Consistent Bayesian community recovery in multilayer networks

Kalle Alaluusua and Lasse Leskelä

Department of Mathematics and Systems Analysis Aalto University Espoo, Finland

kalle.alaluusua@aalto.fi; lasse.leskela@aalto.fi

February 14, 2022

Abstract

Revealing underlying relations between nodes in a network is one of the most important tasks in network analysis. Using tools and techniques from a variety of disciplines, many community recovery methods have been developed for different scenarios. Despite the recent interest on community recovery in multilayer networks, theoretical results on the accuracy of the estimates are few and far between. Given a multilayer, e.g. temporal, network and a multilayer stochastic block model, we derive bounds for sufficient separation between intra- and inter-block connectivity parameters to achieve posterior exact and almost exact community recovery. These conditions are comparable to a well known threshold for community detection by a single-layer stochastic block model. A simulation study shows that the derived bounds translate to classification accuracy that improves as the number of observed layers increases.

Keywords: multilayer network, dynamic network, stochastic block model, community detection, planted bisection model, information-theoretic threshold, Bayesian consistency, tensor-valued data

MSC2020: 05C80, 60B10, 62F15, 62H30, 90B15, 94C15

1 Introduction

Data sets in many application domains, such as physics, sociology, computer science, economics, epidemiology and neuroscience, consist of pairwise interactions. An important unsupervised learning problem is to infer latent community memberships from observed pair interactions, when nodes in the same community are, in some sense, more similar to each other than to the other nodes. This task is commonly known as community recovery, community detection, or clustering.

Community recovery is typically approached by fitting a generative model, such as a stochastic block model (SBM) [1], on the observed network data. The stochastic block model is a probability distribution on the space of adjacency matrices, where the link probability between two nodes is solely determined by the assignment of the nodes into communities, also referred to as blocks. Thus, once the community assignment and the matrix of link probabilities is known, it is easy to sample networks from the model or evaluate the likelihood of given data.

For any given network, the performance of a community detection algorithm can be evaluated by inspecting metrics such as the classification accuracy. This leads to procedures that seem to work well in practice while no claims about their asymptotic properties are made. This is in particular the case with Bayesian methods [2, 3] that have gained popularity due to their flexibility and easy adaptability into various modelling contexts and data types. However, without exercising proper care or judgment, assigning a prior probability distribution on a large parameter space typical for community recovery problems risks failure. Quantifying how well a particular prior combined with a large data set succeeds in outputting a posterior distribution well concentrated near the corrected parameter value, is referred to as Bayesian consistency. In the context of single-layer networks, recent studies include [4, 5, 6, 7, 8].

The stochastic block model has been expanded to model vector-valued or equivalently multilayer, e.g. temporal, interactions between nodes. SBM variants with overlapping communities include those of [9, 10, 11, 12]. Dynamic community recovery combines aspects of time series, where the time increases, and machine learning, where the number of observations increases. These problems are typically characterized by multilayer data, e.g. a three-way adjacency tensor indexed by nodes and a time parameter. Despite recent interest, the number of theoretical results in this direction has so far been rather limited [13, 14, 15, 16, 17, 18]. In particular, there exists little research on Bayesian consistency for multilayer network models.

This paper contributes to the field by deriving bounds for sufficient separation between intra- and inter-block connectivity parameters for consistent Bayesian community recovery in multilayer networks. The theoretical results are demonstrated on simulated networks of small to moderate size. The simulation study shows that the classification accuracy improves as more network layers are observed.

1.1 Notation

We denote by [n] the set $\{1, 2, ..., n\}$. The notation Pf is an abbreviation of $\int f dP$. When Π denotes a prior distribution, $P\Pi$ is the expected posterior probability when data are sampled from P. For probability measures F, G with densities f, g relative to measure μ , define $\rho_{\alpha}(f || g) = \int f^{\alpha}g^{1-\alpha}d\mu$, $\alpha \in (0, 1)$. Denote by $D_{\alpha}(f || g) =$ $(\alpha - 1)^{-1}\log \rho_{\alpha}(f || g)$ the Rényi divergence of order α between f and g (see, e.g., [19]). We consider intra- and inter-block interaction distributions $f = \bigotimes_{t=1}^{T} \operatorname{Ber}_{p_t}$ and $g = \bigotimes_{t=1}^{T} \operatorname{Ber}_{q_t}$, respectively, where $\operatorname{Ber}_p(x) = (1-p)^{1-x}p^x$ denotes the Bernoulli distribution with mean p. Finally, $I_T \coloneqq D_{1/2}(f,g) = \sum_{t=1}^{T} I(p_t, q_t)$, where we define $I(p_t, q_t) = D_{1/2}(\operatorname{Ber}_{p_t}, \operatorname{Ber}_{q_t})$.

