SUPPLEMENT TO "MIMIMAX RATES OF COMMUNITY DETECTION IN STOCHASTIC BLOCK MODELS"

BY Anderson Y. Zhang and Harrison H. Zhou

Yale University

APPENDIX A: ADDITIONAL PROOFS

In this appendix we provide the proofs of Lemma 5.2, Proposition 5.1, Proposition 5.2, Theorem 2.1 and Theorem 3.1.

A.1. Proof of Lemma 5.2. Let $p(z)$ be the probability mass function of Z_i, and $M(t)$ be the moment generating function of Z_i. That is

$$M(t) = \mathbb{E}e^{tX_i}\mathbb{E}e^{-tY_i} = \big(e^t\frac{b}{n} + 1 - \frac{b}{n}\big)\big(e^{-t}\frac{a}{n} + 1 - \frac{a}{n}\big).$$

The minimum of $M(t)$ is achieved at $t^\star = \frac{1}{2}\log\frac{a(1-b/n)}{b(1-a/n)}$, with

$$M(t^\star) = \left(\sqrt{\frac{a}{n}\frac{b}{n}} + \sqrt{\Big(1-\frac{a}{n}\Big)\Big(1-\frac{b}{n}\Big)}\right)^2. \tag{A.1}$$

This gives $I = -\log M(t^\star) = \max_t(-\log M(t))$. Let δ be a positive number which may depend on n. Denote $S_{n'} = \sum_{i=1}^{n'} Z_i$ and $S_{n'}(z) = \sum_{i=1}^{n'} z_i$. Then

$$\begin{aligned}\mathbb{P}(S_{n'} \geq 0) &\geq \sum_{\{z: S_{n'}(z)\in[0,n'\delta]\}}\prod_{i=1}^{n'} p(z_i)\\ &\geq \frac{M^{n'}(t^\star)}{\exp(n't^\star\delta)}\sum_{\{z: S_{n'}(z)\in[0,n'\delta]\}}\prod_{i=1}^{n'}\frac{\exp(t^\star z_i)p(z_i)}{M(t^\star)},\end{aligned}$$

where we use the fact that $\exp(n't^\star\delta) \geq \exp(t^\star\sum_i z_i) \geq \prod\exp(t^\star z_i)$ when $\sum_i z_i < n'\delta$. Denote $q(w) = \frac{\exp(t^\star w)p(w)}{M(t^\star)}$. Then

$$\begin{aligned}\mathbb{P}(S_{n'} \geq 0) &\geq \frac{M^{n'}(t^\star)}{\exp(n't^\star\delta)}\sum_{\{z: S_{n'}(z)\in[0,n'\delta]\}}\prod_{i=1}^{n'} q(z_i)\\ &= \exp\big(-n'I\big)\exp(-n't^\star\delta)\sum_{\{z: S_{n'}(z)\in[0,n'\delta]\}}\prod_{i=1}^{n'} q(z_i).\end{aligned}$$

Note that $q(w)$ is a probability mass function, as $\sum_w \frac{\exp(t^\star w)p(w)}{M(t^\star)} = 1$. Let $W_1, W_2, \ldots, W_{n'}$ be i.i.d random variable with probability mass function $q(w)$, then

$$\mathbb{P}(S_{n'} \geq 0) \geq \exp\big(-n'I\big)\exp(-n't^\star\delta)\mathbb{P}\Big(\delta > \frac{1}{n'}\sum_{i=1}^{n'} W_i \geq 0\Big).$$

A closer look on W_1 gives

$$\mathbb{P}(W_1 = 1) = \mathbb{P}(W_1 = -1) = \frac{1}{M(t^\star)}\sqrt{\frac{a}{n}\frac{b}{n}(1-\frac{a}{n})(1-\frac{b}{n})},$$

and $\mathbb{P}(W_1 = 0) = 1 - \mathbb{P}(W_1 = 1) - \mathbb{P}(W_1 = -1)$. Thus $\mathbb{E}W_1 = 0$ and

$$\mathrm{Var}(W_1) = \frac{2}{M(t^\star)}\sqrt{\frac{a}{n}\frac{b}{n}(1-\frac{a}{n})(1-\frac{b}{n})}.$$

Denote $V = \mathrm{Var}(\sum_{i=1}^{n'} W_i/n') = \mathrm{Var}(W_1)/n'$. By Proposition A.1 we have $I/(t^\star\sqrt{V}) \asymp \sqrt{nI/K}$. Consider the following two cases.

