Detecting Communities on Colourful and Weighted Graphs

January 13, 2016

Abstract

A graph is known, edges and nodes,
communities hide within its folds.
Stochastic blocks the model is,
weighted edges a novel goal.

1 Preliminary

Suppose we have a graph of n nodes and each edge may take on any of L colors. The graph is generated by a stochastic block model with K clusters. Each within-cluster edge takes on color $l \in \{1, ..., L\}$ with probability P_l and each between-cluster edge with probability Q_l. We suppose that $P_l, Q_l \to 0$ so that the graph is sparse. Let n_k be the size of cluster k, we will suppose that $\frac{n}{\beta K} \leq n_k \leq \frac{\beta n}{K}$ for some $\beta \geq 1$.

The goal is to recover the clustering. We want to estimate $\widehat{\sigma} : \{1, ..., n\} \to \{1, ..., K\}$ such that $\min_{\tau \in S_K} d_H(\widehat{\sigma}, \tau \circ \sigma)$ is small where S_K is the permutation group over K elements.

We suppose that the number of colors L is finite and that the probabilities P_l, Q_l are unknown to us. We refer to the setting where P_l, Q_l are known as the *oracle setting*.

We define the Renyi divergence between P, Q as

$$I_{tot} = -2 \log \sum_l \sqrt{P_l Q_l}$$

Our goal is to show that all results (i.e., weak consistency rates, strong consistency thresholds) that hold under the oracle setting also hold when the distribution $\{P_l, Q_l\}$ is unknown.

2 Weak Recovery Under the Oracle Setting

[3] characterizes the minimax rate of weak recovery for Bernoulli P, Q. Their results and proofs can be extended in a straightforward manner to general discrete P, Q under the oracle setting.

Proposition 2.1. *(Oracle Setting Upper bound) Assume $\frac{nI_{tot}}{K \log K} \to \infty$. The maximum likelihood estimator $\widehat{\sigma}$ in the oracle setting achieves:*

$$\sup_{\Theta(n,K,\beta,P,Q)} \mathbb{E} r(\widehat{\sigma}, \sigma) \leq \begin{cases} \exp\left(-(1+o(1))\frac{nI_{tot}}{2}\right), & K = 2, \\ \exp\left(-(1+o(1))\frac{nI_{tot}}{\beta K}\right), & K \geq 3 \end{cases}$$

Proof. Proof of this proposition follows that of Theorem 3.2 in [3]. We describe only the parts that need to be modified.

Because we have a different likelihood function, our $T(\sigma)$ takes on a different form from that of [3] at the bottom of page 8:

$$T(\sigma) = \sum_{i<j} \mathbf{1}_{\sigma(i)=\sigma(j)} \sum_{l=0}^{L} \log \frac{P_l}{Q_l} \mathbf{1}_{A_{ij}=l}$$

Let σ_0 denote the true community assignment. We make a mistake if for some other community assignment σ, we get $T(\sigma) > T(\sigma_0)$. The key part of the proof is to bound

$$P_m \equiv P\left(\exists \sigma \, : \, T(\sigma) > T(\sigma_0), \, d_H(\sigma, \sigma_0) = m\right)$$

To that end, we bound the probability of error of a fixed σ m-distant from σ_0 in Hamming distance. We will prove an analogue of Proposition 5.1 in [3]:

Let σ be an arbitrary assignment satisfying $d(\sigma, \sigma_0) = m$. Let X_i, Y_i be random variables such that

$$X_i = \log \frac{P_l}{Q_l} \text{ w.p. } P_l \qquad Y_i = \log \frac{P_l}{Q_l} \text{ w.p. } Q_l$$

and α, γ be integers where

$$\alpha = |\{(i,j) : \sigma_0(i) = \sigma_0(j) \wedge \sigma(i) \neq \sigma(j)\}| \quad \gamma = |\{(i,j) : \sigma_0(i) \neq \sigma_0(j) \wedge \sigma(i) = \sigma(j)\}|$$

Then

$$P(T(\sigma) \geq T(\sigma_0)) \leq P\left(\sum_{i=1}^{\alpha} X_i - \sum_{i=1}^{\gamma} Y_i < 0\right) \leq \exp(-\frac{\gamma+\alpha}{2} I) \tag{2.1}$$

Lemma 5.3 from [3] bounds α, γ and Proposition 5.2 bounds the number of σ's (up to equivalent classes) such that $d_H(\sigma, \sigma_0) = m$. These pieces together bounds P_m. The rest of the proof follows [3] exactly starting from Page 16.

We devote the rest of the proof toward proving equation 2.1.

$$\begin{aligned}
T(\sigma_0) - T(\sigma') &= \sum_{i<j} \mathbf{1}_{\sigma_0(i)=\sigma_0(j) \wedge \sigma'(i) \neq \sigma'(j)} \sum_{l=1}^{L} \mathbf{1}_{A_{ij}=l} \log \frac{P_l}{Q_l} \\
&\quad - \sum_{i<j} \mathbf{1}_{\sigma_0(i)\neq\sigma_0(j) \wedge \sigma'(i) = \sigma'(j)} \sum_{l=1}^{L} \mathbf{1}_{A_{ij}=l} \log \frac{P_l}{Q_l} \\
&= \sum_{i=1}^{\alpha} X_i - \sum_{i=1}^{\gamma} Y_i
\end{aligned}$$

$$\begin{aligned}
P(\sum_{i=1}^{\gamma} Y_i - \sum_{i=1}^{\alpha} X_i > 0) &\leq \mathbb{E}\left(e^{-t\sum_{i=1}^{\alpha} X_i} e^{t\sum_{i=1}^{\gamma} Y_i}\right) \\
&\leq \mathbb{E}\left(e^{-tX_1\alpha} e^{tY_1\gamma}\right) \\
&\leq \left(\mathbb{E}e^{-tX_i}\mathbb{E}e^{tY_1}\right)^{(1-w)\alpha+w\gamma} \frac{\left(\mathbb{E}e^{tY_1}\right)^{(1-w)(\gamma-\alpha)}}{\left(\mathbb{E}e^{-tX_1}\right)^{w(\gamma-\alpha)}}
\end{aligned}$$

We will show that when $t = 1/2$ and $w = 1/2$, the fraction term equals 1 and the first term equals $\exp\left(-(1/2\alpha + 1/2\gamma)I\right)$.

