \in \mathbb{N}$, $\mu > 0$, $A \in \mathbb{R}^{n_1 \times n_2}$, $\delta \in (0,1)$ with $1/\delta, n_1\delta \in \mathbb{N}$.
2: (Partition) Partition $[n_1]$ into $1/\delta$ subsets S_k of size $n_1\delta$ randomly.
3: (Approximate recovery) For each $k = 1, \dots, 1/\delta$, let A_k denote the restriction of A to the rows with index in S_k^c, run an estimator capable of weak recovery of C_2^* with input $(n_1(1-\delta), n_2, \lceil K_1(1-\delta) \rceil, K_2, \mu, A_k)$ which outputs $\widehat{C}_{2k}$.
4: (Clean up) For each $k = 1, \dots, 1/\delta$ compute $r_i = \sum_{j \in \widehat{C}_{2k}} A_{ij}$ for all $i \in S_k$ and return C_1', the set of K indices in $[n_1]$ with the largest values of r_i.

Theorem 14 (Exact recovery thresholds for Gaussian biclustering)
If for some small $\delta > 0$, C_2^* can be weakly recovered even if a fraction δ of the rows of the matrix are hidden, and if

$$\liminf_{n\to\infty} \frac{\sqrt{K_2}\mu}{\sqrt{2\log K_1} + \sqrt{2\log n_1}} > 1, \tag{39}$$

then C_1^* can be exactly recovered by the voting procedure. Conversely, if C_1^* can be exactly recovered by some estimator, then

$$\liminf_{n\to\infty} \frac{\sqrt{K_2}\mu}{\sqrt{2\log K_1} + \sqrt{2\log n_1}} \geq 1. \tag{40}$$

Similarly, if for some small $\delta > 0$, C_1^* can be weakly recovered even if a fraction δ of the columns of the matrix are hidden, and if

$$\liminf_{n\to\infty} \frac{\sqrt{K_1}\mu}{\sqrt{2\log K_2} + \sqrt{2\log n_2}} > 1, \tag{41}$$

then C_2^* can be exactly recovered by the voting procedure. Conversely, if C_2^* can be exactly recovered by some estimator, then

$$\liminf_{n\to\infty} \frac{\sqrt{K_1}\mu}{\sqrt{2\log K_2} + \sqrt{2\log n_2}} \geq 1. \tag{42}$$

The proofs of Theorems 13 and 14 are given in Appendix D. The sufficient conditions involving δ in Theorem 14 require a certain robustness of the estimator for weak recovery. If the rows indexed by a set S, with $S \subset [n_1]$ and $|S| = \delta n_1$, are hidden, then the observed matrix has dimensions $n_1(1-\delta) \times n_2$ and the planted submatrix has $K_1 - |S \cap C_1^*| \approx K_1(1-\delta)$ rows and K_2 columns. It is shown in (Hajek et al., 2017, Section IV.B) that the MLE has this robustness property for weak recovery of a principal submatrix, and a similar extension can be established for weak recovery for biclustering. The estimator used is the MLE based on the assumption that the submatrix to be found has shape $K_1(1-\delta) \times K_2$. With that extension in hand, the following corollary is a consequence of the two theorems, and it recovers the main result of (Butucea et al., 2015).

Corollary 15 *If* (37), (39), *and* (41) *hold, then* C_1^* *and* C_2^* *can both be exactly recovered by the MLE. Conversely, if exact recovery is possible, then* (38), (40), *and* (42) *hold.*

We conclude this subsection with a few remarks on Theorems 13 and 14:

1. If $n_1 = n_2$ and $K_1 = K_2$, the sufficient conditions and the necessary conditions for weak and for exact recovery, respectively, are identical to those in (Hajek et al., 2017) for the recovery of a $K \times K$ principal submatrix with elevated mean, in a symmetric $n \times n$ Gaussian matrix. Basically, in the bicluster problem the data matrix provides roughly twice the information (because the matrix is not symmetric) and there is twice the information to be learned, namely C_1^* and C_2^* instead of only C^*, and the factors of two cancel to yield the same conditions. It therefore follows from (Hajek et al., 2017, Remark 7), that if $n_1 = n_2$ and $K_1 = K_2 \le n_1^{1/9}$, then (37) implies (39) and (41); in this regime, (37) alone is the sharp condition for both weak and exact recovery.

2. If $\lambda_i = \frac{K_i^2 \mu^2}{n_i}$ are two fixed positive constants and if $K_1 \asymp K_2$, then (37) holds for all sufficiently large n, so weak recovery is information theoretically possible. In contrast, our proof that the optimized message passing algorithm provides weak recovery in this regime requires $(\lambda_1, \lambda_2) \in \mathcal{G}$, where $\mathcal{G}$ is defined in (52) in the next subsection.

5.2 Message passing algorithm for the Gaussian biclustering model

Suppose $n_i \to \infty$ and $\Omega(\sqrt{n_i}) \le K_i \le o(n_i)$ for $i \in \{0, 1\}$, as $n \to \infty$. The belief propagation algorithm and our analysis of it for recovery of a single set of indices can be naturally adapted to the biclustering model.

Let $f(\cdot, t) : \mathbb{R} \to \mathbb{R}$ be a scalar function for each iteration t. To be definite, we shall describe the algorithm such that at each iteration, the messages are passed either from the row indices to the column indices, or vice-versa, but not both. The messages are defined as follows for $t \ge 0$:

$$(t \text{ even}) \qquad \theta_{i\to j}^{t+1} = \frac{1}{\sqrt{n_2}} \sum_{\ell \in [n_2]\backslash\{j\}} W_{i\ell} f(\theta_{\ell\to i}^t, t), \quad \forall i \in [n_1], j \in [n_2] \tag{43}$$

$$(t \text{ odd}) \qquad \theta_{j\to i}^{t+1} = \frac{1}{\sqrt{n_1}} \sum_{\ell \in [n_1]\backslash\{i\}} W_{\ell j} f(\theta_{\ell\to j}^t, t), \quad \forall j \in [n_2], i \in [n_1], \tag{44}$$

with the initial condition $\theta_{\ell\to i}^0 = 0$ for $(\ell, i) \in [n_2] \times [n_1]$. Moreover, let the aggregated beliefs be given by

$$(t \text{ even}) \qquad \theta_i^{t+1} = \frac{1}{\sqrt{n_2}} \sum_{\ell \in [n_2]} W_{i\ell} f(\theta_{\ell\to i}^t, t), \quad \forall i \in [n_1] \tag{45}$$

$$(t \text{ odd}) \qquad \theta_j^{t+1} = \frac{1}{\sqrt{n_1}} \sum_{\ell \in [n_1]} W_{\ell j} f(\theta_{\ell\to j}^t, t), \quad \forall j \in [n_2]. \tag{46}$$

Recall $\lambda_i = \frac{K_i^2 \mu^2}{n_i}$ for $i = 1, 2$. Suppose as $n \to \infty$, for t even (odd), θ_i^t is approximately $\mathcal{N}(\mu_t, \tau_t)$ for $i \in C_1^*$ ($i \in C_2^*$) and $\mathcal{N}(0, \tau_t)$ for $i \in [n_1]\backslash C_1^*$ ($i \in [n_2]\backslash C_2^*$). Then similar to

the symmetric case, the update equations of message passing and the fact that $\theta^t_{i\to j} \approx \theta^t_i$ for all i, j suggest the following state evolution equations for $t \geq 0$:

$$\mu_{t+1} = \begin{cases} \sqrt{\lambda_2}\mathbb{E}\left[f(\mu_t + \tau_t Z, t)\right] & t \text{ even} \\ \sqrt{\lambda_1}\mathbb{E}\left[f(\mu_t + \tau_t Z, t)\right] & t \text{ odd} \end{cases} \tag{47}$$

$$\tau_{t+1} = \mathbb{E}\left[f(\tau_t Z, t)^2\right]. \tag{48}$$

The optimal choice of f for maximizing the signal-to-noise ratio $\frac{\mu_{t+1}}{\tau_{t+1}}$ is again $f(x,t) = e^{x\mu_t - \mu_t^2}$. With this optimized f, we have $\tau_{t+1} = 1$ and the state evolution equations reduce to

$$\mu^2_{t+1} = \begin{cases} \lambda_2 e^{\mu_t^2} & t \text{ even} \\ \lambda_1 e^{\mu_t^2} & t \text{ odd} \end{cases} \tag{49}$$

with $\mu_0 = 0$.

To justify the state evolution equations, we rely on the method of moments, requiring f to be polynomial. Thus, we choose $f = f_d(\cdot, t)$ as per Lemma 7, which maximizes the signal-to-noise ratio among all polynomials with degree up to d. With $f = f_d$, we have $\tau_{t+1} = 1$ and the state evolution equations reduce to

$$\widehat{\mu}^2_{t+1} = \begin{cases} \lambda_2 G_d(\widehat{\mu}_t^2) & t \text{ even} \\ \lambda_1 G_d(\widehat{\mu}_t^2) & t \text{ odd} \end{cases} \tag{50}$$

where $G_d(\mu) = \sum_{k=0}^d \frac{\mu^k}{k!}$.

Combining message passing with spectral cleanup, we obtain the following algorithm for estimating C_1^* and C_2^*.

Algorithm 4 Message passing for biclustering

1: Input: $n_1, n_2, K_1, K_2 \in \mathbb{N}$, $\mu > 0$, $W \in \mathbb{R}^{n_1 \times n_2}$, $d^* \in \mathbb{N}$, $t^* \in 2\mathbb{N}$, and $s^* > 0$.
2: Initialize: $\theta^0_{\ell\to i} = 0$ for $(\ell, i) \in [n_2] \times [n_1]$. For $t \geq 0$, define the sequence of degree-d^* polynomials $f_{d^*}(\cdot, t)$ as per Lemma 7 and $\widehat{\mu}_t$ according to (50).
3: Run t^* iterations of message passing as in (43) and (44) with $f = f_{d^*}$ and compute $\theta_i^{t^*}$ for all $i \in [n_1]$ as per (45) and $\theta_j^{t^*+1}$ for all $j \in [n_2]$ as per (46).
4: Find the sets $\widetilde{C}_1 = \{i \in [n_1] : \theta_i^{t^*} \geq \widehat{\mu}_{t^*}/2\}$ and $\widetilde{C}_2 = \{j \in [n_2] : \theta_j^{t^*+1} \geq \widehat{\mu}_{t^*+1}/2\}$.
5: (Cleanup via power method) Denote the restricted matrix $W_{\widetilde{C}_1\widetilde{C}_2}$ by $\widetilde{W}$. Sample u^0 uniformly from the unit sphere in $\mathbb{R}^{|\widetilde{C}_1|}$ and compute $u^{t+2} = \widetilde{W}\widetilde{W}^\top u^t / \|\widetilde{W}\widetilde{W}^\top u^t\|$, for t even and $0 \leq t \leq 2\lceil s^* \log(n_1 n_2)\rceil - 2$. Let $\widehat{u} = u^{2\lceil s^* \log(n_1 n_2)\rceil}$. Return $\widehat{C}_1$, the set of K_1 indices i in $\widetilde{C}_1$ with the largest values of $|\widehat{u}_i|$. Compute the power iteration with $\widetilde{W}^\top\widetilde{W}$ for odd values of t and return $\widehat{C}_2$ similarly.

We now turn to the performance of Algorithm 4. Let

$$\mathcal{G} = \{(\lambda_1, \lambda_2) : \mu_t \to \infty\}, \tag{51}$$

$$\mathcal{G}_d = \{(\lambda_1, \lambda_2) : \widehat{\mu}_t \to \infty\}. \tag{52}$$

As $d \to \infty$, $G_d(\mu) \to e^\mu$ uniformly over bounded intervals. It suggests that if $(\lambda_1, \lambda_2) \in \mathcal{G}$, then there exists a $d^*(\lambda_1, \lambda_2)$ such that $(\lambda_1, \lambda_2) \in \mathcal{G}_{d^*}$ and hence $\widehat{\mu}_t \to \infty$ as $t \to \infty$. The following lemma confirms this intuition.

Lemma 16 *For $d \geq 1$, $\mathcal{G}_d \subset \mathcal{G}_{d+1}$ with $\mathcal{G}_1 = \{(\lambda_1, \lambda_2) : \lambda_1\lambda_2 \geq 1\}$, and $\cup_{d=1}^\infty \mathcal{G}_d = \mathcal{G}$.*

Proof By definition, $G_1(x) = 1 + x$ and thus for t even, $\widehat{\mu}_{t+2}^2 = \lambda_1(1 + \lambda_2(1 + \widehat{\mu}_t^2))$. As a consequence, $\widehat{\mu}_t \to \infty$ if and only if $\lambda_1\lambda_2 \geq 1$, proving the claim for $\mathcal{G}_1$. Let $\phi_d(x) \triangleq \lambda_1 G_d(\lambda_2 G_d(x))$ so that $\widehat{\mu}_{t+2}^2 = \phi_d(\widehat{\mu}_t^2)$ for t even . The fact $\mathcal{G}_d \subset \mathcal{G}_{d+1} \subset \mathcal{G}$ follows from the fact $\phi_d(x)$ is increasing in d and $\phi_d(x) < \phi(x)$, where ϕ is defined in Remark 17. To prove $\cup_{d=1}^\infty \mathcal{G}_d = \mathcal{G}$, fix $(\lambda_1, \lambda_2) \in \mathcal{G}$. It suffices to show that $(\lambda_1, \lambda_2) \in \mathcal{G}_d$ for d sufficiently large. Since $\phi_2(x)/x^2 \to \infty$ as $x \to \infty$, there exists an absolute constant $x_0 > 1$ such that $\phi_d(x) \geq x^2$ whenever $x \geq x_0$ and $d \geq 2$. Let t_0 be an even number such that $\mu_{t_0}^2 > x_0$. Since $\phi_d(x)$ converges to $\phi(x)$ uniformly on bounded intervals, it follows that the first $t_0/2$ iterates using ϕ_d converge to the corresponding iterates using ϕ. So, for d large enough, $\widehat{\mu}_{t_0}^2 > x_0$, and hence, for such d, $\widehat{\mu}_t^2 \to \infty$ as $t \to \infty$, so $(\lambda_1, \lambda_2) \in \mathcal{G}_d$. ■

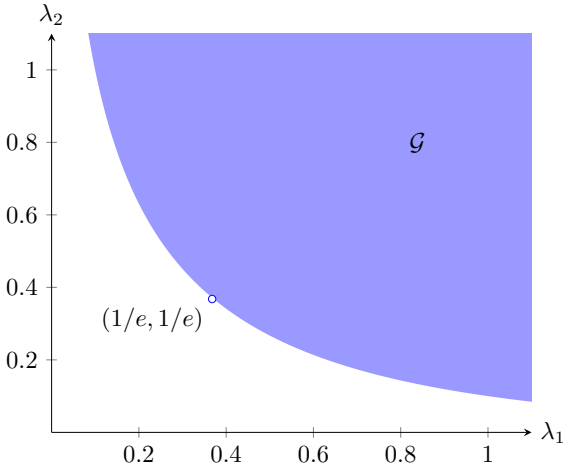

Figure 1: Required signal-to-noise ratios by Algorithm 4 for biclustering.

