Supplementary Material for "Generic Sample Splitting for Refined Community Recovery in Degree Corrected Stochastic Block Models"

Jing Lei and Lingxue Zhu

Carnegie Mellon University

Summary

In this note we provide technical proofs for "Generic Sample Splitting for Refined Community Recovery in Degree Corrected Stochastic Block Models". Theorem numbering follows the original paper.

S1 Proofs for stochastic block models

We first introduce several more notations. For any $i \in \{1,2\}, k \in \{1,2,...,K\}$, we denote

$$\mathcal{I}_k = \{v : g_v = k\} \ , \ \mathcal{I}_k^{(i)} = \left\{v : v \in \mathcal{V}_i, \, g_v^{(i)} = k\right\} \ , \ \hat{\mathcal{I}}_k^{(i)} = \left\{v : v \in \mathcal{V}_i, \, \hat{g}_v^{(i)} = k\right\} \ .$$

Usually the true membership g and estimated membership $\hat{g}$ agree on most entries up to a label permutation. For simplicity we will assume, without loss of generality, that the permutation is identity.

The main theorem for stochastic block models follows from two simple applications of the accuracy of cross clustering (Lemma 3).

In the proof we will frequently use the assumption that $f(n\alpha_n/2, K)\gamma(K) \geq CK^{5/2}$, $\alpha_n \geq CK^3 \log n/(\gamma(K)^2 n)$ and $Cn \geq K^3$ for a sufficiently large C that depends only on π_0 and V.

Proof of Theorem 2. Without loss of generality, assume that the memberships $\{g^{(1)}, \cdots, g^{(V)}\}$ agree under an identity permutation, denoted as σ_I. Note that

$$\mathbb{P}\left(\hat{g} = g\right) \geq \mathbb{P}\left(\hat{\sigma}_v = \sigma_I,\ \forall v \geq 2 \mid \hat{g}^{(v)} = g^{(v)},\ \forall v\right) \mathbb{P}\left(\hat{g}^{(v)} = g^{(v)},\ \forall v\right) .$$

For any v, we have $|\mathcal{V}_v| = \frac{n}{V}$, $|\mathcal{V}_{(-v)}| = (1 - \frac{1}{V})n \geq \frac{n}{2}$. By Lemma 8, $g^{(-v)}$ is $(\frac{\pi_0}{4})$-proper and $g^{(v)}$ is $(\frac{\pi_0}{2V})$-proper with high probability. Then by Lemma 3, when $\alpha_n \geq C\frac{K^3 \log n}{\gamma(K)^2 n}$ for some constant C, $\hat{g}^{(v)} = g^{(v)}$ with high probability, which implies that

$$\mathbb{P}\left(\hat{g}^{(v)} = g^{(v)},\ \forall v\right) \geq 1 - \sum_{v=1}^{V} \mathbb{P}\left(\hat{g}^{(v)} \neq g^{(v)}\right) \geq 1 - O(n^{-1}) .$$

The final conclusion follows by Lemma 7, the consistency of merge algorithm. ∎

Proof of Lemma 3. If for all nodes $v \in \mathcal{V}_2$, for some $\delta > 0$,

$$\|\hat{h}_v - B(g_v, \cdot)\| \leq \delta\alpha_n , \tag{S1.1}$$

then we have the following separation conditions

$$\sup_{v,v' \in \mathcal{V}_2, g_v = g_{v'}} \|\hat{h}_v - \hat{h}_{v'}\| \leq 2\delta\alpha_n ,$$

$$\inf_{v,v' \in \mathcal{V}_2, g_v \neq g_{v'}} \|\hat{h}_v - \hat{h}_{v'}\| \geq \inf_{1 \leq k < k' \leq K} \|B(k, \cdot) - B(k', \cdot)\| - 2\delta\alpha_n \geq (\gamma(K) - 2\delta)\alpha_n .$$

The distance based clustering subroutine $\mathcal{D}$ used in Algorithm 1, such as the minimum spanning tree, can correctly cluster all nodes, i.e., $\hat{g}^{(2)} = g^{(2)}$, if

$$\sup_{v,v' \in \mathcal{V}_2, g_v = g_{v'}} \|\hat{h}_v - \hat{h}_{v'}\| < \inf_{v,v' \in \mathcal{V}_2, g_v \neq g_{v'}} \|\hat{h}_v - \hat{h}_{v'}\| .$$

Therefore, it suffices to show that with high probability, $\|\hat{h}_v - B(g_v, \cdot)\| \leq \delta\alpha_n$ for all $v \in \mathcal{V}_2$, for

$$\delta = \frac{\gamma(K)}{5} . \tag{S1.2}$$

By Lemma 6, the approximation bound (S1.1) and inequality (S1.2) hold with high probability if $\alpha_n \geq C\frac{K^3 \log n}{\gamma(K)^2 n}$ for some constant C depending on π_0. ■

Lemma 6. Given $\mathcal{V}_1$, $\mathcal{V}_2$, $g^{(1)}$, and $\hat{g}^{(1)}$ satisfying the conditions of Lemma 3, there exists a constant $C = C(\pi_0)$, such that for $\alpha_n \geq C\frac{K^3 \log n}{\gamma(K)^2 n}$,

$$\mathbb{P}\left(\|\hat{h}_v - B(g_v, \cdot)\| \leq \delta\alpha_n, \quad \forall\ v \in \mathcal{V}_2\right) \geq 1 - O(n^{-1}) ,$$

where $\delta = \gamma(K)/5$, and the probability is conditional on $\mathcal{V}_1$, $\mathcal{V}_2$ and $\hat{g}^{(1)}$.

