

COMMUNITY DETECTION WITH DEPENDENT CONNECTIVITY

BY YUBAI YUAN¹ AND ANNIE QU²

¹*Department of Statistics, University of California Irvine, yubaiy@uci.edu*

²*Department of Statistics, University of California Irvine, aqu2@uci.edu*

In network analysis, within-community members are more likely to be connected than between-community members, which is reflected in that the edges within a community are intercorrelated. However, existing probabilistic models for community detection such as the stochastic block model (SBM) are not designed to capture the dependence among edges. In this paper, we propose a new community detection approach to incorporate intra-community dependence of connectivities through the Bahadur representation. The proposed method does not require specifying the likelihood function, which could be intractable for correlated binary connectivities. In addition, the proposed method allows for heterogeneity among edges between different communities. In theory, we show that incorporating correlation information can achieve a faster convergence rate compared to the independent SBM, and the proposed algorithm has a lower estimation bias and accelerated convergence compared to the variational EM. Our simulation studies show that the proposed algorithm outperforms the existing multi-network community detection methods assuming conditional independence among edges. We also demonstrate the application of the proposed method to agricultural product trading networks from different countries and to brain fMRI imaging networks.

1. Introduction. Network data has arisen as one of the most common forms of information collection. This is due to the fact that the scope of study not only focuses on subjects alone, but also on the relationships among subjects. Networks consist of two components: (1) nodes or vertices corresponding to basic units of a system, and (2) edges representing connections between nodes. These two main components can have various interpretations under different contexts of application. For example, nodes might be humans in social networks; molecules, genes, or neurons in biology networks, or web pages in information networks. Edges could be friendships, alliances, URLs, or citations. The combination of the nodes and the edges defines a network, which can be represented by an adjacency matrix to reflect direct connectivities among nodes.

Many applications involve multiple networks interconnected by underlying similarities [40, 75], while each individual network exhibits heterogeneous features through edge weight or edge density. In this paper, our goal is to detect the shared community structure of nodes among multiple networks where the connection patterns within communities are different from connection patterns between communities. One particular application suitable for this type of network structure is from neuroimaging, where each subjects' nodes correspond to regions of interest in the brain, and connectivities between nodes reflect the concordance among the neuron activities in different regions [69, 70]. In general, the connectivities are dense as the activities among neurons are highly correlated [47, 44]. More importantly, although the connectivity intensity varies for different subjects, the network community structures of anatomical regions are rather stable and consistent across all subjects, and are associated with functionally-specialized regions or general cognitive functions [47, 44, 9, 2, 52].

The multiple networks data structure is also standard in the literature regarding international trading data which typically consists of a large number of single trading networks among countries, with each of them corresponding to a specific product [20, 71]. In general, the trading relations across different products are governed by the geographical and socio-economical similarities among countries, hence leading to a shared community structure among different product trading networks [71, 8]. In addition, there are other multiple networks sharing common community network structure such as biological networks and transportation networks [40, 20].

Major multi-network community detection methods can be summarized in two categories. One approach is the spectral method whereby multiple networks are jointly projected into a single space with a distinct representation for each network,

MSC 2020 subject classifications: Primary 62R07; secondary 62E17.

Keywords and phrases: Bahadur Representation, high-order approximation, multiple networks, stochastic block model, variational EM

and the shared community structure is encoded in the latent factors which span the space [47, 74, 23, 10, 68, 46]. The second approach is the maximum likelihood method based on a random graph model [28, 44]. This includes the popular stochastic block model (SBM) [33] and its extensions to incorporate the heterogeneity of nodes' degrees [37, 82], and latent distance modeling [30, 31] to address overlapping communities [1, 7].

A common key assumption for most existing methods is that connectivities are conditional independent given the membership of nodes. This assumption simplifies the complexity of the model, and the likelihood function can be explicitly formulated. However, the network data are likely dependent among connectivities, which are also considered in several random network modelings [32, 43, 39, 17]. For community detection, the conditional independency assumption typically does not hold in practice and therefore could lead to a misspecified model [61, 4, 73]. For example, friendships within a social community or functional connectivities in brain networks tend to be highly correlated.

In addition, under conditional independence, the community structure can only be identified based on the marginal mean discrepancy of connectivities between within-communities and across-communities. Specifically, as a fundamental assumption of the independent SBM, the marginal mean discrepancy is required to be greater than a sharp threshold to guarantee community detectability [51, 53]. However, the marginal mean discrepancy assumption might not hold, while the correlations among edges could be non-negligible and highly informative in identifying community structures. We show that the proposed method is able to incorporate the correlation information to achieve consistent community detection when the marginal mean discrepancy is insignificant.

More recently, the SBM has been extended to address the within-community dependency for multiple network samples. For example, [67, 59] apply a fixed-effect model through an independent intercept without incorporating information from other networks. Alternatively, a random-effects model is proposed to incorporate heterogeneity [60, 81], which borrows information from multiple networks. However, both of these approaches require the specification of a distribution for the random effects, and might be applicable in settings with exchangeable correlation structure among edges. In addition, an EM-type algorithm is implemented to integrate out the random-effects, [60, 81] which could be computationally expensive when the size of the community or the network size is large.

In this paper, we propose a novel community detection method to jointly model community structures among multiple networks. The proposed method can simultaneously incorporate the marginal and correlation information to differentiate within-community and between-community connectivities. The key idea is to approximate the joint distribution of correlated within-community connectivities by using a truncated Bahadur representation [5]. Although the approximate likelihood function is not the true likelihood, it is able to maximize the true community memberships and serves as a tighter lower bound to the true likelihood compared with the independent SBM likelihood. Consequently, we identify communities via maximizing the approximate likelihood function, which also serves as a discriminative function for membership assignments of nodes. In particular, within-community correlations provide an additional community-concordance measurement to capture high-order discrepancy between within-community and across-community networks, and therefore increase discriminative power to identify communities.

The main advantages and contributions of the proposed method can be summarized as follows. The proposed method incorporates correlation information among connectivities to achieve more accurate community detection than the independent likelihood method using marginal information only. The improvement of the proposed community detection method is especially powerful when the marginal information is relatively weak in practice. In addition, compared to the existing random-effects model, the proposed method is more flexible in modeling the heterogeneity of communities for multiple networks and heterogeneity of correlations among edges. Furthermore, it does not require a distribution specification among within-community connectivities.

In addition, we establish the consistency of the community estimation for the proposed approximate likelihood under a general within-community edge correlation structure and show that the proposed method achieves a faster convergence rate of membership estimation compared to the independent likelihood under both dense and sparse network regimes. In terms of computational convergence, the proposed algorithm achieves a lower estimation bias and a faster convergence rate compared to the variational EM algorithm at each iteration via incorporating additional correlation information. The theoretical development in this paper is nontrivial, since establishing membership estimation consistency is more challenging under the framework of conditional dependency among edges compared to the existing ones assuming the conditional independent model.

Computationally, we develop a two-step iterative algorithm which is not sensitive to initial values as in the standard variational EM algorithm. In addition, compared to the existing fixed-effects SBM with independent intercepts or the random-effects SBM, the proposed method has lower computational complexity, as it does not involve integration of random effects as in [60], or estimating the fixed effects for each network as in [59]. Simulation studies and a real data application also confirm that the proposed method outperforms the existing variational EM and multi-network community detection algorithms, especially when the marginal information of observed networks is weak.

This paper is organized as follows: Section 2 introduces the background of the proposed method. Section 3 introduces the proposed method to incorporate correlation information for community detection. Section 4 provides an algorithm and implementation strategies. Section 5 illustrates the theoretical properties of the proposed method. Section 6 demonstrates simulation studies, and Section 7 illustrates an application to world agricultural products trading data and brain fMRI imaging data. The last section provides conclusions and some further discussion.

2. Background and Notation. In this section, we provide background and notation of the proposed community detection. The stochastic block model (SBM) [33] is a form of hierarchical modeling which captures the community structure for networks. Consider M symmetric and unweighted sample networks $\mathbf{Y} = \{\mathbf{Y}^m\}_{m=1}^M = \{(Y_{ij}^m)_{N \times N}\}_{m=1}^M$ with N nodes for K communities. Let $\{z_i\}_{i=1}^N$ be the membership for each node and $z_i \in \{1, 2, \dots, K\}$, and denote the membership assignment matrix $\mathbf{Z} = \{(Z_{iq})_{n \times K}\} \in \{0, 1\}^{N \times K}$, where $Z_{iq} = \mathbb{1}\{z_i = q\}$. Here \mathbf{Z} has exactly one 1 in each row and at least one 1 in each column for no-null communities. The unknown membership $z_i \in \{1, 2, \dots, K\}$ can be modeled as a latent variable from a multinomial distribution:

$$z_i \sim \text{Multinomial}(1, \alpha_i),$$

where $i = 1, \dots, N$, $\alpha_i = \{\alpha_{i1}, \dots, \alpha_{iK}\}$ and $\sum_{k=1}^K \alpha_{ik} = 1$. Given the membership of nodes, the homogeneous stochastic blocks model formulates the observed edges between two nodes $\{(Y_{ij}^m)_{n \times n}\}_{m=1}^M$ as M samples of binary random variables $(Y_{ij})_{n \times n}$ following a Bernoulli distribution:

$$(2.1) \quad P(Y_{ij}|z_i = q, z_j = l) \sim \text{Bern}(\mu_{ql}), \text{ for } i, j \in \{1, \dots, N\}, q, l = 1, \dots, K,$$

where μ_{ql} is the probability of nodes i and j being connected, and we denote $f_{ql}(Y_{ij}^m) := P(Y_{ij}^m|z_i = q, z_j = l)$.

For the heterogeneous stochastic blocks model, the marginal mean μ_{ql} for each block can be modeled as a logistic model to incorporate heterogeneity among edges:

$$(2.2) \quad \mu_{ql}(x_{ij}) = \exp(\beta_{ql}x_{ij}) / \{1 + \exp(\beta_{ql}x_{ij})\},$$

where $(x_{ij})_{N \times N}$ are edge-wise covariates, where node i and node j belong to q th community and l th community, respectively. In practice, the edge-wise covariates $\{x_{ij}\}_{1 \leq i < j \leq N}$ could be similarity measurements between two nodes. The edges within the same community preserve homogeneity by sharing a block-wise parameter β_{ql} . The joint likelihood function can be decomposed into a summation of edge-wise terms following the conditional independence assumption:

$$(2.3) \quad \log P(\mathbf{Y}; \mathbf{Z}) = \sum_{m=1}^M \sum_{q=1}^K \sum_{i=1}^N Z_{iq} \log \alpha_q + \sum_{m=1}^M \sum_{q,l=1}^K \sum_{i < j} Z_{iq} Z_{jl} f_{ql}(Y_{ij}^m; \Theta).$$

The latent membership \mathbf{Z} is estimated by $E(\mathbf{Z}|\mathbf{Y})$ using the maximum likelihood estimator of model parameters $\Theta = \{\mu_{ql}; q, l = 1, \dots, K; \alpha_q; q = 1, \dots, K\}$ for (2.1) and $\Theta = \{\beta_{ql}; q, l = 1, \dots, K; \alpha_q; q = 1, \dots, K\}$ for (2.2). However, the classical EM algorithm is not applicable here, because the conditional distribution $P(\mathbf{Z}|\mathbf{Y}) = \frac{P(\mathbf{Y}; \mathbf{Z})}{\sum_{\mathbf{Z}} P(\mathbf{Y}; \mathbf{Z})}$ becomes intractable in the expectation step.

The variational EM algorithm [50, 35] is one of the most popular inference methods, and can be applied to approximate the likelihood $P(\mathbf{Z}|\mathbf{Y})$ by a complete factorized distribution $R(\mathbf{Z}, \boldsymbol{\tau}) = \prod_{i=1}^N h(Z_i; \tau_i)$, where $h(\cdot)$ denotes a multinomial distribution, $\boldsymbol{\tau} = (\tau_1, \dots, \tau_N)$ and $\tau_i = (\tau_{i1}, \dots, \tau_{iK})$ is a probability vector such that $\sum_{q=1}^K \tau_{iq} = 1$. In the expectation step, the likelihood $\log P(\mathbf{Y}; \mathbf{Z})$ is averaged over $R(\mathbf{Z})$ such that for any $\boldsymbol{\tau}$, $E_{R(\mathbf{Z}, \boldsymbol{\tau})}\{\log P(\mathbf{Y}; \mathbf{Z})\} \leq E_{P(\mathbf{Z}|\mathbf{Y})}\{\log P(\mathbf{Y}; \mathbf{Z})\}$ where,

$$\begin{aligned} E_{R(\mathbf{Z}, \boldsymbol{\tau})}\{\log P(\mathbf{Y}; \mathbf{Z})\} = & - \sum_{m=1}^M \sum_{q=1}^K \sum_{i=1}^N \tau_{iq} \log \tau_{iq} + \sum_{m=1}^M \sum_{q=1}^K \sum_{i=1}^N \tau_{iq} \log \alpha_q + \\ & \sum_{m=1}^M \sum_{q,l=1}^K \sum_{i < j} \tau_{iq} \tau_{jl} f_{ql}(Y_{ij}^m). \end{aligned}$$

Instead of directly maximizing $E_{P(\mathbf{Z}|\mathbf{Y})}\{\log P(\mathbf{Y}; \mathbf{Z})\}$, the variational EM approach alternatively maximizes its lower bound $E_{R(\mathbf{Z}, \boldsymbol{\tau})}\{\log P(\mathbf{Y}; \mathbf{Z})\}$ over model parameters Θ and variational parameters $\boldsymbol{\tau}$, and clusters nodes by $\boldsymbol{\tau}$ through $\hat{z}_i = \text{argmax}_k\{\hat{\tau}_{ik}, k = 1, \dots, K\}$.

Throughout this paper, we consider the conditional version of SBM (CSBM) [11, 64, 18], where the true membership \mathbf{Z}^* is fixed. The conditional stochastic block model framework assumes conditional independence among edges, i.e., $Y_{i_1 j_1}^m$ and $Y_{i_2 j_2}^m$ are independent given nodes' membership $z_{i_1}, z_{i_2}, z_{j_1}, z_{j_2}$, and the corresponding log-likelihood of observed sample networks is:

$$(2.4) \quad \log L_{ind}(\mathbf{Y} | \mathbf{Z}) = \sum_{m=1}^M \sum_{q,l=1}^K \sum_{i < j} Z_{iq} Z_{jl} \left\{ y_{ij}^m \log \mu_{ql} + (1 - y_{ij}^m) \log (1 - \mu_{ql}) \right\}.$$

The above log-likelihood can serve as a discriminant function in clustering membership \mathbf{Z} in that if $\log L_{ind}(\mathbf{Y} | \mathbf{Z}_1) > \log L_{ind}(\mathbf{Y} | \mathbf{Z}_2)$ given two membership assignments \mathbf{Z}_1 and \mathbf{Z}_2 , then \mathbf{Z}_1 is preferred over \mathbf{Z}_2 , since the likelihood for the observed sample networks is higher. Naturally, \mathbf{Z}^* can be estimated by

$$\hat{\mathbf{Z}} = \underset{\mathbf{Z}}{\operatorname{argmax}} \log L_{ind}(\mathbf{Y} | \mathbf{Z}).$$

The SBM in (2.4) allows one to differentiate within-community and between-community nodes via utilizing only the marginal information, in that the average connectivity rates within-communities are higher than those between-communities. However, the underlying conditional independence assumption among edges is too restrictive and practically infeasible. In most community detection problems it is common that edges within communities are more correlated. For example, social connections among friends are highly correlated in social networks. However, the dependency among edges is not captured by the traditional SBM, which could lead to significant information loss of the community structure.

3. Methodology.

3.1. Community detection with dependent connectivity. In this paper, we incorporate within-community correlation to improve accuracy and efficiency in identifying communities, in addition to utilizing the edges' marginal mean information, since within-community dependency contains additional information regarding the membership of nodes. This is especially effective when the marginal mean is not informative in differentiating between and within communities' connectivity.

In this section, we propose an approximate likelihood function to capture the dependency among within-community edges. We assume that each observed sample network $Y_{n \times n}^m$ follows a joint binary distribution $P(Y^m)$ such that the within-community edges are pairwisely correlated. Specifically, for the underlying distribution, the correlation among edges $Y_{i_1 j_1}, Y_{i_2 j_2}$ within a community satisfies: $\text{corr}(Y_{i_1 j_1}, Y_{i_2 j_2}) = \rho_{i_1 i_2, j_1 j_2} \in [-1, 1]$ given nodes $z_{i_1}, z_{i_2}, z_{j_1}$ and z_{j_2} are in the same community q , where $1 \leq i_1 < j_1 \leq N, 1 \leq i_2 < j_2 \leq N, (i_1, j_1) \neq (i_2, j_2)$ and $q = 1, \dots, K$. Note that correlations among each pair of edges could be different. Equivalently, the edges in community k show concordance only when

$$\sum_{i_1 < j_1; i_2 < j_2}^N Z_{i_1 k} Z_{j_1 k} Z_{i_2 k} Z_{j_2 k} \rho_{i_1 j_1, i_2 j_2} \hat{y}_{i_1 j_1}^m \hat{y}_{i_2 j_2}^m \geq 0,$$

where $\hat{y}_{i_1 j_1}^m$ and $\hat{y}_{i_2 j_2}^m$ are normalized binary variables using the marginal mean. Note that both positive and negative correlations among edges have been considered for community detection. For example, [59, 60] utilize random effects to model the heterogeneity of the connectivity rate for an individual network, and infer a positive correlation among the edges within the same community. In addition, positive pairwise correlation among edges is more likely to produce star or triadic relations, which are widely observed in social networks [63, 62]. On the other hand, a negative correlation among edges could arise from competition between stores within local retail networks.

In addition to the SBM, exponential random graph models (ERGMs) are another well-known class of random graph models to incorporate edge dependency. In contrast to the SBM which is a hierarchical network modeling for community structure, the ERGMs focuses on specific types of subgraphs in the network. More importantly, nodes are equivalent or exchangeable under the ERGMs, and realizations of subgraphs are assumed to be independent and serve as individual samples for the model, hence the model parameters could be estimated with a single network. However, nodes in the SBM are not equivalent as they belongs to different communities. In addition, the ERGMs and its variants such as the Markov graph [24] are designed to model a specific configuration of dependent structure such as a triad or a star, and control the probability of realizations of these subgroups in sample networks. However, the main purpose of the proposed method is to take the overall correlation intensity within communities into account, without requiring to specify the dependent structure as long as the number of correlated edges is sufficient large.

3.2. Approximate Likelihood. In this section, we propose an informative approximation of the true log-likelihood to cluster \mathbf{Z} via incorporating interactions among edges within a community in addition to marginal mean information. This is because the exact joint likelihood function of correlated binary distribution $P(\mathbf{Y}^m)$ is computationally intractable. Specifically, we construct an approximate likelihood as a substitute of the true likelihood by facilitating the Bahadur representation [5]. That is, we retain the low-order dependency information among edges within-communities and discard the high-order dependency for computational efficiency. Although the approximate likelihood is not a true likelihood, it still serves the purpose of estimating the membership of nodes.

Consider T dependent binary random variables, then the joint likelihood can be represented through the Bahadur representation:

$$(3.1) \quad P(Y_1 = y_1, \dots, Y_T = y_T) = \prod_{j=1}^T \mu_j^{y_j} (1 - \mu_j)^{1-y_j} \left[1 + \sum_{1 \leq j_1 < j_2 \leq T} \rho_{j_1 j_2} \hat{y}_{j_1} \hat{y}_{j_2} + \sum_{1 \leq j_1 < j_2 < j_3 \leq T} \rho_{j_1 j_2 j_3} \hat{y}_{j_1} \hat{y}_{j_2} \hat{y}_{j_3} + \dots + \rho_{12\dots T} \hat{y}_1 \hat{y}_2 \dots \hat{y}_T \right],$$

where

$$(3.2) \quad \mu_j = E(Y_j), \quad \hat{y}_j = \frac{y_j - E(y_j)}{\sqrt{E(y_j)(1 - E(y_j))}},$$

and

$$\rho_{j_1 j_2} = E(\hat{y}_{j_1} \hat{y}_{j_2}), \quad \rho_{j_1 j_2 j_3} = E(\hat{y}_{j_1} \hat{y}_{j_2} \hat{y}_{j_3}), \dots, \quad \rho_{12\dots T} = E(\hat{y}_1 \hat{y}_2 \dots \hat{y}_T).$$

The Bahadur representation (3.1) is an exact decomposition of the joint distribution of dependent binary random variables using a function of moments with a sequential order. For the community detection problem, the binary random variables represent within-community edges, and the corresponding joint distribution can be explicitly decomposed into a marginal part and a correlation part. The marginal part consists of all the marginal mean μ_{ij} for each edge, which can be directly modeled through the dependency of the mean on covariates as in (2.2). The correlation part consists of interactions among all possible pairwise-associations of normalized edges, which add correlation information beyond a conditional independence likelihood model. Note that the conditional independence model is a special case of the proposed model when the correlation is zero, and the corresponding Bahadur representation collapses to a marginal part only, which is equivalent to the $\log L_{ind}(\mathbf{Y}|\mathbf{Z})$ in (2.4).

There are two major challenges in applying the Bahadur representation to model the interactions among within-community edges. First, the dimension of correlation parameters could be high if all the high-order interactions in (3.1) are incorporated, and this could lead to an increasing computational demand as the size of community grows. To solve this problem, we retain all the second-order interactions, but ignore interactions for higher orders beyond the second order, since the pairwise interactions among edges could be most important.

The second challenge is the range of the correlation coefficient could be constrained by the marginal means [22]. Consequently, the correlation parameter space is more restrictive if the variability of marginal means among edges is large. Nevertheless, our primary goal is to construct an objective function which can incorporate information from the marginal mean and correlations of edges within-community, and the objective function is not necessarily the true likelihood function. In the proposed method, we instead construct an approximate likelihood which is more flexible for incorporating highly dependent communities while still achieving computational efficiency.

Specifically, we construct an approximate likelihood $\tilde{L}(\mathbf{Y}|\mathbf{Z})$ incorporating correlated within-community edges as follows:

$$(3.3) \quad \begin{aligned} \log \tilde{L}(\mathbf{Y}|\mathbf{Z}) &= \sum_{m=1}^M \sum_{q,l=1}^K \sum_{i < j}^N Z_{iq} Z_{jl} \left\{ y_{ij}^m \log \mu_{ql} + (1 - y_{ij}^m) \log (1 - \mu_{ql}) \right\} \\ &\quad + \sum_{m=1}^M \log \left\{ 1 + \sum_{k=1}^K \max \left\{ \sum_{\substack{i < j; u < v \\ (i,j) \neq (u,v)}}^N Z_{ik} Z_{jk} Z_{uk} Z_{vk} \rho_{ij,uv} \hat{y}_{ij}^m \hat{y}_{uv}^m, 0 \right\} \right\}, \end{aligned}$$

where μ_{ql} can be formulated in (2.1) or (2.2), \hat{y}_{ij}^m is formulated in (3.2), and $\rho_{ij,uv}$ is the pairwise correlation between Y_{ij} and Y_{uv} . Notice that the first term in (3.3) is the same as the marginal mean model, and the second term in (3.3) measures the concordance among edges within communities clustering \mathbf{Z} .

We denote the second term of (3.3) as

$$(3.4) \quad \log L_{cor}(\mathbf{Y}|\mathbf{Z}) = \sum_{m=1}^M \log \left\{ 1 + \sum_{k=1}^K \max \left\{ \sum_{\substack{i < j; u < v \\ (i,j) \neq (u,v)}} Z_{ik} Z_{jk} Z_{uk} Z_{vk} \rho_{ij,uv} \hat{y}_{ij}^m \hat{y}_{uv}^m, 0 \right\} \right\}.$$

Compared with $\log L_{ind}(\mathbf{Y}|\mathbf{Z})$ in (2.4), the proposed $\tilde{L}(\mathbf{Y}|\mathbf{Z})$ has more discriminative power over \mathbf{Z} , since it utilizes more information of the observed dependency within communities corresponding to clustering \mathbf{Z} . In addition, the non-negativity of $\log L_{cor}(\mathbf{Y}|\mathbf{Z})$ ensure the fact that $\log \tilde{L}(\mathbf{Y}|\mathbf{Z}) \geq \log L_{ind}(\mathbf{Y}|\mathbf{Z})$ is guaranteed, which implies that adding additional correlation information among edges can be more informative given within-community correlation exists. This leads to higher classification accuracy and estimation efficiency through maximizing (3.4).

The key part of the proposed method is to predict memberships of nodes through the Bayes factor constructed by the proposed $\log \tilde{L}(\mathbf{Y}|\mathbf{Z})$. Suppose the memberships of other nodes \mathbf{Z}_{-i} are known, then we classify node i based on the following Bayes factor:

$$\frac{\tilde{L}(\mathbf{Y}|\mathbf{Z}_{-i}, Z_{iq} = 1)}{\tilde{L}(\mathbf{Y}|\mathbf{Z}_{-i}, Z_{ik} = 1)} = \exp \left\{ \log \tilde{L}(\mathbf{Y}|\mathbf{Z}_{-i}, Z_{iq} = 1) - \log \tilde{L}(\mathbf{Y}|\mathbf{Z}_{-i}, Z_{ik} = 1) \right\}.$$

If the above Bayes factor > 1 , then the probability of node i in community q is larger than that of community k . The Bayes factor can be further decomposed as:

$$(3.5) \quad \frac{\tilde{L}(\mathbf{Y}|\mathbf{Z}_{-i}, Z_{iq} = 1)}{\tilde{L}(\mathbf{Y}|\mathbf{Z}_{-i}, Z_{ik} = 1)} = \frac{L_{ind}(\mathbf{Y}|\mathbf{Z}_{-i}, Z_{iq} = 1)}{L_{ind}(\mathbf{Y}|\mathbf{Z}_{-i}, Z_{ik} = 1)} \frac{L_{cor}(\mathbf{Y}|\mathbf{Z}_{-i}, Z_{iq} = 1)}{L_{cor}(\mathbf{Y}|\mathbf{Z}_{-i}, Z_{ik} = 1)},$$

which contains both the marginal ratio and the correlation ratio. It is clear that when the marginal information is weak in differentiating two communities, the marginal ratio is close to 1, and if the correlation ratio is informative, it can enhance the Bayes factor to improve community detection. In addition, the correlation ratio also serves as a correction to lower the estimation bias.

We illustrate the advantage of the proposed method in (3.4) over the conditional independent likelihood (2.4) using a simple numerical illustration. Specifically, we generate multiple networks based on the SBM with 30 nodes evenly split between two communities. The marginal means of within-community and between-community edges are the same at 0.5, implying that the marginal mean is not informative. We assume a true exchangeable correlation $\rho = 0.6$ for within-community edges. Figure 1 illustrates that the likelihood function changes as memberships of nodes change with some misclassified nodes. The left graph is based on the conditional independent SBM utilizing only marginal information, which does not differentiate the two communities at all due to weak marginal information. However, the proposed approximate likelihood in the right graph has high differentiation power for the nodes' memberships, and reaches maximum when the true memberships are selected.