(1) If $nI/K \to \infty$, we have $I/(t^\star\sqrt{V}) \to \infty$ as well. Define $\delta = V^{\frac{1}{4}} I^{\frac{1}{2}} (t^\star)^{-\frac{1}{2}}$ which yields $\delta^2/V = I/(t^\star\sqrt{V}) \to \infty$. Chebyshev's inequality gives

$$\mathbb{P}\Big(\Big|\frac{1}{n'}\sum_{i=1}^{n'} W_i\Big| \geq \delta\Big) \leq \frac{V}{\delta^2} = o(1).$$

By the fact that the distribution of $\frac{1}{n'}\sum_{i=1}^{n'} W_i$ is symmetric, we have

$$\mathbb{P}\Big(\delta > \frac{1}{n'}\sum_{i=1}^{n'} W_i \geq 0\Big) = \frac{1}{2}\Big(1 - \mathbb{P}\Big(\Big|\frac{1}{n'}\sum_{i=1}^{n'} W_i\Big| \geq \delta\Big)\Big) \to \frac{1}{2}.$$

Together with $t^\star\delta = o(I)$ we get $\mathbb{P}(S_{n'} > 0) \geq \exp(-(1+o(1))n'I)$.

(2) If $nI/K = O(1)$ is in a constant order then there exist constants $c_1, c_2 > 0$ such that $nI/K \leq c_1$ and $I/(t^\star\sqrt{V}) \geq c_2$. Define $\delta = c_3K/(nt^\star)$ for some constant $c_3 > 0$, then

$$\frac{\sqrt{V}}{\delta} = \frac{n\sqrt{V}t^\star}{c_3K} \leq \frac{c_1\sqrt{V}t^\star}{c_3I} \leq \frac{c_1}{c_2c_3}.$$

We can choose c_3 large enough such that $V/\delta^2 \leq c_4 < 1$ for some constant $c_4 > 0$. Then by applying the Chebyshev's inequality as in case (1) we obtain

$$\mathbb{P}\Big(\delta > \frac{1}{n'}\sum_{i=1}^{n'} W_i \geq 0\Big) = \frac{1}{2}\Big(1 - \mathbb{P}\Big(\Big|\frac{1}{n'}\sum_{i=1}^{n'} W_i\Big| \geq \delta\Big)\Big) \geq \frac{1}{2}(1 - c_4) > 0.$$

Thus $\mathbb{P}(S_{n'} > 0)$ is lower bounded by some positive constant.

(3) If $nI/K = o(1)$, we can use coupling to convert it into case (2). By Lemma B.1 we have $nI/K \asymp (a-b)^2/(aK)$. We can find an $a' > a$ such that $(a'-b)^2/(a'K) \asymp 1$ or a $b' < b$ such that $(a-b')^2/(aK) \asymp 1$. Thus by coupling the probability

$$\mathbb{P}(\sum_{i=1}^{n'} X_i \geq \sum_{i=1}^{n'} Y_i) \geq \mathbb{P}(\sum_{i=1}^{n'} X_i \geq \sum_{i=1}^{n'} Y_i')$$

is lower bounded by a positive constant, where $\{Y_i'\}_{i=1}^{n'} \overset{iid}{\sim} \text{Ber}(\frac{a'}{n})$. Similar coupling trick can also be applied on b'.

PROPOSITION A.1. *Define I, $t^\star$ and $M(t^\star)$ as in Equation (1.2), Equation (3.3) and Equation (A.1). Let*

$$V = \frac{2K}{nM(t^*)}\sqrt{\frac{a}{n}\frac{b}{n}(1-\frac{a}{n})(1-\frac{b}{n})}.$$

Under the assumption that $0 < b < a < (1-c)n$ for any constant $c > 0$ we have $I^2/((t^\star)^2 V) \asymp \frac{nI}{K}$.

PROOF. By the definition of $M(t^\star)$ we have $M(t^\star) \asymp 1$, which implies $V \lesssim \frac{K}{n}\frac{\sqrt{ab}}{n} \leq \frac{Ka}{n^2}$. For $t^\star$ we have

$$\begin{aligned} t^\star &= -\frac{1}{2}\log\frac{b(1-a/n)}{a(1-b/n)} \\ &= -\frac{1}{2}\log\Big(\Big(1-\frac{a-b}{a}\Big)\Big(1-\frac{a-b}{n-b}\Big)\Big) \\ &\asymp -\log\Big(1-\frac{a-b}{a}\Big). \end{aligned}$$

By Lemma B.1, we have $I \asymp (a-b)^2/(na)$. We consider the following two cases.