1.2 Multilayer stochastic block model

Consider a multilayer SBM with N nodes, T layers, K = 2 blocks, intra-block link probabilities p_1, \ldots, p_T , and inter-block link probabilities q_1, \ldots, q_T . Denote the set of block structures by $\mathcal{Z} = \{z : [N] \to [K]\}$. Let $Q^{(t)}$ be a $K \times K$ matrix such that $Q_{z(i)z(j)}^{(t)} = p_t$ for z(i) = z(j) and $Q_{z(i)z(j)}^{(t)} = q_t$ otherwise. Denote the space of observations by

$$\mathfrak{X} = \Big\{ X : [N] \times [N] \times [T] \to \{0, 1\} : X_{ij}^{(t)} = X_{ji}^{(t)}, \ X_{ii}^{(t)} = 0 \text{ for all } i, j, t \Big\}.$$

Given a node labelling z, the observation is distributed according to the probability measure P_z on \mathfrak{X} defined by

$$P_{z}(X) = \prod_{1 \le i < j \le N} \left(F_{z(i)z(j)}(X_{ij}) \right) = \prod_{1 \le i < j \le N} \prod_{1 \le t \le T} \left(\text{Ber}_{Q_{z(i)z(j)}^{(t)}}(X_{ij}^{(t)}) \right)$$
(1)

where F is be the matrix of intra- and inter-block interaction distributions $f = \bigotimes_{t=1}^{T} \text{Ber}(p_t)$ and $g = \bigotimes_{t=1}^{T} \text{Ber}(q_t)$ such that $F_{z(i)z(j)} = f$ for z(i) = z(j) and $F_{z(i)z(j)} = g$ otherwise.

1.3 Bayesian inference

Given a prior probability distribution Π on \mathcal{Z} and an observation $X \in \mathfrak{X}$, denote the corresponding posterior distribution by

$$\Pi_X(w) = \frac{\Pi(w)P_w(X)}{\sum_{w'}\Pi(w')P_{w'}(X)}, \qquad w \in \mathcal{Z}.$$

We consider Π_X as a Bayesian distributional estimate of an unknown block structure, from which point estimates can be derived for example by taking a mode. The accuracy of such estimates can be analysed using a frequentist viewpoint where we assume that the observed data tensor X is sampled from a model P_z with true block structure z, and we compute the expected mass that the posterior distribution assigns near the true value according to

$$\operatorname{Err}_{z}(r) = P_{z} \Pi_{X} \{ w : d_{\operatorname{ACE}}(w, z) > r \}.$$

$$\tag{2}$$

Here $d_{ACE}(w, z) = \min\{\operatorname{Ham}(w, z), \operatorname{Ham}(w, \tilde{z})\}$ denotes the absolute classification error computed using the Hamming distance $\operatorname{Ham}(w, z) = \sum_{i=1}^{N} (1 - \delta_{w(i)z(i)})$, when \tilde{z} is the modification of z obtained by swapping the labels 1 and 2.

1.4 Large-scale recovery

Large-scale data regimes can be modelled using a sequence of models in which the model parameters (N, T, p_t, q_t) as well as the spaces $\mathfrak{X}, \mathcal{Z}$, the true block structure z, and the distributions Π and Π_X all depend on a scale parameter $\nu = 1, 2, \ldots$ which is omitted from the notation for clarity. In a large-scale regime, we say that the posterior distribution *exactly recovers* z if the error defined in (2) satisfies

$$\operatorname{Err}_{z}(0) = o(1).$$

Note that $d_{ACE}(w, z) = 0$ if and only if $w \in \{z, \tilde{z}\}$. Posterior exact recovery hence means that most of the posterior mass is concentrated exactly at the set $\{z, \tilde{z}\}$ corresponding to the true unlabelled block structure. Similarly, we say that the posterior *almost exactly recovers z* if

$$\operatorname{Err}_{z}(\epsilon N) = o(1)$$

for every scale-independent constant $\epsilon > 0$. Almost exact recovery means that with high probability, most of the posterior mass is concentrated on the set of block structures w for which the relative classification error $N^{-1}d_{ACE}(w, z)$ is at most ϵ .

2 Information-theoretic thresholds

To gain understanding on how the increase in the number of network snapshots (or layers) affects the difficulty of a community recovery problem, we derive sufficient conditions for consistent community recovery, which we compare to existing literature.