Note that

$$\begin{aligned}
\mathbb{E}e^{-tX_1} &= \sum_l P_l e^{-t\log\frac{P_l}{Q_l}} \\
&= \sum_l P_l \left(\frac{Q_l}{P_l}\right)^t \\
&= \sum_l \sqrt{P_l Q_l} \quad (\text{if } t = 1/2)
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}e^{tY_1} &= \sum_l Q_l e^{t\log\frac{P_l}{Q_l}} \\
&= \sum_l Q_l \left(\frac{P_l}{Q_l}\right)^t \\
&= \sum_l \sqrt{P_l Q_l} \quad \text{if } t = 1/2
\end{aligned}$$

$$P(\sum_{i=1}^{\gamma} Y_i - \sum_{i=1}^{\alpha} X_i > 0) \leq \left(\sum_l \sqrt{P_l Q_l} \right)^{\alpha+\gamma}$$
$$\leq \exp(-\frac{1}{2} I)^{\alpha+\gamma}$$
$$\leq \exp\left(-\frac{(\alpha+\gamma)}{2} I \right)$$

□

3 Weak Recovery in the General Setting

Proposition 3.1. *If weak recovery (consistency) is possible under the oracle setting, then it is possible when P, Q are unknown.*

Proof. From proposition 2.1, we know that weak recovery is possible iff $\frac{nI_{tot}}{K \log K} \to \infty$.

From Lemma 9.1, we have

$$I_{tot} = (1 + o(1)) \sum_{l=1}^{L} (\sqrt{P_l} - \sqrt{Q_l})^2$$

Since $I_{tot} = \omega(\frac{K \log K}{n})$ by hypothesis, it must be that, for some l, $(\sqrt{P_l} - \sqrt{Q_l})^2 = \omega(\frac{K \log K}{n})$. We choose such an l and consider an estimator $\widehat{\sigma}_l$ that uses only the information $\mathbf{1}_{A_{ij}=l}$. Since the Renyi-divergence I_l of $Ber(P_l)$ and $Ber(Q_l)$ is

$$I_l = (1 + o(1)) \left((\sqrt{P_l} - \sqrt{Q_l})^2 + (\sqrt{1 - P_l} - \sqrt{1 - Q_l})^2 \right)$$

We have that $\frac{nI_l}{K \log K} \to \infty$ and weak consistency is thus achievable with $\widehat{\sigma}_l$.

□

Although weak consistency is achievable with the estimator $\widehat{\sigma}_l$ that considers only $\mathbf{1}_{A_{ij}=l}$, the estimator converges at $\exp(-\frac{nI_l}{\beta K})$ and therefore does not converge at the same rate. It is easy to see that $I_l \leq I_{tot} \leq L I_l (1 - o(1))$ where the second inequality holds as equality under some cases.

4 Rate Optimal Recovery

We propose the following algorithm to recover the clusters in the general setting. We proceed in two stages: first, we identify a color l that can provide consistent recovery by itself (such l must exist by proposition 3.1), and second, we use σ^l – a clustering based on l – to estimate $\{\widehat{P}_l, \widehat{Q}_l\}$ and then use the estimates to refine the clustering. The second stage closely follows the algorithm from [1].

Stage 1. Identify a consistent color

1. For each color l:
   (a) Perform spectral clustering on $\widetilde{A}_{ij} \equiv \mathbf{1}(A_{ij} = l)$ to get σ^l.

(b) Estimate $\widehat{P}_l, \widehat{Q}_l$ via counts from σ^l.

(c) Estimate $\sqrt{I_l}$ via $\sqrt{\widehat{I}_l} \equiv \frac{|\widehat{P}_l - \widehat{Q}_l|}{\sqrt{\widehat{P}_l \vee \widehat{Q}_l}}$.

2. Discard from L all colors l such that $\sqrt{\widehat{I}_l} \leq \sqrt{\frac{1}{n}}$.

3. Output l^* for which $\sqrt{\widehat{I}}$ is maximized

Stage 2. Refine clusters

1. For each node u:

(a) Use previously computed l^* to perform spectral clustering on $\mathcal{G}_{-u}$, get σ_u.

(b) Use σ_u to estimate $\widehat{P}_l, \widehat{Q}_l$.

(c) Assign $\widehat{\sigma}(u) = \arg\max_k \sum_{v\,:\,\sigma_u(v)=k} \sum_l \log \frac{\widehat{P}_l}{\widehat{Q}_l} \mathbf{1}(A_{uv} = l)$.