Proof. Because

$$\mathbb{P}\left(\|\hat{h}_v - B(g_v, \cdot)\| \geq \delta\alpha_n\right) \leq \sum_{k=1}^{K} \mathbb{P}\left(\left|\hat{h}_{v,k} - B(g_v, k)\right| \geq \frac{\delta}{\sqrt{K}}\alpha_n\right) ,$$

it suffices to bound all K coordinates individually. Now, for any $k \in \{1, 2, .., K\}$, note that

$\pi_0 \le 1, \gamma(K) \le \sqrt{K}$, we have

$$K \le \frac{\pi_0^2 \gamma(K)}{60K^{3/2}} f(n\alpha_n/2, K) \le \frac{f(n\alpha_n/2, K)}{2}\,.$$

Therefore, when n is large enough, we have

$$\sum_{l\neq k} \left|\hat{\mathcal{I}}_l^{(1)} \cap \mathcal{I}_k^{(1)}\right| \le |\{v' : \hat{g}_{v'}^{(1)} \neq g_{v'}^{(1)}\}| \le \frac{|\mathcal{V}_1|}{f(|\mathcal{V}_1|\alpha_n, K)} \le \frac{n}{f(\alpha_n n/2, K)}\,, \tag{S1.3}$$

$$\frac{\pi_0}{K} n \le \left|\mathcal{I}_k^{(1)}\right| \le \left(1 - (K-1)\frac{\pi_0}{K}\right) n = \left((1-\pi_0) + \frac{\pi_0}{K}\right) n\,, \tag{S1.4}$$

$$\left|\hat{\mathcal{I}}_k^{(1)}\right| \ge \left|\mathcal{I}_k^{(1)}\right| - \sum_{l\neq k} \left|\hat{\mathcal{I}}_l^{(1)} \cap \mathcal{I}_k^{(1)}\right| \ge \left[\frac{\pi_0}{K} - \frac{1}{f(n\alpha_n/2, K)}\right] n \ge \frac{\pi_0 n}{2K}\,. \tag{S1.5}$$

For any $1 \le k \le K$, we have

$$\begin{aligned}
\left|\hat{h}_{v,k} - B(g_v, k)\right| &\le \left|\frac{\sum_{v'\in\hat{\mathcal{I}}_k^{(1)}} A_{v,v'} - \sum_{v'\in\mathcal{I}_k^{(1)}} A_{v,v'}}{|\hat{\mathcal{I}}_k^{(1)}|}\right| + \left|\frac{\sum_{v'\in\mathcal{I}_k^{(1)}} A_{v,v'}}{|\hat{\mathcal{I}}_k^{(1)}|} - \frac{\sum_{v'\in\mathcal{I}_k^{(1)}} A_{v,v'}}{|\mathcal{I}_k^{(1)}|}\right| \\
&\quad + \left|\frac{\sum_{v'\in\mathcal{I}_k^{(1)}} A_{v,v'}}{|\mathcal{I}_k^{(1)}|} - B(g_v, k)\right| \\
&\le \frac{\left|\sum_{v'\in\hat{\mathcal{I}}_k^{(1)}} A_{v,v'} - \sum_{v'\in\mathcal{I}_k^{(1)}} A_{v,v'}\right|}{\pi_0 n/(2K)} + \frac{\left||\mathcal{I}_k^{(1)}| - |\hat{\mathcal{I}}_k^{(1)}|\right|}{\pi_0^2 n^2/(2K^2)} \sum_{v'\in\mathcal{I}_k^{(1)}} A_{v,v'} \\
&\quad + \left|\frac{\sum_{v'\in\mathcal{I}_k^{(1)}} A_{v,v'}}{|\mathcal{I}_k^{(1)}|} - B(g_v, k)\right| \\
&= T_1 + T_2 + T_3\,.
\end{aligned}$$

Thus

$$\mathbb{P}\left(\left|\hat{h}_{v,k} - B(g_v, k)\right| \ge \frac{\delta}{\sqrt{K}}\alpha_n\right) \le \sum_{j=1}^{3} \mathbb{P}\left(T_j \ge \frac{\delta}{3\sqrt{K}}\alpha_n\right)\,.$$

Now we only need to bound the three terms individually. First, by inequality (S1.3), we

know $|\{v' : \hat{g}_{v'}^{(1)} \neq g_{v'}^{(1)}\}| \leq n/f(\alpha_n n/2, K)$. Then using Bernstein's inequality, and note that

$$K^{3/2} \leq \frac{\pi_0^2}{60K} f(n\alpha_n/2, K)\gamma(K) \leq \frac{\pi_0}{60} f(n\alpha_n/2, K)\gamma(K)\,,$$

we have,

$$\begin{aligned}
\mathbb{P}\left(T_1 \geq \frac{\delta}{3\sqrt{K}}\alpha_n\right) &\leq \mathbb{P}\left(\sum_{v':\hat{g}_{v'}^{(1)} \neq g_{v'}^{(1)}} A_{v,v'} \geq \frac{\pi_0 n}{2K}\frac{\delta}{3\sqrt{K}}\alpha_n\right) \\
&\leq \exp\left(-\frac{(\frac{\pi_0\gamma(K)n\alpha_n}{30K^{3/2}} - \frac{\alpha_n n}{f(\alpha_n n/2,K)})^2/2}{\frac{\alpha_n n}{f(\alpha_n n/2,K)} + (\frac{\pi_0\gamma(K)n\alpha_n}{30K^{3/2}} - \frac{\alpha_n n}{f(\alpha_n n/2,K)})/3}\right) \\
&\leq \exp\left(-\frac{3}{16}\frac{\pi_0\gamma(K)n\alpha_n}{30K^{3/2}}\right) = n^{-\frac{\pi_0}{160}\frac{\gamma(K)n\alpha_n}{K^{3/2}\log n}}\,.
\end{aligned}$$

To control T_2, note that if $K^{5/2} \leq (1/60)\pi_0^2 f(n\alpha_n/2, K)\gamma(K)$ and use inequalities (S1.3)-(S1.4), similarly we have

$$\begin{aligned}
\mathbb{P}\left(T_2 \geq \frac{\delta}{3\sqrt{K}}\alpha_n\right) &\leq \mathbb{P}\left(\sum_{v'\in\mathcal{I}_k^{(1)}} A_{v,v'} \geq \frac{\pi_0^2 n^2}{2K^2\left||\mathcal{I}_k^{(1)}| - |\hat{\mathcal{I}}_k^{(1)}|\right|}\frac{\delta}{3\sqrt{K}}\alpha_n\right) \\
&\leq \mathbb{P}\left(\sum_{v'\in\mathcal{I}_k^{(1)}} A_{v,v'} \geq \frac{\pi_0^2\delta n r_n f(n\alpha_n/2, K)}{6K^{5/2}}\right) \\
&\leq \exp\left(-\frac{(\frac{\pi_0^2\gamma(K)n\alpha_n f(n\alpha_n/2,K)}{30K^{5/2}} - \alpha_n n)^2/2}{\alpha_n n + (\frac{\pi_0^2\gamma(K)n\alpha_n f(n\alpha_n/2,K)}{30K^{5/2}} - \alpha_n n)/3}\right) \\
&\leq \exp\left(-\frac{3}{16}\frac{\pi_0^2\gamma(K)n\alpha_n f(n\alpha_n/2, K)}{30K^{5/2}}\right) \\
&\leq \exp\left(-\frac{3\alpha_n n}{8}\right) = n^{-\frac{3\alpha_n n}{8\log n}}\,.
\end{aligned}$$