Remark 17 *Clearly $\mathcal{G}$ is an open subset of $\mathbb{R}_+^2$ and $\mathcal{G}$ is an upper closed set. Let $\partial\mathcal{G}$ denote its boundary and let $\phi(x) \triangleq \lambda_1 e^{\lambda_2 e^x}$, so that $\mu_{t+2}^2 = \phi(\mu_t^2)$ for t even. Note that $(\lambda_1, \lambda_2) \in \partial\mathcal{G}$ if and only if the function is such that for some $x > 0$, $\phi(x) = x$ and $\phi'(x) = 1$. Since $\phi'(x) = \phi(x)y$, where $y = \lambda_2 e^x$, it follows that $xy = 1$ where $x = \lambda_1 e^y$. Therefore, it is convenient to express the boundary of $\mathcal{G}$ in the parametric form*

$$\partial\mathcal{G} = \{(xe^{-1/x}, x^{-1}e^{-x}) : x > 0\}.$$

It follows that $(1/e, 1/e) \in \partial G$ and $\{(\lambda_1, \lambda_2) \in \mathbb{R}_+^2 : \lambda_1\lambda_2 \geq e^{-2}\}\backslash\{(1/e, 1/e)\} \subset \mathcal{G}$ (see Fig. 1 for an illustration). Boundaries of $\mathcal{G}_d$ can be determined similar to (25) (see Fig. 2 for plots).

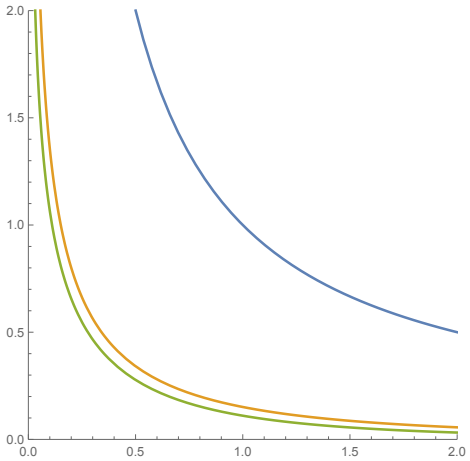

Figure 2: Boundaries of the regions $\mathcal{G}_d$ for $d = 1, 2, 3$; as d increases, $\mathcal{G}_d$ converges to $\mathcal{G}$ in Fig. 1.

The correctness proof for the spectral clean-up procedure in Algorithm 4 is given by Lemma 18 below with s^* defined by (56); it is similar to Lemma 12 used in Theorem 1 but applies to rectangular matrices and uses singular value decomposition.

Lemma 18 *Suppose*

$$\frac{\mu\sqrt{K_1 K_2}}{\sqrt{n_1}+\sqrt{n_2}} \geq \frac{1}{c_0} \tag{53}$$

for some $c_0 > 0$. For $i = 1, 2$, suppose that $\frac{|C_i^*|}{K_i} \to 1$ in probability and $\widetilde{C}_i$ is a set (possibly depending on W) such that

$$\frac{1}{K_i}|\widetilde{C}_i \cap C_i^*| \geq 1 - \epsilon \tag{54}$$

$$K_i(1-\epsilon) \leq |\widetilde{C}_i| \leq n_i\epsilon \tag{55}$$

hold for some $0 < \epsilon < \epsilon_0$, where ϵ_0 depends only on c_0. Let

$$s^* = \left(\log \frac{1-\epsilon-3c_0\sqrt{h(\epsilon)+\epsilon}}{3c_0\sqrt{h(\epsilon)+\epsilon}}\right)^{-1} \tag{56}$$

where $h(\epsilon) \triangleq \epsilon \log \frac{1}{\epsilon} + (1-\epsilon)\log\frac{1}{1-\epsilon}$ is the binary entropy function. Then $\widehat{C}_i$ returned by Algorithm 4 satisfies $|\widehat{C}_i \Delta C_i^*| \leq \eta(\epsilon) K_i$ for $i = 1, 2$, with probability converging to one as $n \to \infty$, where

$$\eta(\epsilon) = 2\epsilon + 650c_0^2\,\frac{h(\epsilon)+\epsilon}{(1-\epsilon)^2}. \tag{57}$$

With Lemma 18, we are ready to show that the bicluster message passing algorithm (Algorithm 4) approximately recovers C_1^* and C_2^*, provided that $(\lambda_1, \lambda_2) \in \mathcal{G}$.

Theorem 19 *Fix $\lambda_1, \lambda_2 > 0$. Suppose $\frac{K_i^2 \mu^2}{n_i} \to \lambda_i$, $K_1 \asymp K_2$, and $\Omega(\sqrt{n_i}) \le K_i \le o(n_i)$ as $n \to \infty$, for $i = 1, 2$. Consider the model (1) with $|C_i^*|/K_i \to 1$ in probability as $n \to \infty$. Suppose $(\lambda_1, \lambda_2) \in \mathcal{G}$ and define $d^*(\lambda_1, \lambda_2)$ as in (59). For every $\eta \in (0, 1)$, there exist explicit positive constants t^*, s^* depending on $(\lambda_1, \lambda_2, \eta)$ such that Algorithm 4 returns $|\widehat{C}_i \cap C_i^*| \ge (1 - \eta)K_i$ for $i = 1, 2$ with probability converging to 1 as $n \to \infty$, and the total running time is bounded by $c(\eta, \lambda_1, \lambda_2) n_1 n_2 \log(n_1 n_2)$, where $c(\eta, \lambda_1, \lambda_2) \to \infty$ as either $\eta \to 0$ or (λ_1, λ_2) approaches $\partial\mathcal{G}$.*

Remark 20 (Exact biclustering via message passing) *If the assumptions of Theorem 19 hold and the voting condition* (39) *(respectively,* (41)*) holds, then C_1^* (respectively, C_2^*) can be exactly recovered by message passing plus a voting procedure as described in Algorithm 3. Similar to the analysis in the symmetric case, whenever information-theoretic sufficient conditions for exact recovery* (39)–(41) *imply the sufficient condition of message passing for weak recovery, i.e., $(\lambda_1, \lambda_2) \in \mathcal{G}$ defined in (52), there is no computational gap for exact recovery.*

To be more precise, consider $K_i = \frac{\rho_i n}{\log n}$ for $i = 1, 2$. Then (39) *and* (41) *are equivalent to $\lambda_i > 8\rho_i$. Thus, whenever K_1 and K_2 are large enough so that $(8\rho_1, 8\rho_2)$ lies in the closure $\boldsymbol{cl}(\mathcal{G})$, or more generally,*

$$\left(\liminf_{n\to\infty} \frac{K_1 \log n_1}{n_1}, \ \liminf_{n\to\infty} \frac{K_2 \log n_2}{n_2} \right) \in \frac{1}{8} \boldsymbol{cl}(\mathcal{G}), \tag{58}$$

then Algorithm 4 plus voting achieves the information-theoretic exact recovery threshold with optimal constants. This result can be viewed as a two-dimensional counterpart of (20) obtained for the symmetric case.

Proof [Proof of Theorem 19] The proof follows step-by-step that of Theorem 1; we shall point out the minor differences. Given λ_1 and λ_2, define

$$d^*(\lambda_1, \lambda_2) = \inf\{d \in \mathbb{N} : (\lambda_1, \lambda_2) \in \mathcal{G}_d\}. \tag{59}$$

By the assumptions of Theorem 19, there exists $c_0 > 0$ so that (53) holds. Given any $\eta \in (0, 1)$, choose an arbitrary $\epsilon \in (0, \epsilon_0)$ such that $\eta(\epsilon)$ defined in (57) is at most η. Notice that ϵ_0 is determined by c_0. Let $M = 8 \log(1/\epsilon)$ and choose

$$t^*(\lambda_1, \lambda_2, M) = \inf \left\{ t : \min\{\widehat{\mu}_t, \widehat{\mu}_{t+1}\} > M \right\}. \tag{60}$$

In view of Lemma 16 and the assumption that $(\lambda_1, \lambda_2) \in \mathcal{G}$, d^* is finite. Since $(\lambda_1, \lambda_2) \in \mathcal{G}_{d^*}$, it follows that $\widehat{\mu}_t \to \infty$ and thus $t^*(\lambda_1, \lambda_2, M)$ is finite.

The assumptions of Theorem 19 imply that $n_1 \asymp n_2$. Lemmas 27 - 29 therefore go through as before, with n in the upper bounds taken to be $\min\{n_1, n_2\}$, so that $\frac{1}{\sqrt{n_i}} \le \frac{1}{\sqrt{n}}$. This modification then implies that Lemma 6, justifying the state evolution equations, goes through as before. See Section 6.1 for more details.

Finally, the proof is complete by invoking Lemma 18.

■

6. Justification of state evolution equations

In this section we prove Lemma 6. Let $f(x,t) = \sum_{i=0}^{d} q_i^t x^i$ with $|q_i^t| \le C$ for a constant C. Let $\{A^t, t \ge 1\}$ be i.i.d. matrices distributed as A conditional on C^* and let $A^0 = A$. We now define a sequence of vectors $\{\xi^t, t \ge 1\}$ with $\xi^t \in \mathbb{R}^n$ given by

$$\xi_{i\to j}^{t+1} = \sum_{\ell \in [n]\setminus \{i,j\}} A_{\ell i}^t f(\xi_{\ell\to i}^t, t), \quad \forall j \ne i \in [n] \tag{61}$$

$$\xi_i^{t+1} = \sum_{\ell \in [n]\setminus\{i\}} A_{\ell i}^t f(\xi_{\ell\to i}^t, t)$$

$$\xi_{i\to j}^0 = 0. \tag{62}$$

In the definition of ξ^t, fresh samples, A^t, of A are used at each iteration, and thus the moments of ξ^t in the asymptotic limit are easier to compute than those of θ^t. Use of the fresh samples A^t does not make the messages $(\xi_{i\to\ell}^t : i \in [n]\setminus\ell)$ independent for fixed $\ell \in [n]$ and fixed $t \ge 2$, because at $t = 1$ the messages sent by any one vertex to all other vertices are statistically dependent, so at $t = 2$ the messages sent by all vertices are statistically dependent. However, we can take advantage of the fact that the contribution of each individual message is small in the limit as $n \to \infty$. Hence, we first prove that ξ^t and θ^t have the same moments of all orders as $n \to \infty$, and then prove the lemma using the method of moments.

The first step is to represent $(\theta_{i\to j}^t, \theta_i^t)$ and $(\xi_{i\to j}^t, \xi_i^t)$ as sums over a family of finite rooted labeled trees as shown by (Deshpande and Montanari, 2015, Lemma 3.3). We next introduce this family in detail. We shall consider rooted trees T of the following form. All edges are directed towards the root. The set of vertices and the set of (directed) edges in a tree T are denoted by $V(T)$ and $E(T)$, respectively. Each vertex has at most d children. The set of leaf vertices of T, denoted by $L(T)$, is the set of vertices with no children. Every vertex in the tree has a *label* which includes the *type* of the vertex, where the types are selected from $[n]$. The label of the root vertex consists of the type of the root vertex, and for every non-root vertex the label has two arguments, where the first argument in the label is the type of the vertex (in $[n]$), and the second one is the *mark* (in $\{0,\dots,d\}$). For a vertex v in T, let $\ell(v)$ denote its type, $r(v)$ its mark (if v is not the root), and $|v|$ its distance from the root in T. For clarity, we restate the definition of family of rooted labeled trees introduced in (Deshpande and Montanari, 2015, Definition 3.2).

Definition 21 *Let $\mathcal{T}^t$ denote the family of labeled trees T with exactly t generations satisfying the conditions:*

1. *The root of T has degree 1.*
2. *Any path $(v_1, v_2, \dots, v_k)$ in the tree is non-backtracking, i.e., the types $\ell(v_i), \ell(v_{i+1}), \ell(v_{i+2})$ are distinct for all i, k.*
3. *For a vertex u that is not the root or a leaf, the mark $r(u)$ is set to the number of children of v.*
4. *Note that $t = \max_{v\in L(T)} |v|$. All leaves u with $|u| \le t-1$ have mark 0.*

Let $\mathcal{T}^t_{i\to j} \subset \mathcal{T}^t$ be the subfamily satisfying the following additional conditions:

1. *The type of the root is i.*
2. *The root has a single child with type distinct from i and j.*

Similarly, let $\mathcal{T}^t_i \subset \mathcal{T}^t$ be the subfamily satisfying the following:

1. *The type of the root is i.*
2. *The root has a single child with type distinct from i.*

We point out that under the above definition, a vertex of a tree in $\mathcal{T}^t$ can have siblings of the same type and mark. Also two trees in $\mathcal{T}^t$ are considered to be the same if and only if the labels of all vertices are the same, with the understanding that the order of the children of any given vertex matters. In addition, the mark of a leaf u with $|u| = t$ is not specified and can possibly take any value in $\{0, \ldots, d\}$. The following lemma is proved by induction on t and the proof can be found in (Deshpande and Montanari, 2015, Lemma 3.3).

Lemma 22

$$\theta^t_{i\to j} = \sum_{T\in\mathcal{T}^t_{i\to j}} A(T)\Gamma(T,\mathbf{q},t)\theta(T),$$
$$\theta^t_i = \sum_{T\in\mathcal{T}^t_i} A(T)\Gamma(T,\mathbf{q},t)\theta(T),$$

where[9]

$$A(T) \triangleq \prod_{u\to v\in E(T)} A_{\ell(u),\ell(v)},$$
$$\Gamma(T,\mathbf{q},t) \triangleq \prod_{u\to v\in E(T)} q^{t-|u|}_{r(u)},$$
$$\theta(T) \triangleq \prod_{u\to v\in E(T):u\in L(T)} (\theta^0_{\ell(u)\to\ell(v)})^{r(u)}.$$

Similarly,

$$\xi^t_{i\to j} = \sum_{T\in\mathcal{T}^t_{i\to j}} \bar{A}(T)\Gamma(T,\mathbf{q},t)\theta(T),$$
$$\xi^t_i = \sum_{T\in\mathcal{T}^t_i} \bar{A}(T)\Gamma(T,\mathbf{q},t)\theta(T),$$

where

$$\bar{A}(T) \triangleq \prod_{u\to v\in E(T)} A^{t-|u|}_{\ell(u),\ell(v)}.$$

9. Often the initial messages for message passing are taken, with some abuse of notation, to have the form $\theta^0_{i\to j} = \theta^0_i$ for all j, and then only the n variables θ^0_i need to be specified. In that case, the expression for $\theta(T)$ simplifies to $\theta(T) \triangleq \prod_{u\in L(T)} (\theta^0_{\ell(u)})^{r(u)}$.