(1) If $a \asymp b$, we have $t^\star \asymp (a-b)/a$. Then

$$\frac{I^2}{(t^\star)^2 V} \gtrsim \frac{\big(\frac{(a-b)^2}{na}\big)^2}{\big(\frac{a-b}{a}\big)^2 \frac{Ka}{n^2}} = \frac{(a-b)^2}{aK} \asymp \frac{nI}{K}.$$

(2) If $a \gg b$, we have $t^\star \asymp \log a/b$. Note that $\frac{b}{a}\big(\log\frac{a}{b}\big)^4 = o(1)$, we have

$$\frac{I^2}{(t^\star)^2 V} \gtrsim \frac{\big(\frac{(a-b)^2}{na}\big)^2}{\big(\log\frac{a}{b}\big)^2 \frac{K\sqrt{ab}}{n^2}} = \frac{(a-b)^4}{a^3 K}\frac{1}{\sqrt{\frac{b}{a}}\big(\log\frac{a}{b}\big)^2} \gtrsim \frac{(a-b)^2}{aK} \asymp \frac{nI}{K}.$$

□

A.2. Proof of Proposition 5.1. Together with Equation (5.4) and Equation (5.5) we have

$$\mathbb{P}(T(\sigma) \geq T(\sigma_0)) \leq \mathbb{P}\Big(\sum_{i=1}^{\gamma} U_i - \sum_{i=1}^{\alpha} V_i \geq \lambda(\gamma - \alpha)\Big),$$

where $\{U_i\}_{i=1}^{\gamma}$, $\{V_i\}_{i=1}^{\alpha}$ are independent random variables and $U_i \sim \text{Ber}(q_i)$, $V_i \sim \text{Ber}(p_i)$ with some $\{q_i\}_{i=1}^{\gamma}$, $\{p_i\}_{i=1}^{\alpha}$ such that $\min p_i \geq a/n$ and $\max q_i \leq b/n$. By coupling, we have

$$\mathbb{P}(T(\sigma) \geq T(\sigma_0)) \leq \mathbb{P}\Big(\sum_{i=1}^{\gamma} X_i - \sum_{i=1}^{\alpha} Y_i \geq \lambda(\gamma - \alpha)\Big),$$

with independent variables $\{X_i\}_{i=1}^{\gamma} \overset{iid}{\sim} \text{Ber}(\frac{b}{n})$ and $\{Y_i\}_{i=1}^{\alpha} \overset{iid}{\sim} \text{Ber}(\frac{a}{n})$. As an application of Markov inequality,

$$\begin{aligned}
\mathbb{P}(T(\sigma) \geq T(\sigma_0)) &\leq \mathbb{P}\Big(\exp\Big(t\sum^{\gamma} X_i - t\sum^{\alpha} Y_i\Big) \geq \exp(t\lambda(\gamma-\alpha))\Big) \\
&\leq e^{-t\lambda(\gamma-\alpha)}\Big(\mathbb{E}e^{tX_1}\Big)^{\gamma}\Big(\mathbb{E}e^{-tY_1}\Big)^{\alpha} \\
&= \Big(\mathbb{E}e^{tX_1}\mathbb{E}e^{-tY_1}\Big)^{(1-w)\alpha+w\gamma}\Big(\frac{(\mathbb{E}e^{tX_1})^{1-w}}{(\mathbb{E}e^{-tY_1})^{w}}e^{-t\lambda}\Big)^{\gamma-\alpha}
\end{aligned}$$

holds for any $t > 0$ and any $w \in [0,1]$. Choose $t = t^\star$. Then by the definition of $t^\star$ in Equation (3.3), we obtain $\mathbb{E}e^{t^\star X_1}\mathbb{E}e^{-t^\star Y_1} = e^{-I}$, and by the definition of λ in Equation (3.4), we have $\frac{(\mathbb{E}e^{t^\star X_1})^{1-w}}{(\mathbb{E}e^{-t^\star Y_1})^{w}}e^{-t^\star\lambda}$ exactly equal to 1. Thus $\mathbb{P}(T(\sigma) \geq T(\sigma_0)) \leq e^{-(\alpha\wedge\gamma)I}$.