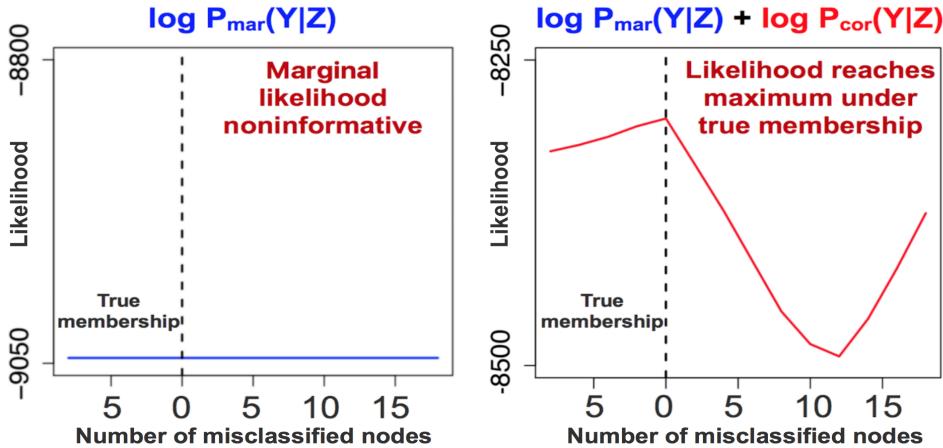


Fig 1: Likelihoods of multiple networks when the memberships of nodes differ from their true memberships. The networks share 30 nodes from two underlying communities. *Left:* Traditional SBM likelihood is insensitive to misclassified nodes. *Right:* The proposed approximate likelihood incorporating correlation information reaches maximum when no node is misclassified.

Remark 1: The proposed method is applicable for continuously weighted edges through modification on the objective function in (3.3). For example, if the edge weights y_{ij} are real values, we can model it as a normal distribution. Based on the weighted stochastic block model, the edgewise marginal likelihood in the first term of (3.3) is modified as

$$\sum_{m=1}^M \sum_{q,l=1}^K \sum_{i < j}^N \left\{ -Z_{iq} Z_{jl} \frac{(y_{ij}^m - \mu_{ql})^2}{\sigma_{ql}^2} \right\},$$

where μ_{ql} and σ_{ql} are the block-wise mean and variance. In addition, we incorporate the within-community dependency information in the second term as

$$\sum_{m=1}^M \sum_{k=1}^K \max \left\{ \sum_{\substack{i < j; u < v \\ (i,j) \neq (u,v)}}^N Z_{ik} Z_{jk} Z_{uk} Z_{vk} \rho_{ij,uv} \hat{y}_{ij}^m \hat{y}_{uv}^m, 0 \right\},$$

where $\hat{y}_{ij}^m = \frac{y_{ij}^m - \mu_{kk}}{\sigma_{kk}}$. Following the principle in (3.3), the proposed objective function can be adapted for other types of edge weights through adjusting the marginal likelihood and edge dependency accordingly.

Remark 2: We can further reduce the number of parameters in (3.3) via a homogeneous correlation structure in that for any $Y_{i_1 j_1}$ and $Y_{i_2 j_2}$ in the q th community,

$$\rho_q := \frac{1}{\#\{\rho_{i_1 j_1, i_2 j_2} \neq 0\}} \sum_{\{\rho_{i_1 j_1, i_2 j_2}\} \neq 0} |\rho_{i_1 j_1, i_2 j_2}|,$$

where $\#\{A\}$ denotes the cardinality of a set A . We can replace the correlation $\rho_{i_1 j_1, i_2 j_2}$ by $\text{sign}(\rho_{i_1 j_1, i_2 j_2})\rho_q$. The rationales of this simplification is based on the following. First, the pairwise correlation parameter $\rho_{i_1 j_1, i_2 j_2}$ is a nuisance correlation parameter to enhance clustering. Second, both the numerical experiments and theoretical findings show that the density of pairwise correlation among within-community edges plays a more important role than the intensity of the correlation in affecting clustering performance.

4. Algorithm and Implementation. In this section, we propose a two-step algorithm to maximize the proposed approximate likelihood function. In addition, we provide implementation strategies to improve the stability and efficiency of the algorithm.

4.1. Algorithm. To estimate the true membership \mathbf{Z}^* of nodes, we can ideally search through all the possible \mathbf{Z} and choose the one with the largest $\log \tilde{L}(\mathbf{Y}|\mathbf{Z})$. However, this becomes infeasible when the number of nodes N and the number of communities K increases. In the following, we propose an iterative two-step algorithm to maximize $\log \tilde{L}(\mathbf{Y}|\mathbf{Z})$ in (3.3) based on the homogeneous stochastic block model (2.1).

Algorithm 1 for homogeneous stochastic block model

Step 1: (i) Input an initial membership probability for each node: $\alpha_{iq}^{(0)}$, $1 \leq i \leq N$, $1 \leq q \leq K$

through spectral clustering on individual sample networks.

(ii) Estimate pairwise correlations $\{\hat{\rho}_{ij,uv}\}$ through the empirical estimator by $(\{Y_{ij}^m, Y_{uv}^m\}_{m=1}^M)$.

Step 2: At the s th iteration, given $\{\mu_{ql}^{(s-1)}\}_{q,l=1}^K$ and $\{\alpha_i^{(s-1)}\}_{i=1}^N$ from the $(s-1)$ th iteration:

(i) **Maximization:** block-wise update $\mu_{ql}^{(s)}$ $q, l \in \{1, \dots, K\}$;

$$(a) \text{Obtain } \mu_{ql}^{(s)} \text{ through } \frac{\sum_{m=1}^M \sum_{i \neq j} \alpha_{iq}^{(s-1)} \alpha_{jl}^{(s-1)} Y_{ij}^m}{\sum_{m=1}^M \sum_{i \neq j} \alpha_{iq}^{(s-1)} \alpha_{jl}^{(s-1)}}.$$

(ii) **Expectation:** given $\{\mu_{ql}^{(s)}\}_{q,l=1}^K$ and $\rho_{ij,uv}$, update $\{\alpha_i^{(s)}\}_{i=1}^N$:

$$\alpha_{iq}^{(s)} = \frac{\alpha_{iq}^{(s-1)} \tilde{L}(\mathbf{Y}|\alpha_{-i}^{(s-1)}, Z_{iq}=1)}{\sum_{k=1}^K \alpha_{ik}^{(s-1)} \tilde{L}(\mathbf{Y}|\alpha_{-i}^{(s-1)}, Z_{ik}=1)}, \quad i = 1, \dots, N, \quad q = 1, \dots, K.$$

Step 3: Iterate until $\max_{1 \leq i \leq N} |\alpha_i^{(s)} - \alpha_i^{(s-1)}| < \epsilon$.

Step 4: Obtain the membership z_i of clusters by $\{\alpha_i^{(s)}\}_{i=1}^N$:

$$z_i = \max_k \{\alpha_{i1}^{(s)}, \dots, \alpha_{iK}^{(s)}\}, \quad i = 1, \dots, N.$$

Here we directly maximize the approximate likelihood instead of a true likelihood as in the EM algorithm. In the expectation step, we alternatively update membership of each node while fixing other nodes, where $\tilde{L}(\mathbf{Y}|\alpha_{-i}; Z_{ik})$ has the same formulation as $\tilde{L}(\mathbf{Y}|Z)$ in (3.3) with $\{Z_{iq}\}_{N \times K}$ replaced by its expectation $\{\alpha_{iq}\}_{N \times K}$, except Z_{ik} . Note that α_{iq} is not the expectation under the true underlying joint distribution $P(Y, Z) = P(Y|Z)P(Z)$. Instead, it corresponds to the distribution defined by the approximate likelihood in (3.4). In the expectation step, the memberships are updated through the Bayes factor in (3.5) with the proposed $\tilde{L}(\mathbf{Y}|Z)$. Note that the variational EM is a special case of the proposed algorithm if the correlation information is ignored and the conditional independent model in (2.4) is assumed. The heterogeneous stochastic block model algorithm differs Algorithm 1 only at the maximization step where we estimate the community-wise parameters β_{ql} through the generalized estimating equation.

4.2. Computation and Implementation. To ensure computational stability, the community-wise parameters β_{ql} could be estimated through a simplified generalized estimation equation assuming an independent working correlation in algorithm 1. This is because the primary interest of community detection is classification accuracy, and the empirical studies show that correlation information plays a relatively minor role in parameter estimation.

We can achieve a better approximation to the true likelihood if higher-order moments are incorporated in the Bahadur representation in (3.2), which also increases its discrimination power. However, higher-order correlation could also increase the computational cost. Alternatively, we can recover partial higher-order interactions (e.g., the fourth order) derived from low order interactions (e.g., the second order). For example, consider the nonnegative exchangeable correlation structure for within-community edges such that $\rho_k := \rho_{i_1 j_1, i_2 j_2} > 0$ for $i_1, j_1, i_2, j_2 \in k$ th community, and four normalized edges $\hat{Y}_{i_1 j_1}^m, \hat{Y}_{i_2 j_2}^m, \hat{Y}_{i_3 j_3}^m, \hat{Y}_{i_4 j_4}^m$ within the community k with a positive fourth-order correlation among them, we have

$$(4.1) \quad E(\hat{Y}_{i_1 j_1}^m \hat{Y}_{i_2 j_2}^m \hat{Y}_{i_3 j_3}^m \hat{Y}_{i_4 j_4}^m) \geq E(\hat{Y}_{i_1 j_1}^m \hat{Y}_{i_2 j_2}^m)E(\hat{Y}_{i_3 j_3}^m \hat{Y}_{i_4 j_4}^m) = \rho_k^2.$$

Denote $(Z_{1k} Z_{2k} \hat{Y}_{12}^m, Z_{1k} Z_{3k} \hat{Y}_{13}^m, \dots, Z_{(N-1)k} Z_{Nk} \hat{Y}_{(N-1)N}^m)$ as $(\gamma_1^m, \gamma_2^m, \dots, \gamma_{N_0}^m)$, where $N_0 = \frac{N^2 - N}{2}$. Then the second-order interaction term for the community k in $L_{cor}(\mathbf{Y}|Z)$ is

$$\rho_k \sum_{\substack{i < j, u < v \\ (i,j) \neq (u,v)}}^N Z_{ik} Z_{jk} Z_{uk} Z_{vk} \hat{y}_{ij}^m \hat{y}_{uv}^m = 2\rho_k \sum_{s < t}^{N_0} \gamma_s^m \gamma_t^m.$$

Based on (4.1) and given Z , we can approximate the fourth-order interaction for community k under the exchangeable correlation structure by its lower bound:

$$(4.2) \quad \sum_{\substack{s_1 < t_1, s_2 < t_2 \\ (s_1, t_1) \neq (s_2, t_2)}}^{N_0} \frac{E(\gamma_{s_1}^m \gamma_{t_1}^m \gamma_{s_2}^m \gamma_{t_2}^m)}{2} \gamma_{s_1}^m \gamma_{t_1}^m \gamma_{s_2}^m \gamma_{t_2}^m \geq \left(\rho_k \sum_{s < t}^{N_0} \gamma_s^m \gamma_t^m \right)^2 - \rho_k^2 \sum_{s < t}^{N_0} (\gamma_s^m \gamma_t^m)^2.$$

Note that the above lower bound of the fourth-order interaction can be calculated by the second-order interaction term in $L_{cor}(\mathbf{Y}|Z)$. Therefore, we can still incorporate higher-order terms in $\log \tilde{L}(\mathbf{Y}|Z)$ without additional computational cost. For other types of non-exchangeable correlation structures, we can incorporate partial higher-order correlation similarly as above. The main difference is that each pair of edges is associated with a specific correlation given a dependency structure. Therefore, the simplified lower bound for higher-order correlations such as (4.2) does not hold in general, and could have a more complex form depending on the specific correlation structure.

In the following, we also provide some guidelines for determining the number of communities K and initial membership of nodes. For a single network, the criterion-based methods choose K to maximize a certain probabilistic criterion such as the integrated likelihood [26, 19, 42], composite likelihood BIC [66] or modularity criterion [12]. In addition, spectral methods estimate K through the spectral property of the transformed adjacent matrix, such as a Laplacian matrix [55], non-backtracking matrix [13] or Bethe Hessian matrix [65]. In the hierarchical Bayesian framework, the number of communities is treated as a model parameter given a certain prior distribution and is jointly estimated with nodes' memberships using the MCMC [26, 56, 57]. For multiple networks, we can extend the above techniques to estimate a consensus number of communities combining observed realizations of the SBM from each individual network.

In the context of the proposed within-community dependent modeling, we can first perform the modularity-maximizing method or spectral clustering on each individual network to obtain K , then take the average of these individual estimated K , which can be treated as a consensus number of communities. The above procedure is sensible under two considerations. First, each sample network is a realization of the SBM so that the individual estimation of K is randomly distributed

around the true underlying K . Thus the average of individual estimations provides an estimation of K with low-bias and low-variance. Second, the spectral clustering or modularity methods are more favorable than other methods, due to their relatively low computational cost in estimating K . This is especially effective when the sample size of networks is large.

As an EM-type algorithm, the proposed optimization procedure can only guarantee the local maximum and requires multiple initializations to find the global maximum. In this paper, we adopt spectral clustering on sample networks to obtain multiple initials, which is a common strategy for multiple network community detection [44, 59, 60]. Spectral clustering is a model-free clustering algorithm and is able to provide a warm start for nodes' memberships.

5. Theoretical Results. In this section, we establish the consistency of the estimated nodes' membership based on the independent likelihood and the approximate likelihood approaches. In addition, we provide the computational convergence theorem for the proposed iterative algorithm in section 4. Compared to the independent likelihood approach, we show that the approximate likelihood approach leads to a computationally faster convergence rate regarding nodes' membership estimation.

5.1. Consistency of Nodes' Membership Estimation. In this subsection, we study the consistency of the maximization likelihood estimator for both the independent likelihood and the approximate likelihood at the population level. With the independence assumption among within-community edges, the consistency and convergence rate of the MLE estimator can be obtained by [15, 80]. However, the convergence property of the MLE remains unknown if there exists a local dependence among edges.

One significant distinction using the independence assumption if the edges are correlated is that the increasing number of nodes and number of edges do not necessarily guarantee a lower misclassification rate and computationally faster convergence. This is because the discrepancy between marginal means from within-community and between-community is not accumulated due to the pairwise correlation, though it can be accumulated through increasing the number of sample networks. However, we show that the proposed approximated approach is able to benefit from the increasing number of nodes, and therefore achieves a faster computational convergence compared to the independent likelihood approach.

In the following theorems, we consider the homogeneous stochastic block model that edges within the same block have the same marginal mean such that $\mu_{z_iz_j} := E(Y_{ij}^m | i \in q, j \in l) = \eta_N c_{ij}$, where $\eta_N \in (0, 1]$ is a sparsity parameter controlling the average node degree. We denote that the true marginal means as $\Theta = \{\mu_{ql}, 1 \leq q < l \leq K\}$, and assume the following two regularity conditions regarding identifiability:

(C1). Suppose for every $q \neq q', 1 \leq q, q' \leq K$, there exists at least one $l \in \{1, \dots, K\}$ such that $\mu_{ql} \neq \mu_{q'l}$. In addition, all the c_{ql} are bounded such that $c_{ql} \in [\zeta, 1 - \zeta], q, l = 1, \dots, K$ with $\zeta > 0$.

Remark 3: Theorem 5.1 and Theorem 5.2 can be extended to heterogeneous stochastic block model (2.2) where the edgewise marginal means within a community could be different due to varying covariates of the corresponding edges. Accordingly, we can replace Condition (C1) by a similar community identifiability condition (C1.b):

(C1.b). Let $C_q (q = 1, \dots, K)$ denote a node set for the q th community. For two different communities $C_q, C_{q'}, q \neq q'$, there exists at least one community C_l such that

$$\sup_{i_1 \in C_q, j_1 \in C_l} E(Y_{i_1 j_1}) \leq \inf_{i_2 \in C_{q'}, j_2 \in C_l} E(Y_{i_2 j_2}) \text{ or } \sup_{i_2 \in C_{q'}, j_2 \in C_l} E(Y_{i_2 j_2}) \leq \inf_{i_1 \in C_q, j_1 \in C_l} E(Y_{i_1 j_1}).$$

Or equivalently, nodes from two different communities are differentiable in that their probabilities of connecting nodes from other communities are unique.

(C2). Community sizes from all sample networks are bounded above and below by $\kappa_1 N \leq |\{i \in \{1, 2, \dots, N\} : Z_{iq}^* = 1\}| \leq \kappa_2 N$, $q = 1, \dots, K$, where κ_1 and κ_2 are constants such that $0 < \kappa_1 < \kappa_2 < 1$.

In the following, we establish the consistency of membership estimation for both the independent likelihood approach and the proposed approximate likelihood approach. For the within-community edges, we define the edgewise second-order pairwise correlation density as

$$(5.1) \quad \lambda = \lambda_{ij}^m := \frac{\#\{(u, v) : |\text{corr}(Y_{ij}^m, Y_{uv}^m)| > 0, Z_u^* = Z_v^* = k\}}{N_k(N_k - 1)/2 - 1} \text{ for edge } Y_{ij}^m \text{ in community k}$$

where $k = 1, 2, \dots, K$ and $N_k(N_k - 1)/2 - 1$ is the number of edges within community k for the sample network \mathbf{Y}^m . For simplicity, we assume the homogeneous second-order correlation density such that $\lambda_{ij}^m = \lambda$ for all the within-community edges. Here $\lambda \in [0, 1]$ serves as a counterpart of sparsity parameter η_N commensurate with edge correlation density, and determines the intensity of local dependency within a community. Specifically, $\lambda = 0$ indicates that within-community edges are all independent, while $\lambda = 1$ indicates that all edges within a community are pair-wisely correlated. In addition, correlation density λ is allowed to depend on the number of nodes, and increases such that it can model a more general

class of correlation structure. For example, in a hub structure, an edge is only correlated with those sharing the same hub nodes and the density $\lambda = \frac{N_k - 1}{N_k^2 - 1} = O_N(\frac{1}{N_k})$.

To establish asymptotic consistency for the proposed likelihood, we assume the sparsity of high-order correlation among within-community edges.

(C3). The number of third and fourth-order correlations defined in (3.1) among within-community edges do not exceed the order of the size of second-order correlations. Specifically, for edge Y_{ij}^m in community k , $\#\{(u_1, v_1), (u_2, v_2) : E(\hat{Y}_{ij}\hat{Y}_{u_1v_1}\hat{Y}_{u_2v_2}) \neq 0\}| \leq O_N(\lambda(N_k^2))$. In addition, $\#\{(u_1, v_1), (u_2, v_2), (u_3, v_3) : E(\hat{Y}_{ij}\hat{Y}_{u_1v_1}\hat{Y}_{u_2v_2}\hat{Y}_{u_3v_3}) \neq 0\}| \leq O_N(\lambda(N_k^3))$, $k = 1, 2, \dots, K$.

In general, assume that the pairwise correlations among the within-community edges are sufficient to cover a broad class of Markov dependence modeling under the general exponential random graph model. This includes the most commonly used edge dependence configurations such as a star, a triangular shape subnetwork [54] and the k-triangles shape [58]. Although considering that the additional higher-order edge correlation improves the model's complexity, it could increase higher computational cost and instability. Empirically, it is sensible to assume that higher-order correlation only exists when second-order correlation already exists among edges, for the sake of identifiability and interpretability of the model. Otherwise, it could lead to the 'near degeneracy' [29] when a higher-order dependency masks a lower-order dependency.

Let $P_{Z^*} := \mathbb{P}(\cdot | Z = z^*; \Theta)$ denote the conditional distribution of edges given the true membership of nodes z^* and true parameters.

THEOREM 5.1. *Under the regularity conditions (C1)-(C3), we establish the convergence rate of the membership estimator z using the independent likelihood approach. That is, for every $t > 0$ and $z \neq z^*$,*

$$(5.2) \quad P_{Z^*} \left\{ \frac{L_{ind}(\mathbf{Y} | \mathbf{Z} = z; \Theta)}{L_{ind}(\mathbf{Y} | \mathbf{Z} = z^*; \Theta)} > t \right\} = \mathcal{O} \left(\exp \left\{ - C_1 \frac{c^* r \eta_N N M}{1 + \rho \kappa_2 \eta_N N \min(r, \kappa_2 \lambda N)} \right\} \right),$$

where $r = \|z - z^*\|_0$ is the number of misclassified nodes up to the permutation labeling, ρ is the largest pairwise correlation among within-community edges, C_1 is a positive constant, and $c^* = \min_{(q,l),(q',l')} \{D_{KL}(c_{ql} || c_{q'l'}) : c_{ql} \neq c_{q'l'}\}$, where D_{KL} denotes the KullbackLeibler divergence distance.

Given the convergence rate based on the independent likelihood ratio, we can characterize the convergence of its estimated node membership as following:

COROLLARY 5.1. *Under the same conditions given in Theorem 5.1, using the independent likelihood approach, for every $t > 0$*

$$(5.3) \quad P_{Z^*} \left\{ \sup_{\{z \neq z^*\}} \frac{L_{ind}(\mathbf{Y} | \mathbf{Z} = z; \Theta)}{L_{ind}(\mathbf{Y} | \mathbf{Z} = z^*; \Theta)} > t \right\} = \mathcal{O} \left(N \exp \left\{ - \frac{c^* \eta_N N M}{1 + \eta_N \lambda N^2} \right\} \right).$$

For the independent likelihood approach, the convergence rate depends on the number of sample network M , the marginal sparsity parameter η_N and the density of the pairwise correlation λ among within-community edges. If there is no pairwise correlation among edges, e.g., $\lambda = 0$, then the convergence rate in (5.3) increases to $\mathcal{O}_{N,M} \{N \exp(-\eta_N N M)\}$, which degenerates to the convergence rate established in [15] under the conditional independent modeling given constant marginal mean $\eta_N = 1$. In addition, the consistency of nodes' membership estimation from the independent likelihood can be guaranteed given that $\eta_N \geq \mathcal{O} \{(\log N)^c / NM\}$, $c > 1$ based on Corollary 5.1. This result is also consistent with the existing condition on membership recovery for a single sparse network ($M = 1$) where the increase of node degree is faster than a polylogarithmic rate [11, 18].

However, if the underlying within-community dependency exists, e.g., $\lambda > 0$, using an independent model does not guarantee the consistency of membership estimation, neither under sparse nor dense network regimes. For example, in the case of the exchangeable correlation structure for within-community edges when $\lambda = 1$, the convergence rate in (5.2) decreases to the order of $\mathcal{O}_{N,M} \{\exp(-\frac{rM}{\min(r, \kappa_2 N)})\}$, and therefore does not benefit from the increasing number of nodes. In this case, the consistency relies on accumulating independent sample networks. Theorem 5.1 also implies that the independent likelihood approach is unable to fully accumulate discriminative power from the increasing number of nodes when there exists dependency among within-community edges. Indeed, the convergence rate of the independent likelihood approach decreases in terms of network size N as the within-community correlation density λ increases. However, we show that the proposed approximate likelihood approach still benefits from increasing nodes size even under the exchangeable correlation structure among edges.

THEOREM 5.2. *Under the regularity conditions (C1)-(C3), we establish the convergence rate of the estimator \mathbf{z} using the proposed approximate likelihood approach. That is, for every $t > 0$, and $\mathbf{z} \neq \mathbf{z}^*$*

$$(5.4) \quad P_{\mathbf{Z}^*} \left\{ \frac{\tilde{L}(\mathbf{Y}|\mathbf{Z}=\mathbf{z}; \Theta)}{\tilde{L}(\mathbf{Y}|\mathbf{Z}=\mathbf{z}^*; \Theta)} > t \right\} = \mathcal{O} \left(\exp \left\{ - C_2 \frac{r\lambda NM(c^*\eta_N + \lambda N^2)}{1 + \rho\kappa_2 N \min(r, \kappa_2\lambda N)} \right\} \right),$$

where $\frac{1}{\lambda} < O_N(NM)$, $r = \|\mathbf{z} - \mathbf{z}^*\|_0$ is the number of misclassified nodes up to the permutation labeling, C_2 is a positive constant, ρ is the largest within-community correlation, and c^* is defined in Theorem 5.1.

Similarly, we characterize the convergence of the proposed method by the following corollary:

COROLLARY 5.2. *Under the same conditions given in Theorem 5.2, the proposed approximate likelihood approach leads to the following convergence rate, for every $t > 0$*

$$(5.5) \quad P_{\mathbf{Z}^*} \left\{ \sup_{\{\mathbf{z} \neq \mathbf{z}^*\}} \frac{\tilde{L}(\mathbf{Y}|\mathbf{Z}=\mathbf{z}; \Theta)}{\tilde{L}(\mathbf{Y}|\mathbf{Z}=\mathbf{z}^*; \Theta)} > t \right\} = \mathcal{O} \left(N \exp \left\{ - \frac{(c^*\eta_N + \lambda N^2)M}{N} \right\} \right).$$

Given $\lambda > 0$, for the same number of networks M and node size N , the proposed approximate likelihood approach is able to achieve a faster convergence rate in (5.5) compared with (5.3) since the convergence rate in (5.4) has an additional term of $\lambda^2 N^3 M$ on the numerator compared to the convergence rate in (5.2). Specifically, the proposed approach is most superior under the exchangeable correlation structure ($\lambda = 1$), where the convergence rate of the independent likelihood is at the order of $\mathcal{O}_{N,M} \{ \exp(-M/N) \}$, in contrast to the proposed convergence rate of $\mathcal{O}_{N,M} \{ \exp(-NM) \}$. Therefore, based on Corollary 5.2, the consistency of membership estimation from the proposed approach can be guaranteed given correlation density $\lambda \geq \mathcal{O}(\log^c N/(NM))$, $c > 1$ regardless of whether the multiple networks are sparse or dense. Intuitively, incorporating the correlation information increases the effective sample size of within-community edges. Under the sparsity assumption of higher-order correlation among edges, the proposed approach benefits from accumulating information on the second-order interactions among edges, while the independent likelihood approach only accumulates information from the first-order marginal mean of edges. It is noticeable that the marginal sparsity η_N affects the convergence rate of the membership estimator not only through marginal information but also through its constraints on the edge-wise correlation with intensity λ . Although the marginal sparsity η_N does not affect the order of convergence rate directly in Corollary 2.2, it imposes an implicit constraint on the possible within-community dependency intensity λ in that $0 \leq \lambda \leq \mathcal{O}(\eta_N)$, which affects the convergence rate of membership estimation. This is because that the number of correlated edges decreases as the networks become sparse.

5.2. Computational Convergence for the Proposed Algorithm. In this subsection, we provide the computational convergence property of the proposed Algorithm 1 in Section 4. The main difference between the proposed method and the variational EM lies in the Bayes factor of (3.5) in the expectation step from Algorithm 1. If we replace $\tilde{L}(\mathbf{Y}|\mathbf{Z})$ by the conditional independent likelihood $L_{ind}(\mathbf{Y}|\mathbf{Z})$ in (2.4) in the expectation step, the standard variational EM becomes a special case of Algorithm 1. Notice that [80] establishes computational convergence with the minimax rate of misclassification only when the within-community edges are independent. In addition, it assumes that the within-community marginal means are all the same, which is too restrictive in practice.

In the following, we establish the computational convergence for the proposed approximate likelihood based on the homogeneous stochastic block model (2.1). Specifically, we are able to show a faster convergence speed and a lower estimation bias compared to the existing one based on the independent likelihood in [80]. The following Theorem 5.3 also relaxes the marginal mean assumption in [80] and allows the within-community marginal means μ_{qq} , $q \in \{1, \dots, K\}$ to be different to each other, and between-community marginal means μ_{ql} , $q \neq l \in \{1, \dots, K\}$ to be different to each other. We denote the estimated memberships of nodes at the s th iteration as $\boldsymbol{\alpha}^{(s)} = (\alpha_1^{(s)}, \dots, \alpha_N^{(s)})$ from Algorithm 1. In addition to the assumptions (C1-C3) in Section 5.1, we require a regularity condition for the following theorem:

(C4). Suppose the distance between initial membership $\boldsymbol{\alpha}^{(0)}$ and true membership \mathbf{z}^* is bounded: $\|\boldsymbol{\alpha}^{(0)} - \mathbf{z}^*\|_1 \leq cN^{1-\phi}$, where $\phi \in (0, 1)$ is a constant.