Since the initial messages are zero, $f(\theta^0_{i\to j},0)=q^0_0$. *Thus, for notational convenience in what follows, we can assume without loss of generality that* $f(x,0)\equiv q^0_0$, *i.e.,* $f(x,0)$ *is a degree zero polynomial.* With this assumption, it follows that for a labeled tree $T\in\mathcal{T}^t$, $\Gamma(T,\mathbf{q},t)=0$ unless the mark of every leaf of T is zero. If the mark of every leaf is zero, then $\theta(T)=1$, because in this case $\theta(T)$ is a product of terms of the form 0^0, which are all one, by convention. Therefore, $\Gamma(T,\mathbf{q},t)\theta(T)=\Gamma(T,\mathbf{q},t)$ for all $T\in\mathcal{T}_t$. Consequently, the factor $\theta(T)$ can be dropped from the representations of $\theta^t_{i\to j}$, θ^t_i, $\xi^t_{i\to j}$, and ξ^t_i given in Lemma 22. Applying Lemma 22, we can prove that all finite moments of θ^t_i and ξ^t_i are asymptotically the same. Before that, we need two key auxiliary lemmas.

Let $\phi(T)_{rs}$ denote the number of occurrences of edges $(u\to v)$ in the tree T with types $\ell(u),\ell(v)=\{r,s\}$.

Definition 23 *For $m\geq 1$ and given an m-tuple of trees $T_1,\dots,T_m$, let G denote the undirected graph obtained by identifying the vertices of the same type in the trees and removing the edge directions. Let $E(G)$ denote the edge set of G. Then an edge (r,s) is in $E(G)$ if and only if $\sum_{\ell=1}^m\phi(T_\ell)_{rs}\geq 1$, i.e., the number of times covered is at least one. Let G_1 denote the restriction of G to the vertices in C^* and G_2 the restriction of G to the vertices in $[n]\backslash C^*$. Let $E(G_1)$ and $E(G_2)$ denote the edge set of G_1 and G_2, respectively. Let E_J denote the set of edges in G with one endpoint in G_1 and the other endpoint in G_2.*

Lemma 24 *Suppose an m-tuple of trees $T_1,\dots,T_m\in\mathcal{T}^t$ has α edges in total, and there are k different edges (r,s) in $E(G_1)$ which are covered exactly once, i.e., $\sum_{\ell=1}^m\phi(T_\ell)_{rs}=1$. Then*

$$\left|\mathbb{E}\left[\prod_{\ell=1}^m A(T_\ell)\right]\right|\leq c\mu^k n^{-\alpha/2}$$

for a consant c independent of n. The same conclusion also holds when replacing $A(T_\ell)$ by $\bar{A}(T_\ell)$.

Proof By the definition of $A(T)$,

$$\begin{aligned}
\left|\mathbb{E}\left[\prod_{\ell=1}^m A(T_\ell)\right]\right| &= \left|\mathbb{E}\left[\prod_{j<j'}(A_{jj'})^{\sum_{\ell=1}^m\phi(T_\ell)_{jj'}}\right]\right| \leq \prod_{j<j'}\left|\mathbb{E}\left[(A_{jj'})^{\sum_{\ell=1}^m\phi(T_\ell)_{jj'}}\right]\right| \\
&= \left(\frac{\mu}{\sqrt{n}}\right)^k \prod_{j<j':\sum_{\ell=1}^m\phi(T_\ell)_{jj'}\geq 2}\left|\mathbb{E}\left[(A_{jj'})^{\sum_{\ell=1}^m\phi(T_\ell)_{jj'}}\right]\right| \\
&\leq \left(\frac{\mu}{\sqrt{n}}\right)^k \prod_{j<j':\sum_{\ell=1}^m\phi(T_\ell)_{jj'}\geq 2}\mathbb{E}\left[|A_{jj'}|^{\sum_{\ell=1}^m\phi(T_\ell)_{jj'}}\right] \\
&\leq c\left(\frac{\mu}{\sqrt{n}}\right)^k \prod_{j<j':\sum_{\ell=1}^m\phi(T_\ell)_{jj'}\geq 2}\left(\frac{1}{\sqrt{n}}\right)^{\sum_{\ell=1}^m\phi(T_\ell)_{jj'}} \\
&= c\mu^k n^{-\alpha/2},
\end{aligned}$$

where the last inequality follows because Z_{ij} are zero-mean subgaussian random variables with proxy variance γ and consequently for $1 \le p \le \alpha$, $\mathbb{E}\left[\left|\frac{Z_{ij}}{\sqrt{n}}\right|^p\right] \le cn^{-p/2}$ and $\mathbb{E}\left[\left|\frac{Z_{ij}+\mu}{\sqrt{n}}\right|^p\right] \le cn^{-p/2}$ where c is a finite constant depending on γ, α, and $\mu_{\max}$, where $\mu_{\max}$ is an upper bound on μ for all n, which is finite[10] by the assumptions that $K = \Omega(\sqrt{n})$ and $\lambda = \Theta(1)$. ■

We next define an equivalence relation on the family of m-tuples of trees in $\mathcal{T}^t$, which is useful for enumerating such m-tuples. Fix a set $D \subset [n]$ of distinguished types; the notion of equivalence depends on D. In this paper when we focus on the messages formed by one vertex $\{i\}$ we take $D = \{i\}$ and when we focus on the covariance of messages formed by two vertices, i and j, we take $D = \{i, j\}$.

Definition 25 *For $D \subset [n]$, two m-tuples of trees in $\mathcal{T}^t$ are* equivalent *(relative to D) if there is a permutation of the set of types $[n]$ such that i maps to i for each $i \in D$, C^* maps to C^*, so also $[n]\backslash C^*$ maps to $[n]\backslash C^*$, such that the following is true: the second m-tuple of trees is obtained by applying the permutation to the types of the vertices of the first m-tuple of trees.*

To clarify, if two m-tuples of trees are equivalent, in particular, the marks of the two m-tuples must be the same, and the set of vertices with type $i \in D$ is the same. Recall that for trees to be considered equal, the order of children matters. The same is true when considering trees to be equivalent relative to D. Thus, for example, if (T_1, T_2) is equivalent to (T_1', T_2') and if (T_1, T_2) has the following property: the first child of the first child of the root in T_1 has the same type as the third child of the second child of the root in T_2, then (T_1', T_2') must have the same property. Furthermore, if the common type for those two vertices in (T_1, T_2) is some $i \in D$, then those two vertices in (T_1', T_2') must also have type i. If the common type for those two vertices in (T_1, T_2) is some $k \in C^*\backslash D$, then those two vertices in (T_1', T_2') must also have some common type $k' \in C^*\backslash D$.

Lemma 26 *For a given set of distinguished types D, let $\mathcal{S}$ denote the set of equivalence classes on the family of m-tuples of trees in $\mathcal{T}^t$. Then $|\mathcal{S}| \le c$ for a constant c dependent on only $m, t, d, |D|$ (not on n). Moreover, fix any equivalence class $S \in \mathcal{S}$ and a representative m-tuple of trees $(T_1, \ldots, T_m)$ which has t_1 and t_2 distinct types in $C^*\backslash D$ and $[n]\backslash(C^* \cup D)$, respectively. Then $|S| \le K^{t_1} n^{t_2}$.*

Proof The total number of vertices of an m-tuple $(T_1, \ldots, T_m)$ is bounded by a function of m, t, d alone and thus independently of n, therefore so are the number of ways to partition these vertices into subsets of vertices with the same type. For each such subset, we need to designate whether the type of the vertices in this subset is one of the distinguished types $i \in D$, or is in $C^*\backslash D$, or is in $[n]\backslash(C^* \cup D)$. The total number of distinct such designations is bounded by a function of $m, t, d, |D|$ independently of n. Finally, we need to assign marks to the vertices of the trees, and the number of distinct assignments is bounded by a function of m, t, d alone. Hence, $|\mathcal{S}|$ is bounded by a function of m, t, d alone.

10. This is where the assumption $K = \Omega(\sqrt{n})$ is used because $\frac{K^2\mu^2}{n}$ is assumed to be a constant λ.

Fix a given equivalence class $S \in \mathcal{S}$ and a representative m-tuple of trees $(T_1, \ldots, T_m)$ in S. Consider all the types from $[n]\backslash D$ that appear at least once for some vertex of some tree in the m-tuple. Then t_1 is the number of such types in $C^*\backslash D$ and t_2 is the number of such types in $[n]\backslash(C^* \cup D)$. The cardinality of S is at most the product of the number of partial permutations of length t_1 of elements chosen from $C^*\backslash D$, times the number of partial permutations of length t_2 of elements chosen from $[n]\backslash(C^* \cup D)$. The conclusion follows. ■

Combining Lemmas 22, 24, and 26, we have the following lemma, showing that all finite moments of θ_i^t and ξ_i^t are asymptotically close.

Lemma 27 *For any $t \geq 1$, there exists a constant c independent of n and dependent on m, t, d, C such that for any $i \in [n]$:*

$$\left|\mathbb{E}\left[(\theta_i^t)^m\right] - \mathbb{E}\left[(\xi_i^t)^m\right]\right| \leq cn^{-1/2}.$$

Proof As explained right after Lemma 22, the assumption that $f(x, 0) \equiv q_0^0$ implies that the factor $\theta(T)$ can be dropped in the representations given in Lemma 22. Therefore, it follows from Lemma 22 that for $t \geq 1$,

$$\mathbb{E}\left[(\theta_i^t)^m\right] = \sum_{T_1,\ldots,T_m \in \mathcal{T}_i^t} \prod_{\ell=1}^m \Gamma(T_\ell, \mathbf{q}, t)\mathbb{E}\left[\prod_{\ell=1}^m A(T_\ell)\right],$$
$$\mathbb{E}\left[(\xi_i^t)^m\right] = \sum_{T_1,\ldots,T_m \in \mathcal{T}_i^t} \prod_{\ell=1}^m \Gamma(T_\ell, \mathbf{q}, t)\mathbb{E}\left[\prod_{\ell=1}^m \bar{A}(T_\ell)\right].$$

Because the coefficients in the polynomial are bounded by C and there are m trees with each tree containing at most $1+d+\cdots+d^{t-1} \leq (d+1)^t$ edges, $|\prod_{\ell=1}^m \Gamma(T_\ell, \mathbf{q}, t)| \leq C^{m(d+1)^t}$. Therefore, it suffices to show

$$\sum_{T_1,\ldots,T_m \in \mathcal{T}_i^t} \left|\mathbb{E}\left[\prod_{\ell=1}^m A(T_\ell)\right] - \mathbb{E}\left[\prod_{\ell=1}^m \bar{A}(T_\ell)\right]\right| \leq cn^{-1/2}.$$

In the following, let c denote a constant only depending on m, t, d and its value may change line by line. Recall (G, G_1, G_2) obtained from a given m-tuple of trees $T_1, \ldots, T_m$ as defined in Definition 23. We partition set $\{(T_1, \ldots, T_m) : T_\ell \in \mathcal{T}_i^t\}$ as a union of four disjoint sets $Q \cup R_1 \cup R_2 \cup R_3$, where

1. Q consists of m-tuples of trees $(T_1, \ldots, T_m)$ such that there exists an edge (r, s) in $E(G_2) \cup E_J$ which is covered exactly once.

2. R_1 consists of m-tuples of trees $(T_1, \ldots, T_m)$ such that all edges in $E(G_2) \cup E_J$ are covered at least twice and at least one of them is covered at least 3 times.

3. R_2 consists of m-tuples of trees $(T_1, \ldots, T_m)$ such that each edge in $E(G_2) \cup E_J$ is covered exactly twice and the graph G contains a cycle.

4. R_3 consists of m-tuples of trees $(T_1, \ldots, T_m)$ such that each edge in $E(G_2) \cup E_J$ is covered exactly twice and the graph G is a tree.

Fix any $(T_1, \ldots, T_m) \in Q$ and let (r, s) be an edge in $E(G_2) \cup E(J)$ which is covered exactly once. Since $\mathbb{E}[A_{rs}] = 0$ and A_{rs} appears in the product $\prod_{\ell=1}^m A(T_\ell)$ once, it follows that $\mathbb{E}[\prod_{\ell=1}^m A(T_\ell)] = 0$. Similarly, $\mathbb{E}\left[\prod_{\ell=1}^m \bar{A}(T_\ell)\right] = 0$. Therefore, it is sufficient to show that for $j = 1, 2, 3$,

$$\sum_{(T_1,\ldots,T_m)\in R_j} \left| \mathbb{E}\left[\prod_{\ell=1}^m A(T_\ell)\right] - \mathbb{E}\left[\prod_{\ell=1}^m \bar{A}(T_\ell)\right] \right| \le cn^{-1/2}.$$

First consider R_1. Further, divide R_1 according to the total number of edges in $T_1, \ldots, T_m$ and the number of edges in $E(G_1)$ which are covered exactly once. In particular, for $\alpha = 1, \ldots, m(d+1)^t$ and $k = 0, 1, \ldots, \alpha$, let $R_{1,\alpha,k}$ denote the subset of R_1 consisting of m-tuples of trees $T_1, \ldots, T_m$ such that there are α edges in $T_1, \ldots, T_m$ and there are k edges in $E(G_1)$ which are covered exactly once. It suffices to show that

$$\sum_{(T_1,\ldots,T_m)\in R_{1,\alpha,k}} \left| \mathbb{E}\left[\prod_{\ell=1}^m A(T_\ell)\right] - \mathbb{E}\left[\prod_{\ell=1}^m \bar{A}(T_\ell)\right] \right| \le cn^{-1/2}. \tag{63}$$

Fix α, k and an m-tuple of trees $(T_1, \ldots, T_m) \in R_{1,\alpha,k}$. It follows from Lemma 24 that

$$\left| \mathbb{E}\left[\prod_{\ell=1}^m A(T_\ell)\right] \right| \le c\mu^k n^{-\alpha/2}. \tag{64}$$

Throughout the proof of this lemma, let $D = \{i\}$ be the set of distinguished types for defining equivalence classes as specified in Definition 25. We consider breaking $R_{1,\alpha,k}$ down into a large number of smaller sets, where each set is an equivalence class. While large, it follows from Lemma 26 that the number of these smaller sets depends on m, t, d, but not on n. Hence, it suffices to upper bound $|S|$ for any given equivalence class $S \subset R_{1,\alpha,k}$. It follows from Lemma 26 that $|S| \le K^{n_1} n^{n_2}$, where n_1 is the number of vertices in G_1 with types in $C^* \backslash \{i\}$ and n_2 is the number of vertices in G_2 with types in $[n] \backslash (C^* \cup \{i\})$.