Directly applying Bernstein's inequality to T_3 and using inequality (S1.4), we have

$$\begin{aligned}\mathbb{P}\left(T_3 \geq \frac{\delta}{3\sqrt{K}}\alpha_n\right) \leq \mathbb{P}&\left(\left|\sum_{v'\in\mathcal{I}_k^{(1)}}[A_{v,v'} - B(g_v,k)]\right| \geq \frac{\delta}{3\sqrt{K}}\alpha_n\frac{\pi_0}{K}n\right)\\ \leq& 2\exp\left(-\frac{(\frac{\delta\pi_0\alpha_n n}{3K^{3/2}})^2/2}{\alpha_n n + (\frac{\delta\pi_0\alpha_n n}{3K^{3/2}})/3}\right)\\ \leq& 2\exp\left(-\frac{\gamma^2(K)\pi_0^2 n\alpha_n/(450K^3)}{1+\gamma(K)\pi_0/(45K^{3/2})}\right) \leq 2\exp^{-\gamma^2(K)\pi_0^2 n\alpha_n/(900K^3)} = 2n^{-\frac{\gamma^2(K)\pi_0^2\alpha_n n}{900\log nK^3}}.\end{aligned}$$

where in the last inequality we used the fact that $\gamma(K) \leq \sqrt{K}$ and $\pi_0 \leq 1$, so that $1+\gamma(K)\pi_0/(45K^{3/2}) \leq 2$. ∎

Lemma 7 (Consistency of Merge)**.** Let $\{\mathcal{V}_v\}_{v=1,...,V}$ be subsets such that $|\mathcal{V}_v| = n/V$ and are $(\pi_0/(2V))$-proper. Then under the same assumptions as Theorem 2, and condition on $\hat{g}^{(v)} = g^{(v)}$ for all $v = 1, ..., V$, Algorithm 3 (Merge) outputs $\hat{g} = g$ with high probability.

Proof. Without loss of generality, assume that the memberships $\{g^{(1)}, \cdots, g^{(V)}\}$ agree under an identity permutation, denoted as σ_I. Note that

$$\mathbb{P}\left(\hat{\sigma}_v = \sigma_I,\ \forall v \geq 2 \mid \hat{g}^{(v)} = g^{(v)},\ \forall v\right) \geq 1 - \sum_{v=2}^{V}\mathbb{P}\left(\hat{\sigma}_v \neq \sigma_I \mid \hat{g}^{(v)} = g^{(v)},\ \forall v\right),$$

it suffices to show that for $\forall v = 2, ..., V$, $\mathbb{P}\left(\hat{\sigma}_v \neq \sigma_I \mid \hat{g}^{(v)} = g^{(v)},\ \forall v\right) \leq O(n^{-1})$. Now, if for all $v = 1, \cdots, V$, for some $\delta > 0$,

$$\|\hat{B}^{(v)} - B\| \leq \delta, \tag{S1.6}$$

then we have the following separation conditions:

$$\|\hat{B}^{(v)} - \hat{B}^{(1)}\| \leq 2\delta\,,$$
$$\min_{\sigma \neq \sigma_I} \|\sigma(\hat{B}^{(v)}) - \hat{B}^{(1)}\| \geq \|\sigma(B) - B\| - 2\delta \geq \alpha_n \gamma(K) - 2\delta\,,$$

where for a permutation σ on {1, ..., K}, $\sigma(B)$ is a short hand for $\sigma(B) = \left(B_{\sigma(k)l}\right)_{1\leq k,l\leq K}$, and the second inequality uses the fact that $\|\sigma(\hat{B}^{(v)}) - \sigma(B)\| = \|\hat{B}^{(v)} - B\|$ for any σ. Therefore, we only need to show that inequality (S1.6) holds with high probability for

$$\delta = \frac{\alpha_n \gamma(K)}{5}\,.$$

To show this, we follow the similar arguments in Lemma 6. Note that

$$\frac{\pi_0 n}{2VK} \leq \left|\mathcal{I}_k^{(v)}\right| \leq \frac{n}{V}\left[1 - \frac{(K-1)\pi_0}{2K}\right]\,, \quad \forall\ 1 \leq v \leq V,\ 1 \leq k \leq K\,, \tag{S1.7}$$

and use Bernstein inequality, we get

$$\begin{aligned}
\mathbb{P}\left(\|\hat{B}^{(v)} - B\| \geq \frac{\alpha_n \gamma(K)}{5}\right) &\leq \sum_{k,l=1}^{K} \mathbb{P}\left(\left|\hat{B}^{(v)}_{kl} - B_{kl}\right| \geq \frac{\alpha_n \gamma(K)}{5K}\right) \\
&\leq \sum_{k,l=1}^{K} \mathbb{P}\left(\left|\sum_{e\in\mathcal{I}^{(v)}_k, e'\in\mathcal{I}^{(1)}_l} A_{e,e'} - B_{kl}\right| \geq \frac{\alpha_n \gamma(K)}{10K}\left|\mathcal{I}^{(v)}_k\right|\left|\mathcal{I}^{(1)}_l\right|\right) \\
&\leq \sum_{k,l=1}^{K} \mathbb{P}\left(\left|\sum_{e\in\mathcal{I}^{(v)}_k, e'\in\mathcal{I}^{(1)}_l} A_{e,e'} - B_{kl}\right| \geq \frac{\alpha_n \gamma(K)}{10K}\frac{\pi_0^2 n^2}{4V^2K^2}\right) \\
&\leq 2K^2 \exp\left(-\frac{(\frac{\pi_0^2\gamma(K)n^2\alpha_n}{40V^2K^3})^2/2}{\frac{\alpha_n n^2}{V^2} + (\frac{\pi_0^2\gamma(K)n^2\alpha_n}{40V^2K^3})/3}\right) \\
&\leq 2K^2 \exp\left(-\frac{\pi_0^4\gamma^2(K)n^2\alpha_n/(3200V^2K^6)}{1+\pi_0^2\gamma(K)/(120K^3)}\right) \\
&\leq 2K^2 \exp\left(-\frac{\pi_0^4\gamma^2(K)n^2\alpha_n}{(6400V^2K^6)}\right) \leq 2\exp\left(-\frac{C\pi_0^2 n\log n}{6400V^2K^3} + 2\log K\right),
\end{aligned}$$