2.1 Main results

Recovering an unknown block structure is possible only if the link probabilities p_t and q_t differ sufficiently from each other. For learning from single-layer observations [20, 21], a sharp information quantity for characterising recoverability is the Rényi divergence of order 1/2 between Bernoulli distributions Ber_{p_t} and Ber_{q_t} given by

$$I(p_t, q_t) = (1 - p_t)^{1/2} (1 - q_t)^{1/2} + p_t^{1/2} q_t^{1/2}.$$

Sparse networks are often modelled by assuming that $p_t = a_t \rho$ and $q_t = b_t \rho$ for scale-independent constants $a_t \neq b_t$ and overall link density $\rho = o(1)$, for example $\rho = N^{-1}$ (constant average degree) or $\rho = \frac{\log N}{N}$ (logarithmic average degree). In such case Taylor expansions show that

$$I(p_t, q_t) = \left(\sqrt{a_t} - \sqrt{b_t}\right)^2 \rho + O(\rho^2).$$

The following theorems characterise posterior recovery from multilayer network data in terms of

$$I_T = \sum_{t=1}^{T} I(p_t, q_t).$$
 (3)

Theorem 1. If $I_T \gg N^{-1}$, then the posterior distribution corresponding to the uniform prior on the set of all block structures almost exactly recovers any particular block structure z.

For $p_t = p$ and $q_t = q$, Theorem 1 shows that $I(p,q) \gg (NT)^{-1}$ suffices for almost exact recovery. Especially, we see that almost exact recovery may be achievable for a bounded number of nodes if the number of layers T is large. Reference [16] arrives to a similar conclusion in the context of latent space models. When we view T as time, we see that the product NT indicates that, from an information-theoretic point of view, observing one new node in the network is equally informative to observing one new time slot.

The following theorem characterises exact recovery of posterior distributions corresponding to noninformative priors.

Theorem 2. If $I_T \ge (2+\delta)\frac{\log N}{N}$ for some scale-independent constant $\delta > 0$, then the posterior distribution corresponding to the uniform prior on the set of all block structures exactly recovers any particular block structure z.

We believe that the sufficient condition in Theorem 2 is sharp because for T = 1, it is known [21] that exact recovery (in a frequentist sense) is impossible when $I_T \leq (2 - \delta) \frac{\log N}{N}$. A related result by [17] shows that observing multiple network layers allows for consistent community detection (by a least squares estimator) from a sparser network. Specialised to T = 1, Theorem 2 also improves the result of [6] who showed that $I_1 \ge (4 + \delta) \frac{\log N}{N}$ is sufficient for posterior exact recovery in single-layer networks. In a frequentist setting, [18] presents consistency thresholds for multilayer SBMs, which are comparable to those of Theorems 1 and 2.

3 Simulation study

We perform a simulation study to examine the effect of the number of observed network layers on the classification accuracy of a community recovery algorithm. In particular, we study the performance of an extension of the Gibbs sampler by [22] that takes as an input a tensor of independent and identically distributed adjacency matrices. The source code for replicating these experiments is available at github.com/kalaluusua.

3.1 Posterior sampler

The dynamic SBM introduced in this section adopts the conjugate priors by [22],

$$Q_{ab} \stackrel{\text{iid}}{\sim} U[0,1], \ 1 \le a \le b \le K,$$

$$z(i) \stackrel{\text{iid}}{\sim} MN_K(1;\theta), \ i \in [N]$$

$$\theta \sim \text{Dir}(K;\alpha).$$
(4)

where Q_{ab} is the link probability between the blocks a and b, $MN_k(n; p_1, \ldots, p_k)$ is the k-variate multinomial distribution with n trials, and $Dir(k; \alpha_1, \ldots, \alpha_k)$ is the kdimensional Dirichlet distribution. The associated network follows the distribution (1), where $Q^{(t)} = Q$ for all t.

To take advantage of the increased number of observed layers, we adjust the likelihood function accordingly, which by the independence of $X^{(t)}$ yields $\Pi(X \mid z, Q) = \prod_{t=1}^{T} \mathcal{L}(z, X^{(t)})$, where

$$\mathcal{L}(z, X^{(t)}) = \prod_{1 \le a \le b \le K} Q_{ab}^{O_{ab}(z, X^{(t)})} (1 - Q_{ab})^{n_{ab}(z) - O_{ab}(z, X^{(t)})}$$
(5)

where O_{ab} is the number of links between communities a and b, and n_{ab} is the maximum number of links that can be formed between communities a and b.