A.3. Proof of Proposition 5.2. Without loss of generality we assume that $d_H(\sigma, \sigma_0) = d(\sigma, \sigma_0)$. Then σ assigns m nodes with different values from σ_0, and there are K possible values for each node. Thus

$$\Big|\Big\{\Gamma : \exists \sigma \in \Gamma \text{ s.t. } d(\sigma, \sigma_0) = m\Big\}\Big| \leq \binom{n}{m} K^m \leq \Big(\frac{enK}{m}\Big)^m.$$

In addition, since each node has at most K possible choices, we have a naive bound for the cardinality of Γ as $|\{\Gamma\}| \leq K^n$.

A.4. Proof of Theorem 2.1.

(1) For $K = 2$, the least favorable case for Θ is still Θ^0. The proof is identical to that of Theorem 2.2.

(2) For $K = 3$, it is always possible to have $\sigma \in \Theta$ such that a constant proportion of communities have size $\lfloor \frac{n}{\beta K} \rfloor$, and another constant proportion have size $\lceil \frac{n}{\beta K} \rceil$, with the rest communities have much larger size. Define Θ^L to contain all such σ. Then with identical arguments used to establish Lemma 5.1 and Lemma 5.2 we have

$$\begin{aligned}\inf_{\hat{\sigma}} \sup_{\Theta} \mathbb{E} r(\sigma, \hat{\sigma}) &\geq \inf_{\hat{\sigma}} \sup_{\sigma \in \Theta^L} B_\tau(\hat{\sigma}(1)) \\ &\geq c\mathbb{P}\Big(\sum_{u=1}^{\lfloor n/\beta K \rfloor} X_u \geq \sum_{u=1}^{\lfloor n/\beta K \rfloor} Y_u \Big) \\ &\geq \exp(-(1+o(1))nI/\beta K).\end{aligned}$$

A.5. Proof of Theorem 3.1 ($K = 2$). Without loss of generality we assume $\frac{n}{2} = \lfloor \frac{n}{2} \rfloor$ throughout this section. For arbitrary $\sigma, \sigma_0 \in \Theta$ with $d(\sigma, \sigma_0) = m$, we can define $\alpha(\sigma; \sigma_0)$ and $\gamma(\sigma; \sigma_0)$ the same way as in Section 5.2. Note that $m \leq \frac{n}{2}$ since $d(\sigma, \sigma_0) = \min\{d_H(\sigma, \sigma_0), n - d_H(\sigma, \sigma_0)\}$. Let $\{X_i\} \overset{iid}{\sim} \text{Ber}(\frac{b}{n})$ and $\{Y_i\} \overset{iid}{\sim} \text{Ber}(\frac{a}{n})$, and $\{X_i\} \perp \{Y_i\}$. By the proof of Proposition 5.1, we have

$$\mathbb{P}(T(\sigma) \geq T(\sigma)) \leq \mathbb{P}\left(\sum_{i=1}^{\gamma} X_i - \sum_{i=1}^{\alpha} Y_i \geq \lambda(\gamma - \alpha) \right).$$

Note that in $K = 2$ we have a specific equality as $\alpha + \gamma = m(n-m)$. Recall that $\lambda = -\frac{1}{2t^\star} \log\big(\frac{\frac{a}{n}\exp(-t^\star)+1-\frac{a}{n}}{\frac{b}{n}\exp(t^\star)+1-\frac{b}{n}} \big)$. By the Chernoff bound,

$$\begin{aligned}\mathbb{P}(T(\sigma) \geq T(\sigma_0)) &\leq \big(\mathbb{E}e^{t^\star X_i}\big)^\gamma \big(\mathbb{E}e^{-t^\star Y_i}\big)^\alpha e^{-t^\star \lambda(\gamma-\alpha)} \\ &= \big(\mathbb{E}e^{t^\star X_i}\mathbb{E}e^{-t^\star Y_i}\big)^{\frac{m(n-m)}{2}} \Big(\frac{\mathbb{E}e^{t^\star X_i}}{\mathbb{E}e^{-t^\star Y_i}} e^{-2t^\star \lambda'} \Big)^{\gamma - \frac{m(n-m)}{2}} \\ &= \exp\Big(-\frac{m(n-m)I}{2} \Big),\end{aligned}$$

