2019 57th Annual Allerton Conference on Communication, Control, and Computing
(Allerton)
Allerton Park and Retreat Center
Monticello, IL, USA, September 24-27, 2019

Mutual Information in Community Detection with Covariate Information and Correlated Networks

Vaishakhi Mayya* and Galen Reeves*†
*Department of Electrical and Computer Engineering, Duke University
†Department of Statistical Science, Duke University

***Abstract*—We study the problem of community detection when there is covariate information about the node labels and one observes multiple correlated networks. We provide an asymptotic upper bound on the per-node mutual information as well as a heuristic analysis of a multivariate performance measure called the MMSE matrix. These results show that the combined effects of seemingly very different types of information can be characterized explicitly in terms of formulas involving low-dimensional estimation problems in additive Gaussian noise. Our analysis is supported by numerical simulations.**

I. Introduction

Networks model relational data between various nodes, e.g. friendship networks in schools or social media. The community detection problem aims to classify the nodes of a network based on those relationships into various communities. The stochastic block model (SBM) is a generative model for a network where each node belongs to exactly one of k communities and the probability of an edge between two nodes is exclusively a function of their community memberships. In this setting, the goal of community detection is to recover the community labels from the observed network.

A recent line of work has studied the information-theoretic limits of recovery. Most of this work has focused on either the two-community SBM [1]–[8] or the so-called k-community symmetric SBM [6], [9]–[11]. In all of these cases, performance is summarized in terms of a single numerical value, which is often referred to as the effective signal-to-noise ratio of the problem. General SBMs have been considered by Abbe and Sandon [9] who characterize conditions for weak recovery, Lesieuir et al. [6] who analyze the performance of an approximate message passing algorithm, and Reeves et. al [12] who study the asymptotic per-node mutual information and MMSE in degree-balanced SBMs.

The contribution of this paper is to extend the analysis in [12] to the setting where one observes:

1) covariate information about the node labels; and
2) multiple networks that are conditionally independent given the same underlying node labels.

Section II gives the problem formulation and describes connections with previous work. Section III provides the main theoretical results, which are upper bounds on mutual information. Numerical simulations are provided in Section IV.

Notation: We use $\mathbb{S}^d$, $\mathbb{S}^d_+$ to denote the space of $d \times d$ symmetric matrices and symmetric positive semi-definite matrices, respectively. Given a positive semi-definite matrix S, we use $S^{1/2}$ to denote the unique positive semi-definite square root. Given matrices $A, B \in \mathbb{S}^d$, the relation $A \preceq B$ means that $B - A \in \mathbb{S}^d_+$.

II. Problem Formulation and Related Work

A. Node labels and covariate information

The labels and covariate information associated with a collection of n nodes are modeled in terms of an i.i.d. sequence of tuples $\{(X_i, Y_i, \tilde{Y}_i)\}_{i=1}^n$ where X_i is the unknown node label and $(Y_i, \tilde{Y}_i)$ is observed covariate information associated with the i-th node.

We focus on the problem of community detection where each label takes exactly one of k values with probability vector $p = (p_1, \dots, p_k)$. Without loss of generality these labels can be embedded into finite dimensional Euclidean space. To facilitate the exposition of our results, we use the whitened representation described in [12], where the labels are supported on a set of k points in $\{\mu_1, \dots, \mu_k\}$ in $\mathbb{R}^{k-1}$ with the property that

$$\sum_{a=1}^k p_a \mu_a = 0, \qquad \sum_{a=1}^k p_a \mu_a \mu_a^T = I. \tag{1}$$

A unique specification of this whitened representation is described in [12, Remark 1].

There are two types of the covariate information. The terms Y_i are supported on a set $\mathcal{Y}$ and are used to model general information about the nodes. The terms $\tilde{Y}_i$ correspond to the output of linear Gaussian channel described by

$$Y_i = S^{1/2} X_i + N_i \tag{2}$$

where $S \in \mathbb{S}^{k-1}_+$ is known and $N_i \sim \mathcal{N}(0, I_{k-1})$ is independent Gaussian noise. These terms play a fundamental role in our proof technique.