2. Run consensus. Output $\widehat{\sigma}$.

Proposition 4.1. *Suppose K is fixed and that $n\frac{I_{tot}}{K} \to \infty$. Suppose that $P_l \asymp Q_l$ for all l. Then the above procedure has error rate satisfying*

$$\lim_{n\to\infty} \sup_{\sigma_0, \{P_l, Q_l\}} P\left(l(\widehat{\sigma}, \sigma_0) \geq \exp\left(-(1-\eta)\frac{nI_{tot}}{\beta K} \right) \right) = 0$$

In particular, this shows that the threshold behavior (for the symmetric $K = 2$ case and finite colors L) that [2] demonstrates hold even if the P_l, Q_l's are not known.

The rest of the write-up constitute the proof of proposition 4.1. The proof proceeds in three steps. In the first step (section 5, we prove a general result that controls the estimation quality $|\widehat{P}_l - P_l|$ and $|\widehat{Q}_l - Q_l|$ where $\widehat{P}_l, \widehat{Q}_l$ are constructed from a consistent clustering algorithm. In the second step (section 7), we provide guarantee for the first stage of the our proposed algorithm. In the third step, we analyze the second stage of the proposed algorithm.

5 Estimation

Let σ_0 be the true clustering and $\widehat{\sigma}$ be a clustering algorithm.

Proposition 5.1. *Let $\sigma = \widehat{\sigma}(G)$ be a clustering of the graph with error rate γ. That is, $d_H(\sigma, \sigma_0) = \gamma n$. Let $\Delta_l = |P_l - Q_l|$. Suppose*

$$\gamma K \log K \to 0. \tag{5.1}$$

Let $\widehat{P}_l = \frac{\sum_{u,v\,:\,\sigma(u)=\sigma(v)} \mathbf{1}(A_{uv}=l)}{\sum_{u,v\,:\,\sigma(u)=\sigma(v)} 1}$ *and* $\widehat{Q}_l = \frac{\sum_{u,v\,:\,\sigma(u)\neq\sigma(v)} \mathbf{1}(A_{uv}=l)}{\sum_{u,v\,:\,\sigma(u)\neq\sigma(v)} 1}$ *be the MLE of* P_l *and* Q_l *based on* σ*. Then, uniformly for all* l *satisfying* $n\frac{\Delta_l^2}{(P_l \vee Q_l)} \to \infty$ *(i.e., consistency), we have that*

$$|\widehat{P}_l - P_l| = \eta\Delta_l \tag{5.2}$$

$$|\widehat{Q}_l - Q_l| = \eta\Delta_l \tag{5.3}$$

with probability at least $1 - Ln^{-(3+\delta)}$ *and for some* $\eta = o(1)$

Proposition 5.2. *Suppose*

$$\gamma K \log K = o\left(\frac{\Delta_l}{P_l \vee Q_l}\right) \tag{5.4}$$

$$\frac{n\Delta_l^3}{(P_l \vee Q_l)^2} \to 0 \tag{5.5}$$

or, alternatively, suppose that

$$\gamma K \log K = o\left(\frac{\Delta_l}{P_l \vee Q_l}\right)^2 \tag{5.6}$$

Then, we have that, uniformly for all l *satisfying the above assumptions,*

$$|\widehat{P}_l - P_l| = \eta\frac{\Delta_l^2}{P_l \vee Q_l} \tag{5.7}$$

$$|\widehat{Q}_l - Q_l| = \eta\frac{\Delta_l^2}{P_l \vee Q_l} \tag{5.8}$$

with probability at least $1 - Ln^{-(3+\delta)}$ *and for some* $\eta = o(1)$.

Note that 5.6 implies 5.5 since

$$\frac{(\sqrt{P_l} - \sqrt{Q_l})^2}{\Delta_l} \leq \frac{\Delta_l}{P_l \vee Q_l} \leq 1$$

Special attention has to be paid to P_0, Q_0. In this case, we will apply our assumptions on $1 - P_0, 1 - Q_0$ because $(\sqrt{P_0} - \sqrt{Q_0})^2 = o(\sqrt{1 - P_0} + \sqrt{1 - Q_0})^2$. Clearly, any analysis on the MLE's of $1 - P_0, 1 - Q_0$ automatically applies to the MLE's of P_0, Q_0.

5.1 Proof

We want to use a rough clustering σ to estimate the parameters. σ itself is a random variable dependent on the edge variables but we will analyze a fixed σ and then take the union bound.

Let σ_0 be the true clustering and σ be a rough one. Suppose that $d_H(\sigma, \sigma_0) = \gamma n$.

There are at most $\binom{n}{\gamma n}$ possible assignment σ's that satisfy the distance constraint.

$$\log\binom{n}{\gamma n} \leq \log\left(\frac{n(n-1)...(n-\gamma n+1)}{(\gamma n)!}\right) \leq \log\left(\frac{n^{\gamma n}e^{\gamma n}}{(\gamma n)^{\gamma n}}\frac{1}{\sqrt{2\pi\gamma n}}\right) \leq C\gamma n \log\frac{1}{\gamma}$$

5.1.1 Bias of $\widehat{P}_l$

Our estimator of P_L is

$$\widehat{P}_l = \frac{\sum_{i,j\,:\,\sigma(i)=\sigma(j)} \mathbf{1}(A_{ij} = l)}{\sum_{i,j\,:\,\sigma(i)=\sigma(j)}}$$

Because σ is an imperfect clustering, $\widehat{P}_l$ will be biased.

$$\mathbb{E}\widehat{P}_l = \frac{\sum_{i,j\,:\,\sigma(i)=\sigma(j)} \mathbf{1}(\sigma_0(i) = \sigma_0(j))P_l + \mathbf{1}(\sigma_0(i) \neq \sigma_0(j))Q_l}{\sum_{i,j\,:\,\sigma(i)=\sigma(j)}}$$