where the last two inequalities use the fact that $\gamma(K) \leq \sqrt{K}$ so $1 + \pi_0^2\gamma(K)/(120K^3) \leq 2$, and $\alpha_n \geq CK^3 \log n/(\gamma(K)^2 n)$. The final result follows by $K^3 \leq Cn$ for C large enough. ■

Lemma 8 (Probability of having proper split subsets)**.** If the true membership vector g on $\{1, ..., n\}$ is π_0 proper, and $\{1, ..., n\}$ is randomly split into two subsets $\mathcal{V}_1, \mathcal{V}_2$ with corresponding $g^{(1)}$, $g^{(2)}$, where $|\mathcal{V}_1| \geq cn$ for some constant $c \in (0, 1)$. Then $g^{(1)}$ is $(c\pi_0/2)$-proper with high probability when $n > K^3$.

Proof. The claimed result follows easily from an exponential tail probability bound for hypergeometric random variables (see, e.g., Skala, 2013),

$$\mathbb{P}\left(|\mathcal{I}^{(1)}_k| < c\pi_0 n/(2K)\right) \leq \mathbb{P}\left(|\mathcal{I}^{(1)}_k| - \mathbb{E}|\mathcal{I}^{(1)}_k| < -c\pi_0 n/(2K)\right) \leq e^{-c^2\pi_0^2 n/(2K^2)}$$

for all $1 \le k \le K$. The claimed results follow by union bound on $k = 1, \cdots, K$. ∎

S2 Proofs for degree corrected block models

In the following proofs, we denote $\tilde{B}$ as the $K \times K$ weighted connectivity matrix, where

$$\tilde{B}(i,j) = \frac{\sum_{v' \in \mathcal{I}_j^{(1)}} \psi_{v'}}{|\mathcal{I}_j^{(1)}|} B(i,j)\,. \tag{S2.8}$$

Proof of Theorem 4. Without loss of generality, assume that the memberships $\{g^{(1)}, \cdots, g^{(V)}\}$ agree under an identity permutation, denoted as σ_I. Note that

$$\mathbb{P}\left(\hat{g} = g\right) \ge \mathbb{P}\left(\hat{\sigma}_v = \sigma_I,\ \forall v \ge 2 \mid \hat{g}^{(v)} = g^{(v)},\ \forall v\right) \mathbb{P}\left(\hat{g}^{(v)} = g^{(v)},\ \forall v\right)\,.$$

For any v, we have $|\mathcal{V}_v| = \frac{n}{V}$, $|\mathcal{V}_{-v}| = (1 - \frac{1}{V})n \ge \frac{n}{2}$. By Lemma 8, $g^{(-v)}$ is $(\frac{\pi_0}{4})$-proper and $g^{(v)}$ is $(\frac{\pi_0}{2V})$-proper with high probability. Then by Lemma 5, when $\alpha_n > C\frac{K^3 \log n}{\tilde{\gamma}(K)^2 L(K)^2 n}$ for some constant C, $\hat{g}^{(v)} = g^{(v)}$ with high probability, which implies that

$$\mathbb{P}\left(\hat{g}^{(v)} = g^{(v)},\ \forall v\right) \ge 1 - \sum_{v=1}^{V} \mathbb{P}\left(\hat{g}^{(v)} \ne g^{(v)}\right) \ge 1 - O(n^{-1})\,.$$

The final conclusion follows by Lemma 11, the consistency of spherical merge algorithm. ∎

Proof of Lemma 5. If for all nodes $v \in \mathcal{V}_2$, for some $\delta > 0$,

$$\left\| \frac{\hat{h}_v}{\|\hat{h}_v\|} - \frac{\tilde{B}(g_v, \cdot)}{\|\tilde{B}(g_v, \cdot)\|} \right\| \le \delta\,, \tag{S2.9}$$

then we have the following separation conditions

$$\sup_{v,v'\in\mathcal{V}_2,g_v=g_{v'}}\left\|\frac{\hat{h}_v}{\|\hat{h}_v\|}-\frac{\hat{h}_{v'}}{\|\hat{h}_{v'}\|}\right\|\le 2\delta\,,$$
$$\inf_{v,v'\in\mathcal{V}_2,g_v\neq g_{v'}}\left\|\frac{\hat{h}_v}{\|\hat{h}_v\|}-\frac{\hat{h}_{v'}}{\|\hat{h}_{v'}\|}\right\|\ge \inf_{1\le k<k'\le K}\left\|\frac{\tilde{B}(k,\cdot)}{\|\tilde{B}(k,\cdot)\|}-\frac{\tilde{B}(k',\cdot)}{\|\tilde{B}(k',\cdot)\|}\right\|-2\delta\,.$$

We know from Lemma 9 that

$$\inf_{1\le k<k'\le K}\left\|\frac{\tilde{B}(k,\cdot)}{\|\tilde{B}(k,\cdot)\|}-\frac{\tilde{B}(k',\cdot)}{\|\tilde{B}(k',\cdot)\|}\right\|\ge \psi_0\tilde{\gamma}(K)\,.$$

Thus the distance based clustering subroutine $\mathcal{D}$ used in Algorithm 1', such as the minimum spanning tree, can correctly cluster all nodes, i.e., $\hat{g}^{(2)}=g^{(2)}$, if

$$\sup_{v,v'\in\mathcal{V}_2,g_v=g_{v'}}\left\|\frac{\hat{h}_v}{\|\hat{h}_v\|}-\frac{\hat{h}_{v'}}{\|\hat{h}_{v'}\|}\right\|<\inf_{v,v'\in\mathcal{V}_2,g_v\neq g_{v'}}\left\|\frac{\hat{h}_v}{\|\hat{h}_v\|}-\frac{\hat{h}_{v'}}{\|\hat{h}_{v'}\|}\right\|\,.$$