We further upper bound n_1 and n_2. The graph G is connected (because all the trees have a root of type i), so $n_1 + n_2$ (the number of vertices of G minus one) is less than or equal to the number of edges in G. The number of edges in G is at most $k + \frac{\alpha-k-1}{2}$ because there are k edges in G covered once, and the rest are covered at least twice, with one edge covered at least three times. So $n_1 + n_2 \le k + \frac{\alpha-k-1}{2}$. Also, since k of the edges in G have both endpoints in C^*, and the vertices of G_2 have types in $[n] - C^*$, there are at most $\frac{\alpha-k-1}{2}$ edges in G with at least one endpoint in G_2. Moreover, since G is connected, each connected component in G_2 is connected by at least one edge to a vertex in G_1. Therefore, $n_2 \le |V(G_2)| \le \frac{\alpha-k-1}{2}$. The bound $K^{n_1} n^{n_2}$ is maximized subject to $n_1 + n_2 \le k + \frac{\alpha-k-1}{2}$ and $n_2 \le \frac{\alpha-k-1}{2}$ by letting equality hold in both constraints, yielding $|S| \le K^k n^{\frac{\alpha-k-1}{2}}$. Combining with (64) shows that

$$\sum_{(T_1,\ldots,T_m)\in S} \left| \mathbb{E}\left[\prod_{\ell=1}^m A(T_\ell)\right] \right| \le c\mu^k n^{-\alpha/2} K^k n^{\frac{\alpha-k-1}{2}} = c\left(\frac{\mu K}{\sqrt{n}}\right)^k n^{-1/2} \le cn^{-1/2}, \tag{65}$$

where we've used the fact that $\frac{\mu K}{\sqrt{n}}$ is bounded independently of n. In a similar way, it can be shown that

$$\sum_{(T_1,\ldots,T_m)\in S} \left| \mathbb{E}\left[\prod_{\ell=1}^{m} \bar{A}(T_\ell)\right] \right| \leq cn^{-1/2}$$

and thus

$$\sum_{(T_1,\ldots,T_m)\in S} \left| \mathbb{E}\left[\prod_{\ell=1}^{m} A(T_\ell)\right] - \mathbb{E}\left[\prod_{\ell=1}^{m} \bar{A}(T_\ell)\right] \right| \leq cn^{-1/2}. \tag{66}$$

Since the number of equivalence classes S does not depend on n, (63) follows.

Next consider R_2. The previous argument carries over with a minor adjustment on the step of upper bounding n_1 and n_2. In particular, define $R_{2,\alpha,k}$ accordingly as $R_{1,\alpha,k}$ and then consider an equivalence class $S \subset R_{2,\alpha,k}$ corresponding to some representative m-tuple in $R_{2,\alpha,k}$. The number of edges in G is at most $k + \frac{\alpha-k}{2}$ because there are k edges in G covered once, and the rest are covered at least twice. Since G has $n_1 + n_2 + 1$ vertices, is connected, and has a cycle, $n_1 + n_2$ is less than or equal to the number of edges of G minus one, so $n_1 + n_2 \leq k + \frac{\alpha-k-2}{2}$. Also, since k of the edges in G have both endpoints with types in C^*, and V_2 has types in $[n] - C^*$, there are at most $\frac{\alpha-k}{2}$ edges in G with at least one endpoint in V_2. Moreover, since G is connected, each connected component in G_2 is connected by at least one edge to a vertex in G_1. Therefore, $n_2 \leq |V(G_2)| \leq \frac{\alpha-k}{2}$. The bound $K^{n_1} n^{n_2}$ is maximized subject to these constraints by letting equality hold in both constraints, yielding $|S| \leq K^{k-1} n^{\frac{\alpha-k}{2}}$. So $|S|\mu^k n^{-\alpha/2} \leq \left(\frac{\mu K}{\sqrt{n}}\right)^k / K \leq c/K \leq cn^{-1/2}$, and the reminder of the proof for bounding the contribution of R_2 is the same as for R_1 above.

Finally, consider R_3. It suffices to establish the following claim. The claim is that for any m-tuple such that G has no cycles, if two directed edges $(a \to b)$ and $(c \to d)$ map to the same edge in G, then they are at the same level in their respective trees (their trees might be the same). Indeed, if the claim is true, then for any m-tuple $(T_1, \ldots, T_m)$ in R_3 and any pair $\{r, s\} \subset [n]$, A^t_{rs} appears in $\prod_{\ell=1}^m \bar{A}(T_\ell)$ for at most one value of t, so that $\mathbb{E}\left[\prod_{\ell=1}^m A(T_\ell)\right] = \mathbb{E}\left[\prod_{\ell=1}^m \bar{A}(T_\ell)\right]$.

We now prove the claim. Let $\{r, s\}$ denote the edge in G covered by both $(a \to b)$ and $(c \to d)$, i.e. $\{\ell(a), \ell(b)\} = \{\ell(c), \ell(d)\} = \{r, s\}$. First consider the case that $\ell(b) = \ell(d)$. Let $u_1, \ldots, u_k$ denote the directed path in the tree containing b that goes from b to the root of that tree, so $b = u_1$ and u_k is the root of the tree. Since there are no cycles in G, and hence no cycles in the set of edges $\{\{\ell(u_1), \ell(u_2)\}, \ldots, \{\ell(u_{k-1}), \ell(u_k)\}\}$, (viewed as a simple set, i.e. with duplications removed) it follows from the non-backtracking property that $\ell(u_1), \ldots, \ell(u_k)$ are distinct vertices in G. That is, $(\ell(u_1), \ldots, \ell(u_k))$ is a simple path in G. Similarly, let $v_1, \ldots, v_{k'}$ denote the path in the tree containing d that goes from d to the root of that tree, so $d = v_1$ and $v_{k'}$ is the root of that tree. As for the first path, $(\ell(v_1), \ldots, \ell(v_{k'}))$ is also a simple path in G. Since the roots of all m trees have the same type, $\ell(u_k)$ and $\ell(v_{k'})$ are the same vertex in G. Therefore, $(\ell(u_1), \ldots, \ell(u_k), \ell(v_{k'-1}), \ldots, \ell(v_1))$ is a closed walk in G that is the concatenation of two simple paths. Since G has no cycles those two paths must be reverses of each other. That is, $k = k'$ and $\ell(u_j) = \ell(v_j)$ for all j, and hence $(a \to b)$ and $(c \to d)$ are at the same level in their trees.

Consider the remaining case, namely, that $\ell(b) = \ell(c)$. Let $u_1, \ldots, u_k$ be defined as before, and let $v_1, \ldots, v_{k'}$ denote the path in the tree containing c that goes from c to the root of that tree, so $c = v_1$, $d = v_2$, and $v_{k'}$ is the root of that tree. Arguing as before yields that $k = k'$ and $\ell(u_j) = \ell(v_j)$ for $1 \le j \le k$. Note that $k' \ge 2$ and so $k \ge 2$ and $\ell(u_2) = \ell(v_2) = \ell(d) = \ell(a)$. Thus, the types along the directed path $a \to u_1 \to u_2$ within one of the trees violates the non-backtracking property, so the case $\ell(b) = \ell(c)$ cannot occur. The claim is proved. This completes the proof of Lemma 27. ■

The second step is to compute the moments of ξ^t in the asymptotic limit $n \to \infty$. We need the following lemma to ensure that all moments of ξ^t are bounded by a constant independent of n.

Lemma 28 *For any $t \ge 1$, there exists a constant c independent of n and dependent on m, t, d, C, γ such that for any $i, j \in [n]$*

$$|\mathbb{E}\left[(\xi^t_{i\to j})^m\right]| \le c, \quad |\mathbb{E}\left[(\xi^t_i)^m\right]| \le c.$$

Proof We prove the claim for ξ^t_i; the claim for $\xi^t_{i\to j}$ follows by the similar argument. Since $\xi^0_{i\to j} = \theta^0_{i\to j} = 0$ for all $i \in [n]$, it follows from Lemma 22 that

$$\mathbb{E}\left[(\xi^t_i)^m\right] = \sum_{T_1,\ldots,T_m \in \mathcal{T}^t_i} \prod_{\ell=1}^m \Gamma(T_\ell, \mathbf{q}, t) \mathbb{E}\left[\prod_{\ell=1}^m \bar{A}(T_\ell)\right].$$

Recalling (G, G_1, G_2) defined as Definition 23 and following the same argument as used for proving Lemma 27, we can partition set $\{(T_1, \ldots, T_m) : T_\ell \in \mathcal{T}^t_i\}$ as a union of four disjoint sets $Q \cup R_1 \cup R_2 \cup R_3$, and show that

$$\sum_{T_1,\ldots,T_m \in Q} \prod_{\ell=1}^m \Gamma(T_\ell, \mathbf{q}, t) \mathbb{E}\left[\prod_{\ell=1}^m \bar{A}(T_\ell)\right] = 0,$$

and

$$\sum_{T_1,\ldots,T_m \in R_1 \cup R_2} \left|\prod_{\ell=1}^m \Gamma(T_\ell, \mathbf{q}, t)\right| \left|\mathbb{E}\left[\prod_{\ell=1}^m \bar{A}(T_\ell)\right]\right| \le cn^{-1/2}.$$

Hence, we only need to check R_3. Again divide R_3 according to the total number of edges in $T_1, \ldots, T_m$ and the number of edges in $E(G_1)$ which are covered exactly once. In particular, $R_3 = \cup_{1\le\alpha\le m(d+1)^t, 0\le k\le\alpha} R_{3,\alpha,k}$, where $R_{3,\alpha,k}$ is defined in the similar way as $R_{1,\alpha,k}$. Furthermore, similar to $R_{1,\alpha,k}$, consider dividing $R_{3,\alpha,k}$ into a number of equivalence classes (defined relative to $D = \{i\}$), the number of which depends only on m, t, d, as shown in Lemma 26. To prove the lemma, it suffices to show that for any such equivalence class S,

$$\sum_{(T_1,\ldots,T_m)\in S} \left|\mathbb{E}\left[\prod_{\ell=1}^m \bar{A}(T_\ell)\right]\right| \le c.$$

Invoking Lemma 24, we have that

$$\left| \mathbb{E}\left[\prod_{\ell=1}^{m} \bar{A}(T_\ell) \right] \right| \leq c\mu^k n^{-\alpha/2},$$

so

$$\sum_{(T_1,\ldots,T_m)\in S} \left| \mathbb{E}\left[\prod_{\ell=1}^{m} \bar{A}(T_\ell) \right] \right| \leq c\mu^k n^{-\alpha/2}|S|.$$

Fix a representative m-tuple $(T_1,\ldots,T_m)$ for S. It follows from Lemma 26 that $|S| \leq K^{n_1}n^{n_2}$, where n_1 is the number of vertices in G_1 with types in $C^*\backslash\{i\}$ and n_2 is the number of vertices in G_2 with types in $[n]\backslash(C^*\cup\{i\})$.

We can further bound n_1 and n_2 in the similar way as we did for $|R_{1,\alpha,k}|$, with the only adjustment being we cannot use the assumption that there exists at least one edge which is covered at least three times. There are n_1+n_2+1 vertices in the connected graph G and, since the m-tuple is in $R_{3,\alpha,k}$, there are at most $k+\frac{\alpha-k}{2}$ edges in G, so $n_1+n_2 \leq k+\frac{\alpha-k}{2}$. Also, at most $\frac{\alpha-k}{2}$ edges of G have at least one endpoint in V_2 so $n_2 \leq \frac{\alpha-k}{2}$. Therefore, $|S| \leq K^{n_1}n^{n_2} \leq K^k n^{\frac{\alpha-k}{2}}$. It follows that

$$\sum_{(T_1,\ldots,T_m)\in S} \left| \mathbb{E}\left[\prod_{\ell=1}^{m} \bar{A}(T_\ell) \right] \right| \leq c\mu^k n^{-\alpha/2} K^k n^{\frac{\alpha-k}{2}} = c\left(\frac{K\mu}{\sqrt{n}}\right)^k \leq c,$$

and the proof is complete. ■

We also need the following lemma to show the convergence of $\frac{1}{|C^*|}\sum_{i\in C^*}(\xi_i^t)^m$ in probability using the Chebyshev inequality.

Lemma 29 *For any $t \geq 1$, $m \geq 1$ and $i \in [n]$,*

$$\lim_{n\to\infty} \mathsf{var}\left(\frac{1}{K} \sum_{i\in C^*} (\xi_i^t)^m \right) = 0$$

$$\lim_{n\to\infty} \mathsf{var}\left(\frac{1}{K} \sum_{\ell\in C^*\backslash\{i\}} (\xi_{\ell\to i}^t)^m \right) = 0$$

$$\lim_{n\to\infty} \mathsf{var}\left(\frac{1}{n} \sum_{i\in [n]\backslash C^*} (\xi_i^t)^m \right) = 0$$

$$\lim_{n\to\infty} \mathsf{var}\left(\frac{1}{n} \sum_{\ell\in [n]\backslash(C^*\cup\{i\})} (\xi_{\ell\to i}^t)^m \right) = 0,$$

where the same also holds when replacing ξ^t by θ^t.

Proof We prove the first claim; the other claim follows by a similar argument. Notice that

$$\mathsf{var}\left(\frac{1}{K}\sum_{i\in C^*}(\xi_i^t)^m\right) = \frac{1}{K^2}\sum_{i,j\in C^*}\left(\mathbb{E}\left[(\xi_i^t)^m(\xi_j^t)^m\right] - \mathbb{E}\left[(\xi_i^t)^m\right]\mathbb{E}\left[(\xi_j^t)^m\right]\right).$$

There are K diagonal terms with $i=j$ in the last displayed equation and each diagonal term is bounded by a constant independent of n in view of Lemma 28. Hence, to prove the claim, it suffices to consider the cross terms. Since there are $\binom{K}{2}$ cross terms, we only need to show that for each cross term with $i\neq j$, $\mathbb{E}\left[(\xi_i^t)^m(\xi_j^t)^m\right] - \mathbb{E}\left[(\xi_i^t)^m\right]\mathbb{E}\left[(\xi_j^t)^m\right]$ converges to 0 as $n\to\infty$. Using the tree representation as shown by Lemma 22 yields

$$\begin{aligned}&\left|\mathbb{E}\left[(\xi_i^t)^m(\xi_j^t)^m\right] - \mathbb{E}\left[(\xi_i^t)^m\right]\mathbb{E}\left[(\xi_j^t)^m\right]\right| \\ &\leq c\sum_{T_1,\ldots,T_m\in\mathcal{T}_i^t,T_1',\ldots,T_m'\in\mathcal{T}_j^t}\left|\mathbb{E}\left[\prod_{\ell=1}^m \bar{A}(T_\ell)\bar{A}(T_\ell')\right] - \mathbb{E}\left[\prod_{\ell=1}^m \bar{A}(T_\ell)\right]\mathbb{E}\left[\prod_{\ell=1}^m \bar{A}(T_\ell')\right]\right|,\end{aligned}$$

where c is a constant independent of n and dependent of m,t,d,γ.

Let (G,G_1,G_2) denote the undirected simple graphs obtained from $2m$-tuple of trees $(T_1,\ldots,T_m,T_1',\ldots,T_m')$ as defined in Definition 23. Notice that roots of $T_1,\ldots,T_m$ have type i and roots of $T_1',\ldots,T_m'$ have type j, so either G is disconnected with one component containing i and the other component containing j, or G is connected. In the former case, there is no edge $(r,s)\in E(G)$ which is covered by $T_1,\ldots,T_m$ and $T_1',\ldots,T_m'$ simultaneously and thus $\mathbb{E}\left[\prod_{\ell=1}^m \bar{A}(T_\ell)\bar{A}(T_\ell')\right] = \mathbb{E}\left[\prod_{\ell=1}^m \bar{A}(T_\ell)\right]\mathbb{E}\left[\prod_{\ell=1}^m \bar{A}(T_\ell')\right]$. In the latter case, i.e., G is connected. We partition set $\{(T_1,\ldots,T_m,T_1',\ldots,T_m') : T_\ell\in\mathcal{T}_i^t, T_\ell'\in\mathcal{T}_j^t\}$ as a union of two disjoint sets $Q\cup R$, where

1. Q consists of $2m$-tuples of trees such that G is connected and there exists an edge (r,s) in $E(G_2)\cup E_J$ which is covered exactly once.
2. R consists of $2m$-tuples of trees such that G is connected and all edges in $E(G_2)\cup E_J$ are covered at least twice.