Furthermore, we define the information function $\mathcal{I}(S) : \mathbb{S}^{k-1}_+ \to \mathbb{R}$ and MMSE function $M(S) : \mathbb{S}^{k-1}_+ \to \mathbb{S}^{k-1}_+$ according to

$$\mathcal{I}(S) \triangleq I(X_1; Y_1, \tilde{Y}_1) \tag{3}$$

$$M(S) \triangleq \mathbb{E}\Big[\mathrm{Cov}(X_1 \mid Y_1, \tilde{Y}_1)\Big], \tag{4}$$

where S appears in the definition of $\tilde{Y}_i$. The matrix version of the I-MMSE relation [13] states that

$$\nabla_S \mathcal{I}(S) = \frac{1}{2} M(S).$$

978-1-7281-3151-1/19/$31.00 ©2019 IEEE

Finally, the collection of node labels is represented by an $n\times(k-1)$ matrix $\boldsymbol{X} = (X_1, \dots X_n)^T$. Similarly, the covariate information is denoted by matrices $\boldsymbol{Y} = (Y_1, \dots, Y_n)^T$ and $\tilde{\boldsymbol{Y}} = (\tilde{Y}_1, \dots, \tilde{Y}_n)^T$ with $\underline{\boldsymbol{Y}} = (\boldsymbol{Y}, \tilde{\boldsymbol{Y}})$.

B. Correlated networks

We consider the setting where one observes multiple networks $\boldsymbol{G}_1, \dots, \boldsymbol{G}_L$ that are conditionally independent given the labels $\boldsymbol{X}$. Each network is represented by an $n\times n$ binary adjacency matrix $\boldsymbol{G}_\ell = (G_{\ell ij})$ where $G_{\ell ij} = G_{\ell ji} = 1$ if there is an edge between nodes i and j and zero otherwise. Following [12], each network is drawn according to a degree-balanced SBM of the form

$$G_{\ell ij} \sim \text{Ber}\left(\frac{d_\ell}{n} + \frac{\sqrt{d_\ell(1-d_\ell/n)}}{n} X_i^T R_\ell X_j\right), \quad i<j, \tag{5}$$

where d_ℓ is a positive real number that parameterizes the expected degree of each node in the network and R_ℓ is a symmetric $(k-1)\times(k-1)$ matrix that describes the relationship between the community labels and the probability of an edge. We assume that the parameters (d_ℓ, R_ℓ) are known and we use $\underline{\boldsymbol{G}} = (\boldsymbol{G}_1, \dots, \boldsymbol{G}_L)$ to denote the collection of networks.

C. Multivariate performance metric

The ability to recover the labels $\boldsymbol{X}$ from the observations $(\underline{\boldsymbol{Y}}, \underline{\boldsymbol{G}})$ is assessed in terms of the MMSE matrix:

$$\mathsf{MMSE}(\boldsymbol{X} \mid \underline{\boldsymbol{Y}}, \underline{\boldsymbol{G}}) \triangleq \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\text{Cov}(\boldsymbol{X} \mid \underline{\boldsymbol{Y}}, \underline{\boldsymbol{G}})], \tag{6}$$

where the expectation is taken with respect to $(\underline{\boldsymbol{Y}}, \underline{\boldsymbol{G}})$. By the matrix I-MMSE relation [13], this matrix can also be expressed as the gradient of the mutual information with respect to the matrix SNR:

$$\mathsf{MMSE}(\boldsymbol{X} \mid \underline{\boldsymbol{Y}}, \underline{\boldsymbol{G}}) = 2\nabla_S I(\boldsymbol{X}; \underline{\boldsymbol{Y}}, \underline{\boldsymbol{G}}).$$

Moreover, by the data processing inequality for covariance and the assumption that the rows of $\boldsymbol{X}$ drawn from the whitened representation, $0 \preceq \mathsf{MMSE}(\boldsymbol{X} \mid \underline{\boldsymbol{G}}, \underline{\boldsymbol{Y}}) \preceq \mathrm{I}_{k-1}$.

Notice that in the absence of network observations $\underline{\boldsymbol{G}}$, the problem of estimating $\boldsymbol{X}$ from the covariate information $\underline{\boldsymbol{Y}}$ decouples into n independent problems and we have:

$$\frac{1}{n} I(\boldsymbol{X}; \underline{\boldsymbol{Y}}) = \mathcal{I}(S) \tag{7}$$

$$\mathsf{MMSE}(\boldsymbol{X} \mid \underline{\boldsymbol{Y}}) = M(S). \tag{8}$$

These terms involve $(k-1)$-dimensional integrals that can be approximated numerically for small values of k. The problem of estimating the node labels in the presence of network observations is more difficult to analyze because the networks induce dependence in the conditional distribution of the labels.

D. Relation to prior work

In recent years, there has been significant interest in community detection in the fields of statistical physics and information theory. The research focuses on quantifying the performance of community detection as a function of the graph parameters, and identifying algorithms that achieve these limits. In [1], the authors conjectured that if community sizes are equal and $R = r\mathrm{I}_{k-1}$, for any $|r| > 1$, it is possible to recover the community labels using a polynomial time algorithm better than random chance. Since then, a number of works have proven the conjecture, and extended the results to general SBMs with two communities. In [12], the authors extended the results to obtain upper bounds to the performance for degree balanced SBMs for $k \geq 2$.