Thus, we have that, for any γ,

$$\begin{aligned} P_l \wedge Q_l \leq \mathbb{E}\widehat{P}_l \leq P_l \vee Q_l \\ P_l \wedge Q_l \leq \mathbb{E}\widehat{Q}_l \leq P_l \vee Q_l \end{aligned} \tag{5.9}$$

We will assume that $P_l \geq Q_l$ first. Then, $\mathbb{E}\widehat{P}_l \leq P_l$.
To get a lower bound of the bias, observe that

$$\begin{aligned} \frac{\sum_{i,j} \mathbf{1}(\sigma(i) = \sigma(j))\mathbf{1}(\sigma_0(i) \neq \sigma_0(j))}{\sum_{i,j} \mathbf{1}(\sigma(i) = \sigma(j))} &= \frac{\sum_k \sum_{i,j\,:\,\sigma(i)=\sigma(j)=k} \mathbf{1}(\sigma_0(i) \neq \sigma_0(j))}{\sum_k \widehat{n}_k(\widehat{n}_k - 1)} \\ &\leq \frac{\sum_k \gamma_k n \widehat{n}_k}{\sum_k \widehat{n}_k(\widehat{n}_k - 1)} \quad \left(\text{for } \sum_k \gamma_k = \gamma\right) \\ &\leq \frac{\max_k \widehat{n}_k \sum_k \gamma_k n}{\min_k \widehat{n}_k \sum_k (\widehat{n}_k - 1)} \\ &\leq \frac{\max_k \widehat{n}_k}{\min_k \widehat{n}_k} \gamma \end{aligned}$$

We define $\widehat{n}_k = \sum_i \mathbf{1}(\sigma(i) = k)$. We can bound the ratio term as follows:

$$\begin{aligned} \frac{\max_k \widehat{n}_k}{\min_k \widehat{n}_k} &\leq \frac{\frac{\beta n}{k} + \gamma n}{\frac{n}{\beta k} - \gamma n} \\ &\leq \frac{\beta^2 + k\gamma\beta}{1 - k\gamma\beta} \\ &\leq (\beta^2 + k\gamma\beta)(1 + o(1)) \quad \text{(assuming } \gamma \to 0) \end{aligned}$$

Therefore,

$$P_l - \beta^2\gamma(P_l - Q_l)(1 + o(1)) \leq \mathbb{E}\widehat{P}_l \leq P_l$$

In the event that $Q_l \geq P_l$, it is straightforward to check that

$$P_l \leq \mathbb{E}\widehat{P}_l \leq P_l + \beta^2\gamma(Q_l - P_l)(1 + o(1))$$

In any case, we have the following bound:

$$|\mathbb{E}\widehat{P}_l - P_l| \leq \beta^2\gamma\Delta_l(1 + o(1))$$

where $\Delta_l = |Q_l - P_l|$.

Under the assumption that $\gamma K \log K \to 0$, we immediately have that

$$|\mathbb{E}\widehat{P}_l - P_l| = o(\Delta_l). \tag{5.10}$$

Using our stronger assumption on γ (equation 5.5), we have that

$$|\mathbb{E}\widehat{P}_l - P_l| = o\left(\frac{\Delta_l^2}{P_l \vee Q_l}\right) \tag{5.11}$$

5.1.2 Variance of $\widehat{P}_l$

Having handled the bias, we now bound the deviation.

Let $\widetilde{A}_{ij} = \mathbf{1}(A_{ij} = l)$. Then, by Bernstein's inequality,

$$P\left(\left|\sum_{i,j\,:\,\sigma(i)=\sigma(j)} (\widetilde{A}_{ij} - \mathbb{E}\widetilde{A}_{ij})\right| > t\right) \leq 2\exp\left(-\frac{t^2}{2\sum_{i,j\,\sigma(i)=\sigma(j)} \mathbb{E}\widetilde{A}_{ij} + \frac{2}{3}t}\right)$$

We first bound $\sum_{i,j\,\sigma(i)=\sigma(j)} \mathbb{E}\widetilde{A}_{ij}$:

$$\begin{aligned}\sum_{i,j\,\sigma(i)=\sigma(j)} \mathbb{E}\widetilde{A}_{ij} &= \sum_k \widehat{n}_k(\widehat{n}_k - 1)\mathbb{E}\widehat{P}_l \\ &\leq (P_l \vee Q_l)\sum_k \widehat{n}_k(\widehat{n}_k - 1) \quad \text{(by Equation 5.9)}\end{aligned}$$

Therefore,

$$P\left(\left|\sum_{i,j\,:\,\sigma(i)=\sigma(j)} (\widetilde{A}_{ij} - \mathbb{E}\widetilde{A}_{ij})\right| > t\right) \leq 2\exp\left(-\frac{t^2}{2(P_l \vee Q_l)\sum_k \widehat{n}_k(\widehat{n}_k - 1) + \frac{2}{3}t}\right)$$

We want to bound the probability by $\exp(-C_1\gamma n \log\frac{1}{\gamma} - (3+\delta)\log n)$. Assuming that $\gamma \geq \frac{1}{n}$, we have that $\gamma \log \gamma^{-1} \geq \frac{1}{n}\log n$.