Therefore, we only need to show that with high probability, $\left\|\frac{\hat{h}_v}{\|\hat{h}_v\|}-\frac{\tilde{B}(g_v,\cdot)}{\|\tilde{B}(g_v,\cdot)\|}\right\|\le\delta$ for all nodes $v\in\mathcal{V}_2$, where

$$\delta=\frac{\psi_0\tilde{\gamma}(K)}{5}\,. \tag{S2.10}$$

By Lemma 10, the approximation bound (S2.9) and inequality (S2.10) hold with high probability if $\alpha_n\ge C\frac{K^3\log n}{\tilde{\gamma}(K)^2L(K)^2n}$ for some constant C depending on (π_0,ψ_0). ■

Lemma 9 (Lower bound of the distances between normalized rows of $\tilde{B}$)**.** If a degree corrected block model satisfies Assumptions A1' and A4, and $\tilde{B}$ is defined as in equation

(S2.8), then

$$\min_{1<k<k'\leq K}\left\|\frac{\tilde{B}(k,\cdot)}{\|\tilde{B}(k,\cdot)\|}-\frac{\tilde{B}(k',\cdot)}{\|\tilde{B}(k',\cdot)\|}\right\|\geq\psi_0\tilde{\gamma}(K)\,.$$

Proof. For simplicity let $\tilde{\gamma}=\tilde{\gamma}(K)$. Define matrix

$$\Psi=\mathrm{diag}\left(\frac{\sum_{v'\in\mathcal{I}_1^{(1)}}\psi_{v'}}{|\mathcal{I}_1^{(1)}|},...,\frac{\sum_{v'\in\mathcal{I}_K^{(1)}}\psi_{v'}}{|\mathcal{I}_K^{(1)}|}\right)\,.$$

We only need to prove that $\left\|\frac{\Psi B_0(k,\cdot)^T}{\|\Psi B_0(k,\cdot)^T\|}-\frac{\Psi B_0(k',\cdot)^T}{\|\Psi B_0(k',\cdot)^T\|}\right\|\geq\psi_0\tilde{\gamma}$, for any $k\neq k'$.

Now we define

$$w=\frac{B_0(k,\cdot)^T}{\|\Psi B_0(k,\cdot)^T\|}-\frac{B_0(k',\cdot)^T}{\|\Psi B_0(k',\cdot)^T\|}=u/s-v/t\,,$$

where $u=\frac{B_0(k,\cdot)^T}{\|B_0(k,\cdot)^T\|}$, $v=\frac{B_0(k',\cdot)^T}{\|B_0(k',\cdot)^T\|}$, $s=\frac{\|\Psi B_0(k,\cdot)^T\|}{\|B_0(k,\cdot)^T\|}$, and $t=\frac{\|\Psi B_0(k',\cdot)^T\|}{\|B_0(k',\cdot)^T\|}$. By Assumption A4, we have

$$\psi_0\leq\frac{\|\Psi B_0(k,\cdot)^T\|}{\|B_0(k,\cdot)^T\|}\leq 1,\quad\forall\ k\,.$$

Thus,

$$\|w\|\geq\min_{\psi_0\leq s,t\leq 1}\left\|\frac{u}{s}-\frac{v}{t}\right\|\,.$$

Because u and v are two unit vectors with $u^Tv\geq 0$, it is straightforward to check that the function

$$f(t,s)=\left\|\frac{u}{s}-\frac{v}{t}\right\|^2=\frac{1}{t^2}+\frac{1}{s^2}-\frac{2}{ts}u^Tv,\quad\psi_0\leq t,s\leq 1$$

reaches its minimum $\|u-v\|^2$, when $t=s=1$. Therefore,

$$\|w\| \geq \|u-v\| = \left\| \frac{B_0(k,\cdot)^T}{\|B_0(k,\cdot)^T\|} - \frac{B_0(k',\cdot)^T}{\|B_0(k',\cdot)^T\|} \right\| \geq \tilde{\gamma}\,.$$

Using the fact that smallest eigenvalue of Ψ satisfies $\lambda_{\min}(\Psi) \geq \psi_0$, we have

$$\left\| \frac{\Psi B_0(k,\cdot)^T}{\|\Psi B_0(k,\cdot)^T\|} - \frac{\Psi B_0(k',\cdot)^T}{\|\Psi B_0(k',\cdot)^T\|} \right\| = \|\Psi w\| \geq \psi_0 \|w\| \geq \psi_0 \tilde{\gamma}\,. \qquad \blacksquare$$

Lemma 10. Given $\mathcal{V}_1$, $\mathcal{V}_2$, $g^{(1)}$, and $\hat{g}^{(1)}$ satisfying the conditions of Lemma 5, then there exists a constant $C = C(\pi_0, \psi_0)$, such that if $\alpha_n \geq CK^3 \log n/(\tilde{\gamma}(K)^2 L(K)^2 n)$,

$$\mathbb{P}\left(\left\| \frac{\hat{h}_v}{\|\hat{h}_v\|} - \frac{\tilde{B}(g_v,\cdot)}{\|\tilde{B}(g_v,\cdot)\|} \right\| \leq \delta, \quad \forall\, v \in \mathcal{V}_2 \right) \geq 1 - O(n^{-1})\,,$$

where $\delta = \psi_0 \tilde{\gamma}(K)/5$, and the probability is conditional on $\mathcal{V}_1, \mathcal{V}_2$ and $\hat{g}^{(1)}$.