In networks with two communities, it has been shown that revealing additional information about the nodes can measurably improve community detection. The effect of node-wise i.i.d. covariate information on community detection has been studied in [14]–[17].

Community detection with multiple networks has been studied in [18], [19]. In [18], the authors use the central limit theorem to predict the performance in settings where the communities are well separated (when the eigenvalues of R are very large). To the best of our knowledge, the information theoretic limits of optimal algorithms with multiple correlated networks have not been addressed. With our work, we can provide theoretical performance limits over a broad range of parameters for the degree-balanced setting.

III. Formulas for Mutual information and MMSE

A. Upper bound on the mutual information

Our analysis focuses on a sequence of problem settings where the number of nodes n scales to infinity. We assume that node labels and covariate information are drawn i.i.d. according to the distribution on $(X_1, Y_1, \tilde{Y_1})$ and the matrices $\{R_\ell\}$ are fixed. We make two additional assumptions.

Assumption 1 (Diverging Average Degree)**.** The average degree of each network d_ℓ increases with n such that both d_ℓ and $(n-d_\ell)$ tend to infinity.

Assumption 2 (Definite Matrix)**.** Each matrix R_ℓ is either positive definite or negative definite.

Our results are stated in terms of a potential function. Let $\mathcal{U} = \{U \in \mathbb{S}_+^{k-1} : U \preceq I\}$ and let $\mathcal{F} : \mathcal{U} \to [0, \infty)$ be defined as

$$\mathcal{F}(U) \triangleq \mathcal{I}\Big(S + \sum_{\ell=1}^{L} R_\ell(I-U)R_\ell\Big) + \frac{1}{4}\sum_{\ell=1}^{L} \text{tr}((R_\ell U)^2). \tag{9}$$

The following result provides an asymptotic upper bound on the per-node mutual information between $\boldsymbol{X}$ and the observations $(\underline{\boldsymbol{Y}}, \underline{\boldsymbol{G}})$. The proof is given in Section V.

Theorem 1. *Under Assumptions 1 and 2,*

$$\limsup_{n\to\infty} \frac{1}{n} I(\boldsymbol{X}; \underline{\boldsymbol{Y}}, \underline{\boldsymbol{G}}) \leq \min_{U\in\mathcal{U}} \mathcal{F}(U). \tag{10}$$

Theorem 1 provides an extension of [12], which focused on the setting of a single network ($L = 1$) without the covariate information provided by $\boldsymbol{Y}$. In this setting, [12, Theorem 1] shows that the upper bound is asymptotically tight when $S = 0$, that is

$$\lim_{n\to\infty} \frac{1}{n} I(\boldsymbol{X}; \underline{\boldsymbol{G}}) = \min_{U\in\mathcal{U}} \left\{ \mathcal{I}(R(I-U)R) + \frac{1}{4}\operatorname{tr}((RU)^2) \right\}. \tag{11}$$

B. Partially revealed labels

As a specific example of covariate information, consider the setting where a fraction of the true node labels are revealed. This is also referred to as the semi-supervised setting [14]. Using the setup introduced in Section II-A, partially revealed labels can be modeled using an erasure channel, where Y_i is equal to X_i with probability α and is equal to an erasure symbol with probability $1-\alpha$. In this setting, the mutual information function is given by

$$\mathcal{I}(S) = \alpha H(X_1) + (1-\alpha) I(X_1; \tilde{Y}_1) \tag{12}$$

where $H(X_1) = \sum_{a=1}^{k} -p_a \log p_a$ is the entropy of the community labels.

C. Heuristic analysis of MMSE matrix

The MMSE matrix is related to the mutual information via the matrix I-MMSE relation [13], which implies

$$I(\boldsymbol{X}; \underline{\boldsymbol{Y}}, \underline{\boldsymbol{G}}) - I(\boldsymbol{X}; \boldsymbol{Y}, \underline{\boldsymbol{G}}) = \frac{n}{2} \int_0^1 \operatorname{tr}\left(\mathsf{MMSE}(\boldsymbol{X} \mid \underline{\boldsymbol{G}}, \underline{\boldsymbol{Y}}) \Big|_{S=S_\gamma} \frac{\mathrm{d}}{\mathrm{d}\gamma} S_\gamma \right) \mathrm{d}\gamma,$$

for any differentiable path S_γ with $S_0 = 0$ and $S_1 = S$. Following the approach outlined in [12, Appendix A.3], it can be shown that upper and lower bounds on the asymptotic per-node mutual information lead to asymptotic bounds on the MMSE matrix. In particular, for the special case of a single network without covariate information, [12, Theorem 3] shows that, for any positive definite S,

$$\mathsf{MMSE}(\boldsymbol{X} \mid \tilde{\boldsymbol{Y}}, \boldsymbol{G}) \preceq U^* + o_n(1),$$

where U^* is any minimizer of $\mathcal{F}(U)$ and $o_n(1)$ denotes a sequence of symmetric matrices that converges to zero in the large-n limit.