We choose t such that

$$\begin{aligned}t^2 &= 2(P_l \vee Q_l)\sum_k \widehat{n}_k(\widehat{n}_k - 1)\left(C_1\gamma n \log\frac{1}{\gamma} + (3+\delta)\log n\right) \vee \left(C_1\gamma n \log\frac{1}{\gamma} + (3+\gamma)\log n\right)^2 \\ &\leq 2(P_l \vee Q_l)\sum_k \widehat{n}_k(\widehat{n}_k - 1)C_\delta\gamma n \log\frac{1}{\gamma} \vee \left(C_\delta\gamma n \log\frac{1}{\gamma}\right)^2 \\ &\leq \left(\sqrt{2(P_l \vee Q_l)\sum_k \widehat{n}_k(\widehat{n}_k - 1)C_\delta\gamma n \log\frac{1}{\gamma}} + C_\delta\gamma n \log\frac{1}{\gamma}\right)^2\end{aligned}$$

It easy to check that, regardless of which among $\{2(P_l \vee Q_l)\sum_k \widehat{n}_k(\widehat{n}_k - 1), C_\delta\gamma n \log\frac{1}{\gamma}\}$ is larger, the probability term is at most $\exp(-C_1\gamma n \log\frac{1}{\gamma} - (3+\delta)\log n)$. Thus, with at most that probability:

$$|\widehat{P}_l - \mathbb{E}\widehat{P}_l| = \frac{\sum_{i,j\,\sigma(i)=\sigma(j)}(\widetilde{A}_{ij} - \mathbb{E}\widetilde{A}_{ij})}{\sum_{i,j}\mathbf{1}(\sigma(i)=\sigma(j))} > \frac{t}{\sum_{i,j}\mathbf{1}(\sigma(i)=\sigma(j))}$$

Substituting in the previous bound we had of t:

$$\begin{aligned}
\frac{t}{\sum_{i,j}\mathbf{1}(\sigma(i)=\sigma(j))} &\leq \frac{\sqrt{2(P_l\vee Q_l)\sum_k \widehat{n}_k(\widehat{n}_k-1)C_\delta\gamma n\log\frac{1}{\gamma}}+C_\delta\gamma n\log\frac{1}{\gamma}}{\sum_{i,j}\mathbf{1}(\sigma(i)=\sigma(j))}\\
&\leq \frac{\sqrt{2(P_l\vee Q_l)C_\delta\gamma n\log\frac{1}{\gamma}}}{\sqrt{\sum_k \widehat{n}_k(\widehat{n}_k-1)}}+\frac{C_\delta\gamma n\log\frac{1}{\gamma}}{\sum_k \widehat{n}_k(\widehat{n}_k-1)}\\
&\leq 2\frac{\sqrt{P_l\vee Q_l}\sqrt{C_\delta\gamma\log\frac{1}{\gamma}}}{\sqrt{\min_k \widehat{n}_k}}+\frac{C_\delta\gamma\log\frac{1}{\delta}}{\min_k \widehat{n}_k}\\
&\leq 2\sqrt{\frac{P_l\vee Q_l}{n}}\sqrt{C_\delta\beta k\gamma\log\frac{1}{\gamma}}+\frac{C_\delta\beta k\gamma\log\frac{1}{\gamma}}{n}
\end{aligned}$$

Under the assumption that $\frac{n\Delta_l^2}{P_l\vee Q_l}\to\infty$, we have that $\Delta_l\geq\sqrt{\frac{P_l\vee Q_l}{n}}$. Under the same assumption, it must be that $P_l\vee Q_l\geq\frac{1}{n}$. If we further apply the assumption that $\gamma K\log K\to 0$, we have that, with probability $1-Ln^{-(3+\delta)}$, for some $\eta=o(1)$, uniformly for all colors l satisfying $\frac{n\Delta_l^2}{P_l\vee Q_l}\to\infty$,

$$|\widehat{P}_l-\mathbb{E}\widehat{P}_l|=\eta\Delta_l$$

Now we apply the stronger assumptions of proposition 5.2 to get the stronger result.
We note that $k\gamma\log\frac{1}{\gamma}=o\left(\frac{(\sqrt{P}_l-\sqrt{Q}_l)^2}{\Delta_l}\right)$ under assumption 5.5.
The first term is

$$\begin{aligned}
&2\sqrt{\frac{P_l\vee Q_l}{n}}\sqrt{C_\delta\beta k\gamma\log\frac{1}{\gamma}}\\
&\leq 2\sqrt{\frac{P_l\vee Q_l}{n}}\sqrt{\frac{\Delta_l}{P_l\vee Q_l}} \quad \text{(by assumption 5.5)}\\
&\leq 2\sqrt{\frac{\Delta_l}{n}}\\
&= o\left(\frac{\Delta_l^2}{P_l\vee Q_l}\right) \quad \text{(assuming that } \frac{(P_l\vee Q_l)^2}{n\Delta_l^3}\to 0\text{, (5.4))}\\
&= o\left(\sqrt{P_l}-\sqrt{Q_l}\right)^2
\end{aligned}$$

We can repeat the above derivation using assumption 5.6 only instead of using both 5.5 and 5.4.

$$\begin{aligned}
&2\sqrt{\frac{P_l\vee Q_l}{n}}\sqrt{C_\delta\beta k\gamma\log\frac{1}{\gamma}}\\
&\leq 2\sqrt{\frac{P_l\vee Q_l}{n}}\frac{\Delta_l}{P_l\vee Q_l} \quad \text{(by assumption 5.6)}\\
&\leq 2\frac{\Delta_l}{\sqrt{nP_l\vee Q_l}}\\
&= o\left(\sqrt{P_l}-\sqrt{Q_l}\right)^2 \quad \left(\text{since } \frac{1}{\sqrt{n}}=o\left(\frac{\Delta_l}{\sqrt{P_l\vee Q_l}}\right)\right)
\end{aligned}$$

Likewise, we have, for the second term

$$\frac{C_\delta \beta k \gamma \log \frac{1}{\gamma}}{n} \leq \frac{\Delta_l}{n(P_l \vee Q_l)} = o\left(\frac{\Delta_l^2}{P_l \vee Q_l}\right)$$

Therefore, we have that

$$|\widehat{P}_l - \mathbb{E}\widehat{P}_l| = o\left(\sqrt{P_l} - \sqrt{Q_l}\right)^2 \tag{5.12}$$

with probability at least $1 - Ln^{-(3+\delta)}$ uniformly for all σ (satisfying $d_H(\sigma, \sigma_0) = \gamma n$) and for all colors l satisfying our two assumptions.