where we use $\mathbb{E}e^{t^\star X_i}\mathbb{E}e^{-t^\star Y_i} = \exp(-I)$ and $e^{2t^\star \lambda'} = \frac{\mathbb{E}e^{t^\star X_i}}{\mathbb{E}e^{-t^\star Y_i}}$. The proof is similar to that of Theorem 3.2. Here we only include the key quantities and omit the details. Assume $0 < \epsilon < 1/8$. Consider the following three cases: *(1)* If $nI/2 > (1+\epsilon)\log n$, define $m_0 = 1$ and $m' = \epsilon n/2$. Then $P_1 \leq n\exp(-(n-1)I/2)$. Denote $R = n\exp(-(n-1)I/2)$. We have

$$P_m \leq \begin{cases} (\frac{2en}{2})^m \exp(-\frac{m(n-m)I}{2}) \leq Rn^{-\epsilon m/4}, & \text{for } m_0 < m \leq m' \\ (\frac{2en}{\epsilon n})^m \exp(-\frac{nmI}{4}) \leq R\exp(-\frac{n(m-4)I}{8}), & \text{for } m' < m \leq n/2. \end{cases}$$

Then $n\mathbb{E}r(\sigma,\hat{\sigma}) \leq \sum_{m=1}^{n/2} mP_m = (1+o(1))R$.
(2) If $nI/2 < (1-\epsilon)\log n$, define $m_0 = n\exp(-(1-e^{-\epsilon nI/2})nI/2)$ and $m' = n\exp(-nI/8)$. We have

$$P_m \leq \begin{cases} (\frac{2en}{m_0})^m \exp(-\frac{m(n-m')I}{2}) = \exp(-e^{-\frac{\epsilon nI}{2}}\frac{nmI}{4}), \text{ for } m_0 < m \leq m', \\ (\frac{2en}{m'})^m \exp(-\frac{nmI}{4}) \leq \exp(-\frac{nmI}{16}), \text{ for } m' < m \leq n/2. \end{cases}$$

Then $\mathbb{E}r(\sigma,\hat{\sigma}) \leq m_0/n + \sum_{m>m_0}^{n/2} P_m = (1+o(1))m_0/n$.
(3) If $\frac{nI}{2\log n} \to 1$, there exists a positive sequence $w \to 0$ such that $|\frac{nI}{2\log n} - 1| \ll w$ and $\frac{1}{\sqrt{\log n}} \leq w$. Define $m_0 = n\exp(-(1-w)nI/2)$ and $m' = w^2 n$.

$$P_m \leq \begin{cases} (\frac{2en}{m_0})^m \exp(-\frac{m(n-m')I}{2}) \leq \exp(-\frac{wnmI}{4}), \text{ for } m_0 < m \leq m' \\ (\frac{2en}{m'})^m \exp(-\frac{nmI}{4}) \leq \exp(-\frac{nmI}{8}), \text{ for } m' < m \leq n/2. \end{cases}$$

Then $\mathbb{E}r(\sigma,\hat{\sigma}) \leq m_0/n + \sum_{m>m_0}^{n/2} P_m = (1+o(1))m_0/n$.

A.6. Proof of Theorem 3.1 ($K \geq 3$). For the upper bound, we need the following lemma in replace of Lemma 5.3. Other than that, the proof is identical to that for Theorem 3.2 and thus omitted.

LEMMA A.1. *Assume $1 \leq \beta < \sqrt{\frac{5}{3}}$. Let $\sigma \in \Theta$ be an arbitrary assignment satisfying $d(\sigma,\sigma_0) = m$, where $0 < m < n$ is a positive integer. Then*

$$\alpha(\sigma;\sigma_0) \wedge \gamma(\sigma;\sigma_0) \geq \begin{cases} \frac{nm}{K\beta} - m^2, \ \textit{if } m \leq \frac{n}{2K}, \\ \frac{c_\beta nm}{K}, \ \textit{if } m > \frac{n}{2K}, \end{cases}$$

where $c_\beta = \frac{(5-3\beta^2)^2}{2\beta(1+3(5-3\beta^2)^2)}$.