Our next result follows a similar approach for the setting of multiple networks and covariate information. This result requires the additional assumption that the upper bound on the mutual information in Theorem 1 is asymptotically tight for $S = 0$. Because this assumption is unproven, the resulting upper bound is considered to be heuristic.

Theorem 2. *Consider Assumptions 1 and 2. If the upper bound in Theorem 1 is asymptotically tight at $S = 0$, that is*

$$\lim_{n\to\infty} \frac{1}{n} I(\boldsymbol{X}; \underline{\boldsymbol{G}}, \boldsymbol{Y}) = \min_{U\in\mathcal{U}} \left\{ \mathcal{I}\Big(\sum_{\ell=1}^{L} R_\ell (I-U) R_\ell \Big) + \frac{1}{4} \sum_{\ell=1}^{L} \operatorname{tr}((R_\ell U)^2) \right\}$$

then, for any positive definite S, the MMSE matrix satisfies

$$\mathsf{MMSE}(\boldsymbol{X} \mid \underline{\boldsymbol{G}}, \underline{\boldsymbol{Y}}) \preceq U^* + o_n(1), \tag{13}$$

where U^* is any minimizer of $\mathcal{F}(U)$ and $o_n(1)$ denotes a sequence of symmetric matrices that converges to zero in the large-n limit.

IV. Simulation Results

A. Covariate information

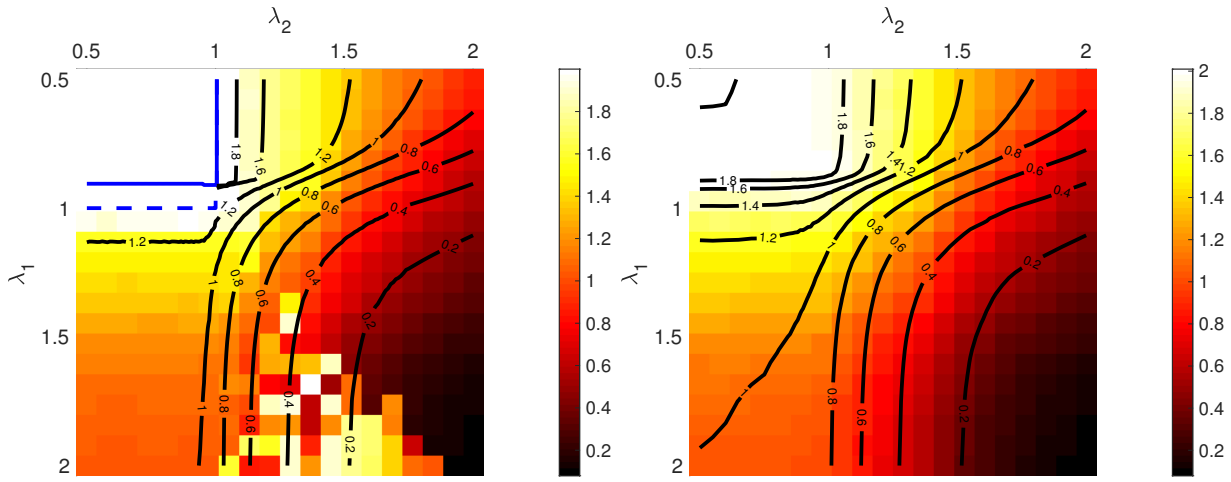
We first consider the effects of partially revealed labels in the setting of a single network observation. Results are obtained on a problem with $n = 10^5$ nodes and $k = 3$ communities with probability vector $p = (0.1, 0.3, 0.6)$. Conditional on the node labels, the network is drawn according to a degree-balanced SBM with average degree $d = 30$ and $R = \operatorname{diag}(\lambda_1, \lambda_2)$. The covariate information in $\boldsymbol{Y}$ consists of the output of an erasure channel, as described in Section III-B.