If $(T_1,\ldots,T_m,T_1',\ldots,T_m')\in Q$, then

$$\mathbb{E}\left[\prod_{\ell=1}^m \bar{A}(T_\ell)\bar{A}(T_\ell')\right] = 0 \quad \text{and} \quad \mathbb{E}\left[\prod_{\ell=1}^m \bar{A}(T_\ell)\right]\mathbb{E}\left[\prod_{\ell=1}^m \bar{A}(T_\ell')\right] = 0.$$

We are left to check R. Following the argument used in Lemma 27, further divide R according to the total number of edges in trees and the number of edges in $E(G_1)$ which is covered exactly once. In particular, define $R_{\alpha,k}$ in the similar manner as $R_{1,\alpha,k}$. Invoking Lemma 24, it can be shown that for any $2m$-tuple in $R_{\alpha,k}$

$$\begin{aligned}\left|\mathbb{E}\left[\prod_{\ell=1}^m \bar{A}(T_\ell)\bar{A}(T_\ell')\right]\right| &\leq c\mu^k n^{-\alpha/2} \\ \left|\mathbb{E}\left[\prod_{\ell=1}^m \bar{A}(T_\ell)\right]\mathbb{E}\left[\prod_{\ell=1}^m \bar{A}(T_\ell')\right]\right| &\leq c\mu^k n^{-\alpha/2},\end{aligned}$$

Furthermore, let $D = \{i, j\}$ to be the set of distinguished vertices in defining equivalence classes as specificed in Definition 25, so that $R_{\alpha,k}$ is divided into a number of equivalence classes, the number of which depends only on m, t, d, by Lemma 26. For any such equivalence class $S \subset R_{\alpha,k}$, it follows from the last displayed equation that

$$\sum_{T_1,\ldots,T_m,T'_1,\ldots,T'_m \in S} \left| \mathbb{E}\left[\prod_{\ell=1}^m \bar{A}(T_\ell)\bar{A}(T'_\ell)\right] \right| + \left| \mathbb{E}\left[\prod_{\ell=1}^m \bar{A}(T_\ell)\right] \mathbb{E}\left[\prod_{\ell=1}^m \bar{A}(T'_\ell)\right] \right| \leq c\mu^k n^{-\alpha/2}|S|.$$

There are two distinguished vertices, i and j, in the graph G, corresponding to the type of the root vertices of the first m trees and the second m trees, respectively. It follows from Lemma 26 that $|S| \leq K^{n_1} n^{n_2}$, where n_1 is the number of vertices in G_1 with types in $C^* \backslash \{i, j\}$ and n_2 is the number of vertices in G_2 with types in $[n] \backslash (C^* \cup \{i, j\})$. There are $n_1 + n_2 + 2$ vertices in the connected graph G and at most $k + \frac{\alpha-k}{2}$ edges, so $n_1 + n_2 \leq k - 1 + \frac{\alpha-k}{2}$. At most $\frac{\alpha-k}{2}$ edges have at least one endpoint in V_2 and G is connected, so $n_2 \leq \frac{\alpha-k}{2}$. Thus, $|S| \leq K^{n_1} n^{n_2} \leq K^{k-1} n^{\frac{\alpha-k}{2}}$. Hence,

$$\sum_{(T_1,\ldots,T_m,T'_1,\ldots,T'_m) \in S} \left| \mathbb{E}\left[\prod_{\ell=1}^m \bar{A}(T_\ell)\bar{A}(T'_\ell)\right] \right| + \left| \mathbb{E}\left[\prod_{\ell=1}^m \bar{A}(T_\ell)\right] \mathbb{E}\left[\prod_{\ell=1}^m \bar{A}(T'_\ell)\right] \right| \leq c\mu^k n^{-\alpha/2} K^{k-1} n^{\frac{\alpha-k}{2}} = c\left(\frac{K\mu}{\sqrt{n}}\right)^k / K \leq c/K.$$

In conclusion, $\mathsf{var}\left(\frac{1}{K}\sum_{i\in C^*}(\xi_i^t)^m\right) \leq c/K$ and the first claim follows. ■

With Lemma 28 and Lemma 29 in hand, we are ready to compute the moments of ξ^t in the asymptotic limit $n \to \infty$.

Lemma 30 *For any $t \geq 0$, $m \geq 1$:*

$$\lim_{n\to\infty} \mathbb{E}\left[(\xi_{i\to j}^t)^m\right] = \mathbb{E}\left[(\mu_t + \tau_t Z_t)^m\right], \quad \forall i \in C^*, j \in [n], j \neq i \tag{67}$$

$$\lim_{n\to\infty} \mathbb{E}\left[(\xi_{i\to j}^t)^m\right] = \mathbb{E}\left[(\tau_t Z_t)^m\right], \quad \forall i \notin C^*, j \in [n], j \neq i. \tag{68}$$

$$\lim_{n\to\infty} \mathbb{E}\left[(\xi_i^t)^m\right] = \mathbb{E}\left[(\mu_t + \tau_t Z_t)^m\right], \quad \forall i \in C^*$$

$$\lim_{n\to\infty} \mathbb{E}\left[(\xi_i^t)^m\right] = \mathbb{E}\left[(\tau_t Z_t)^m\right], \quad \forall i \notin C^*.$$

Proof Below we shall use the following version of the Berry-Esseen central limit theorem. There is an absolute constant C_0 such that if $X_1, X_2, \ldots, X_n$ are independent mean zero random variables and $S_n = X_1 + \cdots + X_n$ then

$$\sup_{x\in\mathbb{R}} \left| \mathbb{P}\left\{ \frac{S_n}{\sqrt{\mathsf{var}(S_n)}} \leq x \right\} - \Phi(x) \right| \overset{(a)}{\leq} \frac{C_0 \sum_{i\in[n]} \mathbb{E}\left[|X_i|^3\right]}{(\mathsf{var}(S_n))^{3/2}} \overset{(b)}{\leq} \frac{C_0 n^{1/4} \left(\sum_{i\in[n]} \mathbb{E}\left[X_i^4\right]\right)^{3/4}}{(\mathsf{var}(S_n))^{3/2}},$$

where Φ is the standard normal CDF; (a) is the original result of Esseen (Esseen, 1942); (b) follows by Jensen's inequality.

We prove the first two claims, (67) and (68). The other two follow similarly. The proof is by induction, so suppose (67) and (68) hold for some t and all $n \geq 1$. We aim to show they also hold for $t+1$. The above identities hold for the base case $t = 0$, because $\xi^0_{i\to j} = 0$ for all $i \neq j$ and $\mu_0 = \tau_0 = 0$. By the induction hypothesis, Lemma 29, and Chebyshev's inequality,

$$\lim_{n\to\infty} \frac{1}{K} \sum_{\ell \in C^* \setminus \{i\}} (\xi^t_{\ell\to i})^m \stackrel{p}{=} \mathbb{E}\left[(\mu_t + \tau_t Z_t)^m\right], \quad \forall i \in [n], \tag{69}$$

$$\lim_{n\to\infty} \frac{1}{n} \sum_{\ell \in [n] \setminus (C^* \cup \{i\})} (\xi^t_{\ell\to i})^m \stackrel{p}{=} \mathbb{E}\left[(\tau_t Z_t)^m\right], \quad \forall i \in [n], \tag{70}$$

where $Z_t \sim \mathcal{N}(0,1)$.

Fix an $i \in C^*$. Then

$$\begin{aligned}
\lim_{n\to\infty} \mathbb{E}\left[\xi^{t+1}_{i\to j} | \mathcal{F}_t\right] &= \lim_{n\to\infty} \mathbb{E}\left[\sum_{\ell \in C^* \setminus \{i,j\}} A^t_{\ell i} f(\xi^t_{\ell\to i}) + \sum_{\ell \in [n] \setminus (C^* \cup \{j\})} A^t_{\ell i} f(\xi^t_{\ell\to i}) | \mathcal{F}_t\right] \\
&= \sqrt{\lambda} \lim_{n\to\infty} \frac{1}{K} \sum_{\ell \in C^* \setminus \{i,j\}} f(\xi^t_{\ell\to i}) \\
&\stackrel{p}{=} \sqrt{\lambda} \lim_{n\to\infty} \frac{1}{K} \sum_{\ell \in C^* \setminus \{i\}} f(\xi^t_{\ell\to i}) \\
&\stackrel{p}{=} \sqrt{\lambda} \mathbb{E}\left[f(\mu_t + \tau_t Z_t)\right] = \mu_{t+1},
\end{aligned} \tag{71}$$

where the first equality follows from the definition of ξ^{t+1} given by (61); the second equality holds because $\mathbb{E}\left[A^t_{\ell i}\right] = \mu/\sqrt{n}$ if $\ell \in C^*$ and $\mathbb{E}\left[A^t_{\ell i}\right] = 0$ otherwise; the third equality holds in view of Lemma 28, the fourth equality holds due to (69) and the fact f is a finite-degree polynomial; the last equality holds due to the definition of μ_{t+1}.

Similarly,

$$\begin{aligned}
\lim_{n\to\infty} \mathsf{var}\left(\xi^{t+1}_{i\to j} | \mathcal{F}_t\right) &= \lim_{n\to\infty} \sum_{\ell \in [n] \setminus \{i,j\}} \mathsf{var}\left(A^t_{\ell i} f(\xi^t_{\ell\to i}) | \mathcal{F}_t\right) \\
&= \lim_{n\to\infty} \frac{1}{n} \sum_{\ell \in [n] \setminus \{i,j\}} f(\xi^t_{\ell\to i})^2 &(72)\\
&= \lim_{n\to\infty} \frac{1}{n} \left\{ \sum_{\ell \in [n] \setminus C^* \cup \{j\}} f(\xi^t_{\ell\to i})^2 + \sum_{\ell \in C^* \setminus \{i,j\}} f(\xi^t_{\ell\to i})^2 \right\} &(73)\\
&\stackrel{p}{=} \lim_{n\to\infty} \frac{1}{n} \sum_{\ell \in [n] \setminus C^*} f(\xi^t_{\ell\to i})^2 &(74)\\
&\stackrel{p}{=} \mathbb{E}\left[f(\tau_t Z_t)^2\right] = \tau^2_{t+1}, &(75)
\end{aligned}$$

where the first equality follows from the conditional independence of $A^t_{\ell i} f(\xi^t_{\ell\to i})$ for $\ell \in [n]$; the second equality holds because $\mathsf{var}(A_{\ell i}) = 1/n$ for all ℓ; the third equality is the result of breaking a sum into two parts, the fourth equality holds in view of Lemma 28 and the

assumption that $K = o(n)$; the fifth equality holds in view of (70) and the fact f is a finite-degree polynomial; the last equality holds due to the definition of τ_{t+1}.

Conditional on $\mathcal{F}_t$, $\xi_{i\to j}^{t+1} - \mathbb{E}\left[\xi_{i\to j}^{t+1}\right]$ is a sum of independent random variables. Therefore, by the form of the Berry-Esseen central limit theorem noted above,

$$\begin{aligned}
&\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left\{\frac{\xi_{i\to j}^{t+1} - \mathbb{E}\left[\xi_{i\to j}^{t+1}|\mathcal{F}_t\right]}{\mathsf{var}(\xi_{i\to j}^{t+1}|\mathcal{F}_t)} \le x\middle|\mathcal{F}_t\right\} - \Phi(x)\right| \\
&\le \frac{C_0 n^{1/4}}{\left(\mathsf{var}(\xi_{i\to j}^{t+1}|\mathcal{F}_t)\right)^{3/2}}\left(\sum_{\ell\in[n]\setminus\{j\}} f(\xi_{\ell\to i}^t)^4 \mathbb{E}\left[(A_{\ell i}^t - \mathbb{E}\left[A_{\ell i}^t\right])^4\right]\right)^{3/4} \\
&\le \frac{C_0 c^{3/4} n^{-1/2}}{\left(\mathsf{var}(\xi_{i\to j}^{t+1}|\mathcal{F}_t)\right)^{3/2}}\left(\frac{1}{n}\sum_{\ell\in[n]\setminus\{j\}} f(\xi_{\ell\to i}^t)^4\right)^{3/4}, \qquad (76)
\end{aligned}$$

where we used the fact $\mathbb{E}\left[(A_{\ell i}^t - \mathbb{E}\left[A_{\ell i}^t\right])^4\right] \le cn^{-2}$, for a constant c depending only on the constant γ appearing in the subgaussian assumption. Taking the limit $n\to\infty$ and noticing that $\mathsf{var}(\xi_{i\to j}^{t+1}|\mathcal{F}_t) \xrightarrow{p} \tau_{t+1}^2$ and $\frac{1}{n}\sum_{\ell\in[n]\setminus\{j\}} f(\xi_{\ell\to i}^t)^4 \xrightarrow{p} \mathbb{E}\left[f(\tau_t Z_t)^4\right]$ (using the same steps as in (72)-(75)), we find the righthand side of (76) converges to zero in probability. Thus, in view of (71), (75), and (76), for any $x\in\mathbb{R}$,

$$\lim_{n\to\infty}\mathbb{P}\left\{\xi_{i\to j}^{t+1} \le x|\mathcal{F}_t\right\} \stackrel{p}{=} \mathbb{P}\left\{\mu_{t+1} + \tau_{t+1}Z_{t+1} \le x\right\}.$$

It follows by the dominated convergence theorem that

$$\lim_{n\to\infty}\mathbb{P}\left\{\xi_{i\to j}^{t+1} \le x\right\} = \mathbb{P}\left\{\mu_{t+1} + \tau_{t+1}Z_{t+1} \le x\right\}.$$

and, since convergence in distribution is preserved under continuous transformations:

$$\lim_{n\to\infty}\mathbb{P}\left\{\left(\xi_{i\to j}^{t+1}\right)^m \le x\right\} = \mathbb{P}\left\{(\mu_{t+1} + \tau_{t+1}Z_{t+1})^m \le x\right\}.$$

Since $\mathbb{E}\left[\left|\xi_{i\to j}^{t+1}\right|^{m+1}\right] \le c$ for some c independent of n, the family of random variables $(\xi_{i\to j}^{t+1})^m$ is uniformly integrable, so that convergence in distribution implies convergence of means to the mean of the limiting distribution. Therefore (67) holds with t replaced by $t+1$.