6 Controlling probability of misclassification

Now that we have control over $\widehat{P}_l$, we study the effect of plugging in these estimates into the refinement stage.

In this section, we will first make the simplifying assumption that all colors l satisfy assumptions 5.4 and 5.5 and furthermore,

$$P_l \asymp Q_l \tag{6.1}$$

Let σ_u be a clustering for all the nodes except u. Suppose that $d_H(\sigma_u, \sigma_0) = \gamma n$.

We assign u based on the criterion:

$$\operatorname*{argmax}_k \sum_{v\, \sigma_u(v)=k} \sum_l \log \frac{\widehat{P}_l}{\widehat{Q}_l} \mathbf{1}(A_{uv} = l)$$

Define $m'_k = \{v : \sigma_u(v) = k, \sigma_0(v) = k\}$, $m'_1 = \{v : \sigma_u(v) = 1, \sigma_0(v) = 1\}$ as the points correctly clustered by σ_u.

Proposition 6.1. *Suppose that σ_u is a clustering of all nodes except for u with error rate γ, that is, $d_H(\sigma_u, \sigma_0) = \gamma n$. Suppose that statements of proposition 5.1 holds.*

Then, we have that, with probability at least $1 - \exp\left(-(1 - o(1))\frac{n}{K} I^*\right)$,

$$\sigma_0(u) = \operatorname*{argmax}_k \sum_{v\, \sigma_u(v)=k} \sum_l \log \frac{\widehat{P}_l}{\widehat{Q}_l} \mathbf{1}(A_{uv} = l)$$

6.1 Proof

Suppose without the loss of generality that $\sigma_0(u) = 1$. We want to then control the probability that for some cluster k,

$$\sum_{v\,:\,\sigma_u(v)=k} \sum_l \log \frac{\widehat{P}_l}{\widehat{Q}_l} \mathbf{1}(A_{uv} = l) \geq \sum_{v\,:\,\sigma_u(v)=1} \sum_l \log \frac{\widehat{P}_l}{\widehat{Q}_l} \mathbf{1}(A_{uv} = l) \quad \text{(iff)}$$

$$\sum_{v\,:\,\sigma_u(v)=k} \bar{A}_{uv} - \sum_{v\,:\,\sigma_u(v)=1} \bar{A}_{uv} \geq 0 \quad \text{(a.e. upper bounded by)}$$

$$\left(\sum_{i=1}^{m'_k} \widetilde{Y}_i + \sum_{i=1}^{m_k - m'_k} \widetilde{X}_i\right) - \left(\sum_{i=1}^{m'_1} \widetilde{X}_i + \sum_{i=1}^{m_1 - m'_1} \widetilde{Y}_i\right) \geq 0 \quad \text{(iff)}$$

$$\exp(t \left(\sum_{i=1}^{m'_k} \widetilde{Y}_i + \sum_{i=1}^{m_k - m'_k} \widetilde{X}_i - \sum_{i=1}^{m'_1} \widetilde{X}_i - \sum_{i=1}^{m_1 - m'_1} \widetilde{Y}_i\right)) \geq 1$$

where $\bar{A}_{uv} \equiv \sum_l \log \frac{\widehat{P}_l}{\widehat{Q}_l} \mathbf{1}(A_{uv} = l)$ and $\widetilde{X}_i = \log \frac{\widehat{P}_l}{\widehat{Q}_l}$ with probability P_l and $\widetilde{Y}_i = \log \frac{\widehat{P}_l}{\widehat{Q}_l}$ with probability Q_l.

$$
\begin{aligned}
&P\left(\exp(t\left(\sum_{i=1}^{m_k'}\widetilde{Y}_i+\sum_{i=1}^{m_k-m_k'}\widetilde{X}_i-\sum_{i=1}^{m_1'}\widetilde{X}_i-\sum_{i=1}^{m_1-m_1'}\widetilde{Y}_i\right))\geq 1\right)\\
&\leq \mathbb{E}\left(\exp(t\left(\sum_{i=1}^{m_k'}\widetilde{Y}_i+\sum_{i=1}^{m_k-m_k'}\widetilde{X}_i-\sum_{i=1}^{m_1'}\widetilde{X}_i-\sum_{i=1}^{m_1-m_1'}\widetilde{Y}_i\right))\right)\\
&\leq \left(\mathbb{E}\exp(t\widetilde{Y}_i)\right)^{m_k'}\exp(t\widetilde{X}_i)^{m_1-m_1'}\left(\mathbb{E}\exp(-t\widetilde{X}_i)\right)^{m_1'}\exp(-t\widetilde{Y}_i)^{m_k-m_k'}\\
&\leq \left(\sum_l e^{t\log\frac{\widehat{P}_l}{\widehat{Q}_l}}Q_l\right)^{m_k'}\left(\sum_l e^{t\log\frac{\widehat{P}_l}{\widehat{Q}_l}}P_l\right)^{m_k-m_k'}\left(\sum_l e^{-t\log\frac{\widehat{P}_l}{\widehat{Q}_l}}P_l\right)^{m_1'}\left(\sum_l e^{-t\log\frac{\widehat{P}_l}{\widehat{Q}_l}}Q_l\right)^{m_1-m_1'}
\end{aligned}
$$