We compare our theoretical results with the empirical performance of belief propagation (BP). For each problem setting the MSE is estimated according to $\frac{1}{n}\sum_{i=1}^{n} \|X_i - \hat{X}_i\|^2$, where $\hat{X}_i$ is the BP estimate of the i-th label. We note that this evaluation of the MSE differs slightly from much of the prior work, which focus on uniform community assignments and include an additional step that minimizes over all permutations of community labels. This additional step is not needed in our setting due to the non-uniformity in community sizes.

Figure 1 provides a comparison of the heuristic upper bound on $\operatorname{tr}(\mathsf{MMSE}(\boldsymbol{X} \mid \underline{\boldsymbol{Y}}, \boldsymbol{G})$ given in Theorem 2 and the empirical MSE of BP, where each pixel is the median of 8 independent trials. The axes correspond to the eigenvalues of R. Figure 1(a) corresponds to the setting without covariate information and Figure 1(b) corresponds to the setting where 1% of the labels are revealed.

Similar to previous work focusing on partially revealed labels [14]–[17], Figure 1 shows that a relatively small amount of extra information can provide significant performance gains. One of main takeaways from Figure 1 is that there is a close qualitative correspondence between the heuristic upper bound given in this paper and the empirical performance.

Finally, we note that there is a region in Figure 1(a) where BP becomes unstable. We suspect that this may be a consequence of asymmetries in the network model.



(a) No covariate information

(b) 1% of labels revealed

Fig. 1: Comparison of the heuristic upper bound on $\mathrm{tr}(\mathsf{MMSE}(\boldsymbol{X} \mid \underline{\boldsymbol{Y}}, \boldsymbol{G})$ given in Theorem 2 (black contour lines) with the empirical MSE of BP (heat map). In the left panel, the solid blue line is the upper bound on the weak recovery threshold given in [12, Theorem 5] and the dashed blue line is the weak recovery threshold for acyclic BP [9].

B. Correlated networks

Next we consider the effects of multiple network observations. Results are obtained for a problem with $n = 10^4$ nodes and $k = 3$ communities with non-uniform probability vector $p = (0.1, 0.3, 0.6)$. Conditional on the labels, two networks are drawn according to the degree-balanced SBM with average degree $d = 30$ and $R_\ell = rI_2$.

In this setting, we found that the BP has convergence issues and so we compare our theoretical results with the empirical performance of a spectral method [20] applied to a linear combination of the adjacency matrices. Specifically, we obtain estimates of the community labels using the following procedure. First, we construct the average of the networks $\boldsymbol{G}_1$ and $\boldsymbol{G}_2$ according to

$$\tilde{\boldsymbol{G}} = \frac{1}{\sqrt{2}}\boldsymbol{G}_1 + \frac{1}{\sqrt{2}}\boldsymbol{G}_2. \tag{14}$$

Note that the conditional expectation of $\tilde{\boldsymbol{G}}$ given $\boldsymbol{X}$ is comparable to that of a single network with $\tilde{R} = \sqrt{2}r\mathrm{I}$. Next, we retain the eigenvectors associated with the second and third leading eigenvalues in the spectral decomposition of $\tilde{\boldsymbol{G}}$. The relationship between these eigenvectors and the node labels is characterized using a Gaussian mixture model (GMM) approach described in [20], evaluated with $\tilde{R}$.

Figure 2 shows the MSE as a function of the SBM parameter r. The solid blue line corresponds to the trace of the heuristic upper bound to the MMSE for two correlated networks computed from Theorem 2, and the red line corresponds to the upper bound for a single network. The black line corresponds to the empirical observations using the method described in this section. With multiple correlated networks, we see that the MSE shows an improvement in the presence of additional information, and our proposed asymptotic upper bound follows the observed performance.

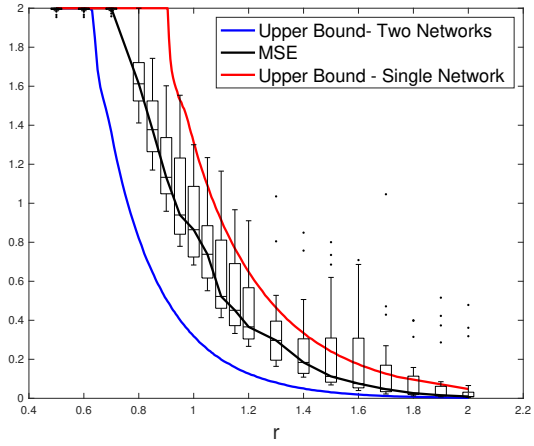


Fig. 2: MSE as a function of the SBM parameter r.