To complete the proof by induction it remains to show (68) holds with t replaced by $t+1$, so fix $i\notin C^*$. Following the previous argument, one can easily check that

$$\begin{aligned}
\mathbb{E}\left[\xi_{i\to j}^{t+1}|\mathcal{F}_t\right] &= 0 \\
\lim_{n\to\infty}\mathsf{var}\left(\xi_{i\to j}^{t+1}|\mathcal{F}_t\right) &\stackrel{p}{=} \tau_{t+1}^2,
\end{aligned}$$

and that, by the central limit theorem, uniformly bounded $m+1^{th}$ moments, and uniform integrability, (68) holds with t replaced by $t+1$. ■

Proof [Proof of Lemma 6] We show the first claim; the second one follows analogously. Fix $t \geq 1$. The convergence property to be proved depends only on the sequence of random empirical distributions of $(\theta_i^t : t \in C^*)$ indexed by n. We may therefore assume without loss of generality that all the random variables $(\theta_i^t : t \in C^*)$ for different n are defined on a single underlying probability space; the joint distribution for different values of n can be arbitrary. To show the convergence in probability, it suffices to show that for any subsequence $\{n_k\}$ there exists a sub-subsequence $\{n_{k_\ell}\}$ such that for $j \neq i$,

$$\lim_{\ell \to \infty} d_{\mathrm{KS}} \left(\frac{1}{K_{k_\ell}} \sum_{i \in C^*} \delta_{\theta_i^t}, \mathcal{N}(\mu_t, \tau_t^2) \right) = 0, \text{ a.s.} \tag{77}$$

Fix a subsequence n_k. In view of Lemmas 27 and 30, for any fixed integer m,

$$\lim_{k \to \infty} \mathbb{E}\left[(\theta_i^t)^m\right] = \mathbb{E}\left[(\mu_t + \tau_t Z_t)^m\right].$$

Combining Lemma 29 with Chebyshev's inequality,

$$\lim_{k \to \infty} \frac{1}{K_k} \sum_{i \in C^*} \left(\theta_i^t\right)^m \stackrel{p}{=} \mathbb{E}\left[(\mu_t + \tau_t Z_t)^m\right], \tag{78}$$

which further implies, by a well-known property of convergence in probability, that there exists a sub-subsequence such that (78) holds almost surely. Using a standard diagonal argument, one can construct a sub-subsequence $\{n_{k_\ell}\}$ such that for all $m \geq 1$,

$$\lim_{\ell \to \infty} \frac{1}{K_{k_\ell}} \sum_{i \in C^*} \left(\theta_i^t\right)^m = \mathbb{E}\left[(\mu_t + \tau_t Z_t)^m\right] \text{ a.s.}$$

Since a Gaussian distribution is determined by its moments, by the method of moments (see, for example, (Chung, 2001, Theorem 4.5.5)), applied for each outcome ω in the underlying probability space (excluding some subset of probability zero), it follows that the sequence of empirical distribution of θ_i^t for $i \in C^*$ weakly converges to $\mathcal{N}(\mu_t, \tau_t^2)$, which, since Gaussian density is bounded, is equivalent to convergence in the KS distance,[11] proving the desired (77). ■

Remark 31 *We discuss the difference between the proof of Lemma 6 and that of (Deshpande and Montanari, 2015, Lemma 2.2). First, a larger K requires modification of bounds from (Deshpande and Montanari, 2015) to calculate the moments of messages in Lemmas 27 - 29. In particular, $(\theta_{i \to j}^t)^m$ can be expanded as a sum of monomials in terms of $A_{k\ell}$ for $k, \ell \in [n]$. A larger K implies that in the expansion, there are more monomials containing $A_{k\ell}$ for $k, \ell \in C^*$. That effect is offset by μ being smaller. Our approach is to balances these two effects by accounting separately the contributions of those $A_{k\ell}$'s which appear only once in a monomial. Such $A_{k\ell}$'s correspond to singly covered edges with both endpoints in C^*. See Lemma 24, as well as $R_{1,\alpha,k}$ in Lemma 27, $R_{3,\alpha,k}$ in Lemma 28, and $R_{\alpha,k}$ in Lemma 29. Finally, Lemma 6 is phrased in terms of KS distance while (Deshpande and Montanari, 2015, Lemma 2.2) is in terms of convergence of expected values of bounded Lipschitz functions.*

11. This follows from the fact that when one of the distributions has bounded density the Lévy distance, which metrizes weak convergence, is equivalent to the KS distance (see, e.g. (Petrov, 1995, 1.8.32)).

6.1 Justification of state evolution equations for biclustering

In this subsection, we briefly describe how to generalize the method of moments from symmetric case to asymmetric biclustering case. Recall $f(x,t) = \sum_{r=0}^{d} q_r^t x^r$ with $|q_r^t| \le C$ for some constant C. Let $\{W^t, t \ge 1\}$ be i.i.d. matrices distributed as W conditional on (C_1^*, C_2^*) and let $W^0 = W$. Similar to (61), we define a sequence of vectors $\{\xi^t, t \ge 1\}$ given by

$$(t \text{ even}) \qquad \xi_{i\to j}^{t+1} = \frac{1}{\sqrt{n_2}} \sum_{\ell \in [n_2]\setminus\{j\}} W_{i\ell}^t f(\xi_{\ell\to i}^t, t), \quad \forall i \in [n_1], j \in [n_2] \tag{79}$$

$$(t \text{ odd}) \qquad \xi_{j\to i}^{t+1} = \frac{1}{\sqrt{n_1}} \sum_{\ell \in [n_1]\setminus\{i\}} W_{\ell j}^t f(\xi_{\ell\to j}^t, t), \quad \forall j \in [n_2], i \in [n_1], \tag{80}$$

with the initial condition $\xi_{\ell\to i}^0 = 0$ for $(\ell, i) \in [n_2] \times [n_1]$. Moreover, let

$$(t \text{ even}) \qquad \xi_i^{t+1} = \frac{1}{\sqrt{n_2}} \sum_{\ell \in [n_2]} W_{i\ell}^t f(\xi_{\ell\to i}^t, t), \quad \forall i \in [n_1] \tag{81}$$

$$(t \text{ odd}) \qquad \xi_j^{t+1} = \frac{1}{\sqrt{n_1}} \sum_{\ell \in [n_1]} W_{\ell j}^t f(\xi_{\ell\to j}^t, t), \quad \forall j \in [n_2]. \tag{82}$$

The first step is to represent $(\theta_{i\to j}^t, \theta_i^t)$ and $(\xi_{i\to j}^t, \xi_i^t)$ as sums over a family of finite rooted labeled trees. Abusing notation slightly, we treat elements from $[n_1]$ and $[n_2]$ as distinct elements. We define $\mathcal{T}^t$ as Definition 21 with an additional constraint: a vertex u must have type from $[n_1]$ if $t - |u|$ is odd and from $[n_2]$ if $t - |u|$ is even. In particular, all leaves u with $|u| = t$ must have type from $[n_2]$, and the root has type from $[n_1]$ if t is odd and from $[n_2]$ if t is even.

The following lemma similar to Lemma 22 can be proved by induction on t similar to (Deshpande and Montanari, 2015, Lemma 3.3).

Lemma 32

$$\theta_{i\to j}^t = \sum_{T \in \mathcal{T}_{i\to j}^t} A(T)\Gamma(T, \mathbf{q}, t),$$
$$\theta_i^t = \sum_{T \in \mathcal{T}_i^t} A(T)\Gamma(T, \mathbf{q}, t),$$

where

$$A(T) \triangleq \prod_{\substack{u\to v \in E(T) \\ \ell(u) \in [n_1]}} \frac{1}{\sqrt{n_1}} W_{\ell(u),\ell(v)} \prod_{\substack{u\to v \in E(T) \\ \ell(u) \in [n_2]}} \frac{1}{\sqrt{n_2}} W_{\ell(v),\ell(u)}$$

and $\Gamma(T, \mathbf{q}, t)$ *is defined as before:*

$$\Gamma(T, \mathbf{q}, t) = \prod_{u\to v \in E(T)} q_{r(u)}^{t-|u|}.$$

Similarly,

$$\xi_{i\to j}^t = \sum_{T\in\mathcal{T}_{i\to j}^t} \bar{A}(T)\Gamma(T,\mathbf{q},t),$$
$$\xi_i^t = \sum_{T\in\mathcal{T}_i^t} \bar{A}(T)\Gamma(T,\mathbf{q},t),$$

where

$$\bar{A}(T) \triangleq \prod_{\substack{u\to v\in E(T)\\ \ell(u)\in[n_1]}} \frac{1}{\sqrt{n_1}} W_{\ell(u),\ell(v)}^{t-|u|} \prod_{\substack{u\to v\in E(T)\\ \ell(u)\in[n_2]}} \frac{1}{\sqrt{n_2}} W_{\ell(v),\ell(u)}^{t-|u|}.$$

Similar to Definition 23, let G denote the undirected *bipartite* graph obtained by identifying the vertices of the same type in the tuple of trees $T_1,\ldots,T_m$ and removing the edge directions. Note that the vertices of type from $[n_1]$ ($[n_2]$) in trees constitute the left (right) part of G. Let $E(G)$ denote the edge set of G. Let G_1 denote the restriction of G to the vertices in (C_1^*, C_2^*) and G_2 the restriction of G to the vertices in $([n_1]\backslash C_1^*, [n_2]\backslash C_2^*)$. Let $E(G_1)$ and $E(G_2)$ denote the edge set of G_1 and G_2, respectively. Let E_J denote the set of edges in G with one endpoint in G_1 and the other endpoint in G_2. Then Lemma 24 goes through as before, with n in the upper bound taken to be $\min\{n_1, n_2\}$. Note that in the definition of $A(T)$, either W_{ij} is divided by $\sqrt{n_1}$ or W_{ji} is divided by $\sqrt{n_2}$ depending on whether $i\in[n_1]$ or $i\in[n_2]$. However, it always holds that $\frac{1}{\sqrt{n_1}} \le \frac{1}{\sqrt{n}}$ and $\frac{1}{\sqrt{n_2}} \le \frac{1}{\sqrt{n}}$.

For a given set of distinguished types D, we can define the equivalence classes on the family of m-tuples of trees in $\mathcal{T}^t$ similar to Definition 25 except that we only allow either permutations of types in $[n_1]$ or permutations of types in $[n_2]$. Then Lemma 26 goes throught as before, with the upper bound in Lemma 26 changes to

$$|S| \le (\max\{K_1, K_2\})^{t_1} (\max\{n_1, n_2\})^{t_2},$$

where we assume a representative m-tuple of trees $(T_1,\ldots,T_m)$ has t_1 and t_2 distinct types in $C_1^*\cup C_2^*\backslash D$ and $[n_1]\cup[n_2]\backslash(C_1^*\cup C_2^*\cup D)$, respectively. By assumptions that $\lambda_1, \lambda_2 = \Theta(1)$ and $K_1 \asymp K_2$, it follows that $n_1 \asymp n_2 \asymp n$. Furthermore, let $K = \min\{K_1, K_2\}$. Then $K_1 \asymp K_2 \asymp K$. Hence, $|S| \le cK^{t_1} n^{t_2}$ for an absolute constant $c > 0$.

With Lemma 32, modified Lemma 24, and modified Lemma 26, Lemmas 27 - 29 go through as before. In particular, we still partition set $\{(T_1,\ldots,T_m) : T_\ell \in \mathcal{T}_i^t\}$ as a union of four disjoint sets $Q\cup R_1\cup R_2\cup R_3$, and introduce $R_{1,\alpha,k}$, $R_{2,\alpha,k}$, $R_{3,\alpha,k}$ and $R_{\alpha,k}$ as before.

Acknowledgments

We would like to acknowledge support for this project from the National Science Foundation (NSF grants IIS-9988642, CCF 14-09106, CCF-1527105, CCF-1755960, OIS 18-23145, and a CAREER award CCF-1651588) and the Multidisciplinary Research Program of the Department of Defense (MURI N00014-00-1-0637).

A. Row-wise thresholding

We describe a simple thresholding procedure for recovering C^*. For simplicity and information theoretic comparisons, we consider the Gaussian case. Let $R_i = \sum_j W_{i,j}$ for $i \in [n]$. Then $R_i \sim \mathcal{N}(K\mu, n)$ if $i \in C^*$ and $R_i \sim \mathcal{N}(0, n)$ if $i \notin C^*$. Let $\widehat{C} = \left\{ i \in [n] : R_i \geq \frac{K\mu}{2} \right\}$. Then $\mathbb{E}\left[|\widehat{C} \triangle C^*| \right] = nQ\left(\frac{K\mu}{2\sqrt{n}} \right)$. Recall that $\lambda = \frac{K^2\mu^2}{n}$. Hence, if

$$\lambda = \omega\left(\log \frac{n}{K} \right), \tag{83}$$

then we have $\mathbb{E}[|\widehat{C} \triangle C^*|] = o(K)$ and hence achieved weak recovery. In the regime $K \asymp n \asymp (n-K)$, $\lambda = \omega(\log \frac{n}{K})$ is equivalent to $\lambda \to \infty$, which is also equivalent to $K\mu^2 \to \infty$ and coincides with the necessary and sufficient condition for the information-theoretic possibility of weak recovery in this regime (Hajek et al., 2017, Corollary 2). (If instead $n - K = o(n)$, weak recovery is trivially provided by $\widehat{C} = [n]$.) Thus, row-wise thresholding provides weak recovery in the regime $K \asymp n \asymp (n-K)$ whenever information theoretically possible. Under the information-theoretic condition (15), an algorithm attaining exact recovery can be built using row-wise thresholding for weak recovery followed by voting, as in Algorithm 2 (see (Hajek et al., 2017, Theorem 3) and its proof). In the regime $\frac{n}{K} \log \frac{n}{K} = o(\log n)$, or equivalently $K = \omega(n \log\log n / \log n)$, condition (15) implies that $\lambda = \omega(\log \frac{n}{K})$, and hence in this regime exact recovery can be attained in linear time $O(n^2)$ whenever information theoretically possible.

B. Proof of Lemma 12

Proof We remind the reader that in this paper we let $A = W/\sqrt{n}$ so that $\mathsf{var}(A_{ij}) = 1/n$ for $i, j \in [n]$ and $\mathbb{E}[A_{ij}] = \mu/\sqrt{n}$ for $i, j \in C^*$.