We will set $t=\frac{1}{2}$, in which case, we have:

$$
\begin{aligned}
&=\left(\sum_l\sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}}Q_l\right)^{m_k'}\left(\sum_l\sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}}P_l\right)^{m_k-m_k'}\left(\sum_l\sqrt{\frac{\widehat{Q}_l}{\widehat{P}_l}}Q_l\right)^{m_1-m_1'}\left(\sum_l\sqrt{\frac{\widehat{Q}_l}{\widehat{P}_l}}P_l\right)^{m_1'}\\
&=\left(\frac{\sum_l\sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}}P_l}{\sum_l\sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}}Q_l}\right)^{m_k-m_k'}\left(\frac{\sum_l\sqrt{\frac{\widehat{Q}_l}{\widehat{P}_l}}Q_l}{\sum_l\sqrt{\frac{\widehat{Q}_l}{\widehat{P}_l}}P_l}\right)^{m_1-m_1'} \qquad (6.2)\\
&\left(\sum_l\sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}}Q_l\right)^{m_k'-(m_1-m_1')}\left(\sum_l\sqrt{\frac{\widehat{Q}_l}{\widehat{P}_l}}P_l\right)^{m_1'-(m_k-m_k')} \qquad (6.3)
\end{aligned}
$$

We will bound term 6.2 and 6.3 separately.

Bound for Term 6.2.

First, note that

$$\frac{P_l}{Q_l}-1=\frac{P_l-Q_l}{Q_l}$$

We will show that $\frac{\widehat{P}}{\widehat{Q}}$ behaves similarly.

$$
\begin{aligned}
\frac{\widehat{P}_l}{\widehat{Q}_l}-1&=\frac{\widehat{P}_l-P_l+P_l}{\widehat{Q}_l-Q_l+Q_l}-1\\
&=\frac{\frac{\widehat{P}_l-P_l}{Q_l}+\frac{P_l}{Q_l}}{\frac{\widehat{Q}_l-Q_l}{Q_l}+1}-1\\
&=\left(\frac{P_l}{Q_l}+\frac{\widehat{P}_l-P_l}{Q_l}\right)\left(1-\frac{\widehat{Q}_l-Q_l}{Q_l}(1+o(1))\right)-1\\
&=\frac{P_l-Q_l}{Q_l}+o\left(\frac{\Delta_l}{Q_l}\right)\quad \text{(assuming } |\widehat{P}_l-P_l|=o(\Delta_l))
\end{aligned}
$$

Therefore,

$$\sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} - 1 = \sqrt{1 + \frac{P_l - Q_l}{Q_l} + o\left(\frac{\Delta_l}{Q_l}\right)} - 1$$
$$\begin{cases} \leq \frac{P_l - Q_l}{Q_l}(1 + o(1)) & (\text{if } P_l \geq Q_l) \\ \geq \frac{P_l - Q_l}{Q_l}(1 + o(1)) & (\text{if } P_l < Q_l) \end{cases}$$

Symmetry yields that

$$\sqrt{\frac{\widehat{Q}_l}{\widehat{P}_l}} - 1 \begin{cases} \leq \frac{Q_l - P_l}{P_l}(1 + o(1)) & (\text{if } Q_l \geq P_l) \\ \geq \frac{Q_l - P_l}{P_l}(1 + o(1)) & (\text{if } Q_l < P_l) \end{cases}$$

Now, we can bound term 6.2:

$$\begin{aligned} \left| 1 - \frac{\sum_l \sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} P_l}{\sum_l \sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} Q_l} \right| &= \frac{\sum_l \sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} (P_l - Q_l)}{\sum_l \sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} Q_l} \\ &= \sum_l \sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} (P_l - Q_l) \\ &= \sum_l \left(\sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} - 1 \right) (P_l - Q_l) \\ &\leq \sum_l \frac{\Delta_l^2}{Q_l}(1 + o(1)) = O(I^*) \quad \text{(assuming that } P_l \asymp Q_l) \end{aligned}$$

Identical analysis shows that

$$\left| 1 - \frac{\sum_l \sqrt{\frac{\widehat{Q}_l}{\widehat{P}_l}} Q_l}{\sum_l \sqrt{\frac{\widehat{Q}_l}{\widehat{P}_l}} P_l} \right| = O(I^*)$$

Therefore, term 6.2 can be bounded as

$$\begin{aligned} &\left(\frac{\sum_l \sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} P_l}{\sum_l \sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} Q_l} \right)^{m_k - m_k'} \left(\frac{\sum_l \sqrt{\frac{\widehat{Q}_l}{\widehat{P}_l}} Q_l}{\sum_l \sqrt{\frac{\widehat{Q}_l}{\widehat{P}_l}} P_l} \right)^{m_1 - m_1'} \\ &\leq \exp(O(I^*)(m_k - m_k' + m_1 - m_1')) \\ &\leq \exp(O(I^*)\gamma n) \\ &\leq \exp\left(\frac{n}{k} o(I^*)\right) \quad (\text{since } \gamma k \to 0) \end{aligned}$$

Bound for Term 6.3.
First, note that $(m'_k - (m_1 - m'_1)) - (m'_1 - (m_k - m'_k)) = m_k - m_1$.
Define $\widehat{I} = -\log\left(\sum_l \frac{\widehat{P}_l}{\widehat{Q}_l} Q_l\right)\left(\sum_l \frac{\widehat{Q}_l}{\widehat{P}_l} P_l\right)$. With this definition,

$$\left(\sum_l \sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} Q_l\right)^{m'_k-(m_1-m'_1)} \left(\sum_l \sqrt{\frac{\widehat{Q}_l}{\widehat{P}_l}} P_l\right)^{m'_1-(m_k-m'_k)}$$

$$= \exp(-\widehat{I})^{\frac{(m'_k-(m_1-m'_1))+(m'_1-(m_k-m'_k))}{2}} \left(\sum_l \sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} Q_l\right)^{\frac{m_k-m_1}{2}} \left(\sum_l \sqrt{\frac{\widehat{Q}_l}{\widehat{P}_l}} P_l\right)^{\frac{m_1-m_k}{2}}$$

We claim that the following three statements are true.