PROOF OF LEMMA A.1. It is sufficient to show the equality for $\gamma(\sigma;\sigma_0)$. First consider the case $m \leq \frac{n}{2\beta K}$. Without loss of generosity, let σ satisfy

$$\sigma(i) = k, \forall i \in \left[\sum_{j=1}^{k-1} n'_j + 1, \sum_{j=1}^{k} n'_j\right].$$

Here $\{n'_k\}$ are sizes of all communities in σ. Assume $d_H(\sigma,\sigma_0) = m$, then $m = |\{i : \sigma(i) \neq \sigma_0(i)\}|$. Define $m_k = |\{i : \sigma(i) = k, \sigma_0(i) \neq k\}|$ then $m = \sum_k m_k$. For $k \in [K]$, define

$$\begin{aligned} \gamma_k(\sigma;\sigma_0) &= |\{(i,j) : \sigma(i) = \sigma(j) = k, \sigma_0(i) \neq \sigma_0(j), i < j\}| \\ &= \left|\left\{(i,j) : \sigma_0(i) \neq \sigma_0(j), \sum_{j=1}^{k-1} n'_j + 1 \leq i < j \leq \sum_{j=1}^{k} n'_j\right\}\right|. \end{aligned}$$

We see that $\gamma(\sigma;\sigma_0) = \sum_{k=1}^{K} \gamma_k(\sigma;\sigma_0)$. We have $m_k \leq \frac{n}{2\beta K} \leq \frac{n'_k}{2}$, and also $\gamma_k(\sigma;\sigma_0) \geq |\{i : \sigma(i) = k, \sigma_0(i) = k\}||\{i : \sigma(i) = k, \sigma_0(i) \neq k\}| = m_k(n_k - m_k)$. Then

$$\gamma(\sigma;\sigma_0) \geq \sum_k m_k(n_k - m_k) \geq \frac{mn}{\beta K} - m^2.$$

Now consider the case that $m > \frac{n}{2\beta K}$. Define $m_{k,k'} = |\{i : \sigma(i) = k, \sigma_0(i) = k'\}|$ for any $k, k' \in [K]$. We see that equations $m_k = \sum_{k' \neq k} m_{k,k'}$ and $n'_k = m_k + m_{k,k}$ and $n_{k'} = \sum_k m_{k,k'}$ hold for all $k, k' \in [K]$.

For each $k \in [K]$, we want to obtain the value of $\gamma_k(\sigma;\sigma_0)$. We divide $k \in [K]$ into the following three categories:

(1) We say $k \in \mathcal{K}_1$ if for all $k' \neq k$, $m_{k,k'} \leq \frac{2}{3}n'_k$. For a given m_k, we have

$$\frac{\gamma_k(\sigma;\sigma_0)}{n'_k m_k} = \frac{\frac{1}{2}(n'^2_k - \sum_{k'} m^2_{k,k'})}{n'_k m_k},$$

with $m_k = \sum_{k' \neq k} m_{k,k'}$. When $m_k \leq \frac{2}{3}n'_k$, it is easy to check

$$\frac{\gamma_k(\sigma;\sigma_0)}{n'_k m_k} \geq \frac{\frac{1}{2}(n'^2_k - (n'_k - m_k)^2 - m^2_k)}{n'_k m_k} = \frac{n'_k - m_k}{n'_k} \geq \frac{1}{3}.$$

When $m_k > \frac{2}{3}n'_k$,

$$\begin{aligned}\frac{\gamma_k(\sigma;\sigma_0)}{n'_k m_k} &\geq \frac{\frac{1}{2}(n'^2_k - (n'_k - m_k)^2 - (m_k - \frac{2}{3}n'_k)^2 - (\frac{2}{3}n'_k)^2)}{n'_k m_k} \\ &\geq \frac{m_k(n'_k - m_k) + \frac{2}{3}n'_k(m_k - \frac{2}{3}n'_k)}{n'_k m_k} \\ &\geq \frac{2}{9}.\end{aligned}$$

Thus $\gamma_k(\sigma;\sigma_0) \geq \frac{2nm_k}{9\beta K}$ in both cases.

(2) We say $k \in \mathcal{K}_2$ if exists $k' \neq k$ such that $m_{k,k'} > \frac{2}{3}n'_k$. Claim $m_{k',k'} > \frac{1}{3}n'_k$. Otherwise, from σ we can exchange the labels k and k' to obtain a new estimator σ'. This helps to correctly recover at least $m_{k,k'} - m_{k,k} - m_{k',k'} > \frac{2}{3}n'_k - \frac{1}{3}n'_k - \frac{1}{3}n'_k > 0$ more nodes. Since $\sigma' \in \Gamma(\sigma)$, this implies $m = d(\sigma_0, \sigma) \leq d_H(\sigma_0, \sigma') < d_H(\sigma_0, \sigma) = m$, which leads to a contradiction.