A common issue for most EM-type algorithms including the one proposed is that they only guarantee convergence to a local optimum. If the likelihood function is unimodal, then the EM-type algorithm converges to the MLE as the unique global optimum. However, the proposed approximate likelihood is non-convex and multi-modal. Therefore, we assume that the initials are in the neighborhood of the MLE to ensure the convergence of the EM algorithm [6, 77]. Condition C4 is a common assumption to guarantee computational convergence for EM-type algorithms [80, 36, 38].

THEOREM 5.3. *Under the regularity conditions (C1)-(C4) and given N is sufficiently large, we establish the convergence property of Algorithm 1 through incorporating correlation information. That is, with the correlation density $\frac{1}{\lambda} = o_N(N^{\frac{\phi}{2}})$, in each iteration for the Algorithm 1, we have*

$$(5.6) \quad \|\boldsymbol{\alpha}^{(s+1)} - \boldsymbol{z}^*\|_1 \leq c_1 N K \exp \left\{ -c_2 \frac{\lambda(c^* \eta_N + \lambda N^2) M}{N} \right\} + \frac{c_3 \|\boldsymbol{\alpha}^{(s)} - \boldsymbol{z}^*\|_1}{(\lambda N M)^{1-\gamma}}, \forall s \geq 0$$

with probability at least $1 - \exp(-c\lambda NM) - (\lambda NM)^{-\gamma} - \epsilon$, where $\gamma \in (0, 1)$, $\epsilon = o_M(1)$, c, c_1, c_2, c_3 are positive constants, and c^* is defined in Theorem 5.1.

In Theorem 5.3, the first term on the right hand side of the inequality represents the estimation bias which measures the discrepancy between the community structure and its realization. Although we do not show that the order of estimation bias in the first term achieves the minimax rate, there is a connection between our result and the minimax rate when $M = 1$. Given that the dependency among edges exists, the bias term in our Theorem 5.3 has the order $\mathcal{O}\{N \exp(-\lambda NM)\}$, where N is the number of nodes and M is the sample size of networks. The minimax rate established in Theorem 3 in [80] under multiple networks with independent edges is $\mathcal{O}\{N \exp(-NM)\}$. Therefore, our Theorem 5.3 indicates that our bias term has the same order as the minimax rate established in [80] given $\lambda = \mathcal{O}(1)$.

The second term provides a decreasing rate of misclassification along each iteration. Theorem 5.3 indicates that the estimated memberships are closer to the true memberships compared to the previous iteration step at a rate of $\frac{1}{(\lambda NM)^{1-\gamma}}$, where a larger sample size M or node size N contribute a faster convergence and a lower estimation bias. In general, Theorem 5.3 guarantees the convergence of the iterative algorithm given that the edge dependency exists, and improves the convergence rate and estimation bias when the intensity of dependency increases.

In contrast to the computational convergence rates in Theorem 3.1 of [80], our Theorem 5.3 shows that incorporating the correlation information enables us to reduce the estimation bias and accelerate the convergence rate. Specifically, if the edges are truly independent, computational convergence using the independent likelihood method under the multiple networks setting can be extended from [80] in that its bias term and convergence rate have the order $\mathcal{O}\{N \exp(-NM)\}$ and $\mathcal{O}\left(\frac{1}{M\sqrt{N}}\right)$, respectively. However, if dependency among edges exists, the order of bias term assuming the independent likelihood becomes larger than $\mathcal{O}\{N \exp(-NM)\}$ because of a smaller effective sample size. Similarly, the convergence rate using the independent method becomes slower than $\mathcal{O}\left(\frac{1}{M\sqrt{N}}\right)$. In contrast, the order of bias term and the convergence rate are $\mathcal{O}\{N \exp(-\lambda NM)\}$ and $\mathcal{O}\left\{\frac{1}{(\lambda MN)^{1-\gamma}}\right\}$ based on Theorem 5.3 under the dependent setting. Therefore, the proposed method achieves a smaller estimation bias given $\lambda = \mathcal{O}(1)$ and a faster convergence rate given $\lambda > \mathcal{O}\{(M^\gamma N^{\gamma-0.5})^{1/(1-\gamma)}\}$ and $\gamma < 0.5$. It is noticeable that Theorem 5.3 can be generalized to the heterogeneous stochastic block model (2.2) with Condition (C1) replaced by (C1.b).

6. Numerical Studies. In this section, we conduct simulation studies to illustrate the performance of the proposed method on community detection in networks for dependent edges within-community. In particular, we compare our method to the existing variational EM method which assumes conditional independence among edges. Besides the comparison between the proposed method and independent likelihood method, we also conduct numerical comparisons between the propose method and existing multiple network community detection methods under different within-community dependency structure. Specifically, [47] proposes a spectral methods based on the optimal weighted average of multiple adjacent matrices, and the weighted average low-rank approximation which replaces an average of adjacent matrices by an average of low-rank approximation to each adjacent matrix. [74] proposes to jointly embed multiple adjacent matrices to a common subspace for clustering. [44] proposes an EM-based algorithm to recover community structure from the multiple noisy realizations of network.

6.1. Study 1: Networks with Dependent Within-community Connectivity. In the first simulation study, we consider networks where edges within the same community are correlated and compare the performance of various methods under different network sample sizes with various magnitudes of marginal means for within-community and between-community.

Suppose the memberships of nodes $\boldsymbol{Z}^* = \{\boldsymbol{Z}_1, \dots, \boldsymbol{Z}_n\}$ in the networks are given with K communities, where \boldsymbol{Z}_i is a binary indicator vector corresponding to the membership of nodes i . Conditional on \boldsymbol{Z}^* , edges in each sample network are generated following the Bernoulli marginal distribution as in (2.1), where within-community edges follow an exchangeable correlation structure as in (3.1). Here we assume that between-community edges are independent from

each other. The block-wise marginal means μ_{ql} ($q, l = 1, \dots, K$) are associated with edgewise covariates through (2.2). In addition, the edgewise covariates follow a uniform distribution, where within-communities covariates

$$(6.1) \quad x_{ij}^m \sim Unif(a_1, a_2) \text{ if } Z_{iq} = Z_{jq} = 1,$$

and between-community covariates

$$(6.2) \quad x_{ij}^m \sim Unif(b_1, b_2) \text{ if } Z_{iq} \neq Z_{jq}, q = 1, \dots, K.$$

Although the probability of each edge is different, the edges within the same community share the same coefficient β_{ql} in (2.2). In the following simulation studies, we generate correlated unweighted edges through the R package "MultiOrd."

Specifically, the sample networks consist of 40 nodes split into two communities. In a balanced community network, each community has 20 nodes. In an unbalanced case, two communities are comprised of 10 and 30 nodes, respectively. We compare the performance under different sample sizes of networks with $M = 20, 40$ and 60 , and different intensities of within-community dependency with correlation coefficient $\rho = 0, 0.3$ and 0.6 .

To simulate a weak marginal signal case, we let the block-wise parameters be $\beta_{11} = 1, \beta_{22} = 1.5$ and $\beta_{12} = \beta_{21} = 0$. The means of within-community and between-community covariates are 0 with $a_1 = b_1 = -0.2$ and $a_2 = b_2 = 0.2$ in (6.1) and (6.2). Here, both the average marginal means of within-community edges and between-community edges are very close to 0.5.

For a strong marginal signal case, the block-wise parameters are $\beta_{11} = 0.3, \beta_{22} = 0.6$ and $\beta_{12} = \beta_{21} = 0.2$. The within-community covariates are generated via (6.1) with $a_1 = 0.9$ and $a_2 = 1.1$, and between-community covariates are generated from (6.2) with $b_1 = -0.8$ and $b_2 = -0.6$. Note that there is a distinct gap between within-community and between-community marginal means, thus the marginal signal is more dominant for nodes within communities. Specifically, the average marginal means for the edges within two communities are 0.57 and 0.64, respectively. The average marginal mean for the between-community edges is 0.46.

We use the Adjusted Rand Index (ARI) to measure the performance of clustering. The ARI takes a value between -1 and 1 , where 1 represents a perfect matching of true memberships and predicted memberships of clustering, 0 indicates a random clustering and a negative value indicates that the agreement is less than the expectation from a random result. In the following simulations, we choose five fixed initial memberships of nodes in both balanced and unbalanced communities. These initials can be obtained from spectral clustering on sample networks. The Adjusted Rand Indices based on these chosen initials range between 0.30 to 0.34 under the unbalanced community case and between 0.25 to 0.29 under the balanced community case, which are far from the true memberships.

We compare the performance of clustering and parameter estimation for the proposed method applying the second-order (Bahadur_{2nd}) and the fourth-order (Bahadur_{4th}) Bahadur approximation, and the variational EM (VEM) approach with only marginal information.

In Table 1 and Table 2, the proposed method with the second-order and fourth-order approximations outperform the variational EM in clustering. Specifically, under the weak marginal signal case in Table 1, the Adjusted Rand Index of the variational EM are 0.34 under different network sizes and correlation strengths, which are similar to the ones calculated by fixed initials. In addition, since the distributions of marginal means from within-community and between-community are similar, the variational EM marginal approach barely improves over the initial memberships as it only utilizes the marginal information. However, the proposed method with the second-order or fourth-order Bahadur representation improves on the ARI by about 280%, compared to the VEM when $\rho = 0.3$ and $\rho = 0.6$. In addition, the performance of the proposed method improves by 1 ~ 5% as the number of sample networks increases from 20 to 60. Furthermore, incorporating the fourth-order interaction can slightly improve the accuracy of clustering.

TABLE 1

Adjusted Rand Index between estimated membership and true membership for networks with two communities and weak marginal signal averaging on 50 replicates.

		Unbalanced community			Balanced community		
		$M = 20$	$M = 40$	$M = 60$	$M = 20$	$M = 40$	$M = 60$
$\rho = 0$	VEM	0.38	0.41	0.48	0.31	0.28	0.28
	Bahadur _{2nd}	0.36	0.41	0.47	0.32	0.29	0.29
	Bahadur _{4th}	0.35	0.37	0.47	0.30	0.29	0.30
$\rho = 0.3$	VEM	0.34	0.34	0.34	0.28	0.28	0.28
	Bahadur _{2nd}	0.94	0.98	0.99	0.96	0.99	1.00
	Bahadur _{4th}	0.96	0.99	1.00	0.99	0.99	1.00
$\rho = 0.6$	VEM	0.34	0.34	0.34	0.29	0.28	0.28
	Bahadur _{2nd}	0.96	0.99	0.99	0.97	1.00	1.00
	Bahadur _{4th}	0.99	1.00	1.00	0.99	1.00	1.00

TABLE 2
Adjusted Rand Index between estimated membership and true membership for networks with two communities and strong marginal signal averaging on 50 replicates.

		Unbalanced community			Balanced community		
		<i>M</i> = 20	<i>M</i> = 40	<i>M</i> = 60	<i>M</i> = 20	<i>M</i> = 40	<i>M</i> = 60
$\rho = 0$	VEM	0.78	0.92	0.98	0.76	0.90	0.97
	Bahadur _{2nd}	0.73	0.91	0.97	0.77	0.92	0.98
	Bahadur _{4th}	0.69	0.86	0.95	0.72	0.92	0.98
$\rho = 0.3$	VEM	0.78	0.81	0.83	0.68	0.79	0.84
	Bahadur _{2nd}	0.99	0.99	1.00	0.98	1.00	1.00
	Bahadur _{4th}	0.99	0.99	1.00	0.99	1.00	1.00
$\rho = 0.6$	VEM	0.78	0.89	0.83	0.84	0.92	0.88
	Bahadur _{2nd}	0.99	1.00	1.00	0.99	1.00	1.00
	Bahadur _{4th}	0.99	1.00	1.00	0.99	1.00	1.00

We notice that when the correlation is as moderate as 0.3, the proposed method still achieves significant improvement over the variational EM and almost fully recovers the true memberships of clustering. We consider this as an intrinsic advantage of the proposed method in capturing the relatively weak dependency among edges to improve the clustering. This is because the proposed method not only captures pairwise dependency but also reflects connectivities among nodes within a community. That is, even a weak dependency among pairwise connectivities can lead to an accumulative information recovery of clustering.

Table 2 illustrates the clustering performance when the marginal signal is strong. In contrast to Table 1, the variational EM significantly improves on clustering because of the large discrepancy between the within-community marginal mean and the between-community marginal mean. Nevertheless, incorporating the correlation among within-community edges still improves the clustering accuracy by 20% to 26% under various sample sizes of networks and intensities of correlation. The clustering accuracy of the proposed method improves when either the sample size or the correlation increases. In general, stronger correlation and a larger sample size lead to better performance when the marginal signal itself is strong.

In addition to clustering, we also provide estimation of the marginal parameters. Tables 3, 4 and 5 compare parameter estimation between the proposed method and the variational EM when the marginal signal is weak. For within-community parameters β_{11} and β_{22} , the estimation of the proposed method consistently reduces bias 30 ~ 99% more than the variational method, except when $M = 20$ and $\rho = 0.6$. This is because the sample size $M = 20$ is not sufficiently large to offset the high variance among highly-correlated within-community edges. For the between-community parameter β_{12} , the estimation bias of the proposed method consistently decreases more than 80% compared to the VEM under all settings. Additionally, the standard errors of the proposed estimator decrease faster than the variational method as the sizes of networks increase.

TABLE 3
Estimation of within-community parameter $\beta_{11} = 1$ for networks with two communities and weak marginal signal.

		Unbalanced community			Balanced community		
		<i>M</i> = 20	<i>M</i> = 40	<i>M</i> = 60	<i>M</i> = 20	<i>M</i> = 40	<i>M</i> = 60
$\rho = 0$	VEM	0.56 _{0.42}	0.59 _{0.29}	0.58 _{0.20}	0.64 _{0.32}	0.57 _{0.16}	0.64 _{0.18}
	Bahadur _{2nd}	0.57 _{0.42}	0.58 _{0.30}	0.57 _{0.21}	0.61 _{0.28}	0.57 _{0.16}	0.66 _{0.20}
	Bahadur _{4th}	0.52 _{0.42}	0.55 _{0.28}	0.57 _{0.19}	0.58 _{0.27}	0.58 _{0.18}	0.65 _{0.19}
$\rho = 0.3$	VEM	0.49 _{0.30}	0.50 _{0.17}	0.52 _{0.14}	0.58 _{0.24}	0.58 _{0.18}	0.59 _{0.12}
	Bahadur _{2nd}	0.81 _{0.48}	0.84 _{0.32}	0.89 _{0.27}	0.95 _{0.24}	0.93 _{0.16}	0.92 _{0.14}
	Bahadur _{4th}	0.85 _{0.47}	0.83 _{0.31}	0.89 _{0.27}	0.96 _{0.24}	0.93 _{0.16}	0.93 _{0.14}
$\rho = 0.6$	VEM	0.56 _{0.22}	0.54 _{0.20}	0.52 _{0.15}	0.61 _{0.27}	0.61 _{0.16}	0.60 _{0.14}
	Bahadur _{2nd}	1.01 _{0.42}	1.04 _{0.35}	1.00 _{0.29}	0.95 _{0.31}	1.00 _{0.19}	0.96 _{0.15}
	Bahadur _{4th}	0.99 _{0.25}	1.05 _{0.15}	1.01 _{0.13}	0.97 _{0.31}	1.01 _{0.19}	0.97 _{0.16}

In addition to comparing to the independent likelihood approach, we also compare the proposed method with other multi-network community detection methods, e.g., weighted networks average (WNA) [47], weighted low-rank average (WLRA) through replacing adjacency matrices by their low-rank approximations [47], the network denoising method [44], and joint embedding [74]. In addition, we extend the dependency structure of the within-community edges to the

TABLE 4

Estimation of within-community parameter $\beta_{22} = 1.5$ for networks with two communities and weak marginal signal.

		Unbalanced community			Balanced community		
		$M = 20$	$M = 40$	$M = 60$	$M = 20$	$M = 40$	$M = 60$
$\rho = 0$	VEM	1.43 _{0.43}	1.42 _{0.34}	1.45 _{0.26}	1.18 _{0.40}	0.94 _{0.16}	0.94 _{0.15}
	Bahadur _{2nd}	1.50 _{0.39}	1.49 _{0.31}	1.45 _{0.25}	1.21 _{0.42}	0.93 _{0.21}	0.97 _{0.22}
	Bahadur _{4th}	1.56 _{0.37}	1.49 _{0.30}	1.46 _{0.23}	1.19 _{0.47}	0.94 _{0.24}	0.96 _{0.22}
$\rho = 0.3$	VEM	1.31 _{0.23}	1.40 _{0.11}	1.37 _{0.11}	1.05 _{0.21}	0.92 _{0.16}	0.92 _{0.16}
	Bahadur _{2nd}	1.56 _{0.19}	1.50 _{0.10}	1.49 _{0.09}	1.48 _{0.22}	1.45 _{0.19}	1.44 _{0.14}
	Bahadur _{4th}	1.55 _{0.19}	1.50 _{0.09}	1.49 _{0.09}	1.48 _{0.22}	1.45 _{0.19}	1.45 _{0.14}
$\rho = 0.6$	VEM	1.46 _{0.16}	1.43 _{0.16}	1.38 _{0.13}	1.16 _{0.21}	1.09 _{0.21}	1.06 _{0.22}
	Bahadur _{2nd}	1.73 _{0.29}	1.60 _{0.15}	1.52 _{0.12}	1.73 _{0.28}	1.60 _{0.29}	1.64 _{0.15}
	Bahadur _{4th}	1.69 _{0.25}	1.60 _{0.15}	1.52 _{0.13}	1.73 _{0.26}	1.61 _{0.29}	1.64 _{0.15}

TABLE 5

Estimation of within-community parameter $\beta_{12} = 0$ for networks with two communities and weak marginal signal.

		Unbalanced community			Balanced community		
		$M = 20$	$M = 40$	$M = 60$	$M = 20$	$M = 40$	$M = 60$
$\rho = 0$	VEM	0.52 _{0.35}	0.57 _{0.24}	0.47 _{0.22}	0.22 _{0.31}	0.39 _{0.14}	0.41 _{0.11}
	Bahadur _{2nd}	0.51 _{0.32}	0.58 _{0.23}	0.48 _{0.21}	0.23 _{0.30}	0.41 _{0.16}	0.39 _{0.15}
	Bahadur _{4th}	0.51 _{0.29}	0.63 _{0.22}	0.48 _{0.20}	0.25 _{0.28}	0.40 _{0.17}	0.41 _{0.13}
$\rho = 0.3$	VEM	0.68 _{0.24}	0.68 _{0.13}	0.69 _{0.10}	0.42 _{0.14}	0.35 _{0.12}	0.40 _{0.10}
	Bahadur _{2nd}	-0.02 _{0.25}	0.00 _{0.15}	0.00 _{0.11}	0.03 _{0.20}	-0.05 _{0.16}	-0.02 _{0.12}
	Bahadur _{4th}	-0.02 _{0.24}	0.00 _{0.14}	0.00 _{0.11}	0.03 _{0.18}	-0.06 _{0.16}	0.03 _{0.12}
$\rho = 0.6$	VEM	0.72 _{0.17}	0.71 _{0.11}	0.70 _{0.09}	0.41 _{0.18}	0.45 _{0.11}	0.48 _{0.11}
	Bahadur _{2nd}	-0.05 _{0.17}	-0.03 _{0.13}	0.02 _{0.11}	0.00 _{0.19}	0.01 _{0.12}	0.03 _{0.12}
	Bahadur _{4th}	-0.04 _{0.17}	-0.03 _{0.13}	-0.02 _{0.11}	-0.02 _{0.18}	0.00 _{0.12}	0.03 _{0.11}

mixture correlation structure and Toeplitz correlation structure, and their submatrix structures are provided as follows:

$$\begin{bmatrix} 1 & \rho & \rho & 0 \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ 0 & \rho & \rho & 1 \end{bmatrix}, \quad \begin{bmatrix} 1 & \rho & -\rho & \rho \\ \rho & 1 & \rho & -\rho \\ -\rho & \rho & 1 & \rho \\ \rho & -\rho & \rho & 1 \end{bmatrix}.$$

Toeplitz correlation structure

Mixture correlation structure.

For the mixture correlation structure, each pair of within-community edges is either positively or negatively correlated, alternatively. For the Toeplitz correlation structure, we choose a specific width of the non-zero diagonal band such that about 20% of all pairs of edges are correlated. In addition, we consider the exchangeable correlation structure with a correlation density λ in (5.1) varying from 10% to 50%. The magnitude of pairwise edge correlation ρ is 0.5 for all the settings. The sample networks consist of two balanced communities. For the simulations on multiple networks with 200 nodes, the connecting probabilities for within-community edges and between-community edges are 0.52 and 0.51, respectively. Figure 2 provides the distributions of within-community connection density in $M = 100$ sample networks under different edge-dependency structures.

Table 6 indicates that the proposed approach achieves the best performance under different edge-dependency structures. Specifically, the proposed method improves on the ARI by about 50% and 15% compared to the best competing method WLRA for the mixture and Toeplitz correlation structures, respectively. Under the exchangeable correlation structure, WLRA and joint embedding have similar performance as the proposed method when the correlation density is larger than 20%, while the proposed method outperforms them when the correlation density is 10%. Different from other competing methods, WLRA and joint embedding utilize the latent space model to represent sample networks and extract shared community structures. The latent space can capture a sufficiently large gap in connection density between within-community edges and between-community edges due to the exchangeable correlation structure. However, the connection density gap is not necessarily significant when the within-community edges follow a Toeplitz or mixture correlation structure. On the other hand, the proposed method is able to identify communities through directly capturing the correlations among edges, and still performs well under non-exchangeable dependency structures.

Furthermore, we investigate the performance of the proposed method on large node size of multi-networks, e.g., 1000 nodes splitting into two balanced communities. The network sample size is 100. Within each community, we let edges from a subgroup of 250 nodes follow a mixture correlation structure with $\rho = 0.5$, while the marginal mean of edges in the subgroup is fixed at 0.52. To model the heterogeneity of edges across sample networks, the marginal mean of within-community edges is 0.53 for half of the sample networks, and 0.49 for the other half of the sample networks for

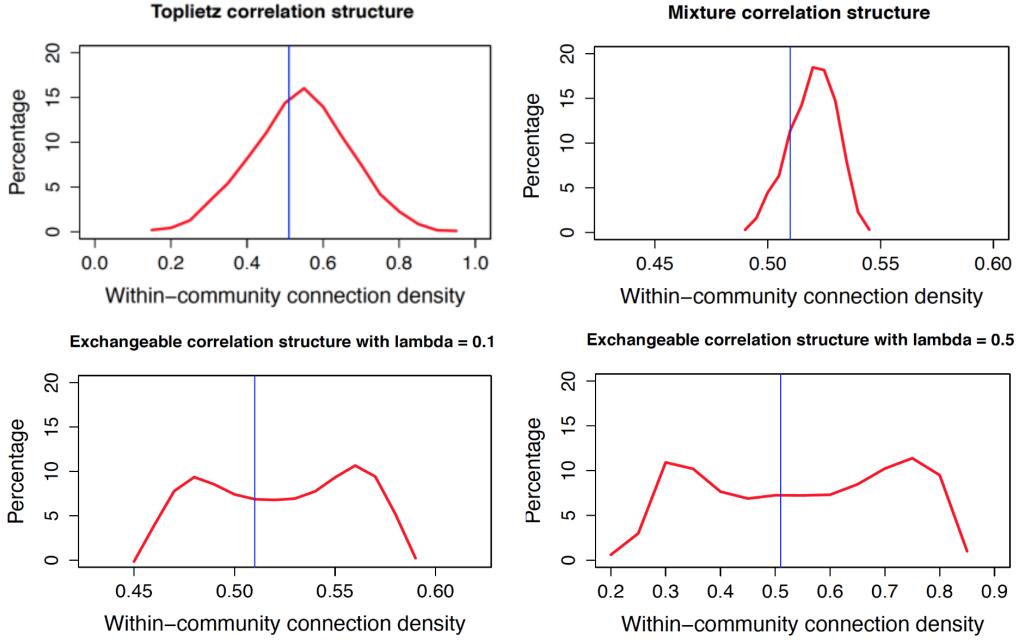


Fig 2: The distribution of within-community connection density over sample networks under different correlation structures. The Y-axis denotes the percentage of $M = 100$ sample networks with a specific within-community connection density. The blue vertical line indicates the between-community connection density at 0.51.

TABLE 6

Comparison of ARI under different correlation structures, varying network sizes and sample sizes. Proposed: the proposed method, VEM: independent likelihood, WNA: weighted networks average [47], WLRA: weighted low rank approximation [47], ND: network denoising [44], JE: joint embedding [74]. Toeplitz and Mixture stand for the Toeplitz and mixture correlation structures. Ex10%, Ex20%, and Ex50% stand for the exchangeable correlation structure with correlation density at 10%, 20%, and 50%.

Corr. struc.	Network size	Sample size	Proposed	VEM	WNA	WLRA	ND	JE
Toeplitz	N=200	M = 50	0.83	0.23	0.42	0.80	0.21	0.57
		M = 100	0.97	0.19	0.43	0.84	0.29	0.64
Ex10%	N=200	M = 50	0.16	0.02	0.00	0.00	0.01	0.02
		M = 100	0.23	0.12	-0.01	0.01	0.04	0.03
Ex20%	N=200	M = 50	1.00	0.52	-0.02	0.88	0.01	0.18
		M = 100	1.00	0.53	-0.01	0.98	0.04	0.35
Ex50%	N=200	M = 50	1.00	0.78	0.28	1.00	0.04	0.96
		M = 100	1.00	0.80	0.19	1.00	0.07	1.00
Mixture	N=200	M = 50	0.83	0.14	0.18	0.56	0.16	0.38
		M = 100	0.97	0.22	0.19	0.63	0.15	0.38
	N=1000	M = 100	0.79	0.50	0.47	0.22	0.00	0.34

uncorrelated edges. The marginal mean of between-community edges is 0.51. Table 6 shows that the proposed method improves on the ARI by about 60% compared with the best performance using the independent likelihood.

We also investigate the clustering performance of the independent likelihood and the proposed approximate likelihood approach given different within-community second-order correlation density λ in (5.1). The setting is similar to the weak marginal signal cases. Specifically, the sample networks contain two communities with identical pairwise within-community correlation $\rho = 0.6$. The sizes of the sample networks and nodes are $M = 40, N = 40$. The density λ increases from 0.01 to 1. The Adjusted Rand Index comparisons are illustrated in Figure 3. In general, the approximate likelihood approach has improving performance when the correlation connectivities among within-community edges increase, in contrast to the independent likelihood approach. Figure 3 shows that the true membership recovery using the approximate likelihood approach is high even when the second-order within-community correlation is relatively sparse ($\lambda = 0.05$), while the independent likelihood approach performs poorly with a constant ARI regardless of λ . This finding supports Theorem 5.1 and 5.2 in that the proposed method produces an accelerated decay in misclassification rate as λ increases.

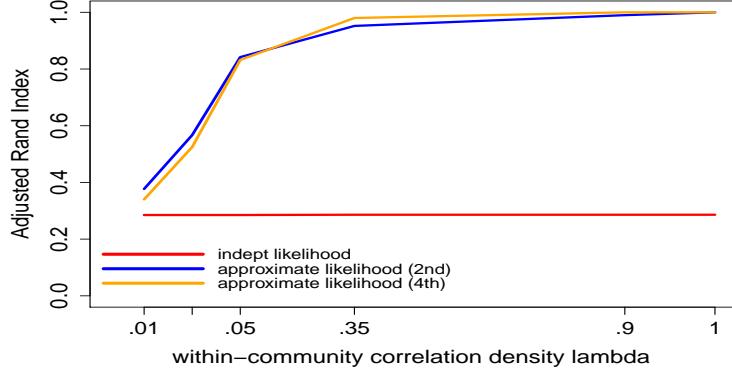


Fig 3: Clustering performance comparisons between independent likelihood and the proposed approximate likelihood approach incorporating the second-order and fourth-order correlations. The power of recovering community structure from the proposed method increases as the intensity of within-community dependency (λ) increases.