Proof. First, by the definition of $\tilde{B}$ in equation (S2.8), we have

$$\max\left\{ \|\hat{h}_v\|, \|\psi_v \tilde{B}(g_v,\cdot)\| \right\} \geq \|\psi_v \tilde{B}(g_v,\cdot)\| \geq \psi_0^2 \alpha_n L(K)\,.$$

Therefore,

$$\begin{aligned}
\left\| \frac{\hat{h}_v}{\|\hat{h}_v\|} - \frac{\tilde{B}(g_v,\cdot)}{\|\tilde{B}(g_v,\cdot)\|} \right\| &= \left\| \frac{\hat{h}_v}{\|\hat{h}_v\|} - \frac{\psi_v \tilde{B}(g_v,\cdot)}{\|\psi_v \tilde{B}(g_v,\cdot)\|} \right\| \\
&\leq 2 \frac{\|\hat{h}_v - \psi_v \tilde{B}(g_v,\cdot)\|}{\max\{\|\hat{h}_v\|, \|\psi_v \tilde{B}(g_v,\cdot)\|\}} \\
&\leq \frac{2}{\psi_0^2 L(K) \alpha_n} \|\hat{h}_v - \psi_v \tilde{B}(g_v,\cdot)\|\,.
\end{aligned}$$

So we only need to bound

$$\mathbb{P}\left(\|\hat{h}_v - \psi_v \tilde{B}(g_v, \cdot)\| \geq \frac{\psi_0^2 L(K)\delta\alpha_n}{2}\right) \leq \sum_{k=1}^{K} \mathbb{P}\left(\left|\hat{h}_{v,k} - \psi_v \tilde{B}(g_v, k)\right| \geq \frac{\psi_0^2 L(K)\delta\alpha_n}{2\sqrt{K}}\right),$$

and the rest of the proof follows by adapting that of Lemma 6. The details are given below.

Since inequalities (S1.3)-(S1.5) in Lemma 6 still hold, for any k, we have

$$\begin{aligned}
\left|\hat{h}_{v,k} - \psi_v \tilde{B}(g_v, k)\right| \leq & \left|\frac{\sum_{v' \in \hat{\mathcal{I}}_k^{(1)}} A_{v,v'} - \sum_{v' \in \mathcal{I}_k^{(1)}} A_{v,v'}}{|\hat{\mathcal{I}}_k^{(1)}|}\right| + \left|\frac{\sum_{v' \in \mathcal{I}_k^{(1)}} A_{v,v'}}{|\hat{\mathcal{I}}_k^{(1)}|} - \frac{\sum_{v' \in \mathcal{I}_k^{(1)}} A_{v,v'}}{|\mathcal{I}_k^{(1)}|}\right| \\
& + \left|\frac{\sum_{v' \in \mathcal{I}_k^{(1)}} A_{v,v'}}{|\mathcal{I}_k^{(1)}|} - \frac{\sum_{v' \in \mathcal{I}_k^{(1)}} \psi_{v'}}{|\mathcal{I}_k^{(1)}|}\psi_v B(g_v, k)\right| \\
\leq & \frac{\left|\sum_{v' \in \hat{\mathcal{I}}_k^{(1)}} A_{v,v'} - \sum_{v' \in \mathcal{I}_k^{(1)}} A_{v,v'}\right|}{\pi_0 n/(2K)} + \frac{\left||\mathcal{I}_k^{(1)}| - |\hat{\mathcal{I}}_k^{(1)}|\right|}{\pi_0^2 n^2/(2K^2)} \sum_{v' \in \mathcal{I}_k^{(1)}} A_{v,v'} \\
& + \left|\frac{\sum_{v' \in \mathcal{I}_k^{(1)}} A_{v,v'}}{|\mathcal{I}_k^{(1)}|} - \frac{\sum_{v' \in \mathcal{I}_k^{(1)}} \psi_{v'}}{|\mathcal{I}_k^{(1)}|}\psi_v B(g_v, k)\right| \\
= & T_1 + T_2 + T_3 .
\end{aligned}$$

Now we only need to bound the three terms individually. First by (S1.3) we know $|\{v' : \hat{g}_{v'}^{(1)} \neq g_{v'}^{(1)}\}| \leq n/f(\alpha_n n/2, K)$. Then using Bernstein's inequality and noticing that

$$K^{3/2} \leq \frac{\pi_0^2}{120K}\psi_0^3 f(\alpha_n n/2, K)\tilde{\gamma}(K)L(K) \leq \frac{\pi_0}{120}\psi_0^3 f(\alpha_n n/2, K)\tilde{\gamma}(K)L(K),$$

we have for n large enough,

$$\begin{aligned}
\mathbb{P}\left(T_1 \geq \frac{\psi_0^2 L(K)\delta\alpha_n}{6\sqrt{K}}\right) &\leq \mathbb{P}\left(\sum_{v':\hat{g}_{v'}^{(1)} \neq g_{v'}^{(1)}} A_{v,v'} \geq \frac{\pi_0 n}{2K}\frac{\psi_0^2 L(K)\delta\alpha_n}{6\sqrt{K}}\right) \\
&\leq \exp\left(-\frac{(\frac{\pi_0\psi_0^3 L(K)\tilde{\gamma}(K)\alpha_n n}{60K^{3/2}} - \frac{n\alpha_n}{f(\alpha_n n/2,K)})^2/2}{\frac{n\alpha_n}{f(\alpha_n n/2,K)} + (\frac{\pi_0\psi_0^3 L(K)\tilde{\gamma}(K)\alpha_n n}{60K^{3/2}} - \frac{n\alpha_n}{f(\alpha_n n/2,K)})/3}\right) \\
&\leq \exp\left(-\frac{3}{16}\frac{\pi_0\psi_0^3 L(K)\tilde{\gamma}(K)\alpha_n n}{60K^{3/2}}\right) = n^{-\frac{\pi_0\psi_0^3}{320}\frac{\tilde{\gamma}(K)L(K)n\alpha_n}{K^{3/2}\log n}}.
\end{aligned}$$