Fix a $\widetilde{C}$ that satisfies (29) – (30), i.e., $|\widetilde{C} \cap C^*| \geq K(1-\epsilon)$ and $K(1-\epsilon) \leq |\widetilde{C}| \leq n\epsilon$. Let $m = |\widetilde{C}|$ and abbreviate the restricted matrix $A_{\widetilde{C}} \in \mathbb{R}^{|\widetilde{C}| \times |\widetilde{C}|}$ by $\widetilde{A}$. Let $\mathbf{1}_{\widetilde{C} \cap C^*} \in \mathbb{R}^{|\widetilde{C}|}$ denote the indicator vector of $\widetilde{C} \cap C^*$. Then the mean of $\widetilde{A}$ is the rank-one matrix $\mathbb{E}[\widetilde{A}] = \frac{\mu}{\sqrt{n}} \mathbf{1}_{\widetilde{C} \cap C^*} \mathbf{1}_{\widetilde{C} \cap C^*}^\top$, whose largest eigenvalue is $\frac{\mu |\widetilde{C} \cap C^*|}{\sqrt{n}}$ with the corresponding eigenvector $v \triangleq \frac{1}{\sqrt{|\widetilde{C} \cap C^*|}} \mathbf{1}_{\widetilde{C} \cap C^*}$. Let $Z = \widetilde{A} - \mathbb{E}[\widetilde{A}]$, and let u denote the principal eigenvector of $\widetilde{A}$ such that $\langle u, v \rangle \geq 0$. Note that $\|u - v\| = \sqrt{2(1 - \langle u, v \rangle)} \leq \sqrt{2(1 - \langle u, v \rangle^2)} = \sqrt{2} \sin\theta$, where θ is the angle between u and v. Combining this observation with the sin theorem of (Davis and Kahan, 1970) yields

$$\begin{aligned} \|u - v\| \leq \sqrt{2} \sin\theta &\leq \sqrt{2} \min\left\{ 1, \frac{\|Z\|}{\mu |\widetilde{C} \cap C| / \sqrt{n} - \|Z\|} \right\} \\ &\leq \frac{2\sqrt{2}\|Z\|}{\mu |\widetilde{C} \cap C| / \sqrt{n}} \leq \frac{2\sqrt{2}\|Z\|}{\sqrt{\lambda}(1-\epsilon)}, \end{aligned} \tag{84}$$

where the last inequality follows from the assumption (29). Observe that Z is a symmetric matrix such that $\{Z_{ij}\}_{i \leq j}$ are independent subgaussian random variables with zero mean

and proxy variance γ/n. To bound $\|Z\|$, we use the following standard concentration inequality, see e.g.. (Deshpande and Montanari, 2015, Lemma A.3): For any $t > 0$,

$$\mathbb{P}\left\{\|Z\| \geq t\right\} \leq 2\exp\left(-m\left(\frac{t^2 n}{16e\gamma m} - \log\frac{5t^2 n}{16\gamma m}\right)\right).$$

Note that if

$$t \geq \sqrt{64e\gamma h(\epsilon) + 160e\gamma m/n},$$

then

$$m\left(\frac{t^2 n}{16e\gamma m} - \log\frac{5t^2 n}{16\gamma m}\right) \geq \frac{t^2 n}{32e\gamma} \geq 2nh(\epsilon).$$

By assumption we have $K(1-\epsilon) \leq m \leq \epsilon n$. Therefore, by setting $\beta = \sqrt{64e\gamma h(\epsilon) + 160e\gamma\epsilon}$, we have for any fixed $\widetilde{C}$,

$$\mathbb{P}\left\{\|Z\| \geq \beta\right\} \leq 2e^{-2nh(\epsilon)}. \tag{85}$$

The number of possible choices of $\widetilde{C}$ that fulfills (30) so that $|\widetilde{C}| \leq \epsilon n$ is at most $\sum_{k\leq n\epsilon}\binom{n}{k}$ which is further upper bounded by $e^{nh(\epsilon)}$ (see, e.g., (Ash, 1965, Lemma 4.7.2)). In view of (85), the union bound yields $\|Z\| \leq \beta$ with high probability as $n \to \infty$.

Throughout the reminder of this proof we assume A and $\widetilde{C}$ are fixed with $\|Z\| \leq \beta$. Combining with (84), we have,

$$\|u - v\| \leq \frac{2\sqrt{2}\beta}{\sqrt{\lambda}(1-\epsilon)}. \tag{86}$$

Next, we argue that either $\widehat{u}$ or $-\widehat{u}$ is close to u, and hence, close to v by the triangle inequality. By the choice of the initial vector u^0, we can write $u^0 = z/\|z\|$ for a standard normal vector $z \in \mathbb{R}^m$. By the tail bounds for Chi-squared distributions, it follows that $\|z\| \leq 2\sqrt{m}$ with high probability. For any fixed u, the random variable $\langle u, z\rangle \sim \mathcal{N}(0,1)$ and thus with high probability, $|\langle u, z\rangle|^2 \geq 1/\log n$, and hence

$$|\langle u, u^0\rangle| = |\langle u, z\rangle|/\|z\| \geq (2\sqrt{n\log n})^{-1}. \tag{87}$$

Replacing u^0 by $-u^0$ would result in replacing u^t by $-u^t$ for each t, and since $\widehat{C}$ returned by Algorithm 1 only depends on $|u_i^t|$, replacing u by $-u$ would have no effect on the ouput $\widehat{C}$ of the algorithm. Thus, we can assume without loss of generality that $\langle u, u^0\rangle \geq 0$, and (87) becomes

$$\langle u, u^0\rangle \geq (2\sqrt{n\log n})^{-1}. \tag{88}$$

By Weyl's inequality, the maximal singular value of $\widetilde{A}$ satisfies $\sigma_1(\widetilde{A}) \geq \frac{\mu K(1-\epsilon)}{\sqrt{n}} - \beta$ and the other singular values are at most β. Let $r = \frac{\sigma_2}{\sigma_1}(\widetilde{A})$. By the assumption that $\epsilon < \epsilon_0$ and $\lambda \geq 1/e$, we have $\sqrt{\lambda}(1-\epsilon) > 2\beta$. As a consequence, $r \leq \frac{2\beta}{\sqrt{\lambda}(1-\epsilon)} < 1$. Since $u^t = \widetilde{A}^t u^0/\|\widetilde{A}^t u^0\|$, it follows that

$$u^t = \frac{\langle u, u^0\rangle u + y}{\|\langle u, u^0\rangle u + y\|}$$

for some $y \in \mathbb{R}^m$, depending on t, such that $\|y\| \le r^t$. Hence,

$$\begin{aligned}\langle u^t, u\rangle &= \frac{\langle u, u^0\rangle + \langle y, u\rangle}{\|\langle u, u^0\rangle u + y\|} \\ &\ge \frac{\langle u, u^0\rangle - r^t}{\langle u, u^0\rangle + r^t} \\ &= 1 - \frac{2r^t}{\langle u, u^0\rangle + r^t},\end{aligned}$$

or, equivalently,

$$\|u^t - u\|^2 = 2(1 - \langle u^t, u\rangle) \le \frac{4r^t}{\langle u, u^0\rangle + r^t}. \tag{89}$$

Recall that $\widehat{u} = u^{\lceil s^* \log n\rceil}$. Thus, choosing $s^* = \frac{2}{\log(\sqrt{\lambda}(1-\epsilon)/(2\beta))}$ as in (33), we obtain $r^{\lceil s^* \log n\rceil} \le n^{-2}$ and consequently in view of (88) and (89),

$$\|\widehat{u} - u\|^2 \le \frac{4n^{-2}}{(2\sqrt{n\log n})^{-1} + n^{-2}} \le \frac{1}{n},$$

for sufficiently large n.

Therefore, by the triangle inequality,

$$\|\widehat{u} - v\| \le \|\widehat{u} - u\| + \|u - v\| \le n^{-1/2} + \frac{2\sqrt{2}\beta}{\sqrt{\lambda}(1-\epsilon)} \overset{(a)}{\le} \frac{3\beta}{\sqrt{\lambda}(1-\epsilon)} \triangleq \beta_o, \tag{90}$$

where (a) holds for sufficiently large n. Let $\widehat{C}_o$ be defined by using a threshold test to estimate C^* based on $\widehat{u}$:

$$\widehat{C}_o = \{i \in \widetilde{C} : |\widehat{u}_i| \ge \tau\},$$

where $\tau = 1 \Big/ \left(2\sqrt{|\widetilde{C} \cap C^*|}\right)$. Note that $v_i = 2\tau \mathbf{1}_{\{i \in \widetilde{C} \cap C^*\}}$. For any $i \in \widehat{C}_o \backslash (\widetilde{C} \cap C^*)$, we have $|\widehat{u}_i| \ge \tau$ and $v_i = 0$; for any $i \in (\widetilde{C} \cap C^*) \backslash \widehat{C}_o$, we have $|\widehat{u}_i| < \tau$ and $v_i = 2\tau$. Therefore $|\widehat{u}_i - v_i| \ge ||\widehat{u}_i| - |v_i|| \ge \tau$ for all $i \in \widehat{C}_o \triangle (\widetilde{C} \cap C^*)$ and thus

$$\|\widehat{u} - v\|^2 \ge |\widehat{C}_o \triangle (\widetilde{C} \cap C^*)| \tau^2.$$

In view of (90), the number of indices in $\widetilde{C}$ incorrectly classified by $\widehat{C}_o$ satisfies

$$|\widehat{C}_o \triangle (\widetilde{C} \cap C^*)| \le 4\beta_o^2 |\widetilde{C} \cap C^*| \le 4\beta_o^2 |C^*|.$$

Since $|C^* \backslash \widetilde{C}| \le \epsilon K$, we conclude that $|C^* \triangle \widehat{C}_o| \le \epsilon K + 4\beta_o^2 |C^*|$. Thus, if the algorithm were to output $\widehat{C}_o$ (instead of $\widehat{C}$) the lemma would be proved.

Rather than using a threshold test in the cleanup step, Algorithm 1 selects the K indices in $\widetilde{C}$ with the largest values of $|\widehat{u}_i|$. Consequently, with probability one, either $\widehat{C}_o \subset \widehat{C}$ or $\widehat{C} \subset \widehat{C}_o$. Therefore, it follows that

$$|C^* \triangle \widehat{C}| \le 2|C^* \triangle \widehat{C}_o| + \big||C^*| - K\big|.$$

By assumption, $|C^*|/K$ converges to one in probability, so that, in probability,

$$\limsup_{n\to\infty} \frac{|C^* \triangle \widehat{C}|}{K} \le 2\epsilon + 8\beta_o^2 \le \eta(\epsilon, \lambda), \tag{91}$$

where η is defined in (34), completing the proof. ■

C. An adaptive variant of Algorithm 1

Recall that the last step of the spectral clean-up of Algorithm 1 involves choosing the K coordinates of the largest magnitude of $\widehat{u}$. In order to be adaptive to the cluster size K, in this section we show that this step can be simply replaced by applying k-means clustering with $k = 2$ to $\{|\widehat{u}_i|\}_{i\in\widetilde{C}}$ so that Theorem 1 continues to hold. Let w denote an optimal solution, i.e., a minimizer of $\|x - |\widehat{u}|_{\widetilde{C}}\|$ over all x in $\mathbb{R}^{|\widetilde{C}|}$ whose coordinates take at most two distinct values. Since $|\widehat{u}_i|$ is a scalar, w can be easily found by sorting $\{|\widehat{u}_i|\}_{i\in\widetilde{C}}$ in descending order, and checking all vectors of the form $(a, \ldots, a, b, \ldots, b)$, where a and b with $a \ge b \ge 0$ are given by the average of the respective set of $|\widehat{u}_i|$'s. Thus w can be found in time $O(n \log n)$.

Define $\widehat{C} = \{i \in \widetilde{C} : w_i = a\}$. To show this $\widehat{C}$ fulfills the same performance guarantee as in Theorem 1, it suffices to modify the proof of Lemma 12 to show that, for any ϵ sufficiently small, if $\widetilde{C} \subset [n]$ satisfies (29) – (30), then $\mathbb{P}\left\{\frac{|C^* \Delta \widehat{C}|}{K} \le \eta\right\} \to 1$ as $n \to \infty$, where η is a function of ϵ and λ such that $\eta \to 0$ as $\epsilon \to 0$ for λ fixed. Without loss of generality we may also assume that

$$|\widetilde{C} \backslash C^*| = \Omega(K). \tag{92}$$

This extra condition is fulfilled by the output of the message-passing algorithm with high probability, because, in view of (32), $|\widetilde{C} \backslash C^*| = \Theta(n)$ with probability tending to one.

Recall that we have shown in (90) that $\min\{\|\widehat{u} - v\|, \|\widehat{u} + v\|\} \le \beta_o$. By the definition of w, since $v \ge 0$ is binary-valued componentwise, we have

$$\||\widehat{u}| - w\| \le \||\widehat{u}| - v\| \le \min\{\|\widehat{u} - v\|, \|\widehat{u} + v\|\} \le \beta_o,$$

and thus

$$\|w - v\| \le \|w - |\widehat{u}|\| + \||\widehat{u}| - v\| \le 2\beta_o.$$

Define

$$S = \{i \in \widetilde{C} : \big|w_i - v_i\big| \ge \tau\}, \qquad \tau \triangleq \frac{1}{2\sqrt{|\widetilde{C} \cap C^*|}}.$$

Then

$$|S|\tau^2 \le \|w - v\|^2 \le 4\beta_0^2,$$

and consequently, $|S| \le 16\beta_0^2 |\widetilde{C} \cap C^*|$. Since β_0 can be made to be sufficiently small by choosing ϵ to be small, we have $|S| < |\widetilde{C} \cap C^*|$. Furthermore, by the assumption that $|C^*|/K \to 1$ in probability and (92), we have $|S| < |\widetilde{C} \backslash C^*|$. Define $T_1 = (\widetilde{C} \cap C^*) \backslash S$ and $T_0 = (\widetilde{C} \backslash C^*) \backslash S$, both of which are non-empty. For each $i \in T_1$ and $j \in T_0$, we have

$$w_i - w_j \ge v_i - v_j - |w_i - v_i| - |w_j - v_j| > 2\tau - \tau - \tau = 0,$$

that is, $w_i = a > b = w_j$. Hence, $\widehat{C}\Delta(\widetilde{C} \cap C^*) \subset S$ and thus

$$|\widehat{C}\Delta(\widetilde{C} \cap C^*)| \leq |S| \leq 16\beta_0^2 |\widetilde{C} \cap C^*| \leq 16\beta_0^2 |C^*|.$$

Since $|C^* \backslash \widetilde{C}| \leq \epsilon K$, we have that $|\widehat{C}\Delta C^*| \leq \epsilon K + 16\beta_0^2 |C^*|$. Therefore,

$$\limsup_{n\to\infty} \frac{|C^*\Delta\widehat{C}|}{K} \leq \epsilon + 16\beta_0^2.$$

Since $\beta_0 \to 0$ as $\epsilon \to 0$, Theorem 1 holds for the adaptive variant of Algorithm 1.

D. Proofs of Theorems 13 and 14

In the proofs below we use the following notation. We write $\mathsf{p}_e(\pi_1, s^2)$ to denote the minimal average error probability for testing $\mathcal{N}(\mu_1, \sigma^2)$ versus $\mathcal{N}(\mu_0, \sigma^2)$ with priors π_1 and $1 - \pi_1$, where $\mu_1 \geq \mu_0$ and $s^2 = \frac{(\mu_0 - \mu_1)^2}{\sigma^2}$. That is,

$$\mathsf{p}_e(\pi_1, s^2) \triangleq \min_x \{\pi_1 Q(s - x) + (1 - \pi_1) Q(x)\}.$$

Proof [Proof of Theorem 13] The proof of sufficiency for weak recovery is closely based on the proof of sufficiency for exact recovery by the MLE given in (Butucea et al., 2015); the reader is referred to (Butucea et al., 2015) for the notation used in this paragraph. The proof in (Butucea et al., 2015) is divided into two sections. In our terminology, (Butucea et al., 2015, Section 3.1) establishes the weak recovery of C_1^* and C_2^* by the MLE under the assumptions (37), (39), and (41). However, the assumption (39) (and similarly, (41)) is used in only one place in the proof, namely for bounding the terms $T_{1,km}$ defined therein. We explain here why (37) alone is sufficient for the proof of weak recovery. Condition (37), in the notation[12] of (Butucea et al., 2015), implies that there exists some sufficiently small $\alpha > 0$ such that

$$\frac{a^2 m}{2\log(N/n)} \geq 1 + \alpha.$$

So (Butucea et al., 2015, (3.4)) can be replaced as: there exist some sufficiently small $\delta_1 > 0$ and $\alpha_1 > 0$ such that

$$\frac{(1-\delta_1)^2}{2} a^2 m \geq (1 + \alpha_1)\log(N/n) \geq (1 + \alpha_1)\log\left(\frac{\delta(N-n)}{n-k}\right),$$

where we use the assumption $0 \leq k < (1 - \delta)n$, or $n - k > \delta n$. Thus, for large enough n,

$$\begin{aligned} T_{1,km} &\leq \exp\left(-\frac{\delta n \alpha_1}{2}\left(\log\left(\frac{N-n}{n-k}\right)\right)\right) \\ &\leq \exp\left(-\frac{\delta n \alpha_1}{2}\log\left(\frac{N-n}{n}\right)\right) = o(1/n), \end{aligned}$$

12. The notation of (Butucea et al., 2015) is mapped to ours as $N \to n_1$, $M \to n_2$, $n \to K_1$, $m \to K_2$, and $a \to \mu$.

from which the desired conclusion, $\sum_{k:(n-k)>\delta n} T_{1,km} = o(1)$, follows. This completes the proof of sufficiency of (37) for weak recovery of both C_1^* and C_2^*, and marks the end of our use of notation from (Butucea et al., 2015).