6.2. Study 2: Robustness to Model Specification and Unbalanced Community Size. In Study 2, we study the robustness of the proposed approach to the model misspecification. We first investigate whether the proposed method holds for a more general dependency structure among edges from different communities, for example, correlation among edges between different communities

$$(6.3) \quad \text{corr}(Y_{i_1 j_1}^m, Y_{i_2 j_2}^m) = \tilde{\rho}, \text{ given } z_{i_1} = z_{j_1} = q, z_{i_2} = z_{j_2} = l, q \neq l,$$

where $\tilde{\rho} \leq \rho_q$ in (3.1) in general. While (3.1) characterizes the concordance of edges within a community, (6.3) also captures the heterogeneity of sample networks. The heterogeneity of multi-layer networks is common in community detection.

In this simulation, we demonstrate that the proposed method is still robust when there is heterogeneity of connectivities among sample networks. To simulate the dependency among inter-community connectivity, we split M sample networks into 10 groups. Within each group, we add the random effects γ_k to the within-community marginal means:

$$\mu_{qq}^m = \frac{\exp(\beta_{ql} x_{ij}^m)}{1 + \exp(\beta_{ql} x_{ij}^m)} + \gamma_k, \quad M \frac{k-1}{10} \leq m \leq M \frac{k}{10},$$

where $\gamma_k \sim N(0, \sigma^2)$, $k = 1, \dots, 10$, $m = 1, \dots, M$, and $q = 1, \dots, K$. The variance σ of the random effect γ_k captures the intensity of dependency among inter-community connectivities, which increases as σ increases. We set $\sigma = 0.5$ to represent a weak inter-community dependency and $\sigma = 1.5$ for a strong inter-community dependency, while the other settings remain the same as in simulation Study 1. Our primary interest is to compare clustering performance between the proposed method and the variational method under the weak marginal signal case.

Tables 7 and 8 illustrate the clustering performance between the variational method and the proposed method under balanced and unbalanced community sizes, respectively. When the within-community correlation is moderate at 0.3, the proposed method improves the clustering accuracy by 170% to 257% for various network sizes and σ . For strong correlation $\rho = 0.6$, the improvement is between 210% to 257%. In particular, the proposed method has better performance when the networks have strong intra-community correlation and large sample sizes under both weak and strong inter-community correlation cases. In addition, using the fourth-order Bahadur representation improves the accuracy by 6% and 14% when $\sigma = 0.5$ and $\sigma = 1.5$ compared to the second-order Bahadur representation, indicating that the higher-order method still enhances the clustering outcome under the misspecified model. It is interesting to note that the performance of the proposed method decreases by 5% to 15% when the inter-community correlation is strong and the number of networks is small, compared to the same setting with weak inter-community correlation. However, the performances under both weak or strong inter-community correlation are similar when the sample size of networks increases.

In addition, we investigate the robustness of the proposed method on misspecifying the number of communities and unbalanced community sizes. For the case of misspecified number of communities, the true $K = 3$ communities, $N = 300$ and $M = 100$. The within-community edges follow a mixture correlation structure with $\rho = 0.5$. The marginal means of within-community edges and between-community edges are 0.52 and 0.51, respectively. Table 9 provides the performances of different methods under the misspecified $K = 2$. The proposed method performs the best, and has an ARI 70% more than the second best method, WLRA.

TABLE 7

Performance comparison given misspecified inter-community correlation with balanced community and weak marginal signal averaging on 50 replicates.

		$\sigma = 0.5$			$\sigma = 1.5$		
		$M = 20$	$M = 40$	$M = 60$	$M = 20$	$M = 40$	$M = 60$
$\rho = 0.3$	VEM	0.28	0.28	0.29	0.28	0.28	0.29
	Bahadur _{2nd}	0.90	0.99	1.00	0.76	0.99	0.99
	Bahadur _{4th}	0.96	1.00	1.00	0.87	0.98	1.00
$\rho = 0.6$	VEM	0.28	0.28	0.29	0.28	0.28	0.29
	Bahadur _{2nd}	0.94	0.99	1.00	0.87	0.99	1.00
	Bahadur _{4th}	0.99	1.00	1.00	0.94	0.99	1.00

TABLE 8

Performance comparison given misspecified inter-community correlation with unbalanced community and weak marginal signal averaging on 50 replicates.

		$\sigma = 0.5$			$\sigma = 1.5$		
		$M = 20$	$M = 40$	$M = 60$	$M = 20$	$M = 40$	$M = 60$
$\rho = 0.3$	VEM	0.32	0.33	0.33	0.33	0.33	0.33
	Bahadur _{2nd}	0.89	0.98	0.99	0.89	0.95	0.97
	Bahadur _{4th}	0.95	0.99	0.99	0.93	0.94	0.94
$\rho = 0.6$	VEM	0.34	0.33	0.34	0.33	0.33	0.33
	Bahadur _{2nd}	0.91	0.96	0.98	0.91	0.95	0.94
	Bahadur _{4th}	0.95	0.96	0.97	0.92	0.93	0.92

For the case of unbalanced community size, the sample networks contain $N = 250$ nodes splitting into three communities. The size of each community is 130, 80 and 40, respectively. The remaining settings are the same as the previous case of misspecified number of communities. Table 9 shows that the proposed method still performs the best, and has an ARI 50% more than the second best WLRA. These results indicate that the proposed method is still robust when the number of communities is misspecified, or community sizes are unbalanced.

TABLE 9

*Comparison of ARI under (1) **misspecified K** case, where the true multiple networks have $K = 3$ with balanced communities, and the misspecified the number of clusters is $K = 2$; (2) **unbalanced size** case, where the true multiple networks consist of three unbalanced communities with the size of 130, 80, 40, and the number of clusters is correctly specified as $K = 3$.*

	Proposed	VEM	WNA	WLRA	ND	JE
misspecified K	0.53	0.14	0.31	0.19	0.11	0.18
unbalanced size	0.73	0.26	0.25	0.49	0.17	0.41

7. Real Data Example. In this section, we analyze two multi-network datasets: food trading networks and brain fMRI imaging networks. We apply the proposed method and other competing methods introduced in Section 6 to detect the community structure within the multiple networks.

7.1. Food Trading Networks. In this subsection, we apply the proposed method to the 2010 Worldwide Food Import/Export Network dataset [20] from the Food and Agriculture Organization of the United Nations (<http://www.fao.org>). We create 364 networks among 214 countries with a total of 318,346 binary edges, where each network captures the trading connections of a specific food product among countries.

The primary goal of the study is to identify the common trading communities among different countries shared by food and agricultural product networks. The phenomenon of common community structure for international food-trade multi-networks has been recently studied and confirmed in [71, 45, 8, 3]. In general, the community structure in trade networks of food products is highly influenced by the geopolitical, socio-economic and political relations among the countries. Therefore, countries tend to be in the same trading community across different food product networks if they have geographical or economic similarity.

One significant feature of these networks is that the average empirical correlation of the pairwise connection among trading countries is 0.29. Therefore, the SBM based on the conditional independent assumption among edges could possibly lead to a biased network clustering of countries.

We first preprocess the data to select nodes corresponding to the trading countries which are most relevant, the number of communities and the initial memberships of countries. Note that several major countries dominate the world economy

and lead a high number of trading connectivities, while the other countries with limited agricultural product categories have fewer trading connections with other countries for specific product networks. Here we focus on the partial trading networks consisting of major countries whose corresponding degrees of nodes are larger than 9, which results in 51 countries with major economic impact in the world, such as the United States, mainland China, Japan and some European countries. The average empirical correlation of the trading connections among these countries is 0.22, indicating that the connectivity dependency should be considered in clustering these countries' trading networks.

In general, there are two major procedures to select the number of communities. First, we can perform the Louvain method for community detection on each individual trading network to obtain the number of communities which maximizes the modularity and the size of the largest community. Next we take the average of the number of communities on networks whose number of communities is smaller than 10 and whose largest community size is larger than 14. This procedure removes the 18% of the product trading networks whose countries are commercially isolated from other countries, as our goal is to detect the commercial communities among the countries which are more connected with other countries. After preprocessing, the average number of communities is 4.9 and we set it to be 4, and there are 296 sample networks remaining in the following analysis. For the proposed method, we apply the homogeneous stochastic block model (2.1) with no edge-wise covariate, and the marginal trading intensity between two nations is modeled based on their community memberships.

Table 10 provides the estimated agricultural products trading communities among 51 countries based on the proposed method and other competing methods including the independent likelihood model using variational EM, WNA, WLRA, the network denoising, and joint embedding. In the following, we focus on the comparison of clustering results between the independent likelihood method and the proposed method as the results from comparisons between the proposed method and other competing methods could be inferred similarly.

The countries in the same community under the proposed method are marked with the same color, while the countries in newly formed communities based on competing methods are colored according to their memberships from the proposed method. In general, the Adjusted Rand Index for clustering between the variational method and the proposed method is 0.43, indicating that the communities detected by the two methods are quite different. The clustering results from the proposed method incorporating within-community dependency are more interpretable compared to the variational EM using only marginal information.

In particular, the proposed method identifies communities 1 and 2 (red and cyan color communities on Table 10) which are highly associated with their geographical and climate environments. However, these features are not detected by the variational method. For example, community 1 based on the variational method mainly consists of two types of countries: one group comprises Nordic and Eastern European countries with the cyan color, and the other group consists of countries in Latin America and Africa with the blue color. In contrast, the proposed method clusters countries from geographically neighboring countries in east Europe, including Austria, Poland and Romania which are clustered with other communities by the variational method. Community 2 based on the variational method contains northern countries such as Canada as well as tropical countries. However, the proposed method identifies community 2 with tropical coastal countries and Arabian Peninsula countries, which provides more meaningful community clusters compared to the variational EM method.

The variational method and proposed method detect the same third community with orange color in Figure Table 10 which contains 7 major countries from the European Union: Belgium, France, Germany, Italy, Netherlands, Spain and the UK.

The fourth community from the variational method consists of 11 Eastern European countries, and all are categorized in community 1 from the proposed method. Community 4 with red color on Table 10 in the proposed method includes countries with large populations or more developed agricultural product trading, such as mainland China, U.S.A, India and Japan.

Similar to the independent likelihood approach, the clustering of countries based on other competing methods (e.g., WNA, WLRA, and joint embedding) do not show clear intrinsic patterns or similarity among nations within the same community. In contrast, the proposed method groups countries based on geographical and climatic similarity. In particular, geographical distance has a significant impact on the tendency that two countries belong to the same trading community across different food products, and this finding is also supported by [71].

In terms of parameter estimation, the average probability of having trading connections for communities 1 and 2 based on the variational method are 0.21 and 0.52, respectively. For the proposed method, the estimated correlations of connectivities within communities 1 and 2 are both 0.22, and the corresponding average within-communities connection rates are 0.28 and 0.22, respectively. The relatively low connection rates and correlations may be related to the low diversity and high overlaps of product categories due to more restrictive geographical and climate environments.

For community 3, the corresponding estimated marginal parameters β_{33} from the proposed method and the variational method are 2.58 and 2.00 respectively, both of which indicate that the trading connection rate within European Union communities is greater than 88% on average. This strong marginal signal of within-community connection explains that the additional correlation information is less influential in clustering. Additionally, the estimated correlation within the third community is 0.58, implying a high connection rate within-community. For community 4, the corresponding average connection rate is 0.49 based on the variational method, and the estimated within-community average connection rate and the correlation are 0.61 and 0.27, respectively. This is because community 4 involves large population countries with more frequent trading on product categories due to their higher food diversity than other countries.

TABLE 10

Clustering of nations for the agricultural products trading networks based on different methods, the countries are colored according to their community memberships from the proposed method.

	Community 1	Community 2	Community 3	Community 4
Proposed method	Austria, Denmark, Finland, Ireland Poland, Russia, Sweden, Switzerland Turkey, Bulgaria, Croatia, Czech Greece, Hungary, Israel, Lithuania Norway, Portugal, Romania, Slovakia Slovenia, Ukraine	Brazil, Hong Kong, Taiwan Indonesia, Lebanon Korea, Argentina, Mexico New Zealand, Qatar South Africa, Chile Philippines	Belgium, France Germany Netherlands Spain Italy United Kingdom	Australia, Canada Mainland, India, Japan Malaysia, Singapore U.S.A, Thailand
VEM	Brazil, Denmark, Finland, Ireland Lebanon, Russia, Sweden, Switzerland Turkey, Ukraine, Israel, Argentina Mexico, Chile, Norway, Portugal South Africa, Qatar	Australia, Canada, Hong Kong Mainland, India, Japan, Taiwan Indonesia, Korea, Malaysia Philippines, Singapore, U.S.A Thailand, New Zealand	Belgium, France Germany Netherlands Spain Italy United Kingdom	Austria, Poland, Bulgaria Croatia, Czech, Romania Hungary, Lithuania Slovakia, Slovenia Greece
WNA	Brazil, Denmark, Finland, Lebanon Russia, Sweden, Turkey, Ukraine Israel, Lithuania, Mexico, Chile South Africa, Qatar, Ireland, Portugal Switzerland, Norway, Argentina	Taiwan, Canada, Mainland Hong Kong, Australia, U.S.A India, Thailand, Indonesia Malaysia, Philippines, Korea New Zealand, Singapore, Japan	Belgium, France Germany Netherlands Spain, Italy United Kingdom	Austria, Poland, Bulgaria Czech, Greece, Hungary Romania, Slovakia Slovenia, Croatia
WLRA	Austria, Denmark, Finland, Ireland Lebanon, Poland, Russia, Switzerland Turkey, Ukraine, Bulgaria, Israel Norway, Czech, Greece, Slovenia Slovakia, Sweden, Portugal, Hungary Lithuania, Romania, Croatia	Australia, Canada, India, Brazil Hong Kong, Taiwan, Chile, Japan Indonesia, New Zealand, Malaysia Philippines, Singapore, Thailand Argentina, South Africa, Mexico Qatar, Korea, U.S.A	Belgium, France Germany Netherlands Spain, Italy United Kingdom	Mainland
Network denoising	Austria, Denmark, Finland Ireland, Poland, Bulgaria Croatia, Czech, Greece Hungary, Lithuania, Romania Slovenia, Slovakia New Zealand, Chile, Norway, Portugal Indonesia, Qatar, Turkey	Brazil, Hong Kong, Taiwan, Japan India, Malaysia, South Africa Korea, Philippines, Russia, Sweden Singapore, Thailand, Switzerland Ukraine, Israel, Argentina, Mexico Lebanon	Belgium, France Germany Netherlands Spain, Italy United Kingdom	Mainland, Australia U.S.A, Canada
Joint embedding	Austria, Denmark, Finland, Poland Russia, Sweden, Turkey, Ukraine Czech, Greece, Hungary, Lithuania Norway, Portugal, Romania Slovenia, Ireland, Switzerland Croatia, Israel, Bulgaria, Slovakia	Australia, Brazil, Taiwan, Indonesia Japan, Malaysia, Philippines Argentina, Chile, New Zealand South Africa, Hong Kong, Lebanon Korea, Mexico, Qatar, Canada Thailand, Singapore	Belgium, France Germany Netherlands Spain, Italy United Kingdom	Mainland, India, U.S.A

7.2. Brain fMRI Imaging Networks. In this subsection, we apply the proposed method to the brain fMRI dataset from the Center for Biomedical Research Excellence (http://fcon_1000.projects.nitrc.org/indi/retro/cobre.html). The data consist of raw anatomical and functional magnetic resonance imaging data from 72 schizophrenia subjects and 75 normal subjects. Schizophrenia is a serious mental disorder which impairs cognitive abilities such as memory, attention and executive function. Although the causes and mechanisms of the disease are still unclear, empirical studies show that schizophrenia is related to abnormal functional connectivity and topological structures in brain neural networks [49, 27].

We preprocess the original imaging data using the Matlab toolbox "CONN" [76] for registration, background subtraction, and normalization. Specifically, the fMRI for each subject is reconstructed as a single network where the vertices represent the pre-defined 132 regions of interest (ROIs) in the brain based on the well-established FSL Harvard-Oxford

brain atlas [21, 25] and the AAL brain atlas [72]; and each edge represents the functional connectivity between a pair of brain regions, measured by Pearson correlations between the processed regional time series of blood oxygen levels when the subject is resting. After preprocessing, we obtain 147 correlation matrices with a dimension of 132×132 . The goal of this analysis is to detect the functional communities among the 132 pre-defined brain regions which could be associated with schizophrenia.

To construct the brain functional connectivity network for the m th subject, where $m = 1, \dots, 147$, we dichotomize elements in the m th correlation matrix $Y_{132 \times 132}^m$ such that $Y_{ij}^m = 1$ if the absolute value of correlation between region i and region j is larger than 0.35 and $Y_{ij}^m = 0$ otherwise. The cutoff correlation of 0.35 is proposed by [44]. To avoid too small a size of estimated communities (e.g., containing less than 5 regions) and facilitate the comparison between the proposed method and the existing methods introduced in Section 6, we set the number of communities as $K = 3$. Both the proposed method and the independent model utilize initial values generated from spectral clustering on randomly selected individual brain networks. We perform community detection on the brain networks from patients and normal subjects separately. One important feature of the networks is that 14% of the absolute values of pairwise correlations between the brain functional connectivities are larger than 0.2, which also justifies the proposed method in incorporating dependent connectivities.

Figure 4 and Figure 5 provide the visualization of the brain network clusterings for patients and normal subjects via the BrainNet Viewer [78] where regions in the same community are marked with the same color. The graphs in the left and right columns of Figure 4 and Figure 5 illustrate the communities in the brain networks of schizophrenic subjects, and normal subjects respectively. We adopt the adjusted rand index (ARI) to quantify the similarity between patients' communities and normal subjects' communities. The proposed method shows that the community memberships from 26 brain regions differ between schizophrenic and normal subjects. These 26 regions mainly consist of the temporal pole involving semantic representation and socio-emotional processing, the middle and inferior temporal gyrus associated with visual information processing, the hippocampus related to memory encoding and retrieval, and the cerebellum involved in attention and emotional control. The detailed memberships of brain regions based on the proposed method and competing methods are provided in the Supplementary material and the exact definition of each region label can be found in the Matlab toolbox "CONN".

One of the significant features of schizophrenia is the relatively low connectivity within patients' brain neural networks in that functional connectivity density among relevant brain regions is significantly reduced for schizophrenic patients compared with normal subjects [14, 48]. Accordingly, Table 11 illustrates the functional connectivity density within patients' brain network communities based on each method where the density is defined as the average within-community connectivity density of each patient network. Specifically, the within-community density from the proposed method is slightly lower than the independent model for all three communities, while the connectivity density in Community 2 from the proposed method is slightly higher than that from WNA, WLRA, network denoising, and joint embedding. However, the proposed method achieves much lower within-community densities for brain Community 1 and Community 3.

TABLE 11
Within-community connectivity density in patient network based on different clustering methods

	Community 1	Community 2	Community 3
Proposed method	0.26	0.25	0.19
VEM	0.28	0.26	0.20
WNA	0.50	0.19	0.45
WLRA	0.50	0.19	0.45
Network Denoising	0.54	0.19	0.26
Joint Embedding	0.52	0.19	0.47

In addition, the empirical studies [79, 27] show that schizophrenic patients have a decreased efficiency of the brain network among certain regions. Network efficiency [41] measures the efficiency of information exchange within a network and is defined as $\frac{1}{n(n-1)} \sum_{i \neq j \in G} \frac{1}{d(i,j)}$, where G is a network or subnetwork and $d(i,j)$ denotes the length of the shortest path between node i and node j in G . Therefore, we calculate the average within-community efficiency of the patients' brain networks based on different clustering methods. The result in Table 12 shows that the average within-community efficiency based on the proposed method is slightly lower than the independent method for all three communities. Compared with other competing methods, the proposed method has a comparable efficiency for Community 2, but a lower efficiency for Community 1 and Community 3 in the patients' brain networks. In summary, the proposed method is able to detect communities among brain regions which better capture the global brain network features associated with schizophrenia.

Furthermore, study [34] indicates that the connectivities between the superior frontal gyrus and the cerebellum regions significantly decrease among schizophrenic patients. Accordingly, the proposed method groups the superior frontal gyrus

Fig 4: Clusters of patient and normal subject brain networks based on the proposed method, independent likelihood method, and WNA. The similarity of brain network clusters between patient and normal subject is measured by the ARI. The regions with different community memberships in patient brain network and normal brain network are displayed with larger size and labels.

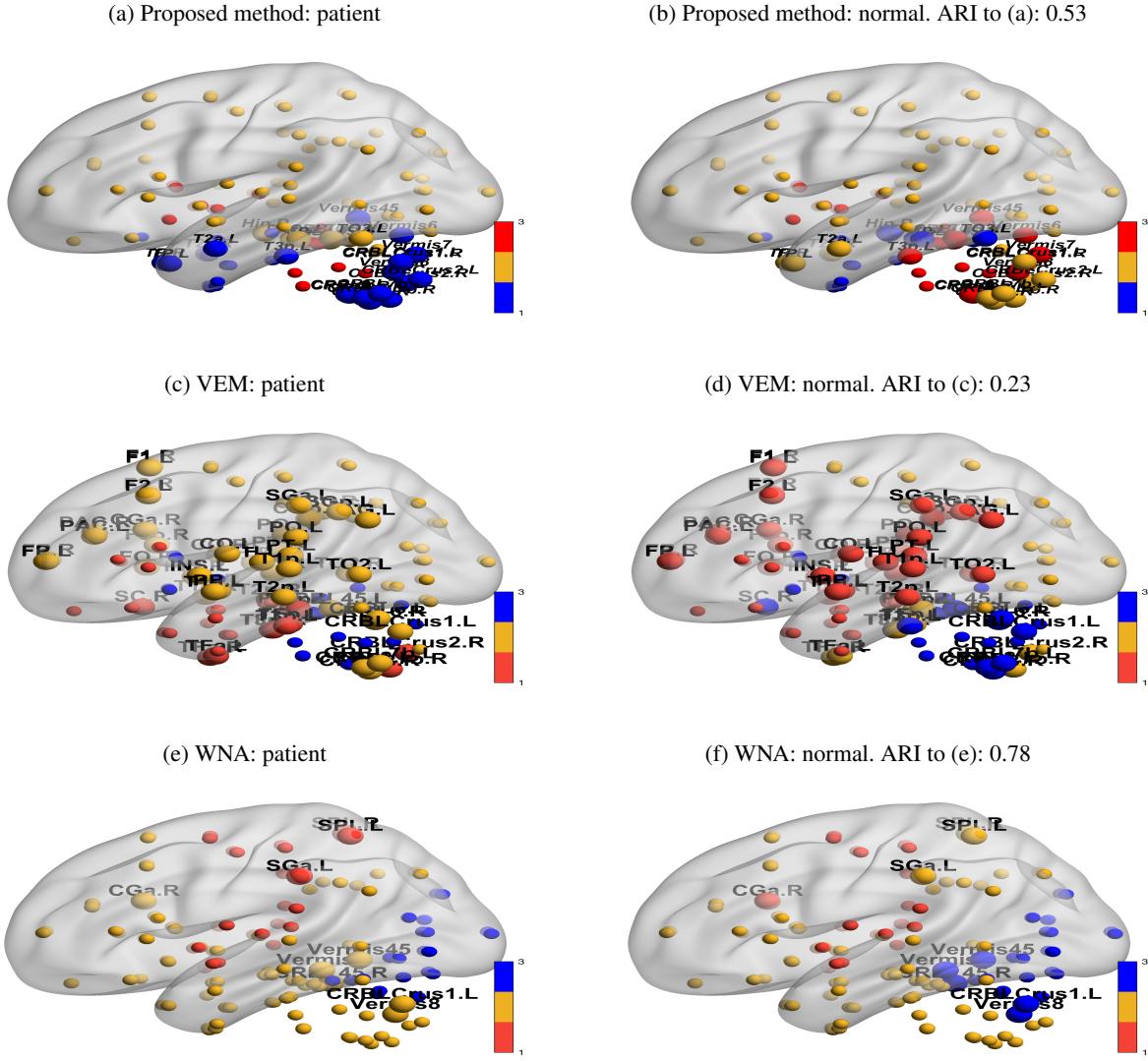


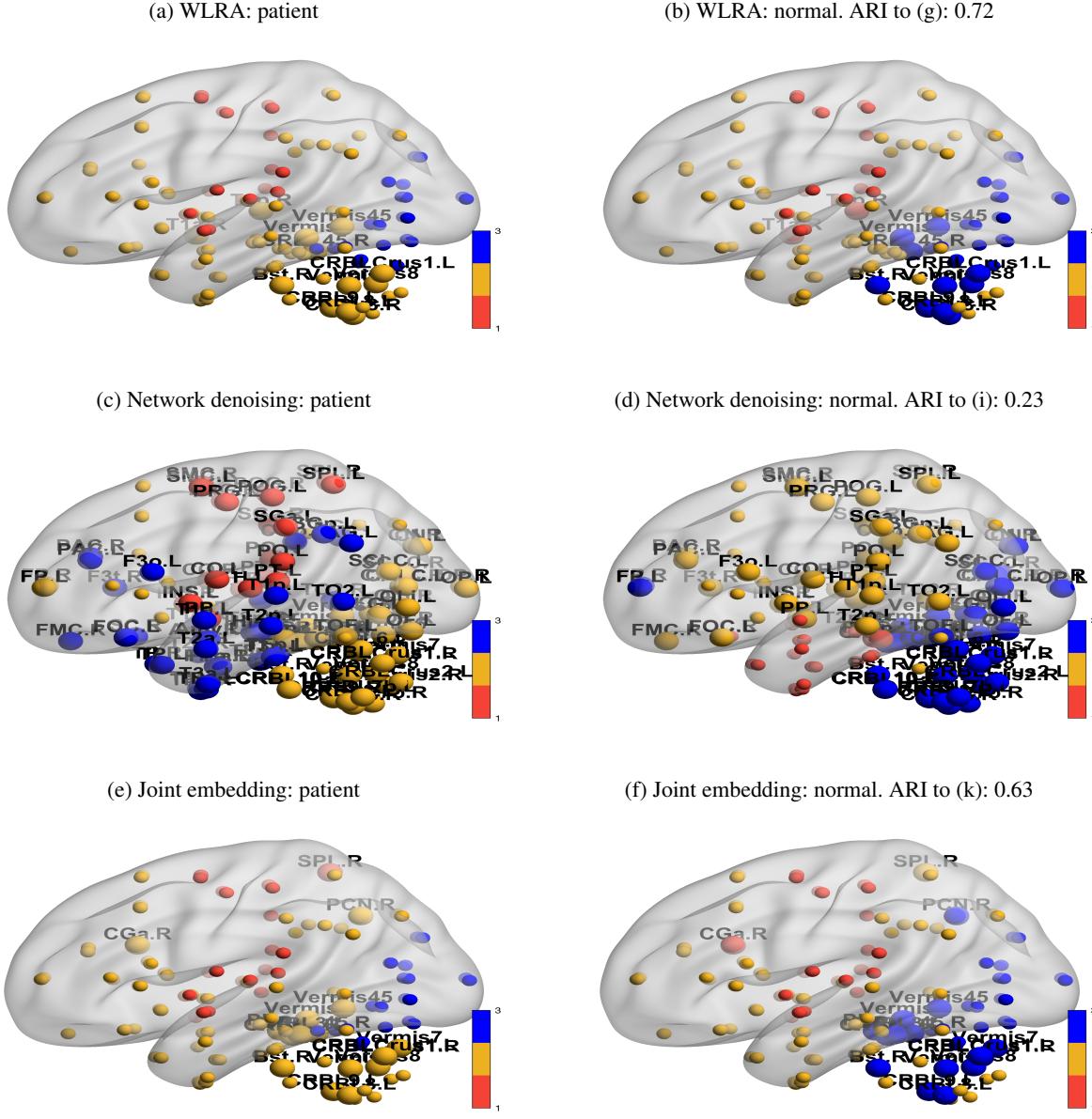
TABLE 12

Within-community patient network efficiency based on different clustering methods

	Community 1	Community 2	Community 3
Proposed method	0.57	0.60	0.40
VEM	0.58	0.61	0.47
WNA	0.74	0.55	0.71
WLRA	0.74	0.55	0.71
Network Denoising	0.77	0.55	0.58
Joint Embedding	0.76	0.55	0.72

and 4 cerebellum regions into the same community, while the superior frontal gyrus is clustered with 9 cerebellum regions by the independent model, 15 cerebellum regions by the WNA and WLRA, 18 cerebellum regions by network denoising, and 15 cerebellum regions by joint embedding. Compared with other methods, the communities detected from the proposed method are more consistent with the phenomenon of decreased connections between the superior frontal gyrus and the cerebellum for schizophrenic patients.