V. Proof of Theorem 1

The proof of Theorem 1 follows the approach in [12] with appropriate modifications to handle the covariate information and multiple networks. The first step of the proof is to establish an asymptotic equivalence between the mutual information in the community detection problem and the mutual information in the symmetric matrix estimation problem defined by

$$\boldsymbol{W}_\ell = \frac{1}{\sqrt{n}}\boldsymbol{X}R_\ell\boldsymbol{X}^T \tag{15}$$

$$\boldsymbol{Z}_\ell = \sqrt{t}\boldsymbol{W}_\ell + \boldsymbol{\xi}_\ell, \tag{16}$$

where $\boldsymbol{\xi}$ is a symmetric matrix with $\xi_{ij} \sim \mathcal{N}(0,1)$ for $i < j$ and $\xi_{ii} \sim \mathcal{N}(0,2)$. We use $\underline{\boldsymbol{Z}} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_L)$ to denote the collection of matrix observations.

Lemma 3 (Channel Universality)**.** *Under Assumption 1,*

$$\lim_{n\to\infty}\frac{1}{n}|I(\boldsymbol{X};\underline{\boldsymbol{Y}},\underline{\boldsymbol{G}})-I(\boldsymbol{X};\underline{\boldsymbol{Y}},\underline{\boldsymbol{Z}})|=0. \tag{17}$$

Proof. To simplify the expression, we will prove the result without $\boldsymbol{Y}$. The result can then be extended to the setting with $\boldsymbol{Y}$ following the approach used in [12, Corollary 7].

To proceed, let us define $a_1 = I(\boldsymbol{X};\underline{\boldsymbol{Z}})$, $a_{L+1} = I(\boldsymbol{X};\underline{\boldsymbol{G}})$, and

$$a_\ell = I(\boldsymbol{X};\boldsymbol{G}_1,\dots,\boldsymbol{G}_{\ell-1},\boldsymbol{Z}_\ell,\dots,\boldsymbol{Z}_L),$$

for $\ell = 2,\dots,L$. By the triangle inequality, we can then write

$$|I(\boldsymbol{X};\underline{\boldsymbol{G}})-I(\boldsymbol{X};\underline{\boldsymbol{Z}})| = \left|\sum_{\ell=1}^{L} a_{\ell+1}-a_\ell\right| \le \sum_{\ell=1}^{L}|a_{\ell+1}-a_\ell|.$$

Next, by the chain rule for mutual information one finds that

$$\begin{aligned} a_{\ell+1}-a_\ell &= I(\boldsymbol{X};\boldsymbol{G}_\ell \mid \mathcal{D}_\ell) - I(\boldsymbol{X};\boldsymbol{Z}_\ell \mid \mathcal{D}_\ell) \\ &= I(\boldsymbol{W}_\ell;\boldsymbol{G}_\ell \mid \mathcal{D}_\ell) - I(\boldsymbol{W}_\ell;\boldsymbol{Z}_\ell \mid \mathcal{D}_\ell), \end{aligned}$$

where $\mathcal{D}_\ell = (\boldsymbol{G}_1,\dots,\boldsymbol{G}_{\ell-1},\boldsymbol{Z}_{\ell+1},\dots\boldsymbol{Z}_L)$. Under the assumed distribution on $\boldsymbol{W}_\ell$, we can apply [12, Theorem 6] to show that $\frac{1}{n}|a_{\ell+1}-a_\ell|$ converges to zero in the large-n limit. □

The next step in our proof is to obtain an upper bound on $I(\boldsymbol{X};\underline{\boldsymbol{Y}},\underline{\boldsymbol{Z}})$. We define the function

$$\mathcal{I}(S,t) \triangleq \frac{1}{n} I(\boldsymbol{X};\underline{\tilde{\boldsymbol{Y}}},\underline{\boldsymbol{Z}}). \tag{18}$$

where we note that $\mathcal{I}(S,0) = \mathcal{I}(S)$ is the information function defined in (3). The function $\mathcal{I}(S,t)$ is concave and differentiable in (S,t) with

$$\nabla_S \mathcal{I}(S,t) = \frac{1}{2}\,\mathsf{MMSE}(\boldsymbol{X} \mid \underline{\boldsymbol{Y}},\underline{\boldsymbol{Z}}). \tag{19}$$

The next result provides an upper bound on the partial derivative with respect to t.