We propose a dynamic posterior sampler that is an extension of the Gibbs sampler introduced in [22]. The sampler approximates the posterior densities of $(\theta, Q), z(1), \ldots, z(N)$, where (θ, Q) is treated as a single random vector with the prior density $\Pi(\theta, Q)$. Given the likelihood function (5), the conditional distribution of z(i) becomes

$$\Pi(z(i) = a \mid X, P, \theta, z_{-i}) = C\theta_a \prod_{t=1}^{T} \mathcal{L}^{(i)}(z_{-i}, X),$$

where $z_{-i} \coloneqq (z(j))_{j \neq i}$ and

$$\mathcal{L}^{(i)}(z_{-i}, X) = \prod_{b=1}^{K} Q_{ab}^{O_b^{(i)}(z_{-i}, X^{(t)})} (1 - Q_{ab})^{n_b^{(i)}(z_{-i}) - O_b^{(i)}(z_{-i}, X^{(t)})}$$

such that C is a constant independent of a, $O_b^{(i)}$ is the number of i-v links such that the node $v \in G$ is assigned to community b, and $n_b^{(i)}$ is the number of nodes $v \neq i$ in community b. The posterior distribution $\Pi(\theta, Q \mid z, X)$ is given by independent Dirichlet distributions with parameters

$$(n_{a} + \alpha_{a})_{a \in [K]} \qquad \text{for } \theta,$$
$$\left(\sum_{t=1}^{T} O_{ab}(X^{(t)}) + 1, Tn_{ab} - \sum_{t=1}^{T} O_{ab}(X^{(t)}) + 1\right) \text{ for } Q_{ab},$$

where $1 \le a \le b \le K$. The posterior mode of Q_{ab} becomes $1/T \sum_t O_{ab}(X^{(t)})/n_{ab}$, the average proportion of *ab*-links over all layers. This is an extension of the block constant least squares estimator used extensively in literature [23, 24, 4].

3.2 Simulation design

We study the community detection performance of the dynamic SBM on synthetically generated networks of K = 2 communities of size N = 100. We choose the following cases for link probabilities:

Case 1: p = 0.3 and q = 0.2;

Case 2: p = 0.15 and q = 0.1,

where p and q are the intra- and inter-block link probabilities, respectively. In both cases we observe 10 synthetically generated networks with $T \in \{1, 3, 5, 7\}$ independent and identically distributed network layers. To control the sources of variation, the 10 synthetically generated networks share a community structure z_0 where the nodes are deterministically and uniformly assigned into K communities. Finally, the networks are generated from (1) with $Q_0 = \begin{bmatrix} p & q \\ q & p \end{bmatrix}$.

To construct a single point estimate of community assignment from the Markov chain generated by the Gibbs sampler, we employ a method presented in [25]. The method uses information from all the community assignments and selects a certain average assignment. In practice, we average over the last 100 members of the sequence of 1100 states, and allow for an initial burn-in period of 1000 initial iterations before stationary is reached.

To evaluate the accuracy of our estimate given the underlying community structure, we use the Hubert-Arabie adjusted Rand index [26, 27], which is a measure of similarity between two community assignments. The index is one when the assignments are identical and zero when they are independent. Since the definition disregards the relative community sizes, it tends to represent the level of agreement among large communities. However, in our experiment the average community sizes are close to being equal, and use of the index is justified.

3.3 Simulation results

Table 1 depicts the averages and the standard deviations of the classification error d_{ACE}/N and the adjusted Rand index of our estimate, given the number of observed network layers T and the link probabilities p, q. For each pair (T, Case i), the 10 simulated networks have N = 100 nodes with K = 2 communities, and connectivity

parameters that vary between Case 1 and Case 2. The priors are described in (4). The number of communities is known and $\alpha = (200, 200)$. From now on will refer to the average adjusted Rand index by accuracy. The table also depicts the standard deviation of each associated vector of estimates. Case 2 is more difficult than Case 1 since it invokes sparser networks where inter-block link probability is very close to the intra-block link probability. The Rényi divergence I values corresponding to cases 1 and 2 are 0.010 and 0.006, respectively. We expect that the community detection performance improves as T increases and is overall worse in Case 2. This is precisely what Table 1 shows.

Table 1: Average classification error (CE) and average adjusted Rand index (AR) of the community assignment estimate with the corresponding standard deviation in parenthesis.