Fig 5: Clusters of patient and normal subject brain networks based on WLRA, network denoising, and joint embedding. The similarity of brain network clusters between patient and normal subject is measured by the ARI. The regions with different community memberships in patient brain network and normal brain network are displayed with larger size and labels.



8. Discussion. In this paper, we propose a new community detection method for networks incorporating the underlying dependency structure among connectivities. To model the correlation without specifying a joint likelihood for correlated edges, we construct an approximate likelihood based on the Bahadur representation which decomposes a joint distribution into a marginal term and high-order interaction terms. The proposed method provides flexible modeling on the correlation structure which can be specified through the interaction term in the approximate likelihood.

In theory, we establish the consistency of the nodes' membership estimator based on the proposed approximate likelihood and show that it achieves a faster convergence than the independent method. In addition, we show that the proposed iterative algorithm possesses desirable convergence properties. In particular, we show that the proposed approximate approach can achieve a faster computational convergence and a lower clustering bias compared to the variational EM algorithm. Furthermore, we show that the variational EM algorithm is a special case of our algorithm under the conditional independent model, which confirms that incorporating correlation information improves the accuracy for community detection.

Our numeric studies indicate that incorporating the within-community correlation among edges can improve the clustering performance compared to the marginal model, even under a moderately misspecified model on inter-community dependency. The improvement of community detection is more significant when the marginal signal is weak, which is less informative for distinguishing between within-community and between-community networks. In addition, the proposed method enables us to achieve more accurate parameter estimation.

In this paper, we only consider incorporating the within-community dependency. It would be worthy of further research to investigate more generalized dependency structures to include between-community dependency as well.

APPENDIX: PROOF OF THE MAIN THEOREMS

In the following, we denote the membership of node as random variable $z_i, i = 1, \dots, N$. Then $\mathbf{Z} = \{z_1, z_2, \dots, z_N\}$. Accordingly, we define the true membership of nodes as $z_i^* \in \{1, 2, \dots, K\}, i = 1, \dots, N$ and $z^* = \{z_1^*, z_2^*, \dots, z_N^*\}$. We denote $P^*(\cdot) = P(\cdot | \mathbf{Z} = z^*)$ as the conditional probability of observed networks given the true nodes' membership z^* . The number of misclassified nodes is denoted as r such that $\|z - z^*\|_0 = r$ for $z \neq z^*$. Define the m -th sample network as $\mathbf{Y}^m = (Y_{ij}^m)_{N \times N}$ and m -th sample network standardized by $\hat{\mu}_{aa}$ as $\hat{\mathbf{Y}}^{m,a} = (\hat{Y}_{ij}^{t,a})_{N \times N}$ where $\hat{Y}_{ij}^{m,a} = \frac{Y_{ij}^m - \hat{\mu}_{aa}}{\sqrt{\hat{\mu}_{aa}(1 - \hat{\mu}_{aa})}}, a = 1, \dots, K, m = 1, \dots, M$. We further define the s -th column of $\hat{\mathbf{Y}}^{m,a}$ as $\hat{Y}_s^{m,a}$.

Denote $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_N)$ as the estimated probability of nodes' memberships. Specifically, let $\alpha_i = (\alpha_{i1}, \dots, \alpha_{iK})_{1 \times K}$ be the probability of nodes i belonging to each community where $\sum_{q=1}^K \alpha_{iq} = 1, i = 1, \dots, N$. For simplicity of notation, if the subscripts indicate the community then $\alpha_q = (\alpha_{1q}, \dots, \alpha_{Nq})_{1 \times N}$ represents the probability of each node belonging to community q , where $q = 1, \dots, K$. Similarly, $z_q^* = \{z_{1q}^*, z_{2q}^*, \dots, z_{Nq}^*\}$ is a binary vector indicating nodes whose true membership belongs to community $q, q = 1, \dots, K$. Let $\text{vec}(\cdot)$ stand for the operation of vectorizing a matrix into a column.

A.1. Proof of Theorem 5.1. For the independent model $\log L_{\text{ind}}(\mathbf{Y} | \mathbf{Z} = z)$, we can simplify the likelihood ratio between a random membership z and the true membership z^* as

$$(A.1) \quad \log \frac{P_{\text{ind}}(\mathbf{Y} | \mathbf{Z} = z)}{P_{\text{ind}}(\mathbf{Y} | \mathbf{Z} = z^*)} = \sum_{m=1}^M \sum_{i < j} \left\{ Y_{ij}^m \log \frac{\mu_{z_i z_j}}{\mu_{z_i^* z_j^*}} + (1 - Y_{ij}^m) \log \frac{1 - \mu_{z_i z_j}}{1 - \mu_{z_i^* z_j^*}} \right\}.$$

We define two transformation functions $f_1(x)$ and $f_2(x)$ as:

$$\begin{aligned} f_1(x) &= \sqrt{\left\{ x \log \frac{\mu_{z_i z_j}}{\mu_{z_i^* z_j^*}} + (1 - x) \log \frac{1 - \mu_{z_i z_j}}{1 - \mu_{z_i^* z_j^*}} \right\}_+}, \\ f_2(x) &= \sqrt{\left\{ x \log \frac{\mu_{z_i z_j}}{\mu_{z_i^* z_j^*}} + (1 - x) \log \frac{1 - \mu_{z_i z_j}}{1 - \mu_{z_i^* z_j^*}} \right\}_-}. \end{aligned}$$

where $\{\cdot\}_+$ and $\{\cdot\}_-$ are positive part and negative part of a random variable. The previous summation can be decomposed as positive part and negative part:

$$\log \frac{P_{\text{ind}}(\mathbf{Y} | \mathbf{Z} = z)}{P_{\text{ind}}(\mathbf{Y} | \mathbf{Z} = z^*)} = \sum_{m=1}^M \sum_{i < j} \{f_1^2(Y_{ij}^m) - f_2^2(Y_{ij}^m)\}.$$

Define the vectorized edges in the t th sample network as:

$$(A.2) \quad X_m^+ = \{f_1(Y_{12}^m), f_1(Y_{13}^m), \dots, f_1(Y_{N-1,N}^m)\}, X_m^- = \{f_2(Y_{12}^m), f_2(Y_{13}^m), \dots, f_2(Y_{N-1,N}^m)\}.$$

Note that each element in X_m^+ or X_m^- is a bounded binary random variable. In addition, as $f_1(Y_{ij}^m)$ or $f_2(Y_{ij}^m)$ only rescale Y_{ij}^m then they preserve the within-community correlation among Y_{ij}^m . Then we consider the following quadratic forms

$$Q_1 = \sum_{m=1}^M \langle X_m^+, X_m^+ \rangle, Q_2 = \sum_{m=1}^M \langle X_m^-, X_m^- \rangle.$$

such that

$$\log \frac{P_{\text{ind}}(\mathbf{Y} | \mathbf{Z} = z)}{P_{\text{ind}}(\mathbf{Y} | \mathbf{Z} = z^*)} = Q_1 - Q_2 \text{ and } E(\log \frac{P_{\text{ind}}(\mathbf{Y} | \mathbf{Z} = z)}{P_{\text{ind}}(\mathbf{Y} | \mathbf{Z} = z^*)}) = EQ_1 - EQ_2.$$

Denote the centralized version quadratic forms Q_1 and Q_2 as \mathcal{Q}_1 and \mathcal{Q}_2 such that

$$\mathcal{Q}_1 = \sum_{m=1}^M \langle X_m^+ - E(X_m^+), X_m^+ - E(X_m^+) \rangle, \quad \mathcal{Q}_2 = \sum_{m=1}^M \langle X_m^- - E(X_m^-), X_m^- - E(X_m^-) \rangle.$$

Denote the following quadratic difference as:

$$\begin{aligned} \Delta(Q_1, \mathcal{Q}_1) &:= (Q_1 - E(Q_1)) - (\mathcal{Q}_1 - E(\mathcal{Q}_1)) = 2 \sum_{m=1}^M \langle E(X_m^+), X_m^+ - E(X_m^+) \rangle, \\ \Delta(Q_2, \mathcal{Q}_2) &:= (Q_2 - E(Q_2)) - (\mathcal{Q}_2 - E(\mathcal{Q}_2)) = 2 \sum_{m=1}^M \langle E(X_m^-), X_m^- - E(X_m^-) \rangle. \end{aligned}$$

For any $t > 0$, we have

$$\begin{aligned} P^* \left\{ \frac{P_{ind}(\mathbf{Y} | \mathbf{Z} = \mathbf{z})}{P_{ind}(\mathbf{Y} | \mathbf{Z} = \mathbf{z}^*)} > t \right\} &= P^* \left\{ (Q_1 - EQ_1) - (Q_2 - EQ_2) > \log t - E(Q_1 - Q_2) \right\} \\ &\leq P^* \left\{ Q_1 - EQ_1 > \frac{\log t - E(Q_1 - Q_2)}{2} \right\} + P^* \left\{ Q_2 - EQ_2 < -\frac{\log t - E(Q_1 - Q_2)}{2} \right\} \\ &= P^* \left\{ Q_1 - EQ_1 > \frac{\log t - E(Q_1 - Q_2)}{2} - \Delta(Q_1, \mathcal{Q}_1) \right\} \\ &\quad + P^* \left\{ Q_2 - EQ_2 < -\frac{\log t - E(Q_1 - Q_2)}{2} - \Delta(Q_2, \mathcal{Q}_2) \right\}, \end{aligned}$$

where

$$\begin{aligned} (A.3) \quad &P^* \left\{ Q_1 - EQ_1 > \frac{\log t - E(Q_1 - Q_2)}{2} - \Delta(Q_1, \mathcal{Q}_1) \right\} \\ &\leq \frac{1}{2} P^* \left\{ |Q_1 - EQ_1| > \frac{\log t - E(Q_1 - Q_2)}{2} - \Delta(Q_1, \mathcal{Q}_1) \right\} \\ &\quad P^* \left\{ Q_2 - EQ_2 > \frac{\log t - E(Q_1 - Q_2)}{2} - \Delta(Q_2, \mathcal{Q}_2) \right\} \\ (A.4) \quad &\leq \frac{1}{2} P^* \left\{ |Q_2 - EQ_2| > \frac{\log t - E(Q_1 - Q_2)}{2} - \Delta(Q_2, \mathcal{Q}_2) \right\}. \end{aligned}$$

Next, we estimate each of the term in (A.3). Given the $\{Y_{ij}^m\}_{m=1}^M$ are binary random variables and the setting that any two within-community edges $Y_{i_1 j_1}$ and $Y_{i_2 j_2}$ have a nonnegative correlation $\text{corr}(Y_{i_1 j_1}, Y_{i_2 j_2}) \geq 0$. Notice that

$$\text{corr}(f_1(Y_{i_1 j_1}), f_1(Y_{i_2 j_2})) = \begin{cases} \text{corr}(Y_{i_1 j_1}, Y_{i_2 j_2}) & \text{if } \mu_{z_i z_j} \geq \mu_{z_i^* z_j^*} \\ -\text{corr}(Y_{i_1 j_1}, Y_{i_2 j_2}) & \text{if } \mu_{z_i z_j} < \mu_{z_i^* z_j^*} \end{cases}.$$

We denote the covariance matrix of X_m^+ and X_m^- as Σ_1 and Σ_2 . Notice that a term in (A.1) is zero only when its corresponding node membership is misclassified. Define the the number of nonzero term in (1) as N_r given $\|\mathbf{z} - \mathbf{z}^*\|_0 = r$. Then we have $N_r = \frac{1}{2}rNM$. According to Lemma 1, X_m^+ is a subgaussian vector with a bounded subgaussian norm $L \leq C_1$ where C_1 is a positive constant and

$$(A.5) \quad L = \inf\{\alpha \geq 0 : E(\exp(\langle \mathbf{z}, X_m^+ - E(X_m^+) \rangle)) \leq \exp\{\alpha^2 \langle \Sigma_1 \mathbf{z}, \mathbf{z} \rangle\}/2\}.$$

Next we estimate $\|\Sigma_1\|_F$, $\|\Sigma_1\|_{op}$ and $\|\Sigma_2\|_F$, $\|\Sigma_2\|_{op}$ where $\|\cdot\|_F$ is the matrix Frobenius norm and $\|\cdot\|_{op}$ is the matrix spectral norm. Denote

$$\Lambda = \text{diag}(\sqrt{\text{Var}\{(X_m^+)_ {12}\}}, \sqrt{\text{Var}\{(X_m^+)_ {13}\}}, \dots, \sqrt{\text{Var}\{(X_m^+)_ {N-1, N}\}}).$$

Then $\|\Sigma_1\|_{op} = \|\Lambda R \Lambda\|_{op} \leq C_2 \|R\|_{op}$ where R is the correlation matrix of X_m^+ and based on (C1),

$$C_2 \leq \max_{1 \leq i < j \leq n} \text{Var}\{(X_m^+)_ {ij}\} \leq \eta_N \max\{\log \frac{\zeta}{1-\zeta}, \log \frac{1-\zeta}{\zeta}\}.$$

Denote the largest eigenvalue of R as λ_R . From the Gershgorin circle theorem, we have

$$\lambda_R \leq 1 + \max_{i=1, \dots, N(N-1)/2} \sum_{j \neq i} |R_{ij}|.$$

Denote the number of node in the largest community is N_k . Note that the misclassification number of node $\|z - z^*\|_0 = r$ and edgewise correlation density λ both affect the sparsity of R , we have for each row in R :

$$\sum_{j \neq i} |R_{ij}| \leq \rho N_k \min(r, \lambda N_k) \leq \rho \kappa_2 N \min(r, \kappa_2 \lambda N),$$

where $\rho = \max_{i,j} R_{ij}$. Therefore, we have

$$\|\Sigma_1\|_{op} \leq C \{1 + \rho \kappa_2 \eta_N N \min(r, \kappa_2 \lambda N)\},$$

for some constant C . Similarly we have a same upper bound for $\|\Sigma_2\|_{op}$. Notice that the dimension of R is $N_r \times N_r$ and $N_r \leq rN$. In each row of R , the number of non-zero elements is less than $1 + N_k \min(r, \lambda N_k)$. Therefore, we have

$$\|\Sigma_1\|_F^2 \leq C_2 \rho^2 r \eta_N N \{1 + \kappa_2 \eta_N N \min(r, \kappa_2 \lambda N)\}.$$

Then we are able to estimate the upper bound for the first term in (A.3). According to the generalized Hanson-Wright inequality in ([16]), we have:

$$(A.6) \quad \frac{1}{2} P^* \left\{ |Q_1 - EQ_1| > s \right\} \leq \exp \left\{ -C \min \left(\frac{s^2}{L^4 \|\Sigma_1\|_F^2 \|A\|_F^2}, \frac{s}{L^2 \|\Sigma_1\|_{op} \|A\|_{op}} \right) \right\}.$$

where $s = \frac{\log t - E(Q_1 - Q_2)}{2} - \Delta(Q_1, Q_1)$, $A = \mathbf{I}_{M \times M}$ and L is subgaussian norm of X_m^+ defined in (A.5). Then we have $L \leq C_1$ and $\|A\|_F^2 = M$, $\|A\|_{op} = 1$. To estimate s , notice

$$\begin{aligned} E(Q_1 - Q_2) &= E \left[\sum_{m=1}^M \sum_{i < j} \left\{ Y_{ij}^m \log \frac{\mu_{z_i z_j}}{\mu_{z_i^* z_j^*}} + (1 - Y_{ij}^m) \log \frac{1 - \mu_{z_i z_j}}{1 - \mu_{z_i^* z_j^*}} \right\} \right] \\ &= -M \sum_{i < j} \left\{ \mu_{z_i^* z_j^*} \log \frac{\mu_{z_i^* z_j^*}}{\mu_{z_i z_j}} + (1 - \mu_{z_i^* z_j^*}) \log \frac{1 - \mu_{z_i^* z_j^*}}{1 - \mu_{z_i z_j}} \right\}, \end{aligned}$$

where there are total N_r non-zero terms in the summation. We introduce the function

$$k(x, y) = x \log(x/y) + (1 - x) \log(1 - x)/(1 - y).$$

Notice that $k(x, y) > 0$ for every $x, y \in (0, 1)$. Then we define:

$$(A.7) \quad c^* := \min \{k(c_{ql}, c_{q'l'})\} > 0,$$

where the minimum are taken over $\{((q, l), (q', l')) | c_{q,l} \neq c_{q',l'}\}$. Given that $\eta_N = o_N(1)$, it can be shown that $k(\mu_{ql}, \mu_{q'l'}) \asymp \eta_N k(c_{ql}, c_{q'l'})$. Combined with $N_r = \frac{1}{2}rNM$, we have $-E(Q_1 - Q_2) > \frac{c^*}{2}r\eta_N NM$. To estimate $\Delta(Q_1, Q_1)$, given elements in X_m^+ are bounded, we denote $\omega_1 = \max_{1 \leq i < j \leq n} E\{(X_m^+)_{ij}\}$, $\omega_2 = \max_{1 \leq i < j \leq n} \text{Var}\{(X_m^+)_{ij}\}$

$$\begin{aligned} P(|\Delta(Q_1, Q_1)| > \frac{c^*}{2}rNM) &\leq P\left(\omega_1 \left| \sum_{m=1}^M \sum_{i=1}^{N_r} (X_{ti}^+ - E(X_{ti}^+)) \right| > \frac{c^*}{2}rNM\right) \\ &\leq \frac{\omega_1^2 M \text{Var}(\sum_{i=1}^{N_r} X_{ti}^+)}{c^{*2} r^2 N^2 M^2 / 4} \leq \frac{\omega_1^2 (\omega_2 r N (2 + \rho \lambda r N))}{c^{*2} r^2 N^2 M} \leq O\left(\frac{\eta_N \lambda}{M}\right). \end{aligned}$$

Therefore, as M or N increases s is dominated by $-E(Q_1 - Q_2)$ with probability approaching 1. Then for any fixed $t > 0$, $s > \mathcal{O}_N(\frac{c^*}{2}rNM)$. Therefore, we have

$$\begin{aligned} &\min \left(\frac{s^2}{L^4 \|\Sigma_1\|_F^2 \|A\|_F^2}, \frac{s}{L^2 \|\Sigma_1\|_{op} \|A\|_{op}} \right) \\ &\geq \min \left(\frac{(\frac{c^*}{2}r\eta_N NM)^2}{C_1^2 M C_2 \rho^2 r N \{1 + \kappa_2 \eta_N N \min(r, \kappa_2 \lambda N)\}}, \frac{\frac{c^*}{2}r\eta_N NM}{C_1 C_2 \{1 + \rho \kappa_2 \eta_N N \min(r, \kappa_2 \lambda N)\}} \right) \\ &\geq C_3 \frac{c^* r \eta_N NM}{1 + \rho \kappa_2 \eta_N N \min(r, \kappa_2 \lambda N)}, \end{aligned}$$

where $C_3 = \frac{c^*}{C_1^2 C_2 \rho^2}$. Hence for (A.6) we have:

$$\frac{1}{2} P^* \left\{ |Q_1 - EQ_1| > s \right\} \leq \exp \left\{ -C \frac{c^* r \eta_N N M}{1 + \rho \kappa_2 \eta_N N \min(r, \kappa_2 \lambda N)} \right\},$$

where C is a positive constant. Follow Lemma 1, X_m^- is also subgaussian vector. Then we can obtain a same upper bound for

$$\frac{1}{2} P^* \left\{ |Q_2 - EQ_2| > \frac{\log t - E(Q_1 - Q_2)}{2} \right\}$$

in (A.3) through the above procedure. Therefore,

$$P^* \left\{ \frac{P_{ind}(\mathbf{Y} | \mathbf{Z} = \mathbf{z})}{P_{ind}(\mathbf{Y} | \mathbf{Z} = \mathbf{z}^*)} > t \right\} \leq \exp \left\{ -C \frac{c^* r \eta_N N M}{1 + \rho \kappa_2 \eta_N N \min(r, \kappa_2 \lambda N)} \right\}.$$

A.2. Proof of Theorem 5.2. We continue use the notations in the previous proof of Theorem 5.1. First decompose the proposed approximate likelihood in two parts:

$$\begin{aligned} \log \frac{\tilde{L}(\mathbf{Y} | \mathbf{Z} = \mathbf{z})}{\tilde{L}(\mathbf{Y} | \mathbf{Z} = \mathbf{z}^*)} &= \log \frac{P_{ind}(\mathbf{Y} | \mathbf{Z} = \mathbf{z})}{P_{ind}(\mathbf{Y} | \mathbf{Z} = \mathbf{z}^*)} \\ &\quad + \sum_{m=1}^M \log \frac{1 + \sum_{k=1}^K \max \left\{ \sum_{\substack{i < j; u < v \\ (i,j) \neq (u,v)}} z_{ik} z_{jk} z_{uk} z_{vk} \rho_{ij,uv} \hat{Y}_{ij}^{m,k} \hat{Y}_{uv}^{m,k}, 0 \right\}}{1 + \sum_{k=1}^K \max \left\{ \sum_{\substack{i < j; u < v \\ (i,j) \neq (u,v)}} z_{ik}^* z_{jk}^* z_{uk}^* z_{vk}^* \rho_{ij,uv} \hat{Y}_{ij}^{m,k} \hat{Y}_{uv}^{m,k}, 0 \right\}}. \end{aligned}$$

Based on the mean value theorem, we have for some constant C_1 that

$$\begin{aligned} (A.8) \quad &\log \frac{1 + \sum_{k=1}^K \max \left\{ \sum_{\substack{i < j; u < v \\ (i,j) \neq (u,v)}} z_{ik} z_{jk} z_{uk} z_{vk} \rho_{ij,uv} \hat{Y}_{ij}^{m,k} \hat{Y}_{uv}^{m,k}, 0 \right\}}{1 + \sum_{k=1}^K \max \left\{ \sum_{\substack{i < j; u < v \\ (i,j) \neq (u,v)}} z_{ik}^* z_{jk}^* z_{uk}^* z_{vk}^* \rho_{ij,uv} \hat{Y}_{ij}^{m,k} \hat{Y}_{uv}^{m,k}, 0 \right\}} \\ &= C_1 \sum_{k=1}^K \left\{ \max \left(\sum_{\substack{i < j; u < v \\ (i,j) \neq (u,v)}} z_{ik} z_{jk} z_{uk} z_{vk} \rho_{ij,uv} \hat{Y}_{ij}^{m,k} \hat{Y}_{uv}^{m,k}, 0 \right) \right. \\ &\quad \left. - \max \left(\sum_{\substack{i < j; u < v \\ (i,j) \neq (u,v)}} z_{ik}^* z_{jk}^* z_{uk}^* z_{vk}^* \rho_{ij,uv} \hat{Y}_{ij}^{m,k} \hat{Y}_{uv}^{m,k}, 0 \right) \right\} \\ &\leq C_1 \sum_{k=1}^K \left\{ \sum_{\substack{i < j; u < v \\ (i,j) \neq (u,v)}} (z_{ik} z_{jk} z_{uk} z_{vk} - z_{ik}^* z_{jk}^* z_{uk}^* z_{vk}^*) \rho_{ij,uv} \hat{Y}_{ij}^{m,k} \hat{Y}_{uv}^{m,k} \right\}. \end{aligned}$$

Notice in summation (A.8), the terms are non-zero only when $z_{ik} z_{jk} z_{uk} z_{vk} \neq z_{ik}^* z_{jk}^* z_{uk}^* z_{vk}^*$. We denote two node sets

$$\begin{aligned} \xi_1 &= \{(i, j, u, v) | z_{ik} z_{jk} z_{uk} z_{vk} = 1, z_{ik}^* z_{jk}^* z_{uk}^* z_{vk}^* = 0, k = 1, \dots, K\}, \\ \xi_2 &= \{(i, j, u, v) | z_{ik}^* z_{jk}^* z_{uk}^* z_{vk}^* = 1, z_{ik} z_{jk} z_{uk} z_{vk} = 0, k = 1, \dots, K\}, \end{aligned}$$

where $\#|\xi_1| = N_1$ and $\#|\xi_2| = N_2$. Given the number of misclassified nodes $\|z - z^*\|_0 = r$, we have $N_1 = \mathcal{O}(rN^3)$ and $N_2 = \mathcal{O}(rN^3)$. In the following, we construct the augmented edge vectors for the t th sample network by incorporating

the vectorized pairwise edge interaction in (A.8) such that:

$$\tilde{X}_m^+ = \left\{ X_m^+, \underbrace{\left(\sqrt{\frac{C_1}{2} \{ \rho_{ij,uv} \hat{Y}_{ij}^{m,k} \hat{Y}_{uv}^{m,k} \}_+} \right)_{1 \times N_1}}_{\substack{(i,j,u,v) \in \xi_1 \\ z_{ik} z_{jk} z_{uk} z_{vk} = 1 \\ k=1, \dots, K}}, \underbrace{\left(\sqrt{\frac{C_1}{2} \{ -\rho_{ij,uv} \hat{Y}_{ij}^{m,k} \hat{Y}_{uv}^{m,k} \}_+} \right)_{1 \times N_2}}_{\substack{(i,j,u,v) \in \xi_2 \\ z_{ik}^* z_{jk}^* z_{uk}^* z_{vk}^* = 1 \\ k=1, \dots, K}} \right\},$$

$$\tilde{X}_m^- = \left\{ X_m^-, \underbrace{\left(\sqrt{\frac{C_1}{2} \{ \rho_{ij,uv} \hat{Y}_{ij}^{m,k} \hat{Y}_{uv}^{m,k} \}_-} \right)_{1 \times N_1}}_{\substack{(i,j,u,v) \in \xi_1 \\ z_{ik} z_{jk} z_{uk} z_{vk} = 1 \\ k=1, \dots, K}}, \underbrace{\left(\sqrt{\frac{C_1}{2} \{ -\rho_{ij,uv} \hat{Y}_{ij}^{m,k} \hat{Y}_{uv}^{m,k} \}_-} \right)_{1 \times N_2}}_{\substack{(i,j,u,v) \in \xi_2 \\ z_{ik}^* z_{jk}^* z_{uk}^* z_{vk}^* = 1 \\ k=1, \dots, K}} \right\},$$

where X_m^+ and X_m^- are defined in (A.2). Denote the covariance matrix for \tilde{X}_m^+ and \tilde{X}_m^- are $\tilde{\Sigma}_1$ and $\tilde{\Sigma}_2$ respectively. Since the second-order terms in X_m^+ and X_m^- such as $\sqrt{\frac{C_1}{2} \{ \rho_{ij,uv} \hat{Y}_{ij}^{m,k} \hat{Y}_{uv}^{m,k} \}_+}$ only rescale the original edgewise interaction $\hat{Y}_{ij}^{m,k} \hat{Y}_{uv}^{m,k}$ then they preserve the third-order and fourth-order correlation within communities such that

$$|E \left\{ f_1(Y_{i_1 j_1}^m) \sqrt{\frac{C}{2} \{ \rho_{ij,uv} \hat{Y}_{i_2 j_2}^{m,k} \hat{Y}_{i_3 j_3}^{m,k} \}_+} \right\}| = |E(\hat{Y}_{i_1 j_1}^{m,k} \hat{Y}_{i_2 j_2}^{m,k} \hat{Y}_{i_3 j_3}^{m,k})|,$$

$$|E \left\{ f_2(Y_{i_1 j_1}^m) \sqrt{\frac{C}{2} \{ \rho_{ij,uv} \hat{Y}_{i_2 j_2}^{m,k} \hat{Y}_{i_3 j_3}^{m,k} \}_-} \right\}| = |E(\hat{Y}_{i_1 j_1}^{m,k} \hat{Y}_{i_2 j_2}^{m,k} \hat{Y}_{i_3 j_3}^{m,k})|,$$

$$|E \left\{ \sqrt{\frac{C}{2} \{ \rho_{ij,uv} \hat{Y}_{i_1 j_1}^{m,k} \hat{Y}_{i_2 j_2}^{m,k} \}_+} \sqrt{\frac{C}{2} \{ \rho_{ij,uv} \hat{Y}_{i_3 j_3}^{m,k} \hat{Y}_{i_4 j_4}^{m,k} \}_+} \right\}| = |E(\hat{Y}_{i_1 j_1}^{m,k} \hat{Y}_{i_2 j_2}^{m,k} \hat{Y}_{i_3 j_3}^{m,k} \hat{Y}_{i_4 j_4}^{m,k})|,$$

$$|E \left\{ \sqrt{\frac{C}{2} \{ \rho_{ij,uv} \hat{Y}_{i_1 j_1}^{m,k} \hat{Y}_{i_2 j_2}^{m,k} \}_-} \sqrt{\frac{C}{2} \{ \rho_{ij,uv} \hat{Y}_{i_3 j_3}^{m,k} \hat{Y}_{i_4 j_4}^{m,k} \}_-} \right\}| = |E(\hat{Y}_{i_1 j_1}^{m,k} \hat{Y}_{i_2 j_2}^{m,k} \hat{Y}_{i_3 j_3}^{m,k} \hat{Y}_{i_4 j_4}^{m,k})|.$$