On the other hand, we have $m_{k'} = n'_{k'} - m_{k',k'} \geq n'_{k'} - (n_{k'} - m_{k,k'}) \geq \frac{n}{\beta K} - \frac{\beta n}{K} + \frac{2n}{3\beta K} \geq \frac{(5-3\beta^2)n}{3\beta K} > 0$. This implies

$$\frac{\gamma_k(\sigma;\sigma_0) + \gamma_{k'}(\sigma;\sigma_0)}{m_k + m_{k'}} \geq \frac{\gamma_{k'}(\sigma;\sigma_0)}{m_k + m_{k'}} \geq \frac{m_{k',k'} m_{k'}}{m_k + m_{k'}} \geq \frac{\frac{1}{3}n'_k}{\frac{m_k}{m_{k'}} + 1} \geq \frac{\frac{n}{3\beta K}}{\frac{\beta}{(5-3\beta)^2/(3\beta)} + 1}.$$

Thus we have $\gamma_k(\sigma;\sigma_0)+\gamma_{k'}(\sigma;\sigma_0) \geq \frac{2c_\beta n(m_k+m_{k'})}{K} \geq \frac{2c_\beta nm_k}{K}$.

Apparently $[K]=\mathcal{K}_1\cup\mathcal{K}_2$ and $\mathcal{K}_1\cap\mathcal{K}_2=\emptyset$. Claim for any $k\in\mathcal{K}_1$, there exists at most one $k'\neq k$ such that $m_{k',k}>\frac{2}{3}n'_{k'}$. Otherwise if there exists another $k''\neq k'$ such that $k''\neq k$ and $m_{k'',k}>\frac{2}{3}n'_{k''}$. Since $k',k''\in\mathcal{K}_2$, this leads to $m_{k,k}\geq\frac{1}{3}(n'_k\vee n'_{k'})$. Then $n_k\geq m_{k',k}+m_{k'',k}+m_{k,k}>n'_{k'}+\frac{2}{3}n'_{k''}\geq \frac{5n}{3\beta K}>\frac{\beta n}{K}$ which leads to a contradiction. Note that $c_\beta\leq\frac{2}{9}$. Thus

$$\begin{aligned}\gamma(\sigma;\sigma_0) &= \frac{1}{2}\sum_{k\in[K]}2\gamma_k(\sigma;\sigma_0)\\ &\geq \frac{1}{2}\left(\sum_{k\in\mathcal{K}_1}\frac{2nm_k}{9\beta K}+\sum_{k\in\mathcal{K}_2}\frac{2c_\beta nm_k}{K}\right)\\ &\geq \frac{c_\beta nm}{K}.\end{aligned}$$

□

APPENDIX B: ASYMPTOTIC EEQUIVALENCE OF I

LEMMA B.1. *Let a and b satisfy $0<a,b<n$ and $|a-b|/n\leq 1-c$ for any constant $1>c>0$. We have*

$$I \asymp \frac{(a-b)^2}{n\min\{a+b,2n-a-b\}}$$

When $b\leq a\leq(1-c)n$ we have $I\asymp(a-b)^2/(na)$. In addition if $a=o(n)$, we have $I=(1+o(1))(\sqrt{a}-\sqrt{b})^2/n$.

PROOF. Without loss of generality we assume $b\leq a$. Write

$$\begin{aligned}I &= -\log\left(\left(\sqrt{\frac{a}{n}\frac{b}{n}}+\sqrt{1-\frac{a}{n}}\sqrt{1-\frac{b}{n}}\right)^2\right)\\ &= -\log\left(1-\left(\frac{a}{n}+\frac{b}{n}\right)+2\sqrt{\frac{ab}{n^2}}\left(\sqrt{\left(1-\frac{a}{n}\right)\left(1-\frac{b}{n}\right)}+\sqrt{\frac{ab}{n^2}}\right)\right)\\ &= -\log\left(1-\left(\sqrt{\frac{a}{n}}-\sqrt{\frac{b}{n}}\right)^2-2\sqrt{\frac{ab}{n^2}}\left(1-\sqrt{\frac{ab}{n^2}}-\sqrt{\left(1-\frac{a}{n}\right)\left(1-\frac{b}{n}\right)}\right)\right).\end{aligned}$$

Note that $\sqrt{\frac{a}{n}\frac{b}{n}}+\sqrt{1-\frac{a}{n}}\sqrt{1-\frac{b}{n}}\geq\frac{b}{n}+1-\frac{a}{n}\geq c$ bounded away from 0, we have

$$I\asymp\left(\sqrt{\frac{a}{n}}-\sqrt{\frac{b}{n}}\right)^2+2\sqrt{\frac{ab}{n^2}}\left(1-\sqrt{\frac{ab}{n^2}}-\sqrt{\left(1-\frac{a}{n}\right)\left(1-\frac{b}{n}\right)}\right). \tag{B.1}$$