	Case 1		Case 2	
T	CE	AR	CE	AR
1	$0.34_{(0.13)}$	$0.16_{(0.20)}$	$0.45_{(0.03)}$	$0.01_{(0.01)}$
3	$0.03_{(0.01)}$	$0.90_{(0.05)}$	$0.30_{(0.15)}$	$0.23_{(0.23)}$
5	$0.01_{(0.01)}$	$0.98_{(0.03)}$	$0.09_{(0.09)}$	$0.70_{(0.23)}$
7	$0.00_{(0.01)}$	$0.99_{(0.03)}$	$0.04_{(0.02)}$	$0.85_{(0.08)}$

When we inspect each case individually, we observe that the classification error decreases and the adjusted Rand index increases as T increases. When T = 1, the sampler misclassifies on average third of the observations in Case 1 (with an accuracy of 0.16) and nearly half of the observations (with an accuracy of 0.01) in Case 2. Recall that due to symmetry the domain of the classification error is [0, 1/2] and a Rand index of 0.01 implies that the assignments are effectively independent. The relatively large standard deviations in Case 1 imply that, despite the bad overall performance, given a favourable initial values the sampler may correctly classify a large proportion of the nodes. Relatively small standard deviations in Case 2 imply that this is unlikely when the problem is more difficult. In Case 1 the accuracy improves rapidly as the number of observed layers increases; when T = 3, the sampler is very likely to classify all but few nodes correctly, while T values of 5 and 7 lead to near perfect accuracy. In Case 2 the increase in accuracy is more muted but nevertheless apparent; when T = 3, the accuracy of the sampler resembles that in Case 1 with T = 1, and when T = 7, it resembles that in Case 1 with T = 3.

4 Final remarks

There are many important questions left unanswered, including the question of whether the sufficient conditions presented in Section 2 are also necessary. Moreover, our analysis is limited to stochastic block models with two communities. Generalizing the results for K > 2 communities, with K possibly unknown, is left for future work. Another interesting research direction would be to further examine community detection in networks that vary over time.

5 Proofs

5.1 Mirkin distance

The Mirkin distance between $z, w : [N] \to [K]$ is defined by

$$d_{\rm Mir}(z,w) = 2(M_{01} + M_{10}),\tag{6}$$

where M_{ab} is the number of unordered pairs $\{i, j\}$ such that $\delta_{z(i)z(j)} = a$ and $\delta_{w(i)w(j)} = b$. The Mirkin distance is one of the common pair-counting based cluster validity indices [28, 29], and it is related to the Rand index by $d_{\text{Mir}}(z, w) = N(N-1)(1-d_{\text{Rand}}(z, w))$.

Lemma 1. For K = 2, the Mirkin distance $M = d_{Mir}(z, w)$, the Hamming distance H = Ham(z, w), and the absolute classification error $A = d_{ACE}(z, w)$ are related by

$$M = 2(N - H)H = 2(N - A)A.$$
 (7)

Proof. Denote the blocks under focus by $C_k = \{i : z(i) = k\}$ and $C'_{\ell} = \{i : w(i) = \ell\}$. Also denote $N_{k\ell} = |C_k \cap C'_{\ell}|$. For K = 2, we find that

$$M_{01} = N_{11}N_{21} + N_{12}N_{22},$$

$$M_{10} = N_{11}N_{12} + N_{21}N_{22}.$$

By summing these, we find that

$$\frac{1}{2}d_{\rm Mir}(z,w) = M_{01} + M_{10} = (N_{11} + N_{22})(N_{12} + N_{21}).$$

The first equality in (7) follows by noting that $H = N_{12} + N_{21}$ and $N_{11} + N_{22} = N - H$. The second equality in (7) follows by noting that for K = 2, the group of permutations only contains the identity map and the transposition τ_{12} which swaps 1 and 2. In this case we find that $\operatorname{Ham}(z, \tau \circ w) = N - \operatorname{Ham}(z, w)$. Therefore, the term N(N - H) in (7) remains invariant if we replace w by $\tau \circ w$.

5.2 Upper bound on posterior mass

The following is a generalised version of [6, Proposition 3.1:(ii)] for multilayer SBMs.

Lemma 2. For any $z \in \mathbb{Z}$ and any $S \subset \mathbb{Z}$ not containing z, the expected posterior mass relative to a prior distribution Π on \mathbb{Z} is bounded by

$$P_{z}\Pi_{X}(S) \leq \sum_{w \in S} \left(\frac{\Pi(w)}{\Pi(z)}\right)^{1/2} e^{-\frac{1}{4}I_{T}d_{\operatorname{Mir}}(z,w)},\tag{8}$$

where I_T is defined by (3) and $d_{\text{Mir}}(z, w)$ by (6).