Lemma 4. *Under Assumption 2,*

$$\partial_t \mathcal{I}(S,t) \le \frac{1}{4}\sum_{\ell=1}^{L} g_\ell(2\nabla_S \mathcal{I}(S,t)) \tag{20}$$

where

$$g_\ell(U) = \frac{1}{n^2}\operatorname{tr}\big(\mathbb{E}\big[(R_\ell \boldsymbol{X}^T\boldsymbol{X})^2\big]\big) - \operatorname{tr}\big((R_\ell(I-U))^2\big).$$

Proof. Suppose that each observation $\boldsymbol{Z}_\ell$ has a separate parameters t_ℓ. By the chain rule for differentiation, we can then write

$$\partial_t \mathcal{I}(S,t) = \sum_{\ell=1}^{L} \partial_{t_\ell}\frac{1}{n} I(\boldsymbol{X};\underline{\boldsymbol{Y}},\underline{\boldsymbol{Z}})\Big|_{t_1=\dots t_L=t} \tag{21}$$

Furthermore, by the chain rule for mutual information and the fact that $\boldsymbol{Z}_\ell$ is conditionally independent of everything else given $\boldsymbol{W}_\ell$, we have

$$\partial_{t_\ell} I(\boldsymbol{X};\underline{\boldsymbol{Y}},\underline{\boldsymbol{Z}}) = \partial_{t_\ell} I(\boldsymbol{W}_\ell;\boldsymbol{Z}_\ell \mid \underline{\boldsymbol{Y}},\boldsymbol{Z}_{\sim\ell}),$$

where the subscript $\sim\ell$ means that the ℓ-th term is omitted.

Following the steps outlined in outlined in [12, Appdendix D] and the proof of [12, Lemma 11], one finds that the $\partial_{t_\ell}\mathcal{I}(S,t) \le \frac{1}{4} g_\ell(\nabla_S \mathcal{I}(S,t)$. Plugging this inequality back into the expression above completes the proof. □

Having established Lemma 4, the rest of the proof follows similarly to the proof of Theorem 8 in [12]. Specifically, we obtain

$$\mathcal{I}(S,1) \le \min_{U\in\mathcal{U}}\left\{\mathcal{I}^*(U) + \frac{1}{2}\operatorname{tr}(SU) + \frac{1}{4}\sum_{\ell=1}^{L} g_\ell(U)\right\}, \tag{22}$$

where

$$\mathcal{I}^*(U) = \sup_{S\succeq 0}\left\{\mathcal{I}(S) - \frac{1}{2}\operatorname{tr}(SU)\right\} \tag{23}$$

is the convex conjugate of $\mathcal{I}(S)$.

For the final step in the proof, observe that

$$g_\ell(U) = \delta_\ell + 2\operatorname{tr}(R_\ell^2 U) - \operatorname{tr}\big((R_\ell U)^2\big)$$

where $\delta_\ell = \frac{1}{n^2}\operatorname{tr}\big(\mathbb{E}\big[(R_\ell(\boldsymbol{X}^T\boldsymbol{X}-I))^2\big]\big)$. For all $\tilde{U}\in\mathcal{U}$, the inequality

$$-\operatorname{tr}((R_\ell U)^2) \le -2\operatorname{tr}(R_\ell U R_\ell \tilde{U}) + \operatorname{tr}((R_\ell\tilde{U})^2),$$

leads to

$$g_\ell(U) \le \delta_\ell + 2\operatorname{tr}(R(I-\tilde{U})RU) + \operatorname{tr}((R_\ell\tilde{U})^2)$$

Combining this inequality with (22), we see that, for all $U,\tilde{U}$ in $\mathcal{U}$,

$$\begin{aligned} \mathcal{I}(S,1) \le \mathcal{I}^*(U) &+ \frac{1}{2}\sum_{\ell=1}^{L}\operatorname{tr}((S+R_\ell(I-\tilde{U})R_\ell)U) \\ &+ \frac{1}{4}\sum_{\ell=1}^{L}\operatorname{tr}((R_\ell\tilde{U})^2) + \frac{1}{4}\sum_{\ell=1}^{L}\delta_\ell. \end{aligned} \tag{24}$$

The minimum of the first two terms with respect to U then leads to

$$\begin{aligned} &\min_{U\in\mathcal{U}}\left\{\mathcal{I}^*(U) + \frac{1}{2}\sum_{\ell=1}^{L}\operatorname{tr}((S+R_\ell(I-\tilde{U})R_\ell)U)\right\} \\ &\qquad = \mathcal{I}\Big(S + \sum_{\ell=1}^{L} R_\ell(I-\tilde{U})R_\ell\Big). \end{aligned}$$

where we have used the fact that $\mathcal{I}(S)$ is concave, and thus equal to its biconjugate. Plugging this expression back into (24) and then taking the the minimum with respect to $\tilde{U}$ yields,

$$\mathcal{I}(S,1) \le \min_{\tilde{U}\in\mathcal{U}} \mathcal{F}(\tilde{U}) + \frac{1}{4}\sum_{\ell=1}^{L}\delta_\ell. \tag{25}$$

Under the assumed distribution on $\boldsymbol{X}$ each term δ_ℓ vanishes in the large-n limit. Combining (25) with Lemma 3 completes the proof of Theorem 1

REFERENCES

[1] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová, "Inference and phase transitions in the detection of modules in sparse networks," *Physical Review Letters*, vol. 107, no. 6, 2011.