Notice that each element in \tilde{X}_m^+ or \tilde{X}_m^- is a bounded binary random variable. Follow the same procedure in Lemma 1, we can show that both \tilde{X}_m^+ and \tilde{X}_m^- are subgaussian random vectors such that $L_1 \leq C_2, L_2 \leq C_2$ for some constant C_2 where L_1, L_2 are subgaussian norm of \tilde{X}_m^+ and \tilde{X}_m^- . Then consider the following quadratic forms:

$$\tilde{Q}_1 = \sum_{m=1}^M \langle \tilde{X}_m^+, \tilde{X}_m^+ \rangle, \tilde{Q}_2 = \sum_{m=1}^M \langle \tilde{X}_m^-, \tilde{X}_m^- \rangle.$$

Therefore, we have

$$\log \frac{\tilde{L}(\mathbf{Y} | \mathbf{Z} = \mathbf{z})}{\tilde{L}(\mathbf{Y} | \mathbf{Z} = \mathbf{z}^*)} \leq \tilde{Q}_1 - \tilde{Q}_2.$$

Denote the centralized version quadratic forms \tilde{Q}_1 and \tilde{Q}_2 as $\tilde{\mathcal{Q}}_1$ and $\tilde{\mathcal{Q}}_2$ such that

$$\tilde{\mathcal{Q}}_1 = \sum_{m=1}^M \langle \tilde{X}_m^+ - E(\tilde{X}_m^+), \tilde{X}_m^+ - E(\tilde{X}_m^+) \rangle, \tilde{\mathcal{Q}}_2 = \sum_{m=1}^M \langle \tilde{X}_m^- - E(\tilde{X}_m^-), \tilde{X}_m^- - E(\tilde{X}_m^-) \rangle.$$

Denote the following quadratic difference as:

$$\Delta(\tilde{Q}_1, \tilde{\mathcal{Q}}_1) := (\tilde{Q}_1 - E(\tilde{Q}_1)) - (\tilde{\mathcal{Q}}_1 - E(\tilde{\mathcal{Q}}_1)) = 2 \sum_{m=1}^M \langle E(\tilde{X}_m^+), \tilde{X}_m^+ - E(\tilde{X}_m^+) \rangle,$$

$$\Delta(\tilde{Q}_2, \tilde{\mathcal{Q}}_2) := (\tilde{Q}_2 - E(\tilde{Q}_2)) - (\tilde{\mathcal{Q}}_2 - E(\tilde{\mathcal{Q}}_2)) = 2 \sum_{m=1}^M \langle E(\tilde{X}_m^-), \tilde{X}_m^- - E(\tilde{X}_m^-) \rangle.$$

Similar to (A.3), for any fixed $t > 0$:

$$\begin{aligned}
(A.9) \quad & P^* \left\{ \frac{\tilde{L}(\mathbf{Y}|\mathbf{Z}=\mathbf{z})}{\tilde{L}(\mathbf{Y}|\mathbf{Z}=\mathbf{z}^*)} > t \right\} \leq P^* \left\{ \tilde{Q}_1 - \tilde{Q}_2 > \log t \right\} \\
& \leq P^* \left\{ \tilde{Q}_1 - E\tilde{Q}_1 > \frac{\log t - E(\tilde{Q}_1 - \tilde{Q}_2)}{2} \right\} + P^* \left\{ \tilde{Q}_2 - E\tilde{Q}_2 < -\frac{\log t - E(\tilde{Q}_1 - \tilde{Q}_2)}{2} \right\} \\
& = P^* \left\{ \tilde{Q}_1 - E\tilde{Q}_1 > \frac{\log t - E(\tilde{Q}_1 - \tilde{Q}_2)}{2} - \Delta(\tilde{Q}_1, \tilde{Q}_1) \right\} \\
& \quad + P^* \left\{ \tilde{Q}_2 - E\tilde{Q}_2 < -\frac{\log t - E(\tilde{Q}_1 - \tilde{Q}_2)}{2} - \Delta(\tilde{Q}_2, \tilde{Q}_2) \right\} \\
& \leq \frac{1}{2} P^* \left\{ |\tilde{Q}_1 - E\tilde{Q}_1| > \frac{\log t - E(\tilde{Q}_1 - \tilde{Q}_2)}{2} - \Delta(\tilde{Q}_1, \tilde{Q}_1) \right\} \\
& \quad + \frac{1}{2} P^* \left\{ |\tilde{Q}_2 - E\tilde{Q}_2| > \frac{\log t - E(\tilde{Q}_1 - \tilde{Q}_2)}{2} - \Delta(\tilde{Q}_2, \tilde{Q}_2) \right\}.
\end{aligned}$$

Next we estimate $\|\tilde{\Sigma}_1\|_F$, $\|\tilde{\Sigma}_1\|_{op}$ and $\|\tilde{\Sigma}_2\|_F$, $\|\tilde{\Sigma}_2\|_{op}$. Denote

$$\tilde{\Lambda} = diag(\Lambda, sd \left(\underbrace{\sqrt{\frac{1}{2} \{ \rho_{ij,uv} \hat{Y}_{ij}^{m,k} \hat{Y}_{uv}^{m,k} \}}_+}_{\substack{(i,j,u,v) \in \xi_1 \\ z_{ik} z_{jk} z_{uk} z_{vk} = 1 \\ k=1, \dots, K}} \right)_{1 \times N_1}, sd \left(\underbrace{\sqrt{\frac{1}{2} \{ -\rho_{ij,uv} \hat{Y}_{ij}^{m,k} \hat{Y}_{uv}^{m,k} \}}_+}_{\substack{(i,j,u,v) \in \xi_2 \\ z_{ik}^* z_{jk}^* z_{uk}^* z_{vk}^* = 1 \\ k=1, \dots, K}} \right)_{1 \times N_2}),$$

then $\|\tilde{\Sigma}_1\|_{op} = \|\tilde{\Lambda} \tilde{R} \tilde{\Lambda}\|_{op} \leq C_3 \|\tilde{R}\|_{op}$ where \tilde{R} is the correlation matrix of \tilde{X}_m^+ and C_3 is the largest variance of elements in \tilde{X}_m^+ . Denote the largest eigenvalue of \tilde{R} as $\lambda_{\tilde{R}}$. From the Gershgorin circle theorem, we have

$$\lambda_{\tilde{R}} \leq 1 + \max_i \sum_{j \neq i} |\tilde{R}_{ij}|.$$

Given that the misclassification number of node $\|z - z^*\|_0 = r$, edgewise correlation density λ and condition C3, for each row in \tilde{R} , there exists some constant $C_4 i_0$ such that:

$$(A.10) \quad \sum_{j \neq i} |R_{ij}| \leq C_4 \rho N_k \min(r, \lambda N_k) = C_4 \rho \kappa_2 N \min(r, \kappa_2 \lambda N),$$

where $\rho = \max_{i,j} \tilde{R}_{ij}$. Therefore, we have

$$\|\tilde{\Sigma}_1\|_{op} \leq C_3 \{1 + C_4 \rho \kappa_2 N \min(r, \kappa_2 \lambda N)\}.$$

Similarly, $\|\tilde{\Sigma}_2\|_{op}$ follows a same upper bound. Notice that the dimension of \tilde{R} is $(N_r + N_1 + N_2) \times (N_r + N_1 + N_2)$. Under the condition C3, in each row of \tilde{R} , the number of non-zero elements is less than $1 + C_4 N_k \min(r, \lambda N_k)$. Therefore, we have for a constant $C' > 0$:

$$\begin{aligned}
\|\tilde{\Sigma}_1\|_F^2 & \leq C_3 \rho^2 (N_r + N_1 + N_2) \{1 + C_4 \kappa_2 N \min(r, \kappa_2 \lambda N)\} \\
& \leq C' \rho^2 (rN + rN^3) \{1 + C_4 \kappa_2 N \min(r, \kappa_2 \lambda N)\}.
\end{aligned}$$

According to the generalized Hanson-Wright inequality in ([16]):

$$(A.11) \quad \frac{1}{2} P^* \left\{ |\tilde{Q}_1 - E\tilde{Q}_1| > s \right\} \leq \exp \left\{ -C \min \left(\frac{s^2}{L_1^4 \|\tilde{\Sigma}_1\|_F^2 \|A\|_F^2}, \frac{s}{L_1^2 \|\tilde{\Sigma}_1\|_{op} \|A\|_{op}} \right) \right\},$$

where $s = \frac{\log t - E(\tilde{Q}_1 - \tilde{Q}_2)}{2} - \Delta(\tilde{Q}_1, \tilde{Q}_1)$, $A = \mathbf{I}_{M \times M}$ and L_1 is subgaussian norm of \tilde{X}_m^+ . Notice $\|A\|_F^2 = M$, $\|A\|_{op} = 1$. Given (A.8), we have

$$\begin{aligned}
E(\tilde{Q}_1 - \tilde{Q}_2) & = E(Q_1 - Q_2) + \\
& C_1 \sum_{m=1}^M \sum_{k=1}^K \left\{ \sum_{\substack{i < j; u < v \\ (i,j) \neq (u,v)}}^N (z_{ik} z_{jk} z_{uk} z_{vk} - z_{ik}^* z_{jk}^* z_{uk}^* z_{vk}^*) \rho_{ij,uv} E(\hat{Y}_{ij}^{m,k} \hat{Y}_{uv}^{m,k}) \right\}.
\end{aligned}$$

Denote ρ_{min} as the lower bound of all non-zero correlation among edges such that $E(\hat{Y}_{ij}^{m,k}\hat{Y}_{uv}^{m,k}) = \rho_{ij,uv} \geq \rho_{min}$. Given the edges from different communities are independent and within-community correlation density λ , we have for some positive constant C_5 ,

$$\#\{(i,j,u,v) : E(\hat{Y}_{ij}^{m,k}\hat{Y}_{uv}^{m,k}) \neq 0, (i,j,u,v) \in \xi_2\} = \lambda N_1 = \lambda C_5 r N^3,$$

$$\#\{(i,j,u,v) : E(\hat{Y}_{ij}^{m,k}\hat{Y}_{uv}^{m,k}) \neq 0, (i,j,u,v) \in \xi_1\} \leq \lambda \binom{r}{4}.$$

Assume that $r \leq cN$ for some constant $0 < c < 1$, we have for some constant $c_0 > 0$:

$$-E(\tilde{Q}_1 - \tilde{Q}_2) \geq \frac{c^*}{2} r NM + \lambda M \frac{C_1 \rho_{min}^2}{2} (C_5 r N^3 - \binom{r}{4}) \geq c_0 r (c^* \eta_N N + \lambda N^3) M.$$

To estimate $\Delta(\tilde{Q}_1, \tilde{Q}_2)$, given all the elements in \tilde{X}_m^+ are bounded, we denote

$$\omega_3 = \max_i E\{(\tilde{X}_m^+)_i\}, \omega_4 = \max_i \text{Var}\{(\tilde{X}_m^+)_i\}.$$

According to the definition of \tilde{X}_m^+ and $N_1 = \mathcal{O}(rN^3), N_2 = \mathcal{O}(rN^3)$, there exists a positive constant C^+ such that $\#\tilde{X}_m^+ = \frac{rN}{2} + C^+ r N^3$, therefore

$$\begin{aligned} (A.12) \quad & P(|\Delta(\tilde{Q}_1, \tilde{Q}_2)| > c_0 r (c^* \eta_N N + \lambda N^3) M) \\ & \leq P(\omega_3 \sum_{m=1}^M \sum_{i=1}^{\#\tilde{X}_m^+} |\tilde{X}_{mi}^+ - E(\tilde{X}_{mi}^+)| > c_0 r (c^* \eta_N N + \lambda N^3) M) \\ & \leq \frac{\omega_3^2 M \text{Var}(\sum_{i=1}^{\#\tilde{X}_m^+} \tilde{X}_{mi}^+)}{c_0^2 r^2 (c^* \eta_N N + \lambda N^3)^2 M^2}. \end{aligned}$$

From the assumption (C3), there exists a positive constant ω_5 such that

$$\begin{aligned} (A.13) \quad & \text{Var}\left(\sum_{i=1}^{\#\tilde{X}_m^+} \tilde{X}_{mi}^+\right) = \sum_{i=1}^{\#\tilde{X}_m^+} \text{Var}(\tilde{X}_{mi}^+) + \sum_{i,j} \text{Cov}(\tilde{X}_{mi}^+, \tilde{X}_{mj}^+) \\ & \leq \omega_4 \left(\frac{rN}{2} + C^+ r N^3 \right) + w_4 \rho \left(\frac{\lambda r^2 N^2}{4} + r N^3 \cdot \omega_5 \lambda N^2 + r N \cdot \omega_5 \lambda N^2 \right). \end{aligned}$$

Through combining (A.12) and (A.13), give $\lambda > 0$ we have

$$\begin{aligned} & P(|\Delta(\tilde{Q}_1, \tilde{Q}_2)| > c_0 r (c^* \eta_N N + \lambda N^3) M) \\ & \leq \frac{\omega_3^2 \omega_4}{c_0^2 M} \left(\frac{1}{2r\lambda^2 N^5} + \frac{C^+}{r\lambda^2 N^3} + \frac{\rho}{4\lambda N^4} + \frac{\rho\omega_5}{r\lambda N} + \frac{\rho\omega_5}{r\lambda N^3} \right). \end{aligned}$$

Therefore, given $\frac{1}{\lambda} < O_N(NM)$ and M, N increasing, s is dominated by $-E(\tilde{Q}_1 - \tilde{Q}_2)$ with probability approaching 1. Given any fixed $t > 0$, $s > \mathcal{O}_N(r(c^* \eta_N N + \lambda N^3) M)$. For the first term in (A.11),

$$\frac{s^2}{L_1^4 \|\tilde{\Sigma}_1\|_F^2 \|A\|_F^2} \geq \frac{r^2 (c^* \eta_N N + \lambda N^3)^2 M^2}{L_1^4 C' \rho^2 (rN + rN^3) \{1 + C_4 \kappa_2 N \min(r, \kappa_2 \lambda N)\} M}.$$

For the second term in (A.11),

$$\frac{s}{L_1^2 \|\tilde{\Sigma}_1\|_{op} \|A\|_{op}} \geq \frac{r(c^* \eta_N N + \lambda N^3) M}{L_1^2 C' \{1 + C_4 \rho \kappa_2 N \min(r, \kappa_2 \lambda N)\}}.$$

Given $\lambda > 0$, we have for some constant $C_6 > 0$

$$(A.14) \quad \min\left(\frac{s^2}{L_1^4 \|\tilde{\Sigma}_1\|_F^2 \|A\|_F^2}, \frac{s}{L_1^2 \|\tilde{\Sigma}_1\|_{op} \|A\|_{op}}\right) \geq C_6 \frac{r \lambda N M (c^* \eta_N + \lambda N^2)}{1 + C_4 \rho \kappa_2 N \min(r, \kappa_2 \lambda N)}.$$

Follow the same procedure we can show a upper bound for $P^*\{|\tilde{Q}_2 - E\tilde{Q}_2| > s\}$ with a same order to (A.14). Combined with (A.9) and (A.11), we have for $\lambda > 0$ and some constant $C > 0$:

$$P_{Z^*} \left\{ \frac{\tilde{L}(\mathbf{Y} | \mathbf{Z} = z; \Theta)}{\tilde{L}(\mathbf{Y} | \mathbf{Z} = z^*; \Theta)} > t \right\} \leq \exp \left\{ -C \frac{r \lambda N M (c^* \eta_N + \lambda N^2)}{1 + C_4 \rho \kappa_2 N \min(r, \kappa_2 \lambda N)} \right\}.$$

A.3. Proof of Theorem 5.3. Part One: we show that given the $\|\boldsymbol{\alpha} - \boldsymbol{z}^*\|_1 = cN^{1-\phi}$, the estimated marginal mean $\hat{\mu}_{ql}$ based on the nodes' memberships $\boldsymbol{\alpha}$ concentrates towards the true μ_{ql} with high probability for $q, l \in \{1, \dots, K\}$. Based on the M -step in Algorithm 1,

$$\begin{aligned} |\hat{\mu}_{ql} - \mu_{ql}| &= \left| \frac{\sum_{m=1}^M \sum_{i \neq j}^N \alpha_{iq} \alpha_{jl} Y_{ij}^m}{\sum_{m=1}^M \sum_{i \neq j}^N \alpha_{iq} \alpha_{jl}} - \mu_{ql} \right| \\ &\leq \left| \frac{\sum_{m=1}^M \sum_{i \neq j}^N \alpha_{iq} \alpha_{jl} (Y_{ij}^m - EY_{ij}^m)}{\sum_{m=1}^M \sum_{i \neq j}^N \alpha_{iq} \alpha_{jl}} \right| + \left| \frac{\sum_{m=1}^M \sum_{i \neq j}^N \alpha_{iq} \alpha_{jl} EY_{ij}^m}{\sum_{m=1}^M \sum_{i \neq j}^N \alpha_{iq} \alpha_{jl}} - \mu_{ql} \right|. \end{aligned}$$

Given $\|\boldsymbol{\alpha} - \boldsymbol{z}^*\| = cN^{1-\phi}$, for each $q = 1, \dots, K$, $\|\boldsymbol{\alpha}_q - \boldsymbol{z}_q^*\|_1 \geq c(N - N^{1-\phi})$ for some constant c . Combined with Condition C2, we have $\sum_{i \neq j}^N \alpha_{iq} \alpha_{jl} \asymp (N - N^{1-\phi})^2$. For the numerator of first term in previous inequality, we chose a $\epsilon = o_M(1)$ such that $\gamma = (M\epsilon)^{-1/2} = o_M(1)$. Notice that $Var(Y_{ij}^m) = O(\eta_N)$ where η_N is the sparsity parameter defined in Section 5.1. Then

$$\begin{aligned} P\left\{ \left| \sum_{m=1}^M \sum_{i \neq j}^N \alpha_{iq} \alpha_{jl} (Y_{ij}^m - EY_{ij}^m) \right| > \gamma M(N - N^{1-\phi})^2 \right\} &\leq \frac{\sum_{m=1}^M Var(\sum_{i \neq j}^N \alpha_{iq} \alpha_{jl} (Y_{ij}^m - EY_{ij}^m))}{\gamma^2 M^2 (N - N^{1-\phi})^4} \\ &= \epsilon O\left\{ \frac{\sum_{m=1}^M \sum_{i \neq j}^N \alpha_{iq}^2 \alpha_{jl}^2 \eta_N + \sum_{m=1}^M \eta_N \sum_{i \neq j, u \neq v} \alpha_{iq} \alpha_{jl} \alpha_{uq} \alpha_{vl} \rho_{ij,uv}}{M(N - N^{1-\phi})^4} \right\} = \epsilon O\left\{ \frac{\eta_N N^2 + \lambda \eta_N N^4}{(N - N^{1-\phi})^4} \right\} < \epsilon. \end{aligned}$$

Therefore, with probability at least $1 - \epsilon$ we have

$$\left| \frac{\sum_{m=1}^M \sum_{i \neq j}^N \alpha_{iq} \alpha_{jl} (Y_{ij}^m - EY_{ij}^m)}{\sum_{m=1}^M \sum_{i \neq j}^N \alpha_{iq} \alpha_{jl}} \right| = O\left\{ \frac{\gamma M(N - N^{1-\phi})^2}{M(N - N^{1-\phi})^2} \right\} = O(\gamma).$$

For the second term, denote $\mathbf{1} = (1, \dots, 1)_{N \times 1}$ we have

$$\begin{aligned} \left| \frac{\sum_{m=1}^M \sum_{i \neq j}^N \alpha_{iq} \alpha_{jl} EY_{ij}^m}{\sum_{m=1}^M \sum_{i \neq j}^N \alpha_{iq} \alpha_{jl}} - \mu_{ql} \right| &= \left| \frac{\sum_{m=1}^M \langle EY^m - \mu_{ql} \mathbf{1} \mathbf{1}^T, \boldsymbol{\alpha}_q \boldsymbol{\alpha}_l^T \rangle}{\sum_{m=1}^M \sum_{i \neq j}^N \alpha_{iq} \alpha_{jl}} \right| \\ &\leq \left| \frac{\sum_{m=1}^M \max_{(q,l),(q',l')} |\mu_{ql} - \mu_{q'l'}| \langle \mathbf{1} \mathbf{1}^T - \boldsymbol{z}_q^* \boldsymbol{z}_l^{*T}, \boldsymbol{\alpha}_q \boldsymbol{\alpha}_l^T - \boldsymbol{z}_q^* \boldsymbol{z}_l^{*T} \rangle}{\sum_{m=1}^M \sum_{i \neq j}^N \alpha_{iq} \alpha_{jl}} \right| \\ &= O\left\{ \frac{\|\boldsymbol{\alpha}_q \boldsymbol{\alpha}_l^T - \boldsymbol{z}_q^* \boldsymbol{z}_l^{*T}\|_1}{(N - N^{1-\phi})^2} \right\} = O\left\{ \frac{\|\boldsymbol{\alpha} - \boldsymbol{z}^*\|_1 (\|\boldsymbol{\alpha}_q\|_1 + \|\boldsymbol{z}_l^*\|_1)}{(N - N^{1-\phi})^2} \right\} = O(N^{-\phi}). \end{aligned}$$

Therefore, given the nodes' memberships satisfying $\|\boldsymbol{\alpha} - \boldsymbol{z}^*\|_1 = cN^{1-\phi}$ we have $\|\hat{\mu}_{ql} - \mu_{ql}\| = o_N(1) + o_M(1)$ with probability at least $1 - \epsilon$.

Part Two: Given the nodes' memberships estimation $\boldsymbol{\alpha}$ satisfying $\|\boldsymbol{\alpha} - \boldsymbol{z}^*\|_1 = cN^{1-\phi}$ and the marginal mean estimation $\hat{\mu}_{ql}$ with $|\hat{\mu}_{ql} - \mu_{ql}| = o_N(1) + o_M(1)$, let \mathbf{E} stands for the operator of E-step in Algorithm 1, we prove that

$$\|\mathbf{E}(\boldsymbol{\alpha}) - \boldsymbol{z}^*\|_1 \leq c_1 N K \exp\left\{ -c_2 \frac{\lambda(c^* \eta_N + \lambda N^2) M}{N} \right\} + \frac{c_3 \|\boldsymbol{\alpha} - \boldsymbol{z}^*\|_1}{(\lambda N M)^{1-\gamma}},$$

with probability at least $1 - \exp(-c\lambda NM) - (\lambda NM)^{-\gamma} - o_M(1)$, where $\gamma \in (0, 1)$, c, c_1, c_2, c_3 are positive constants, and c^* is defined in Theorem 5.1.