Proof. By [30, Proposition D.1], it follows that for any z and w, the likelihood ratio test $\phi_{zw}(X) = \mathbb{1}(\frac{P_w(X)}{P_z(X)} > \frac{\Pi(z)}{\Pi(w)})$ satisfies

$$\Pi(z)P_z\phi_{zw} + \Pi(w)P_w(1-\phi_{zw}) \le \frac{\Pi(z)^{\alpha}}{\Pi(w)^{\alpha-1}}\rho_{\alpha}(P_z \mid\mid P_w)$$

for all $0 < \alpha < 1$. By dividing both sides by $\Pi(z)$, it follows that

$$P_z \phi_{zw} + \frac{\Pi(w)}{\Pi(z)} P_w (1 - \phi_{zw}) \le \left(\frac{\Pi(w)}{\Pi(z)}\right)^{1-\alpha} \rho_\alpha(P_z \mid\mid P_w) \tag{9}$$

Let us define $\phi_z(X) = \max_{w \in S} \phi_{zw}(X)$. By [31, Lemma 2.2], we have

$$P_z \Pi_X(S) \le P_z \phi_z + \sum_{w \in S} \frac{\Pi(w)}{\Pi(z)} P_w (1 - \phi_z).$$

Then $P_z \phi_z \leq \sum_{w \in S} P_z \phi_{zw}$ and $1 - \phi_z \leq 1 - \phi_{zw}$ for all $w \in S$, so it follows by (9) that

$$P_{z}\Pi_{X}(S) \leq \sum_{w \in S} P_{z}\phi_{zw} + \sum_{w \in S} \frac{\Pi(w)}{\Pi(z)} P_{w}(1 - \phi_{zw})$$
$$= \sum_{w \in S} \left(P_{z}\phi_{zw} + \frac{\Pi(w)}{\Pi(z)} P_{w}(1 - \phi_{zw}) \right)$$
$$\leq \sum_{w \in S} \left(\frac{\Pi(w)}{\Pi(z)} \right)^{1-\alpha} \rho_{\alpha}(P_{z} \mid\mid P_{w}).$$
(10)

Recall that by (1), $P_z = \prod_{1 \le i < j \le N} F_{z(i)z(j)}$. Because Rényi divergence is linear with respect to products, it follows that $D_{\alpha}(P_z \mid\mid P_w) = \sum_{1 \le i < j \le N} D_{\alpha}(F_{z(i)z(j)} \mid\mid F_{w(i)w(j)})$. By definition of F, it now follows that

$$D_{\alpha}(P_z || P_w) = M_{01}D_{\alpha}(g || f) + M_{10}D_{\alpha}(f || g).$$

By setting $\alpha = \frac{1}{2}$ and recalling definition (6), we find that

$$D_{1/2}(P_z || P_w) = \frac{1}{2} d_{\text{Mir}}(z, w) I_T.$$
(11)

Inequality (8) follows by combining (10) and (11).

5.3 Preliminary estimates

Denote $B_{z,r} = \{w : d_{ACE}(z, w) \leq r\}$ and $S_{z,k} = \{w : d_{ACE}(z, w) = k\}$. By combining Lemma 2 and Lemma 1, and assuming that Π is uniform¹, it follows that error quantity defined by (2) is bounded by

$$\operatorname{Err}_{z}(r) = P_{z} \Pi_{X}(B_{z,r}^{c})$$

$$\leq \sum_{w \in B_{z,r}^{c}} \left(\frac{\Pi(w)}{\Pi(z)}\right)^{1/2} e^{-\frac{1}{4}d_{\operatorname{Mir}}(z,w)I_{T}}$$

$$= \sum_{w \in B_{z,r}^{c}} e^{-\frac{1}{2}I_{T}(N-d_{\operatorname{ACE}}(z,w))d_{\operatorname{ACE}}(z,w)}$$

$$= \sum_{r < k \le N/2} |S_{z,k}| e^{-\frac{1}{2}I_{T}(N-k)k}.$$

By noting that $|S_{z,k}| \leq 2\binom{N}{k}$, it follows that

$$\operatorname{Err}_{z}(r) \leq 2 \sum_{r < k \leq N/2} \binom{N}{k} e^{-\frac{1}{2}I_{T}(N-k)k}.$$
(12)

¹It suffices to assume that Π restricted to S is uniform, and we might relax this assumption rather easily.