[2] Y. Deshpande, E. Abbe, and A. Montanari, "Asymptotic mutual information for the balanced binary stochastic block model," *Information and Inference: A Journal of the IMA*, vol. 6, no. 2, pp. 125–170, 2016.

[3] F. Caltagirone, M. Lelarge, and L. Miolane, "Recovering asymmetric communities in the stochastic block model," *IEEE Transactions on Network Science and Engineering*, vol. 5, no. 3, pp. 237–246, 2018.

[4] M. Lelarge and L. Miolane, "Fundamental limits of symmetric low-rank matrix estimation," *Probability Theory and Related Fields*, 2018.

[5] J. Barbier, M. Dia, N. Macris, F. Krzakala, T. Lesieur, and L. Zdeborová, "Mutual information for symmetric rank-one matrix estimation: A proof of the replica formula," in *Advances in Neural Information Processing Systems*, vol. 29, (Barcelona, Spain), pp. 424–432, 2016.

[6] T. Lesieur, F. Krzakala, and L. Zdeborová, "Constrained low-rank matrix estimation: Phase transitions, approximate message passing and applications," *Journal of Statistical Mechanics: Theory and Experiment*, July 2017.

[7] F. Krzakala, J. Xu, and L. Zdeborová, "Mutual information in rank-one matrix estimation," in *2016 IEEE Information Theory Workshop (ITW)*, pp. 71–75, IEEE, 2016.

[8] Y. Deshpande, A. Montanari, E. Mossel, and S. Sen, "Contextual stochastic block models," in *NeurIPS*, 2018.

[9] E. Abbe and C. Sandon, "Proof of the achievability conjectures for the general stochastic block model," *Communications on Pure and Applied Mathematics*, vol. 71, no. 7, pp. 1334–1406, 2018.

[10] J. Banks, C. Moore, J. Neeman, and P. Netrapalli, "Information-theoretic thresholds for community detection in sparse networks," in *Conference On Learning Theory*, 2016.

[11] E. Abbe, "Community detection and stochastic block models: Recent developments," *Journal of Machine Learning Research*, vol. 18, no. 177, pp. 1–86, 2018.

[12] G. Reeves, V. Mayya, and A. Volfovsky, "The geometry of community detection via the mmse matrix," *arXiv preprint arXiv:1907.02496*, 2019.

[13] G. Reeves, H. D. Pfister, and A. Dytso, "Mutual information as a function of matrix snr for linear gaussian channels," in *2018 IEEE International Symposium on Information Theory (ISIT)*, pp. 1754–1758, IEEE, 2018.

[14] P. Zhang, C. Moore, and L. Zdeborová, "Phase transitions in semisupervised clustering of sparse networks," *Physical Review E*, vol. 90, no. 5, p. 052802, 2014.

[15] T. T. Cai, T. Liang, and A. Rakhlin, "Inference via message passing on partially labeled stochastic block models," *arXiv preprint arXiv:1603.06923*, 2016.

[16] C. Stegehuis and L. Massoulié, "Efficient inference in stochastic block models with vertex labels," *IEEE Transactions on Network Science and Engineering*, 2019.

[17] H. Saad and A. Nosratinia, "Community detection with side information: Exact recovery under the stochastic block model," *IEEE Journal of Selected Topics in Signal Processing*, vol. 12, no. 5, pp. 944–958, 2018.

[18] K. Levin, A. Athreya, M. Tang, V. Lyzinski, and C. E. Priebe, "A central limit theorem for an omnibus embedding of multiple random dot product graphs," in *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pp. 964–967, IEEE, 2017.

[19] J. Arroyo, A. Athreya, J. Cape, G. Chen, C. E. Priebe, and J. T. Vogelstein, "Inference for multiple heterogeneous networks with a common invariant subspace," *arXiv preprint arXiv:1906.10026*, 2019.

[20] H. Mathews, V. Mayya, A. Volfovsky, and G. Reeves, "Gaussian mixture models for stochastic block models with non-vanishing noise," in *2019 IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, IEEE, 2019. To appear.