We consider the following two cases:
(1) When $1-b/n \geq a/n$, i.e. $n-b \geq a$, we have

$$1-\sqrt{\frac{ab}{n^2}}-\sqrt{\left(1-\frac{a}{n}\right)\left(1-\frac{b}{n}\right)} = \frac{\left(1-\sqrt{\frac{ab}{n^2}}\right)^2-\left(1-\frac{a}{n}\right)\left(1-\frac{b}{n}\right)}{1-\sqrt{\frac{ab}{n^2}}+\sqrt{\left(1-\frac{a}{n}\right)\left(1-\frac{b}{n}\right)}}$$

$$= \frac{\left(\sqrt{\frac{a}{n}}-\sqrt{\frac{b}{n}}\right)^2}{1-\sqrt{\frac{ab}{n^2}}+\sqrt{\left(1-\frac{a}{n}\right)\left(1-\frac{b}{n}\right)}},$$

where the denominator is in a constant order, which implies that there is some constant $c_1>0$,

$$I \asymp \left(1+2c_1\sqrt{\frac{ab}{n^2}}\right)\frac{(\sqrt{a}-\sqrt{b})^2}{n} \asymp \frac{(\sqrt{a}-\sqrt{b})^2}{n}. \tag{B.2}$$

Note that

$$\frac{(\sqrt{a}-\sqrt{b})^2}{n} = \frac{(a-b)^2}{n(\sqrt{a}+\sqrt{b})^2} \asymp \frac{(a-b)^2}{n(a+b)},$$

which yields $I \asymp \frac{(a-b)^2}{n(a+b)}$.
(2) If $b/n > 1-a/n$, i.e. $n-b<a$, we have

$$1-\sqrt{\frac{ab}{n^2}}-\sqrt{\left(1-\frac{a}{n}\right)\left(1-\frac{b}{n}\right)} = \frac{\left(1-\sqrt{\left(1-\frac{a}{n}\right)\left(1-\frac{b}{n}\right)}\right)^2-\frac{ab}{n^2}}{1-\sqrt{\left(1-\frac{a}{n}\right)\left(1-\frac{b}{n}\right)}+\sqrt{\frac{ab}{n^2}}}$$

$$= \frac{\left(\sqrt{1-\frac{a}{n}}-\sqrt{1-\frac{b}{n}}\right)^2}{1-\sqrt{\left(1-\frac{a}{n}\right)\left(1-\frac{b}{n}\right)}+\sqrt{\frac{ab}{n^2}}},$$

where the denominator is in a constant order. This implies that for some constant $c_2>0$,

$$I \asymp \left(1+2c_2\sqrt{\frac{ab}{n^2}}\right)\left(\sqrt{1-\frac{a}{n}}-\sqrt{1-\frac{b}{n}}\right)^2 \asymp \left(\sqrt{1-\frac{a}{n}}-\sqrt{1-\frac{b}{n}}\right)^2.$$

Note that

$$\left(\sqrt{1-\frac{a}{n}}-\sqrt{1-\frac{b}{n}}\right)^2 \asymp \frac{(\frac{a}{n}-\frac{b}{n})^2}{\left(\sqrt{1-\frac{a}{n}}+\sqrt{1-\frac{b}{n}}\right)^2} \asymp \frac{(a-b)^2}{n((n-b)+(n-a))},$$

which yields $I \asymp \frac{(a-b)^2}{n(2n-a-b)}$.

When $b \leq a \leq (1-c)n$, we immediately obtain $I \asymp (a-b)^2/(na)$ since $a \asymp (a+b) \lesssim (2n-a-b)$. In addition if $a = o(n)$, the proof is nearly identical with case (1), except that $(\sqrt{a}-\sqrt{b})^2/n = o(1)$ and $ab/n^2 = o(1)$. Note that I is equal to the right hand side of Equation (B.1) up to a $(1+o(1))$ factor. Then Equation (B.2) leads to $I = (1+o(1))(\sqrt{a}-\sqrt{b})^2/n$. □

Department of Statistics
Yale University
New Haven, CT 06511
E-mail: ye.zhang@yale.edu
E-mail: huibin.zhou@yale.edu
URL: http://www.stat.yale.edu/~hz68/