5.4 Proof of Theorem 1

Recall that $|S_{z,k}| \leq 2\binom{N}{k} \leq 2(\frac{eN}{k})^k \leq 2(\frac{eN}{r})^k$ for all $r < k \leq N/2$. Together with the bound $(N-k) \geq N/2$, we find that by applying (12) that

$$\operatorname{Err}_{z}(r) \leq 2 \sum_{r < k \le N/2} \left(\frac{eN}{r}\right)^{k} e^{-\frac{1}{4}NI_{T}k} \leq 2 \sum_{k=1}^{\infty} b_{r}^{k},$$

where $b_r = \frac{eN}{r}e^{-\frac{1}{4}NI_T}$. For $r = \epsilon N$ for $\epsilon > 0$ being a scale-independent constant, we see that $b_r \to 0$ due to $I_T \gg N^{-1}$. In light of the above inequality, it follows that $\operatorname{Err}_z(\epsilon N) \to 0$ for every scale-independent constant $\epsilon > 0$. Hence Theorem 1 is valid.

5.5 Proof of Theorem 2

To prove Theorem 2, we will conduct a more careful analysis by splitting the sum in (12) at $\ell = N^{2/3}$.

For $1 \le k \le \ell$, we apply the inequalities $N-k \ge N-\ell$ and $\binom{N}{k} \le \frac{N^k}{k!}$ to conclude that $\binom{N}{k}e^{-\frac{1}{2}(N-k)kI_T} \le \frac{a_{\ell}^k}{k!}$, where $a_{\ell} = Ne^{-\frac{1}{2}(N-\ell)I_T}$. For $\ell < k \le N/2$, we apply the same bounds as in the proof of Theorem 1, to conclude that $\binom{N}{k}e^{-\frac{1}{2}(N-k)kI} \le b_{\ell}^k$ where $b_{\ell} = (\frac{eN}{\ell})e^{-\frac{1}{4}NI_T}$. It follows by applying (12) with r = 0 that

$$\operatorname{Err}_{z}(0) \leq 2 \sum_{1 \leq k \leq \ell} \frac{a_{\ell}^{k}}{k!} + 2 \sum_{\ell < k \leq N/2} b_{\ell}^{k}$$
$$\leq 2 \sum_{k \geq 1} \frac{a_{\ell}^{k}}{k!} + 2 \sum_{k \geq 1} b_{\ell}^{k}.$$

Especially, when $b_{\ell} < 1$, we see that

$$\operatorname{Err}_{z}(0) \le 2(e^{a_{\ell}} - 1) + \frac{2b_{\ell}}{1 - b_{\ell}}.$$
 (13)

Due to our choice $\ell = N^{2/3}$, we find that

$$-\log a_{\ell} = \frac{1}{2}(1 - N^{-1/3})NI_T - \log N,$$

$$-\log b_{\ell} = \frac{1}{4}NI_T - \frac{1}{3}\log N - 1.$$

The assumption that $NI_T \ge (2+\delta) \log N$ for some scale-independent constant $\delta > 0$ now implies that $-\log a_\ell \to \infty$ and $-\log b_\ell \to \infty$, and therefore $a_\ell, b_\ell \to 0$. Then (13) shows that $\operatorname{Err}_z(0) \to 0$ and confirms Theorem 2.

²Because $\frac{k^k}{k!} \leq \sum_{s=0}^{\infty} \frac{k^s}{s!} = e^k$, we see that $\binom{N}{k} \leq \frac{N^k}{k!} \leq (\frac{eN}{k})^k$.

References

- P. W. Holland, K. B. Laskey, and S. Leinhardt, "Stochastic blockmodels: First steps," *Social Networks*, vol. 5, no. 2, pp. 109–137, 1983.
- [2] J. M. Hofman and C. H. Wiggins, "Bayesian approach to network modularity," *Physical Review Letters*, vol. 100, no. 25, p. 258701, 2008.
- [3] T. P. Peixoto, "Bayesian stochastic blockmodeling," Advances in network clustering and blockmodeling, pp. 289–332, 2019.
- [4] S. van der Pas and A. van der Vaart, "Bayesian community detection," Bayesian Analysis, vol. 13, no. 3, pp. 767–796, 2018.
- [5] P. Ghosh, D. Pati, and A. Bhattacharya, "Posterior contraction rates for stochastic block models," Sankhya A, vol. 82, no. 2, p. 448–476, 2019.
- [6] B. J. K. Kleijn and J. van Waaij, "Confidence sets in a sparse stochastic block model with two communities of unknown sizes," 2021, arXiv:2108.07078.
- [7] S. Jiang and S. Tokdar, "Consistent Bayesian community detection," 2021, arXiv:2101.06531.
- [8] T. Li, T. Zhou, K.-W. Tsui, L. Wei, and Y. Ji, "Posterior contraction rate of sparse latent feature models with application to proteomics," *Statistical Theory* and Related Fields, pp. 1–11, 2021.
- [9] T. Herlau, M. Mørup, and M. Schmidt, "Modeling temporal evolution and multiscale structure in networks," in *International Conference on Machine Learn*ing. PMLR, 2013, pp. 960–968.
- [10] K. S. Xu and A. O. Hero, "Dynamic stochastic blockmodels for time-evolving social networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 4, pp. 552–562, 2014.
- [11] Q. Han, K. Xu, and E. Airoldi, "Consistent estimation of dynamic and multilayer block models," in *Proceedings of the 32nd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 37. PMLR, 2015, pp. 1511–1520.
- [12] C. Matias and V. Miele, "Statistical clustering of temporal networks through a dynamic stochastic block model," *Journal of the Royal Statistical Society B*, vol. 79, no. 4, pp. 1119–1141, 2017.
- [13] A. Ghasemian, P. Zhang, A. Clauset, C. Moore, and L. Peel, "Detectability thresholds and optimal algorithms for community structure in dynamic networks," *Physical Review X*, vol. 6, no. 3, p. 031005, 2016.
- [14] P. Barucca, F. Lillo, P. Mazzarisi, and D. Tantari, "Disentangling group and link persistence in dynamic stochastic block models," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2018, no. 123407, pp. 1–18, 2018.