To control T_2, note that $K^{5/2} \leq (1/120)\pi_0^2\psi_0^3 f(\alpha_n n/2, K)\tilde{\gamma}(K)L(K)$. Similarly we have, for n large enough, using Bernstein's inequality,

$$\begin{aligned}
\mathbb{P}\left(T_2 \geq \frac{\psi_0^2 L(K)\delta\alpha_n}{6\sqrt{K}}\right) &\leq \mathbb{P}\left(\sum_{v'\in\mathcal{I}_k^{(1)}} A_{v,v'} \geq \frac{\pi_0^2 n^2}{2K^2\left||\mathcal{I}_k^{(1)}| - |\hat{\mathcal{I}}_k^{(1)}|\right|}\frac{\psi_0^2 L(K)\delta\alpha_n}{6\sqrt{K}}\right) \\
&\leq \mathbb{P}\left(\sum_{v'\in\mathcal{I}_k^{(1)}} A_{v,v'} \geq \frac{\pi_0^2\psi_0^2 L(K)\delta\alpha_n n f(\alpha_n n/2,K)}{12K^{5/2}}\right) \\
&\leq \exp\left(-\frac{\left(\frac{\pi_0^2\psi_0^3 L(K)\tilde{\gamma}(K)\alpha_n n f(\alpha_n n/2,K)}{60K^{5/2}} - n\alpha_n((1-\pi_0)+\frac{\pi_0}{K})\right)^2/2}{n\alpha_n((1-\pi_0)+\frac{\pi_0}{K}) + \left(\frac{\pi_0^2\psi_0^3 L(K)\tilde{\gamma}(K)\alpha_n n f(\alpha_n n/2,K)}{60K^{5/2}} - n\alpha_n((1-\pi_0)+\frac{\pi_0}{K})\right)/3}\right) \\
&\leq \exp\left(-\frac{3}{16}\frac{\pi_0^2\psi_0^3 L(K)\tilde{\gamma}(K)\alpha_n n f(\alpha_n n/2,K)}{60K^{5/2}}\right) \leq \exp\left(-\frac{3\alpha_n n}{8}\right) = n^{-\frac{3}{8}\frac{\alpha_n n}{\log n}}.
\end{aligned}$$

Directly applying Bernstein's inequality to T_3 and using inequality (S1.4), we have

$$\begin{aligned}
\mathbb{P}\left(T_3 > \frac{\psi_0^2 L(K)\delta\alpha_n}{6\sqrt{K}}\right) &\le \mathbb{P}\left(\left|\sum_{v'\in\mathcal{I}_k^{(1)}}[A_{v,v'} - \psi_v\psi_{v'}B(g_v,k)]\right| \ge \frac{\psi_0^2 L(K)\delta\alpha_n}{6\sqrt{K}}\frac{\pi_0 n}{K}\right) \\
&\le 2\exp\left(-\frac{(\frac{\pi_0\psi_0^2 L(K)\delta\alpha_n n}{6K^{3/2}})^2/2}{n\alpha_n((1-\pi_0)+\frac{\pi_0}{K})+\frac{\pi_0\psi_0^2 L(K)\delta\alpha_n n}{6K^{3/2}}/3}\right) \\
&\le 2\exp\left(-\frac{\pi_0^2\psi_0^6 L(K)^2\tilde{\gamma}(K)^2 n\alpha_n/(1800K^3)}{1+\pi_0\psi_0^3 L(K)\tilde{\gamma}(K)/(90K^{3/2})}\right) \\
&\le 2\exp\left(-\pi_0^2\psi_0^6 L(K)^2\tilde{\gamma}(K)^2 n\alpha_n/(3600K^3)\right) = 2n^{-\frac{\pi_0^2\psi_0^6}{3600}\frac{\tilde{\gamma}(K)^2 L(K)^2 n\alpha_n}{K^3\log n}},
\end{aligned}$$

where the last inequality uses the fact that $L(K) \le \sqrt{K}, \tilde{\gamma}(K) \le \sqrt{K}$ and $\pi_0, \psi_0 \le 1$, so that $1 + \pi_0\psi_0^3 L(K)\tilde{\gamma}(K)/(90K^{3/2}) \le 2$. ∎

Lemma 11 (Consistency of MergeSphere)**.** Let $\{\mathcal{V}_v\}_{v=1,\dots,V}$ be disjoint subsets such that $|\mathcal{V}_v| = n/V$ and are $(\pi_0/(2V))$-proper. Then under the same assumptions as in Theorem 4, and condition on $\hat{g}^{(v)} = g^{(v)}$ for all $v = 1, ..., V$, Algorithm 3' (MergeSphere) outputs $\hat{g} = g$ with high probability.

Proof of Lemma 11. Without loss of generality, assume that the memberships $\{g^{(1)}, \cdots, g^{(V)}\}$ agree under an identity permutation, denoted as σ_I. Note that

$$\mathbb{P}\left(\hat{\sigma}_v = \sigma_I,\ \forall v \ge 2 \mid \hat{g}^{(v)} = g^{(v)},\ \forall v\right) \ge 1 - \sum_{v=2}^{V}\mathbb{P}\left(\hat{\sigma}_v \ne \sigma_I \mid \hat{g}^{(v)} = g^{(v)},\ \forall v\right),$$

it suffices to show that for $\forall v = 2, ..., V$, $\mathbb{P}\left(\hat{\sigma}_v \ne \sigma_I \mid \hat{g}^{(v)} = g^{(v)},\ \forall v\right) \le O(n^{-1})$. We define $\bar{\psi}_k^{(v)}$ to be the average node activeness in $\mathcal{I}_k^{(v)}$:

$$\bar{\psi}_k^{(v)} = \frac{\sum_{e'\in\mathcal{I}_k^{(v)}}\psi_{e'}}{|\mathcal{I}_k^{(v)}|},$$

and $\tilde{B}^{(1)}$ and its row-normalized version $B_*^{(1)}$ as follows:

$$\tilde{B}^{(1)}(k,l) = \bar{\psi}_l^{(1)} B(k,l)\,, \quad B_*^{(1)}(k,\cdot) = \frac{\tilde{B}^{(1)}(k,\cdot)}{\|\tilde{B}^{(1)}(k,\cdot)\|}\,.$$