Follow the notations introduced in Theorem 5.1 and Theorem 5.2, we further define that $\mathbf{w} = \max \frac{P^{(s)}(Z_i=q)}{P^{(s)}(Z_i=l)}$, $i = 1, \dots, N$, $q, l = 1, \dots, K$. We first consider the misclassification of updated estimated membership for a specific node t , e.g., $\mathbf{E}(z_t)$ from the current estimation $\boldsymbol{\alpha}_t$. We denote that $\boldsymbol{\alpha}_{-t}$ as the estimations of nodes' memberships except node t and denote the true membership of node t as b , i.e., $z_t^* = b$. For the proposed approximate likelihood \tilde{L} , we have

$$\begin{aligned} \|\mathbf{E}(\boldsymbol{\alpha}_t) - z_t^*\|_1 &= \\ &\left| \frac{P(z_t=1)\tilde{L}(\mathbf{Y}|\boldsymbol{\alpha}_{-t}; z_t=1)}{\sum_{q=1}^K P(z_t=q)\tilde{L}(\mathbf{Y}|\boldsymbol{\alpha}_{-t}; z_t=q)} - 0 \right| + \dots + \left| \frac{P(z_t=b)\tilde{L}(\mathbf{Y}|\boldsymbol{\alpha}_{-t}; z_t=b)}{\sum_{q=1}^K P(z_t=q)\tilde{L}(\mathbf{Y}|\boldsymbol{\alpha}_{-t}; z_t=q)} - 1 \right| \end{aligned}$$

$$\begin{aligned}
&\leq 2 \frac{\sum_{q \neq b} P(z_t = q) \tilde{L}(\mathbf{Y} | \boldsymbol{\alpha}_{-t}; z_t = q)}{\sum_{q=1}^K P(z_t = q) \tilde{L}(\mathbf{Y} | \boldsymbol{\alpha}_{-t}; z_t = q)} \leq 2w \sum_{q \neq b}^K \frac{\tilde{L}(\mathbf{Y} | \boldsymbol{\alpha}_{-t}; z_t = q)}{\tilde{L}(\mathbf{Y} | \boldsymbol{\alpha}_{-t}; z_t = b)} \\
(A.15) \quad &= 2w \sum_{q \neq b}^K \min[1, \exp\{\log \tilde{L}(\mathbf{Y} | \boldsymbol{\alpha}_{-t}; z_t = q) - \log \tilde{L}(\mathbf{Y} | \boldsymbol{\alpha}_{-t}; z_t = b)\}].
\end{aligned}$$

Then given that node t belongs to different communities while the estimated membership for other nodes $\boldsymbol{\alpha}_{-t}$ are fixed, we decompose the proposed approximate likelihood into marginal part and correlation part in the following: $\log \tilde{L}(\mathbf{Y} | \boldsymbol{\alpha}_{-t}; z_t) = \log L_{mar}(\mathbf{Y} | \boldsymbol{\alpha}_{-t}; z_t) + \log L_{cor}(\mathbf{Y} | \boldsymbol{\alpha}_{-t}; z_t)$. The marginal likelihood $\log L_{mar}(\mathbf{Y} | \boldsymbol{\alpha}_{-t}; z_t)$ can be expanded as:

$$\begin{aligned}
&\log L_{mar}(\mathbf{Y} | \boldsymbol{\alpha}_{-t}; z_t = a) \\
&= \sum_{m=1}^M \underbrace{\left[\log \prod_{q,l}^K \prod_{i \neq j \neq t}^N \left\{ \mu_{ql}^{Y_{ij}^m} (1 - \mu_{ql})^{(1-Y_{ij}^m)} \right\}^{\alpha_{iq} \alpha_{jl}} + \prod_{q=1}^K \prod_{i \neq t}^N \left\{ \mu_{qa}^{Y_{it}^m} (1 - \mu_{qa})^{(1-Y_{it}^m)} \right\}^{\alpha_{iq}} \right]}_{\text{not depend on } z_t}.
\end{aligned}$$

Therefore, the discrepancy among marginal likelihood is

$$\begin{aligned}
&\log L_{mar}(\mathbf{Y} | \boldsymbol{\alpha}_{-t}; z_t = a) - \log L_{mar}(\mathbf{Y} | \boldsymbol{\alpha}_{-t}; z_t = b) \\
&= \sum_{m=1}^M \sum_{q=1}^K \sum_{i \neq t}^N \left[\alpha_{iq} \{ Y_{it}^m \log \frac{\hat{\mu}_{qa}}{\hat{\mu}_{qb}} + (1 - Y_{it}^m) \log \frac{1 - \hat{\mu}_{qa}}{1 - \hat{\mu}_{qb}} \} \right] \\
&= \sum_{m=1}^M \sum_{q=1}^K \sum_{i \neq t}^N \left[\alpha_{iq} \{ Y_{it}^m \log \frac{\mu_{qa}}{\mu_{qb}} + (1 - Y_{it}^m) \log \frac{1 - \mu_{qa}}{1 - \mu_{qb}} \} \right] \\
&+ \sum_{m=1}^M \sum_{q=1}^K \sum_{i \neq t}^N \left[\alpha_{iq} \{ Y_{it}^m \log \frac{\mu_{qa} \hat{\mu}_{qb}}{\hat{\mu}_{qa} \mu_{qb}} + (1 - Y_{it}^m) \log \frac{(1 - \mu_{qa})(1 - \hat{\mu}_{qb})}{(1 - \hat{\mu}_{qa})(1 - \hat{\mu}_{qb})} \} \right].
\end{aligned}$$

We can decompose the marginal discrepancy into four parts:

$$\begin{aligned}
&\log L_{mar}(\mathbf{Y} | \boldsymbol{\alpha}_{-t}; z_t = a) - \log L_{mar}(\mathbf{Y} | \boldsymbol{\alpha}_{-t}; z_t = b) \\
&= \underbrace{\sum_{m=1}^M \sum_{q=1}^K \sum_{i \neq t}^N (\alpha_{iq} - z_{iq}^*) \{ Y_{it}^m - E(Y_{it}^m) \} (\log \frac{\mu_{qa}}{\mu_{qb}} - \log \frac{1 - \mu_{qa}}{1 - \mu_{qb}})}_{A_1} \\
&+ \underbrace{\sum_{m=1}^M \sum_{q=1}^K \sum_{i \neq t}^N \left[(\alpha_{iq} - z_{iq}^*) \{ EY_{it}^m \log \frac{\mu_{qa}}{\mu_{qb}} + (1 - EY_{it}^m) \log \frac{1 - \mu_{qa}}{1 - \mu_{qb}} \} \right]}_{A_2} \\
&+ \underbrace{\sum_{m=1}^M \sum_{q=1}^K \sum_{i \neq t}^N \left[z_{iq}^* \{ Y_{it}^m \log \frac{\mu_{qa}}{\mu_{qb}} + (1 - Y_{it}^m) \log \frac{1 - \mu_{qa}}{1 - \mu_{qb}} \} \right]}_{A_3} \\
&+ \underbrace{\sum_{m=1}^M \sum_{q=1}^K \sum_{i \neq t}^N \left[\alpha_{iq} \{ Y_{it}^m \log \frac{\mu_{qa} \hat{\mu}_{qb}}{\hat{\mu}_{qa} \mu_{qb}} + (1 - Y_{it}^m) \log \frac{(1 - \mu_{qa})(1 - \hat{\mu}_{qb})}{(1 - \hat{\mu}_{qa})(1 - \hat{\mu}_{qb})} \} \right]}_{A_4}.
\end{aligned}$$

For the correlation part, we consider the pairwise interaction terms in the $\log L_{cor}(\mathbf{Y} | \boldsymbol{\alpha})$. Notice that for any within-community pairwise correlation $|\hat{\rho}_{ij,kg} - \rho_{ij,kg}| = o_M(1)$ with probability at least $1 - o_M(1)$ then we replace $\hat{\rho}_{ij,kg}$ within

L_{cor} in Algorithm 1 by the true $\rho_{ij,kg}$ in the following proof and the results hold with probability at least $1 - o_M(1)$. For $m = 1, \dots, M$ and consider the node t

$$\sum_{\substack{i < j; k < g \\ (i,j) \neq (k,g)}}^N \alpha_{ia} \alpha_{ja} \alpha_{ka} \alpha_{ga} \rho_{ij,kg} \hat{Y}_{ij}^{m,a} \hat{Y}_{kg}^{m,a} = \alpha_{ta} \sum_{i \neq t}^N \alpha_{ia} \sum_{k < g}^N \alpha_{ka} \alpha_{ga} \rho_{it,kg} \hat{Y}_{it}^{m,a} \hat{Y}_{kg}^{m,a} + A_a^m,$$

where A_q^m does not depend on the membership of node t . We define the vectorized pairwise interactions of edges associating with node t as $\mathcal{Y}_t^{m,a} = \{\rho_{it,kg} \hat{Y}_{it}^{m,a} \hat{Y}_{kg}^{m,a} | i \neq t; k < g\}$. For the correlation part $\log L_{cor}(\mathbf{Y}|\boldsymbol{\alpha}_{-t}; z_t)$, if $\alpha_{sq} = 0$, $q \neq a$ and $\alpha_{sa} = 1$:

$$\begin{aligned} \log L_{cor}(\mathbf{Y}|\boldsymbol{\alpha}_{-t}; z_t = a) &= \sum_{m=1}^M \left\{ 1 + \sum_{q=1}^K \max \left(\sum_{\substack{i < j; k < g \\ (i,j) \neq (k,g)}}^N \alpha_{iq} \alpha_{jq} \alpha_{kq} \alpha_{gq} \rho_{ij,kg} \hat{Y}_{ij}^{t,q} \hat{Y}_{kg}^{t,q}, 0 \right) \right\} \\ &= 1 + \underbrace{\sum_{m=1}^M \sum_{q=1}^K A_q^m}_{\mathbf{A}} + \underbrace{\sum_{m=1}^M \sum_{i \neq t}^N \alpha_{ia} \sum_{k < g}^N \alpha_{ka} \alpha_{ga} \rho_{it,kg} \hat{Y}_{it}^{m,a} \hat{Y}_{kg}^{m,a}}_{\mathbf{B}_a}. \end{aligned}$$

Through the Taylor expansion, the discrepancy of correlation information when node t belongs to different communities a and b :

$$\begin{aligned} \log L_{cor}(\mathbf{Y}|\boldsymbol{\alpha}_{-t}; z_t = a) - \log L_{cor}(\mathbf{Y}|\boldsymbol{\alpha}_{-t}; z_t = b) &= \log(1 + \mathbf{A} + \mathbf{B}_a) - \log(1 + \mathbf{A} + \mathbf{B}_b) \\ &= \log\left(1 + \frac{\mathbf{B}_a - \mathbf{B}_b}{1 + \mathbf{A} + \mathbf{B}_b}\right) \leq C_A (\mathbf{B}_a - \mathbf{B}_b), \end{aligned}$$

where C_A is a constant relating to the gradient of function $\log(1 + 1/x)$ at \mathbf{A} . Then we let $\rho = \min \rho_{ij,kg}$ and $\alpha_a = (z_{1a}, z_{2a}, \dots, z_{Na})$, $a \in \{1, \dots, K\}$

$$\begin{aligned} \mathbf{B}_a - \mathbf{B}_b &= \sum_{m=1}^M \sum_{i \neq t}^N \alpha_{ia} \sum_{k < g}^N \alpha_{ka} \alpha_{ga} \rho_{it,kg} \hat{Y}_{it}^{m,a} \hat{Y}_{kg}^{m,a} - \sum_{m=1}^M \sum_{i \neq t}^N \alpha_{ib} \sum_{k < g}^N \alpha_{kb} \alpha_{gb} \rho_{it,kg} \hat{Y}_{it}^{m,b} \hat{Y}_{kg}^{m,b} \\ &\leq \sum_{m=1}^M \left(\langle \alpha_a \otimes \text{vec}(\alpha_a^T \alpha_a), \mathcal{Y}_t^{m,a} \rangle - \langle \alpha_b \otimes \text{vec}(\alpha_b^T \alpha_b), \mathcal{Y}_t^{m,b} \rangle \right). \end{aligned}$$

For the simplicity of notation, we define and decompose the correlation discrepancy as followings:

$$\begin{aligned} \mathbf{B} &:= \sum_{m=1}^M C_A \left(\langle \alpha_a \otimes \text{vec}(\alpha_a^T \alpha_a), \mathcal{Y}_t^{m,a} \rangle - \langle \alpha_b \otimes \text{vec}(\alpha_b^T \alpha_b), \mathcal{Y}_t^{m,b} \rangle \right) \\ &= C_A \underbrace{\sum_{m=1}^M \left(\langle \alpha_a \otimes \text{vec}(\alpha_a^T \alpha_a) - z_a^* \otimes \text{vec}(z_a^{*T} z_a^*), \mathcal{Y}_t^{m,a} \rangle - \langle \alpha_b \otimes \text{vec}(\alpha_b^T \alpha_b) - z_b^* \otimes \text{vec}(z_b^{*T} z_b^*), \mathcal{Y}_t^{m,b} \rangle \right)}_{\text{misclassification error: } \mathbf{B}_1} \\ &\quad + C_A \underbrace{\sum_{m=1}^M \left(\langle z_a^* \otimes \text{vec}(z_a^{*T} z_a^*), \mathcal{Y}_t^{m,a} \rangle - \langle z_b^* \otimes \text{vec}(z_b^{*T} z_b^*), \mathcal{Y}_t^{m,b} \rangle \right)}_{\text{estimation bias: } \mathbf{B}_2}. \end{aligned}$$

Notice that $\min\{1, \exp(x)\} \leq \exp(x_0) + \sum_{l=0}^{m-1} \frac{1-\exp(x_0)}{m} \mathbb{1}\{x \geq (1-l/m)x_0\}$ and set $x_0 = -\alpha' MN$, where $\alpha' = \frac{\lambda(c^* \eta_N + \lambda N^2)}{1+\lambda N^2}$. Combined with (A.15) we have:

$$(A.16) \quad \|\boldsymbol{\alpha} - \mathbf{z}^*\|_1 \leq 2wNK \exp(-\alpha' MN) + 2w \sum_{l=0}^{m-1} \sum_{a=1}^K \sum_{b \neq a} \sum_{i: z_i^* = b} \frac{1 - \exp(-\alpha' MN)}{m} \mathbf{L},$$

where $\mathbf{L} = \mathbb{1}\{\mathbf{A} + \mathbf{B} \geq \frac{m-l}{m}x_0\}$. For some specific $T > 0$,

$$\begin{aligned} \mathbb{1}\left(\mathbf{A} + \mathbf{B} \geq \frac{m-l}{m}x_0\right) &= \mathbb{1}\left(\mathbf{A}_1 + \mathbf{A}_2 + \mathbf{A}_3 + \mathbf{A}_4 + \mathbf{B}_1 + \mathbf{B}_2 \geq \frac{m-l}{m}x_0\right) \\ (A.17) \quad &\leq \mathbb{1}\left(\mathbf{A}_1 + \mathbf{B}_1 \geq T\right) + \mathbb{1}\left(\mathbf{A}_3 + \mathbf{B}_2 \geq \frac{m-l}{m}x_0 - T - \mathbf{A}_2 - \mathbf{A}_4\right). \end{aligned}$$

We then transfer $\mathbf{A}_3 + \mathbf{B}_2$ into a quadratic form. For each community $q, q = 1, \dots, K$ and each sample network $m = 1, \dots, M$ define the transformations:

$$\begin{aligned} f_q^+(x) &= \sqrt{\left[z_{iq}^* \{Y_{it}^m \log \frac{\mu_{qa}}{\mu_{qb}} + (1 - Y_{it}^m) \log \frac{1 - \mu_{qa}}{1 - \mu_{qb}}\}\right]_+}, \\ f_q^-(x) &= \sqrt{\left[z_{iq}^* \{Y_{it}^m \log \frac{\mu_{qa}}{\mu_{qb}} + (1 - Y_{it}^m) \log \frac{1 - \mu_{qa}}{1 - \mu_{qb}}\}\right]_-}, \end{aligned}$$

$$X_m^+ = \{f_1^+(Y_{1t}^m), \dots, f_1^+(Y_{nt}^m), f_2^+(Y_{1t}^m), \dots, f_2^+(Y_{Nt}^m), \dots, f_K^+(Y_{1t}^m), \dots, f_K^+(Y_{Nt}^m)\},$$

$$X_m^- = \{f_1^-(Y_{1t}^m), \dots, f_1^-(Y_{Nt}^m), f_2^-(Y_{1t}^m), \dots, f_2^-(Y_{Nt}^m), \dots, f_K^-(Y_{1t}^m), \dots, f_K^-(Y_{Nt}^m)\}.$$

Notice that the total number of non-zero terms in X_m^+ or X_m^- is N . We define the node sets

$$\tilde{\xi}_a = \{(i_1, i_2, i_3) | z_{i_1 a}^* z_{i_2 a}^* z_{i_3 a}^* = 1\} \quad \tilde{\xi}_b = \{(i_1, i_2, i_3) | z_{i_1 b}^* z_{i_2 b}^* z_{i_3 b}^* = 1\}.$$

Note $\#|\tilde{\xi}_a| = o(N_a^3)$ and $\#|\tilde{\xi}_b| = o(N_b^3)$ where N_a and N_b are number of node in community a and b . We further define augmented edges vectors:

$$\begin{aligned} \bar{X}_m^+ &= \left(X_m^+, \underbrace{\left(\frac{C_A}{4} \sqrt{\{\rho_{i_1 t, i_2 i_3} \hat{Y}_{i_1 t}^{m,a} \hat{Y}_{i_2 i_3}^{m,a}\}}_+ \right)_{1 \times \#\tilde{\xi}_a}}_{(i_1, i_2, i_3) \in \tilde{\xi}_a}, \underbrace{\left(\frac{C_A}{4} \sqrt{\{-\rho_{i_1 t, i_2 i_3} \hat{Y}_{i_1 t}^{m,b} \hat{Y}_{i_2 i_3}^{m,b}\}}_+ \right)_{1 \times \#\tilde{\xi}_b}}_{(i_1, i_2, i_3) \in \tilde{\xi}_b} \right), \\ \bar{X}_m^- &= \left(X_m^-, \underbrace{\left(\frac{C_A}{4} \sqrt{\{\rho_{i_1 t, i_2 i_3} \hat{Y}_{i_1 t}^{m,a} \hat{Y}_{i_2 i_3}^{m,a}\}}_- \right)_{1 \times \#\tilde{\xi}_a}}_{(i_1, i_2, i_3) \in \tilde{\xi}_a}, \underbrace{\left(\frac{C_A}{4} \sqrt{\{-\rho_{i_1 t, i_2 i_3} \hat{Y}_{i_1 t}^{m,b} \hat{Y}_{i_2 i_3}^{m,b}\}}_- \right)_{1 \times \#\tilde{\xi}_b}}_{(i_1, i_2, i_3) \in \tilde{\xi}_b} \right). \end{aligned}$$

Denote the covariance of \bar{X}_m^+ and \bar{X}_m^- as $\bar{\Sigma}_1$ and $\bar{\Sigma}_2$. Note that each element in \bar{X}_m^+ or \bar{X}_m^- is a bounded binary random variable. Similarly, \bar{X}_m^+ and \bar{X}_m^- are subgaussian vectors. Therefore,

$$\mathbf{A}_3 + \mathbf{B}_2 = \sum_{m=1}^M (\langle \bar{X}_m^+, \bar{X}_m^+ \rangle - \langle \bar{X}_m^-, \bar{X}_m^- \rangle) = \bar{Q}_1 - \bar{Q}_2,$$

$$E(\mathbf{A}_3 + \mathbf{B}_2) = E\bar{Q}_1 - E\bar{Q}_2.$$

Denote $s = \frac{m-l}{m}x_0 - T - \mathbf{A}_2 - \mathbf{A}_4 - E(\mathbf{A}_3 + \mathbf{B}_2)$, we estimate $E(\mathbf{A}_3 + \mathbf{B}_2)$, \mathbf{A}_2 and \mathbf{A}_4 in the following. Given $z_t^* = b$ and the result in (A.7), we have for some constant $c > 0$ and $q = 1, \dots, K$:

$$E \left[\{Y_{it}^m \log \frac{\mu_{qa}}{\mu_{qb}} + (1 - Y_{it}^m) \log \frac{1 - \mu_{qa}}{1 - \mu_{qb}}\} \right] = \mu_{qb} \log \frac{\mu_{qa}}{\mu_{qb}} + (1 - \mu_{qb}) \log \frac{1 - \mu_{qa}}{1 - \mu_{qb}} < -c < 0.$$

Then

$$E \mathbf{A}_3 = \sum_{m=1}^M \sum_{q=1}^K \sum_{i \neq t}^N \left[z_{iq}^* \{\mu_{qb} \log \frac{\mu_{qa}}{\mu_{qb}} + (1 - \mu_{qb}) \log \frac{1 - \mu_{qa}}{1 - \mu_{qb}}\} \right] < -c^* \eta_N (N-1) M.$$

Given that edges from different communities are independent and correlation density λ , there exists a constant $C > 0$ such that

$$\begin{aligned} E \mathbf{B}_2 &= \sum_{m=1}^M C_A \left[\langle \alpha_a \otimes \text{vec}(\alpha_a^T \alpha_a), E\{\mathcal{Y}_t^{m,a}\} \rangle - \langle \alpha_b \otimes \text{vec}(\alpha_b^T \alpha_b), E\{\mathcal{Y}_t^{m,b}\} \rangle \right] \\ &= - \sum_{m=1}^M C_A \langle \alpha_b \otimes \text{vec}(\alpha_b^T \alpha_b), E\{\mathcal{Y}_t^{m,b}\} \rangle \leq -C \lambda N_b^3. \end{aligned}$$

Therefore, $-E(\mathbf{A}_3 + \mathbf{B}_2) \geq c'(c^*\eta_N N + \lambda N^3)M$ for some positive constant c' . Based on condition C1 that $\mu_{ql}, q, l = 1, \dots, K$ are bounded and condition C4, it can be shown that $|EY_{it}^m \log \frac{\mu_{qa}}{\mu_{qb}} + (1 - EY_{it}^m) \log \frac{1 - \mu_{qa}}{1 - \mu_{qb}}|$ is bounded then $|\mathbf{A}_2| = \mathcal{O}(N^{1-\phi}M)$. Based on that $|\hat{\mu}_{ql} - \mu_{ql}| = o_N(1) + o_M(1)$, we have

$$\log \frac{\mu_{qa}\hat{\mu}_{qb}}{\hat{\mu}_{qa}\mu_{qb}} = o_N(1) + o_M(1), \log \frac{(1 - \mu_{qa})(1 - \hat{\mu}_{qb})}{(1 - \hat{\mu}_{qa})(1 - \hat{\mu}_{qb})} = o_N(1) + o_M(1).$$

Then we have

$$\begin{aligned} |\mathbf{A}_4| &= \left| \sum_{m=1}^M \sum_{q=1}^K \sum_{i \neq t}^N \left[\alpha_{iq} \{ Y_{it}^m \log \frac{\mu_{qa}\hat{\mu}_{qb}}{\hat{\mu}_{qa}\mu_{qb}} + (1 - Y_{it}^m) \log \frac{(1 - \mu_{qa})(1 - \hat{\mu}_{qb})}{(1 - \hat{\mu}_{qa})(1 - \hat{\mu}_{qb})} \} \right] \right| \\ &= No_M(M) + Mo_N(N). \end{aligned}$$

Therefore we have $|\mathbf{A}_2 + \mathbf{A}_4| = o_{N,M}(NM)$. We choose $T = -\frac{E(\mathbf{A}_3 + \mathbf{B}_2)}{2}$ and $x_0 = -\alpha' MN$ where $\alpha' > 0$. Then $\frac{m-l}{m}x_0 = o_N\{E(\mathbf{A}_3 + \mathbf{B}_2)\}$. Therefore, $E(\mathbf{A}_3 + \mathbf{B}_2)$ is dominant term in s such that $s \geq -C'\lambda N^3$ where $C' > 0$ is a constant. Follow a similar discussion in (A.10) and condition C3, we have the upper bound for $\|\bar{\Sigma}_1\|_{op}$:

$$\|\bar{\Sigma}_1\|_{op} \leq c_0(1 + c_1\lambda N^2).$$

In addition, from $\#|X_m^+| = N$, $\#|\bar{\xi}_a| = o(N_a^3)$, $\#|\bar{\xi}_b| = o(N_b^3)$ and condition C3, we have the upper bound for $\|\bar{\Sigma}_1\|_F^2$:

$$\|\bar{\Sigma}_1\|_F^2 \leq C_1 N(1 + c_1\lambda N^2) + C_2 N^3(1 + c_2\lambda N^2),$$

where C_1, C_2, c_1, c_2 are constants. Then we estimate the upper bound for the second term in (A.17) following the similar decentralized quadratic decomposition in Theorem 5.1 and Theorem 5.2:

$$\begin{aligned} \mathbb{P}(\mathbf{A}_3 + \mathbf{B}_2 \geq \frac{m-l}{m}x_0 - t - \mathbf{A}_2 - \mathbf{A}_4) &= \mathbb{P}\left\{(\bar{Q}_1 - E\bar{Q}_1) - (\bar{Q}_2 - E\bar{Q}_2) > Ms\right\} \\ &\leq \frac{1}{2}\mathbb{P}\left\{|\bar{Q}_1 - E\bar{Q}_1| > \frac{Ms}{2}\right\} + \frac{1}{2}\mathbb{P}\left\{|\bar{Q}_2 - E\bar{Q}_2| > \frac{Ms}{2}\right\}. \end{aligned}$$

According to the generalized Hanson-Wright inequality in ([16]):

$$(A.18) \quad \frac{1}{2}\mathbb{P}\left\{|\bar{Q}_1 - E\bar{Q}_1| > s\right\} \leq \exp\left\{-C \min\left(\frac{s^2 M^2}{\bar{L}_1^4 \|\bar{\Sigma}_1\|_F^2 \|A\|_F^2}, \frac{sM}{\bar{L}_1^2 \|\bar{\Sigma}_1\|_{op} \|A\|_{op}}\right)\right\},$$

where $A = \mathbf{I}_{M \times M}$ and \bar{L}_1 is subgaussian norm of \bar{X}_m^+ . Notice that

$$\begin{aligned} \frac{s^2 M^2}{\bar{L}_1^4 \|\bar{\Sigma}_1\|_F^2 \|A\|_F^2} &\geq \frac{(C'\lambda N^3)^2 M^2}{\bar{L}_1^4 \{C_1 N(1 + c_1\lambda N^2) + C_2 N^3(1 + c_2\lambda N^2)\} M}, \\ \frac{sM}{\bar{L}_1^2 \|\bar{\Sigma}_1\|_{op} \|A\|_{op}} &\geq \frac{C'\lambda N^3 M}{\bar{L}_1^2 c_0(1 + c_3\lambda N^2)}. \end{aligned}$$

Given $\lambda N^{\frac{\phi}{2}} > 1$, we have for some constant $C^* > 0$

$$C \min\left(\frac{s^2 M^2}{\bar{L}_1^4 \|\bar{\Sigma}_1\|_F^2 \|A\|_F^2}, \frac{sM}{\bar{L}_1^2 \|\bar{\Sigma}_1\|_{op} \|A\|_{op}}\right) \geq C^* \alpha' MN.$$

The upper bound for $\mathbb{P}\left\{|\bar{Q}_2 - E\bar{Q}_2| > \frac{Ms}{2}\right\}$ can be similarly obtained. Therefore, there exist constant C'

$$\begin{aligned} &E\left\{ \sum_{a=1}^K \sum_{b \neq a} \sum_{i:z_i^*=b} \mathbb{1} \left(\mathbf{A}_3 + \mathbf{B}_2 \geq \frac{m-l}{m}x_0 - T - \mathbf{A}_2 - \mathbf{A}_4 \right) \right\} \\ &= \sum_{a=1}^K \sum_{b \neq a} \sum_{i:z_i^*=b} \mathbb{P} \left(\mathbf{A}_3 + \mathbf{B}_2 \geq \frac{m-l}{m}x_0 - T - \mathbf{A}_2 - \mathbf{A}_4 \right) \\ &= O\{KN \exp(-C^* \alpha' MN)\} = O\{\exp(-C' \alpha' MN)\}. \end{aligned}$$

By applying Markov inequality, we have

$$\mathbb{P}\left\{\sum_{a=1}^K \sum_{b \neq a} \sum_{i:z_i^*=b} \mathbb{1}\left(\mathbf{A}_3 + \mathbf{B}_2 \geq \frac{m-l}{m}x_0 - T - \mathbf{A}_2 - \mathbf{A}_4\right) \geq \exp(-C'\alpha' MN/2)\right\} \leq \exp(-C'\alpha' MN/2).$$

Therefore, with probability at least $1 - \exp(-C'\alpha' MN/2)$,

$$(A.19) \quad \sum_{a=1}^K \sum_{b \neq a} \sum_{i:z_i^*=b} \mathbb{1}\left(\mathbf{A}_3 + \mathbf{B}_2 \geq \frac{m-l}{m}x_0 - T - \mathbf{A}_2 - \mathbf{A}_4\right) \leq \exp(-C'\alpha' MN/2).$$

Next, we estimate the term $\mathbb{P}(\mathbf{A}_1 + \mathbf{B}_1 \geq t)$. Notice

$$\begin{aligned} E(\mathbf{A}_1) &= E\left[\sum_{m=1}^M \sum_{q=1}^K \sum_{i \neq t} (\alpha_{iq} - z_{iq}^*) \{Y_{it}^m - E(Y_{it}^m)\} \left(\log \frac{\mu_{qa}}{\mu_{qb}} - \log \frac{1 - \mu_{qa}}{1 - \mu_{qb}}\right)\right] = 0, \\ E(\mathbf{B}_1) &= C_A \sum_{m=1}^M \left[\langle \alpha_a \otimes \text{vec}(\alpha_a^T \alpha_a) - z_a^* \otimes \text{vec}(z_a^{*T} z_a^*), E\{\mathcal{Y}_t^{m,a}\} \rangle - \right. \\ &\quad \left. \langle \alpha_b \otimes \text{vec}(\alpha_b^T \alpha_b) - z_b^* \otimes \text{vec}(z_b^{*T} z_b^*), E\{\mathcal{Y}_t^{m,b}\} \rangle \right]. \end{aligned}$$