- [15] S. Bhattacharyya and S. Chatterjee, "General community detection with optimal recovery conditions for multi-relational sparse networks with dependent layers," 2020, arXiv:2004.03480.
- [16] D. Durante, D. B. Dunson, and J. T. Vogelstein, "Nonparametric bayes modeling of populations of networks," *Journal of the American Statistical Association*, vol. 112, no. 520, pp. 1516–1530, 2017.
- [17] J. Lei, K. Chen, and B. Lynch, "Consistent community detection in multi-layer network data," *Biometrika*, vol. 107, no. 1, pp. 61–73, 2020.
- [18] K. Avrachenkov, M. Dreveton, and L. Leskelä, "Community recovery in nonbinary and temporal stochastic block models," 2022, arXiv:2008.04790.
- [19] T. van Erven and P. Harremoës, "Rényi divergence and Kullback-Leibler divergence," *IEEE Transactions on Information Theory*, vol. 60, no. 7, pp. 3797– 3820, 2014.
- [20] E. Abbe, A. S. Bandeira, and G. Hall, "Exact recovery in the stochastic block model," *IEEE Transactions on information theory*, vol. 62, no. 1, pp. 471–487, 2015.
- [21] E. Mossel, J. Neeman, and A. Sly, "Consistency thresholds for the planted bisection model," in *Proceedings of the 47th Annual ACM symposium on Theory* of Computing, 2015, pp. 69–75.
- [22] K. Nowicki and T. A. B. Snijders, "Estimation and prediction for stochastic blockstructures," *Journal of the American Statistical Association*, vol. 96, no. 455, p. 1077–1087, 2001.
- [23] C. Gao, Y. Lu, and H. H. Zhou, "Rate-optimal graphon estimation," Annals of Statistics, vol. 43, no. 6, pp. 2624–2652, 2015.
- [24] O. Klopp, A. B. Tsybakov, and N. Verzelen, "Oracle inequalities for network models and sparse graphon estimation," *Annals of Statistics*, vol. 45, no. 1, pp. 316–354, 2017.
- [25] D. B. Dahl, "Model-based clustering for expression data via a dirichlet process mixture model," *Bayesian inference for gene expression and proteomics*, vol. 4, pp. 201–218, 2006.
- [26] W. M. Rand, "Objective criteria for the evaluation of clustering methods," Journal of the American Statistical Association, vol. 66, no. 336, pp. 846–850, 1971.
- [27] L. Hubert and P. Arabie, "Comparing partitions," Journal of Classification, vol. 2, no. 1, pp. 193–218, 1985.
- [28] M. M. Gösgens, A. Tikhonov, and L. Prokhorenkova, "Systematic analysis of cluster similarity indices: How to validate validation measures," in *Proceedings* of the 38th International Conference on Machine Learning, vol. 139. PMLR, 18–24 Jul 2021, pp. 3799–3808.

- [29] Y. Lei, J. C. Bezdek, S. Romano, N. X. Vinh, J. Chan, and J. Bailey, "Ground truth bias in external cluster validity indices," *Pattern Recognition*, vol. 65, pp. 58–70, 2017.
- [30] S. Ghosal and A. van der Vaart, *Fundamentals of nonparametric Bayesian inference.* Cambridge University Press, 2017.
- [31] B. Kleijn, "Frequentist validity of Bayesian limits," Annals of Statistics, vol. 49, no. 1, pp. 182–202, 2021.