If for all $v = 1, \cdots, V$, for some $\delta > 0$,

$$\|\hat{B}_*^{(v)} - B_*^{(1)}\| \le \delta\,, \tag{S2.11}$$

then use Lemma 9, we have the following separation conditions:

$$\|\hat{B}_*^{(v)} - \hat{B}_*^{(1)}\| \le 2\delta\,,$$
$$\min_{\sigma \neq \sigma_I} \|\sigma(\hat{B}_*^{(v)}) - \hat{B}_*^{(1)}\| \ge \|\sigma(B_*^{(1)}) - B_*^{(1)}\| - 2\delta \ge \psi_0 \tilde{\gamma}(K) - 2\delta\,,$$

where for a permutation σ on $\{1, ..., \mathrm{K}\}$, $\sigma(B_*)$ is a short hand for $\sigma(B_*) = (B_*(\sigma(k), l))_{1 \le k,l \le K}$, and the second inequality uses the fact that $\|\sigma(\hat{B}_*^{(v)}) - \sigma(B_*^{(1)})\| = \|\hat{B}_*^{(v)} - B_*^{(1)}\|$ for any σ. Therefore, we only need to show that inequality (S2.11) holds with high probability for

$$\delta = \frac{\psi_0 \tilde{\gamma}(K)}{5}\,.$$

Note that by assumption, for $\forall\ 1 \le v \le V,\ 1 \le k \le K$, we have $\hat{\mathcal{I}}_k^{(v)} = \mathcal{I}_k^{(v)}$ and

$$\frac{\pi_0 n}{2VK} \le \left|\mathcal{I}_k^{(v)}\right| \le \frac{n}{V}\left[1 - \frac{(K-1)\pi_0}{2K}\right]\,,$$
$$\max\{\|\hat{B}^{(v)}(k,\cdot)\|, \|\bar{\psi}_k^{(v)} \tilde{B}^{(1)}(k,\cdot)\|\} \ge \|\bar{\psi}_k^{(v)} \tilde{B}^{(1)}(k,\cdot)\| \ge \psi_0^2 \alpha_n L(K)\,.$$

The second inequality further implies that

$$\begin{aligned}\left|\left|\hat{B}_*^{(v)}(k,\cdot) - B_*^{(1)}(k,\cdot)\right|\right| &= \left|\left|\frac{\hat{B}^{(v)}(k,\cdot)}{\|\hat{B}^{(v)}(k,\cdot)\|} - \frac{\bar{\psi}_k^{(v)}\tilde{B}^{(1)}(k,\cdot)}{\|\bar{\psi}_k^{(v)}\tilde{B}^{(1)}(k,\cdot)\|}\right|\right| \\ &\leq 2\frac{\|\hat{B}^{(v)}(k,\cdot) - \bar{\psi}_k^{(v)}\tilde{B}^{(1)}(k,\cdot)\|}{\max\{\|\hat{B}^{(v)}(k,\cdot)\|, \|\bar{\psi}_k^{(v)}\tilde{B}^{(1)}(k,\cdot)\|} \\ &\leq \frac{2}{\psi_0^2\alpha_n L(K)}\|\hat{B}^{(v)}(k,\cdot) - \bar{\psi}_k^{(v)}\tilde{B}^{(1)}(k,\cdot)\|\,.\end{aligned}$$

Then using Bernstein inequality, we get

$$\begin{aligned}\mathbb{P}\left(\|\hat{B}_*^{(v)} - B_*^{(1)}\| \geq \frac{\psi_0\tilde{\gamma}(K)}{5}\right) &\leq \sum_{k=1}^K \mathbb{P}\left(\left|\left|\hat{B}_*^{(v)}(k,\cdot) - B_*^{(1)}(k,\cdot)\right|\right| \geq \frac{\psi_0\tilde{\gamma}(K)}{5\sqrt{K}}\right) \\ &\leq \sum_{k=1}^K \mathbb{P}\left(\left|\left|\hat{B}^{(v)}(k,\cdot) - \bar{\psi}_k^{(v)}\tilde{B}^{(1)}(k,\cdot)\right|\right| \geq \frac{\psi_0\tilde{\gamma}(K)}{5\sqrt{K}}\frac{\psi_0^2\alpha_n L(K)}{2}\right) \\ &\leq \sum_{k,l=1}^K \mathbb{P}\left(\left|\hat{B}^{(v)}(k,l) - \bar{\psi}_k^{(v)}\tilde{B}^{(1)}(k,l)\right| \geq \frac{\psi_0^3\tilde{\gamma}(K)L(K)\alpha_n}{10K}\right) \\ &= \sum_{k,l=1}^K \mathbb{P}\left(\left|\frac{\sum_{e\in\mathcal{I}_k^{(v)},\ e'\in\mathcal{I}_l^{(1)}} A_{e,e'} - \psi_e\psi_{e'}B(k,l)}{\left|\mathcal{I}_k^{(v)}\right|\left|\mathcal{I}_l^{(1)}\right|}\right| \geq \frac{\psi_0^3\tilde{\gamma}(K)L(K)\alpha_n}{10K}\right) \\ &\leq \sum_{k,l=1}^K \mathbb{P}\left(\left|\sum_{e\in\mathcal{I}_k^{(v)},\ e'\in\mathcal{I}_l^{(1)}} A_{e,e'} - \psi_e\psi_{e'}B(k,l)\right| \geq \frac{\psi_0^3\tilde{\gamma}(K)L(K)\alpha_n}{10K}\left(\frac{\pi_0 n}{2VK}\right)^2\right) \\ &\leq 2K^2\exp\left(-\frac{(\frac{\psi_0^3\pi_0^2\tilde{\gamma}(K)L(K)n^2\alpha_n}{40V^2K^3})^2/2}{\frac{\alpha_n n^2}{V^2} + (\frac{\psi_0^3\pi_0^2\tilde{\gamma}(K)L(K)n^2\alpha_n}{40V^2K^3})/3}\right) \\ &\leq 2K^2\exp\left(-\frac{\psi_0^6\pi_0^4\tilde{\gamma}^2(K)L^2(K)n^2\alpha_n/(3200V^2K^6)}{1+\psi_0^3\pi_0^2\tilde{\gamma}(K)L(K)/(120K^3)}\right) \\ &\leq 2K^2\exp\left(-\frac{\psi_0^6\pi_0^4\tilde{\gamma}^2(K)L^2(K)n^2\alpha_n}{6400V^2K^6}\right) \\ &\leq 2\exp\left(-\frac{C\psi_0^6\pi_0^4 n\log n}{6400V^2K^3} + 2\log K\right)\end{aligned}$$

where the last two inequalities use the fact that $\gamma(K) \leq \sqrt{K}, L(K) \leq \sqrt{K}$ so

$$1 + \psi_0^3 \pi_0^2 \tilde{\gamma}(K) L(K)/(120K^3) \leq 2\,,$$

and $\alpha_n \geq C \frac{K^3 \log n}{\tilde{\gamma}(K)^2 L(K)^2 n}$. The final result follows by $K^3 \leq Cn$ for C large enough. ■