Given condition C4 such that $\|\alpha - z^*\|_1 = cN^{1-\phi}$, $0 < \phi < 1$,

$$\begin{aligned} \mathbf{B}_1 &= C_A \sum_{m=1}^M \left\{ \langle \alpha_a \otimes \text{vec}(\alpha_a^T \alpha_a) - z_a^* \otimes \text{vec}(z_a^{*T} z_a^*), \mathcal{Y}_t^{m,a} \rangle \right. \\ &\quad \left. - \langle \alpha_b \otimes \text{vec}(\alpha_b^T \alpha_b) - z_b^* \otimes \text{vec}(z_b^{*T} z_b^*), \mathcal{Y}_t^{m,b} \rangle \right\}. \end{aligned}$$

Notice that for any community $a = 1, \dots, K$,

$$\begin{aligned} \|(\text{vec}(\alpha_a^T \alpha_a) - \text{vec}(z_a^{*T} z_a^*))\|_2 &\leq \|\alpha_a \otimes (\alpha_a - z_a^*)\|_2 + \|(\alpha_a - z_a^*) \otimes z_a^*\|_2 \\ &\leq \|\alpha_a\|_2 \|(\alpha_a - z_a^*)\|_2 + \|(\alpha_a - z_a^*)\|_2 \|z_a^*\|_2, \\ \|E(\mathcal{Y}_t^{m,a})\|_2 &\leq \frac{N^{3/2}}{\hat{\mu}_{aa}(1 - \hat{\mu}_{aa})}. \end{aligned}$$

Therefore, we have

$$\begin{aligned} &\langle \alpha_a \otimes \text{vec}(\alpha_a^T \alpha_a) - z_a^* \otimes \text{vec}(z_a^{*T} z_a^*), E\{\mathcal{Y}_t^{m,a}\} \rangle \\ &\leq \|\alpha_a \otimes \text{vec}(\alpha_a^T \alpha_a) - z_a^* \otimes \text{vec}(z_a^{*T} z_a^*)\|_2 \|E\{\mathcal{Y}_t^{m,a}\}\|_2 \\ &\leq (\|\alpha_a \otimes \text{vec}(\alpha_a^T \alpha_a) - \text{vec}(z_a^{*T} z_a^*)\|_2 + \|(\alpha_a - z_a^*) \otimes \text{vec}(z_a^{*T} z_a^*)\|_2) \|E\{\mathcal{Y}_t^{m,a}\}\|_2 \\ &\leq \|\alpha_a - z_a^*\|_2 \cdot (\|\alpha_a\|_2^2 + \|z_a^*\|_2^2 + \|\alpha_a\|_2 \|z_a^*\|_2) \cdot \|E(\hat{\mathcal{Y}}_t^{m,a})\|_2 \cdot \|E(\hat{\mathbf{Y}}^{m,a})\|_2 \\ &\leq \frac{3N * N^{3/2}}{\hat{\mu}_{aa}(1 - \hat{\mu}_{aa})} \|\alpha_a - z_a^*\|_2. \end{aligned}$$

Since $\|\alpha_a - z_a^*\|_2 = \sqrt{\|\alpha_a - z_a^*\|_2^2} \leq \sqrt{\|\alpha - z^*\|_1}$ for any $a = 1, \dots, K$, then based on condition C4 for some constant $C > 0$,

$$|E(\mathbf{B}_1)| \leq CN^{3-\frac{\phi}{2}}.$$

We define edge vectors \tilde{Y}_m , $m = 1, \dots, M$ and membership vector $\theta_{a,b}$ as:

$$\begin{aligned} \tilde{Y}_m &= \left\{ \underbrace{Y_{\cdot t}^m - E(Y_{\cdot t}^m), \dots, Y_{\cdot t}^m - E(Y_{\cdot t}^m)}_{NK}, \mathcal{Y}_t^{m,a} \mathcal{Y}_t^{m,b} \right\}, \\ \theta_{a,b} &= \left[\underbrace{(\alpha_{iq} - z_{iq}^*) \left(\log \frac{\mu_{qa}}{\mu_{qb}} - \log \frac{1 - \mu_{qa}}{1 - \mu_{qb}} \right), \dots, (\alpha_{iK} - z_{iK}^*) \left(\log \frac{\mu_{Ka}}{\mu_{Kb}} - \log \frac{1 - \mu_{Ka}}{1 - \mu_{Kb}} \right)}_{i=1 \dots, N} \right] \end{aligned}$$

$$C_A\{\alpha_a \otimes \text{vec}(\alpha_a^T \alpha_a) - z_a^* \otimes \text{vec}(z_a^{*T} z_a^*)\}, C_A\{\alpha_b \otimes \text{vec}(\alpha_b^T \alpha_b) - z_b^* \otimes \text{vec}(z_b^{*T} z_b^*)\}\Big].$$

Notice for $a, b = 1, \dots, K$, we have

$$\begin{aligned} \|\boldsymbol{\theta}_{a,b}\|_2^2 &\leq \mu_2 \|\boldsymbol{\alpha} - z^*\|_2^2 + \|\alpha_a \otimes \text{vec}(\alpha_a^T \alpha_a) - z_a^* \otimes \text{vec}(z_a^{*T} z_a^*)\|_2^2 \\ &\quad + \|\alpha_b \otimes \text{vec}(\alpha_b^T \alpha_b) - z_b^* \otimes \text{vec}(z_b^{*T} z_b^*)\|_2^2 \\ &\leq \mu_2 \|\boldsymbol{\alpha} - z^*\|_1 + C_1 N^2 (\|\alpha_a - z_a^*\|_1 + \|\alpha_b - z_b^*\|_1), \end{aligned}$$

where $\mu_2 := \max\{\log \frac{\mu_{qa}}{\mu_{qb}} - \log \frac{1-\mu_{qa}}{1-\mu_{qb}}\}$, $q = 1, \dots, K$ and $C_1 > 0$ is a constant. Then we can transform $\text{Var}(\mathbf{A}_1 + \mathbf{B}_1)$ into

$$\text{Var}(\mathbf{A}_1 + \mathbf{B}_1) = \sum_{m=1}^M \text{Var}(\boldsymbol{\theta}_{a,b} \tilde{Y}_m) = \sum_{m=1}^M \boldsymbol{\theta}_{a,b}^T \text{Cov}(\tilde{Y}_m, \tilde{Y}_m) \boldsymbol{\theta}_{a,b} \leq \|\text{Cov}(\tilde{Y}_m, \tilde{Y}_m)\|_{op} \|\boldsymbol{\theta}_{a,b}\|_2^2.$$

From the condition C3 and same discussion in (A.10), we have for some constant $C > 0$ and $c > 0$:

$$\|\text{Cov}(\tilde{Y}_m, \tilde{Y}_m)\|_{op} \leq C(1 + c\lambda N^2).$$

Given $\frac{1}{\lambda} = o(N^{\frac{\phi}{2}})$, we have $E(\mathbf{A}_1 + \mathbf{B}_1) = o_N(E(\mathbf{A}_3 + \mathbf{B}_2))$ then the $E(\mathbf{A}_3 + \mathbf{B}_2)$ is dominating in the term $\{T - E(\mathbf{A}_1 + \mathbf{B}_1)\}^2$. Based on the Markov inequality, for some constant $C_2 > 0$

$$\begin{aligned} \mathbb{P}(\mathbf{A}_1 + \mathbf{B}_1 \geq T) &\leq \frac{M \text{Var}(\mathbf{A}_1 + \mathbf{B}_1)}{\{T - E(\mathbf{A}_1 + \mathbf{B}_1)\}^2} \leq \frac{\sum_{m=1}^M \|\text{Cov}(\tilde{Y}_m, \tilde{Y}_m)\|_{op} \|\boldsymbol{\theta}_{a,b}\|_2^2}{M^2 \{c'(N + \lambda N^3)\}^2} \\ &\leq \frac{C(1 + c\lambda N^2) \{\mu_2 \|\boldsymbol{\alpha} - z^*\|_1 + C_1 N^2 (\|\alpha_a - z_a^*\|_1 + \|\alpha_b - z_b^*\|_1)\}}{(c'(N + \lambda N^3))^2 M} \\ &\leq \frac{2Cc \{\mu_2 \|\boldsymbol{\alpha} - z^*\|_1 + C_1 N^2 (\|\alpha_a - z_a^*\|_1 + \|\alpha_b - z_b^*\|_1)\}}{c'^2 (1 + \sqrt{\lambda} N^2)^2 M} \\ &\leq C_2 \frac{(\|\alpha_a - z_a^*\|_1 + \|\alpha_b - z_b^*\|_1)}{\lambda N^2 M}. \end{aligned}$$

For $\gamma \in (0, 1)$ by applying Markov inequality we have

$$\begin{aligned} \mathbb{P}\left\{\sum_{a=1}^K \sum_{b \neq a} \sum_{i: z_i^* = b} \mathbb{1}(\mathbf{A}_1 + \mathbf{B}_1 \geq T) > \frac{\|\boldsymbol{\alpha} - z^*\|_1}{(\lambda NM)^{1-\gamma}}\right\} \\ \leq \frac{\sum_{a=1}^K \sum_{b \neq a} \sum_{i: z_i^* = b} C_2 \frac{(\|\alpha_a - z_a^*\|_1 + \|\alpha_b - z_b^*\|_1)}{\lambda N^2 M}}{\frac{\|\boldsymbol{\alpha} - z^*\|_1}{(\lambda NM)^{1-\gamma}}} = \frac{\frac{C_2 \|\boldsymbol{\alpha} - z^*\|_1}{\lambda NM}}{\frac{\|\boldsymbol{\alpha} - z^*\|_1}{(\lambda NM)^{1-\gamma}}} = C_2 (\lambda NM)^{-\gamma}. \end{aligned}$$

Therefore, with probability at least $1 - (\lambda NM)^{-\gamma}$ we have

$$(A.20) \quad \sum_{a=1}^K \sum_{b \neq a} \sum_{i: z_i^* = b} \mathbb{1}(\mathbf{A}_1 + \mathbf{B}_1 \geq T) \leq \frac{\|\boldsymbol{\alpha} - z^*\|_1}{(\lambda NM)^{1-\gamma}}.$$

Combine (A.19), (A.20), (A.16) and $\alpha' = \frac{\lambda(c^* \eta_N + \lambda N^2)}{1 + \lambda N^2}$, we have with probability at least $1 - \exp(-c\lambda NM) - (\lambda NM)^{-\gamma} - o_M(1)$

$$\|\mathbf{E}(\boldsymbol{\alpha}) - z^*\|_1 \leq c_1 NK \exp\{-c_2 \alpha' NM\} + \frac{c_3 \|\boldsymbol{\alpha} - z^*\|_1}{(\lambda NM)^{1-\gamma}},$$

where c, c_1, c_2, c_3 are positive constants. A combination of results from **Part One** and **Part Two** immediately implies the theorem.

Acknowledgements. The authors would like to thank two anonymous referees, the Associate Editor and the Editor for their constructive comments and help feedback that improved the quality of this paper. The work is supported by NSF grants DMS 1952406 and DMS 1821198.

SUPPLEMENTARY MATERIAL

Supplementary material: “Community detection with dependent connectivity”.

([Supplementary material to Community detection with dependent connectivity.pdf](#)). All remaining proofs and memberships of brain regions based on different methods in Section 7.2 are provided in the supplementary material.

REFERENCES

- [1] AIROLDI, E. M., BLEI, D. M., FIENBERG, S. E. and XING, E. P. (2008). Mixed membership stochastic blockmodels. *Journal of Machine Learning Research* **9** 1981–2014.
- [2] ALEXANDER-BLOCH, A., LAMBIOTTE, R., ROBERTS, B., GIEDD, J., GOVTAY, N. and BULLMORE, E. (2012). The discovery of population differences in network community structure: New methods and applications to brain functional networks in schizophrenia. *Neuroimage* **59** 3889–3900.
- [3] AMELIO, A., MANGIONI, G. and TAGARELLI, A. (2019). Modularity in multilayer networks using redundancy-based resolution and projection-based inter-layer coupling. *IEEE Transactions on Network Science and Engineering* 1-1.
- [4] ANAGNOSTOPOULOS, A., KUMAR, R. and MAHDIAN, M. (2008). Influence and correlation in social networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data mining* 7–15. ACM.
- [5] BAHADUR, R. R. (1959). A representation of the joint distribution of responses to n dichotomous items Technical Report.
- [6] BALAKRISHNAN, S., WAINWRIGHT, M. J., YU, B. et al. (2017). Statistical guarantees for the EM algorithm: From population to sample-based analysis. *The Annals of Statistics* **45** 77–120.
- [7] BALL, B., KARRER, B. and NEWMAN, M. E. (2011). Efficient and principled method for detecting communities in networks. *Physical Review E* **84**.
- [8] BARIGOZZI, M., FAGIOLI, G. and GARLASCHELLI, D. (2010). Multinetwork of international trade: A commodity-specific analysis. *Physical Review E* **81** 046104.
- [9] BETZEL, R. F., BERTOLERO, M. A., GORDON, E. M., GRATTON, C., DOSENBACH, N. U. and BASSETT, D. S. (2018). The community structure of functional brain networks exhibits scale-specific patterns of variability across individuals and time. *bioRxiv* 413278.
- [10] BHATTACHARYYA, S. and CHATTERJEE, S. (2018). Spectral clustering for multiple sparse networks: I. *arXiv preprint arXiv:1805.10594*.
- [11] BICKEL, P. J. and CHEN, A. (2009). A nonparametric view of network models and Newman–Girvan and other modularities. *Proceedings of the National Academy of Sciences*.
- [12] BLONDEL, V. D., GUILLAUME, J.-L., LAMBIOTTE, R. and LEFEBVRE, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* **2008**.
- [13] BORDENAVE, C., LELARGE, M. and MASSOULIÉ, L. (2015). Non-backtracking spectrum of random graphs: community detection and non-regular Ramanujan graphs. In *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on* 1347–1357. IEEE.
- [14] CALHOUN, V. D., EICHELE, T. and PEARLSON, G. (2009). Functional brain networks in schizophrenia: A review. *Frontiers in human neuroscience* **3** 17.
- [15] CELISSE, A., DAUDIN, J.-J., PIERRE, L. et al. (2012). Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electronic Journal of Statistics* **6** 1847–1899.
- [16] CHEN, X. and YANG, Y. (2018). Hanson-Wright inequality in Hilbert spaces with application to K-means clustering for non-Euclidean data. *arXiv preprint arXiv:1810.11180*.
- [17] CHENG, J., LEVINA, E., WANG, P. and ZHU, J. (2014). A sparse ising model with covariates. *Biometrics* **70** 943–953.
- [18] CHOI, D. S., WOLFE, P. J. and AIROLDI, E. M. (2012). Stochastic blockmodels with a growing number of classes. *Biometrika* **99** 273–284.
- [19] DAUDIN, J.-J., PICARD, F. and ROBIN, S. (2008). A mixture model for random graphs. *Statistics and Computing* **18** 173–183.
- [20] DE DOMENICO, M., NICOSIA, V., ARENAS, A. and LATORA, V. (2015). Structural reducibility of multilayer networks. *Nature Communications* **6** 1–9.
- [21] DESIKAN, R. S., SÉGONNE, F., FISCHL, B., QUINN, B. T., DICKERSON, B. C., BLACKER, D., BUCKNER, R. L., DALE, A. M., MAGUIRE, R. P., HYMAN, B. T. et al. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage* **31** 968–980.
- [22] DIGGLE, P. and KENWARD, M. G. (1994). Informative drop-out in longitudinal data analysis. *Applied Statistics* 49–93.
- [23] DONG, X., FROSSARD, P., VANDERGHEYNST, P. and NEFEDOV, N. (2013). Clustering on multi-layer graphs via subspace analysis on Grassmann manifolds. *IEEE Transactions on Signal Processing* **62** 905–918.
- [24] FRANK, O. and STRAUSS, D. (1986). Markov graphs. *Journal of the American Statistical Association* **81** 832–842.
- [25] FRAZIER, J. A., CHIU, S., BREEZE, J. L., MAKRIS, N., LANGE, N., KENNEDY, D. N., HERBERT, M. R., BENT, E. K., KONERU, V. K., DIETERICH, M. E. et al. (2005). Structural brain magnetic resonance imaging of limbic and thalamic volumes in pediatric bipolar disorder. *American Journal of Psychiatry* **162** 1256–1265.
- [26] GENG, J., BHATTACHARYA, A. and PATI, D. (2019). Probabilistic community detection with unknown number of communities. *Journal of the American Statistical Association* **114** 893–905.
- [27] HADLEY, J. A., KRAGULJAC, N. V., WHITE, D. M., VER HOEF, L., TABORA, J. and LAHTI, A. C. (2016). Change in brain network topology as a function of treatment response in schizophrenia: A longitudinal resting-state fMRI study using graph theory. *npj Schizophrenia* **2** 1–7.
- [28] HAN, Q., XU, K. and AIROLDI, E. (2015). Consistent estimation of dynamic and multi-layer block models. In *International Conference on Machine Learning* 1511–1520.
- [29] HANDCOCK, M. S. (2003). Statistical models for social networks. *Dynamic social network modeling and analysis*. National Academies Press, Washington.
- [30] HANDCOCK, M. S., RAFTERY, A. E. and TANTRUM, J. M. (2007). Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **170** 301–354.

- [31] HOFF, P. (2008). Modeling homophily and stochastic equivalence in symmetric relational data. In *Advances in Neural Information Processing Systems* 657–664.
- [32] HOFF, P. D. (2018). Additive and multiplicative effects network models. *arXiv preprint arXiv:1807.08038*.
- [33] HOLLAND, P. W., LASKEY, K. B. and LEINHARDT, S. (1983). Stochastic blockmodels: First steps. *Social Networks* **5** 109–137.
- [34] HONEY, G. D., POMAROL-CLOTET, E., CORLETT, P. R., HONEY, R. A., MCKENNA, P. J., BULLMORE, E. T. and FLETCHER, P. C. (2005). Functional dysconnectivity in schizophrenia associated with attentional modulation of motor function. *Brain* **128** 2597–2611.
- [35] JAAKKOLA, T. (2001). Tutorial on variational approximation methods. *Advanced Mean Field Methods: Theory and Practice*.
- [36] JAIN, P., NETRAPALLI, P. and SANGHAVI, S. (2013). Low-rank matrix completion using alternating minimization. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing* 665–674. ACM.
- [37] KARRER, B. and NEWMAN, M. E. (2011). Stochastic blockmodels and community structure in networks. *Physical Review E* **83**.
- [38] KESHAVAN, R. H., MONTANARI, A. and OH, S. (2010). Matrix completion from a few entries. *IEEE Transactions on Information Theory* **56** 2980–2998.
- [39] KIM, N., WILBURNE, D., PETROVIĆ, S. and RINALDO, A. (2016). On the geometry and extremal properties of the edge-degeneracy model. *arXiv preprint arXiv:1602.00180*.
- [40] KIVELÄ, M., ARENAS, A., BARTHELEMY, M., GLEESON, J. P., MORENO, Y. and PORTER, M. A. (2014). Multilayer networks. *Journal of Complex Networks* **2** 203–271.
- [41] LATORA, V. and MARCHIORI, M. (2001). Efficient behavior of small-world networks. *Physical review letters* **87** 198701.
- [42] LATOUCHE, P., BIRMELE, E. and AMBROISE, C. (2012). Variational Bayesian inference and complexity control for stochastic block models. *Statistical Modelling* **12** 93–115.
- [43] LAURITZEN, S., RINALDO, A. and SADEGHI, K. (2018). Random networks, graphical models and exchangeability. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **80** 481–508.
- [44] LE, C. M., LEVIN, K., LEVINA, E. et al. (2018). Estimating a network from multiple noisy realizations. *Electronic Journal of Statistics* **12** 4697–4740.
- [45] LEE, K.-M., YANG, J.-S., KIM, G., LEE, J., GOH, K.-I. and KIM, I.-M. (2011). Impact of the topology of global macroeconomic network on the spreading of economic crises. *PLoS One* **6**.
- [46] LEVIN, K., ATHREYA, A., TANG, M., LYZINSKI, V., PARK, Y. and PRIEBE, C. E. (2017). A central limit theorem for an omnibus embedding of multiple random graphs and implications for multiscale network inference. *arXiv preprint arXiv:1705.09355*.
- [47] LEVIN, K., LODHIA, A. and LEVINA, E. (2019). Recovering low-rank structure from multiple networks with unknown edge distributions. *arXiv preprint arXiv:1906.07265*.
- [48] LIANG, M., ZHOU, Y., JIANG, T., LIU, Z., TIAN, L., LIU, H. and HAO, Y. (2006). Widespread functional disconnectivity in schizophrenia with resting-state functional magnetic resonance imaging. *Neuroreport* **17** 209–213.
- [49] LYNALL, M.-E., BASSETT, D. S., KERWIN, R., MCKENNA, P. J., KITZBICHLER, M., MULLER, U. and BULLMORE, E. (2010). Functional connectivity and brain networks in schizophrenia. *Journal of Neuroscience* **30** 9477–9487.
- [50] MARIADASSOU, M., ROBIN, S., VACHER, C. et al. (2010). Uncovering latent structure in valued graphs: a variational approach. *The Annals of Applied Statistics* **4** 715–742.
- [51] MASSOULIÉ, L. (2014). Community detection thresholds and the weak Ramanujan property. In *Proceedings of the forty-sixth annual ACM symposium on Theory of Computing* 694–703. ACM.
- [52] MEUNIER, D., LAMBIOTTE, R. and BULLMORE, E. T. (2010). Modular and hierarchically modular organization of brain networks. *Frontiers in Neuroscience* **4** 200.
- [53] MOSSEL, E., NEEMAN, J. and SLY, A. (2015). Consistency thresholds for the planted bisection model. In *Proceedings of the forty-seventh annual ACM symposium on Theory of Computing* 69–75. ACM.
- [54] NEWMAN, M. E. (2003). The structure and function of complex networks. *SIAM Review* **45** 167–256.
- [55] NEWMAN, M. E. (2006). Finding community structure in networks using the eigenvectors of matrices. *Physical Review E* **74**.
- [56] NEWMAN, M. E. and REINERT, G. (2016). Estimating the number of communities in a network. *Physical Review Letters* **117**.
- [57] NOBILE, A. and FEARNSIDE, A. T. (2007). Bayesian finite mixtures with an unknown number of components: The allocation sampler. *Statistics and Computing* **17** 147–162.
- [58] PATTISON, P. and ROBINS, G. (2002). Neighborhood-based models for social networks. *Sociological Methodology* **32** 301–337.
- [59] PAUL, S. and CHEN, Y. (2018). A random effects stochastic block model for joint community detection in multiple networks with applications to neuroimaging. *arXiv preprint arXiv:1805.02292*.
- [60] PAVLOVIC, D. M. (2015). Generalised stochastic blockmodels and their applications in the analysis of brain networks, PhD thesis, University of Warwick.
- [61] RAHMAN, M. S., DEY, L. R., HAIDER, S., UDDIN, M. A. and ISLAM, M. (2017). Link prediction by correlation on social network. In *Computer and Information Technology (ICCIT), 2017 20th International Conference of 1–6*. IEEE.
- [62] ROBINS, G., PATTISON, P., KALISH, Y. and LUSHER, D. (2007). An introduction to exponential random graph (p^*) models for social networks. *Social Networks* **29** 173–191.
- [63] ROBINS, G., SNIJders, T., WANG, P., HANDCOCK, M. and PATTISON, P. (2007). Recent developments in exponential random graph (p^*) models for social networks. *Social Networks* **29** 192–215.
- [64] ROHE, K., CHATTERJEE, S., YU, B. et al. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics* **39** 1878–1915.
- [65] SAADE, A., KRZAKALA, F. and ZDEBOROVÁ, L. (2014). Spectral clustering of graphs with the bethe hessian. In *Advances in Neural Information Processing Systems* 406–414.
- [66] SALDANA, D. F., YU, Y. and FENG, Y. (2017). How many communities are there? *Journal of Computational and Graphical Statistics* **26** 171–181.
- [67] STANLEY, N., SHAI, S., TAYLOR, D. and MUCHA, P. J. (2016). Clustering network layers with the strata multilayer stochastic block model. *IEEE Transactions on Network Science and Engineering* **3** 95–105.

- [68] TANG, W., LU, Z. and DHILLON, I. S. (2009). Clustering with multiple graphs. In *2009 Ninth IEEE International Conference on Data Mining* 1016–1021. IEEE.
- [69] TAYLOR, S. F., CHEN, A. C., TSO, I. F., LIBERZON, I. and WELSH, R. C. (2011). Social appraisal in chronic psychosis: Role of medial frontal and occipital networks. *Journal of Psychiatric Research* **45** 526–538.
- [70] TAYLOR, S. F., DEMETER, E., PHAN, K. L., TSO, I. F. and WELSH, R. C. (2014). Abnormal GABAergic function and negative affect in schizophrenia. *Neuropsychopharmacology* **39** 1000–1008.
- [71] TORREGGIANI, S., MANGIONI, G., PUMA, M. J. and FAGIOLO, G. (2018). Identifying the community structure of the food-trade international multi-network. *Environmental Research Letters* **13** 054026.
- [72] TZOURIO-MAZOYER, N., LANDEAU, B., PAPATHANASSIOU, D., CRIVELLO, F., ETARD, O., DELCROIX, N., MAZOYER, B. and JO-LIOT, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage* **15** 273–289.
- [73] VON DER MALSBURG, C. (1994). The correlation theory of brain function. In *Models of Neural Networks* 95–119. Springer.
- [74] WANG, S., ARROYO, J., VOGELSTEIN, J. T. and PRIEBE, C. E. (2019). Joint embedding of graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- [75] WASSERMAN, S., FAUST, K. et al. (1994). *Social network analysis: Methods and applications* **8**. Cambridge University Press.
- [76] WHITFIELD-GABRIELI, S. and NIETO-CASTANON, A. (2012). Conn: a functional connectivity toolbox for correlated and anticorrelated brain networks. *Brain connectivity* **2** 125–141.
- [77] WU, C. J. et al. (1983). On the convergence properties of the EM algorithm. *The Annals of Statistics* **11** 95–103.
- [78] XIA, M., WANG, J. and HE, Y. (2013). BrainNet Viewer: a network visualization tool for human brain connectomics. *PloS one* **8** e68910.
- [79] YU, Q., ALLEN, E., SUI, J., ARBABSHIRANI, M., PEARLSON, G. and CALHOUN, V. (2012). Brain connectivity networks in schizophrenia underlying resting state functional magnetic resonance imaging. *Current topics in medicinal chemistry* **12** 2415–2425.
- [80] ZHANG, A. Y., ZHOU, H. H. et al. (2020). Theoretical and computational guarantees of mean field variational inference for community detection. *Annals of Statistics* **48** 2575–2598.
- [81] ZHANG, J., SUN, W. W. and LI, L. (2018). Mixed-effect time-varying network model and application in brain connectivity analysis. *arXiv preprint arXiv:1806.03829*.
- [82] ZHAO, Y., LEVINA, E., ZHU, J. et al. (2012). Consistency of community detection in networks under degree-corrected stochastic block models. *The Annals of Statistics* **40** 2266–2292.