The rate distortion argument used in the proof of (Hajek et al., 2017, Theorem 1) shows that (38) must hold if C_1^* and C_2^* are both weakly recoverable.

■

Proof [Proof of Theorem 14] We give the proof for exact recovery of C_1^*; the proof for exact recovery of C_2^* is analogous. For the sufficiency part, Recall that in Algorithm 3, the set $[n_1]$ is partitioned into sets, $S_1, \ldots, S_{1/\delta}$ of size $n_1\delta$. There are $1/\delta$ rounds of the algorithm, and indices in S_k are classified in the k^{th} round. For the k^{th} round, by assumption, given $\epsilon > 0$, there exists an estimator $\widehat{C}_{2k}$ based on observation of W with the rows indexed by S_k hidden such that $|\widehat{C}_{2k} \Delta C_2^*| \leq \epsilon K_2$ with high probability. Then the voting procedure estimates whether $i \in C_1^*$ for each $i \in S_k$ by comparing $\sum_{j \in \widehat{C}_{2k}} W_{i,j}$ to a threshold. This sum has approximately the $\mathcal{N}(K_2\mu, K_2)$ distribution if $i \in C_1^*$ and $\mathcal{N}(0, K_2)$ distribution otherwise ; the discrepancy can be made sufficiently small by choosing ϵ to be small (See (Hajek et al., 2017, Theorem 3) for a proof). Thus, the mean number of classification errors is well approximated by $n_1 \mathsf{p}_e(K_1/n_1, K_2\mu^2)$, which converges to zero under (39), completing the sufficiency proof for exact recovery of C_1^*. The necessity part is proved in (Butucea et al., 2015, Section 4.2). ■

E. Proof of Lemma 18

Proof (Similar to proof of Lemma 12.) We prove the lemma for $\widehat{C}_1$; the proof for $\widehat{C}_2$ is identical. For the first part of the proof we assume that for $i = 1, 2$, $\widetilde{C}_i$ is fixed, and later use a union bound over all possible choices of $\widetilde{C}_i$. Recall that $W_{\widetilde{C}_1\widetilde{C}_2}$, which we abbreviate henceforth as $\widetilde{W}$, is the matrix W restricted to entries in $\widetilde{C}_1 \times \widetilde{C}_2$. Let $Z = \widetilde{W} - \mathbb{E}[\widetilde{W}]$ and

$$\mathbb{E}[\widetilde{W}] = \mu\sqrt{|\widetilde{C}_1 \cap C_1^*||\widetilde{C}_2 \cap C_2^*|}v_1 v_2^\top \tag{93}$$

is a rank-one matrix, where v_i is the unit vector in $\mathbb{R}^{|\widetilde{C}_i|}$ obtained by normalizing the indicator vector of $\widetilde{C}_i \cap C_i^*$. Thus, thanks to (54), the leading singular value of $\mathbb{E}[\widetilde{W}]$ is at least $\mu\sqrt{K_1K_2}(1-\epsilon)$ with left singular vector v_1 and right singular vector v_2.

It is well-known (see, e.g., (Vershynin, 2010, Corollary 5.35)) that if M is an $m_1 \times m_2$ matrix with i.i.d. *standard normal* entries, then $\mathbb{P}\left\{\|M\| \geq \sqrt{m_1} + \sqrt{m_2} + t\right\} \leq 2e^{-t^2/2}$. Applying this result for $m_i = |\widetilde{C}_i|$, which satisfies $m_i \leq \epsilon n_i$ by (55), and $t = 2\sqrt{h(\epsilon)(n_1+n_2)}$, we have for fixed $(\widetilde{C}_1, \widetilde{C}_2)$,

$$\mathbb{P}\left\{\|Z\| \geq (\sqrt{n_1} + \sqrt{n_2})\beta\right\} \leq 2e^{-2(n_1+n_2)h(\epsilon)},$$

where $\beta \triangleq 3\sqrt{\epsilon + h(\epsilon)})$. Similar to the proof of Lemma 12, the number of $(\widetilde{C}_1, \widetilde{C}_2)$ that satisfies (55) is at most $e^{(n_1+n_2)h(\epsilon)}$. By union bound, if we drop the assumption that $\widetilde{C}_i$ is fixed for $i = 1, 2$, we still have that with high probability, $\|Z\| \leq (\sqrt{n_1} + \sqrt{n_2})\beta$.

Denote by u the leading left singular vector of $W_{\widetilde{C}_1\widetilde{C}_2}$ such that $\langle u, v_1\rangle \geq 0$. Then, letting θ denote the angle between u and v_1,

$$\|u - v_1\| \leq \sqrt{2}\sin(\theta) \overset{(a)}{\leq} \sqrt{2}\min\left\{\frac{\|Z\|}{\left(\sigma_1\left(\widetilde{W}\right) - \sigma_2\left(\mathbb{E}\left[\widetilde{W}\right]\right)\right)_+}, 1\right\} \overset{(b)}{\leq} \frac{2\sqrt{2}\|Z\|}{\sigma_1(\mathbb{E}[\widetilde{W}])},$$

where (a) follows from Wedin's sin-θ theorem for SVD (Wedin, 1972), and (b) follows from $\sigma_2\left(\mathbb{E}\left[\widetilde{W}\right]\right) = 0$ and Weyl's inequality $\sigma_1(\widetilde{W}) \geq \sigma_1(\mathbb{E}[\widetilde{W}]) - \|Z\|$. In view of (93), conditioning on the high-probability event that $\|Z\| \leq (\sqrt{n_1} + \sqrt{n_2})\beta$, we have

$$\|u - v_1\| \leq \frac{2\sqrt{2}\beta(\sqrt{n_1} + \sqrt{n_2})}{\mu(1-\epsilon)\sqrt{K_1K_2}} \leq \frac{2\sqrt{2}c_0\beta}{1-\epsilon}, \tag{94}$$

where the last inequality follows from the standing assumption (53).

Next, we argue that $\widehat{u}$ or $-\widehat{u}$ is close to u, and hence, close to v_1 by the triangle inequality. By (88), the initial value $u^0 \in \mathbb{R}^{|\widetilde{C}_1|}$ satisfies $|\langle u, u^0\rangle| \geq (2\sqrt{n_1 \log n_1})^{-1}$ with high probability, and without loss of generality we can assume as in the proof of Lemma 12 that $\langle u, u^0\rangle \geq (2\sqrt{n_1 \log n_1})^{-1}$. By Weyl's inequality, the largest singular value of $\widetilde{W}$ is at least $\mu\sqrt{K_1K_2}(1-\epsilon) - (\sqrt{n_1}+\sqrt{n_2})\beta$, and the other singular values are at most $(\sqrt{n_1}+\sqrt{n_2})\beta$. In view of (53), $\frac{1-\epsilon}{c_0\beta} - 1 > 1$ for all $\epsilon < \epsilon_0$, where $\epsilon_0 > 0$ depends only on c_0. Let λ_1 and λ_2 denote the first and second eigenvalue of $\widetilde{W}\widetilde{W}^\top$ in absolute value, respectively. Let $r = \lambda_2/\lambda_1$. Then $r \leq (\frac{c_0\beta}{1-\epsilon-c_0\beta})^2$. Since for even t, $u^t = (\widetilde{W}\widetilde{W}^\top)^{t/2}u^0/\|(\widetilde{W}\widetilde{W}^\top)^{t/2}u^0\|$, the same analysis of power iteration that leads to (89) yields

$$\|u^t - u\|^2 = 2(1 - \langle u^t, u\rangle) \leq \frac{4r^{t/2}}{\langle u, u^0\rangle + r^{t/2}}.$$

Since $\widehat{u} = u^{2\lceil s^* \log n\rceil}$ and $s^* = (\log\frac{1-\epsilon-c_0\beta}{c_0\beta})^{-1}$, we have $r^{\lceil s^* \log n_1\rceil} \leq n_1^{-2}$ and thus $|\langle\widehat{u}, u\rangle^2| \geq 1 - n_1^{-1}$ and consequently, $\|uu^\top - \widehat{u}(\widehat{u})^\top\|_{\mathrm{F}}^2 = 2 - 2\langle u, \widehat{u}\rangle^2 \leq n_1^{-1}$. Similar to (90), applying (94) and the triangle inequality, we obtain

$$\|\widehat{u} - v\| \leq \|\widehat{u} - u\| + \|u - v\| \leq n^{-1/2} + \frac{2\sqrt{2}c_o\beta}{\sqrt{\lambda}(1-\epsilon)} < \frac{3c_o\beta}{\sqrt{\lambda}(1-\epsilon)} \triangleq \beta_o, \tag{95}$$

By the same argument that proves (91), we have $\limsup_{n\to\infty} |C_1^* \triangle \widehat{C}_1|/K_1 \leq 2\epsilon + 8\beta_0^2 \leq \eta(\epsilon)$ with η defined in (57), completing the proof. ■

References

E. Abbe and C. Sandon. Community detection in general stochastic block models: fundamental limits and efficient recovery algorithms. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 670–688, 2015. arXiv 1503.00609.

E. Abbe, A. S. Bandeira, and G. Hall. Exact recovery in the stochastic block model. *IEEE Transactions on Information Theory*, 62(1):471–487, 2016.

N. Alon, M. Krivelevich, and B. Sudakov. Finding a large hidden clique in a random graph. *Random Structures and Algorithms*, 13(3-4):457–466, 1998.

A. Amini, A. Chen, P. J. Bickel, and E. Levina. Pseudo-likelihood methods for community detection in large sparse networks. *The Annals of Statistics*, 41(4):2097–2122, 2013.

R. B. Ash. *Information Theory*. Dover Publications Inc., New York, NY, 1965.

M. Bayati and A. Montanari. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory*, 57(2): 764–785, 2011.

C. Butucea and Y. I. Ingster. Detection of a sparse submatrix of a high-dimensional noisy matrix. *Bernoulli*, 19(5B):2652–2688, 11 2013. doi: 10.3150/12-BEJ470.

C. Butucea, Y. Ingster, and I. Suslina. Sharp variable selection of a sparse submatrix in a high-dimensional noisy matrix. *ESAIM: Probability and Statistics*, 19:115–134, June 2015.

T. T. Cai, T. Liang, A. Rakhlin, et al. Computational and statistical boundaries for submatrix localization in a large noisy matrix. *The Annals of Statistics*, 45(4):1403–1430, 2017.

Y. Chen and J. Xu. Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices. In *Proceedings of ICML 2014 (Also arXiv:1402.1267)*, Feb 2014.

K. Chung. *A course in probability theory*. Academic press, 2nd edition, 2001.

A. Condon and R. M. Karp. Algorithms for graph partitioning on the planted partition model. *Random Struct. Algorithms*, 18(2):116–140, Mar. 2001.

C. Davis and W. Kahan. The rotation of eigenvectors by a perturbation. III. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

Y. Deshpande and A. Montanari. Finding hidden cliques of size $\sqrt{N/e}$ in nearly linear time. *Foundations of Computational Mathematics*, 15(4):1069–1128, August 2015.

C.-G. Esseen. "on the liapunoff limit of error in the theory of probability". *Arkiv för matematik, astronomi och fysik*, A28:1–19, 1942.

D. Féral and S. Péché. The largest eigenvalue of rank one deformation of large Wigner matrices. *Communications in mathematical physics*, 272(1):185–228, 2007.

B. Hajek, Y. Wu, and J. Xu. Semidefinite programs for exact recovery of a hidden community. In *Proceedings of Conference on Learning Theory (COLT)*, pages 1051–1095, New York, NY, Jun 2016. arXiv:1602.06410.

B. Hajek, Y. Wu, and J. Xu. Information limits for recovering a hidden community. *IEEE Trans. on Information Theory*, 63(8):4729 – 4745, 2017.

J. A. Hartigan. Direct clustering of a data matrix. *Journal of the American Statistical Association*, 67(337):123–129, 1972.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

A. Javanmard and A. Montanari. State evolution for general approximate message passing algorithms, with applications to spatial coupling. *Information and Inference: A Journal of the IMA*, 2(2):115–144, 2013.

A. Knowles and J. Yin. The isotropic semicircle law and deformation of Wigner matrices. *Communications on Pure and Applied Mathematics*, 66(11):1663–1749, 2013.

M. Kolar, S. Balakrishnan, A. Rinaldo, and A. Singh. Minimax localization of structural information in large noisy matrices. In *Advances in Neural Information Processing Systems*, 2011.

Z. Ma and Y. Wu. Computational barriers in minimax submatrix detection. *The Annals of Statistics*, 43(3):1089–1116, 2015.

A. Montanari, D. Reichman, and O. Zeitouni. On the limitation of spectral methods: From the Gaussian hidden clique problem to rank one perturbations of Gaussian tensors. In *Advances in Neural Information Processing Systems*, pages 217–225, 2015. arXiv 1411.6149.

E. Mossel, J. Neeman, and S. Sly. Belief propagation, robust reconstruction, and optimal recovery of block models (extended abstract). In *JMLR Workshop and Conference Proceedings (COLT proceedings)*, volume 35, pages 1–35, 2014.

E. Mossel, J. Neeman, and A. Sly. Consistency thresholds for the planted bisection model. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing*, STOC '15, pages 69–75, New York, NY, USA, 2015. ACM.

V. V. Petrov. *Limit theorems of probability theory: Sequences of independent random variables*. Oxford Science Publications, Clarendon Press, Oxford, United Kingdom, 1995.

A. A. Shabalin, V. J. Weigman, C. M. Perou, and A. B. Nobel. Finding large average submatrices in high dimensional data. *The Annals of Applied Statistics*, 3(3):985–1012, 2009.

G. Szegö. *Orthogonal polynomials*. American Mathematical Society, Providence, RI, 4th edition, 1975.

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *Arxiv preprint arxiv:1011.3027*, 2010.

P. Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111, 1972.

S. Yun and A. Proutiere. Optimal cluster recovery in the labeled stochastic block model. arXiv 1510.05956, Oct. 2015.