Claim 1 $m'_1 - (m_k - m'_k) \geq n_1 - 2\gamma n$ and likewise for $m'_k - (m_1 - m'_1)$

Claim 2 $\widehat{I} - I^* \geq -o(1)I^*$

Claim 3 $\left(\sum_l \sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} Q_l\right)^{\frac{m_k-m_1}{2}} \left(\sum_l \sqrt{\frac{\widehat{Q}_l}{\widehat{P}_l}} P_l\right)^{\frac{m_1-m_k}{2}} = \exp(\frac{n}{k} o(I^*)$

Let us first suppose that these statements are true and see that term 6.3 can be bounded.

$$\exp(-\widehat{I})^{\frac{(m'_k-(m_1-m'_1))+(m'_1-(m_k-m'_k))}{2}} \left(\sum_l \sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} Q_l\right)^{\frac{m_k-m_1}{2}} \left(\sum_l \sqrt{\frac{\widehat{Q}_l}{\widehat{P}_l}} P_l\right)^{\frac{m_1-m_k}{2}}$$

$$\leq \exp(-(I^* + (\widehat{I} - I^*))^{\frac{(m'_k-(m_1-m'_1))+(m'_1-(m_k-m'_k))}{2}} \left(\frac{\sum_l \sqrt{\frac{\widehat{Q}_l}{\widehat{P}_l}} P_l}{\sum_l \sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} Q_l}\right)^{\frac{m_1-m_k}{2}}$$

$$\leq \exp\left(-(1-o(1))I^*(n_1 - \gamma n)\right) \left(\frac{\sum_l \sqrt{\frac{\widehat{Q}_l}{\widehat{P}_l}} P_l}{\sum_l \sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} Q_l}\right)^{\frac{m_1-m_k}{2}} \quad \text{(by claim 1 and 2)}$$

$$\leq \exp\left(-(1-o(1))\frac{n}{\beta k} I^*\right) \quad \text{(by claim 3)}$$

The last inequality holds because $\gamma = o\left(\frac{1}{k \log k}\right)$. We will prove each of the three claims in the remainder of the proof.

Claim 1: This is straightforward. σ_u has at most γn errors and therefore, $m'_1 \geq n_1 - \gamma n$ and $m_k - m'_k \leq \gamma n$.

Claim 2: We show that the estimation error of $\widehat{P}_l, \widehat{Q}_l$ does not make $\widehat{I}$ too small.

$$\widehat{I} - I^* = -\log \frac{\left(\sum_l \sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} Q_l\right)\left(\sum_l \sqrt{\frac{\widehat{Q}_l}{\widehat{P}_l}} P_l\right)}{\left(\sum_l \sqrt{P_l Q_l}\right)^2} \tag{6.4}$$

Let us consider the numerator.

$$\begin{aligned}
&\left(\sum_l \sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} Q_l\right)\left(\sum_l \sqrt{\frac{\widehat{Q}_l}{\widehat{P}_l}} P_l\right) \\
&= \left(\sum_l \sqrt{P_l Q_l}\sqrt{\frac{\widehat{P}_l}{P_l}\frac{Q_l}{\widehat{Q}_l}}\right)\left(\sum_l \sqrt{P_l Q_l}\sqrt{\frac{P_l}{\widehat{P}_l}\frac{\widehat{Q}_l}{Q_l}}\right) \\
&= \sum_l P_l Q_l + \sum_{l,l'} \sqrt{P_l Q_l P_{l'} Q_{l'}} + \sum_{l<l'} \sqrt{P_l Q_l P_{l'} Q_{l'}}\left(\sqrt{T} + \frac{1}{\sqrt{T}} - 2\right) \\
&= \left(\sum_l \sqrt{P_l Q_l}\right)^2 + \sum_{l<l'} \sqrt{P_l Q_l P_{l'} Q_{l'}}\left(\sqrt{T} + \frac{1}{\sqrt{T}} - 2\right)
\end{aligned}$$

where we define $T = \frac{\widehat{P}_l}{P_l}\frac{Q_l}{\widehat{Q}_l}\frac{P_{l'}}{\widehat{P}_{l'}}\frac{\widehat{Q}_{l'}}{Q_{l'}}$. It will be later shown that $T \to 1$ and thus, continuing equation 6.4,

$$\begin{aligned}
\widehat{I} - I^* &= -\log\left(1 + \frac{\sum_{l<l'} \sqrt{P_l Q_l P_{l'} Q_{l'}}\left(\sqrt{T} + \frac{1}{\sqrt{T}} - 2\right)}{\left(\sum_l \sqrt{P_l Q_l}\right)^2}\right) \\
&\geq -\log\left(1 + 4\sum_{l<l'} \sqrt{P_l Q_l P_{l'} Q_{l'}}\left(\sqrt{T} + \frac{1}{\sqrt{T}} - 2\right)\right) \quad \text{(assuming that } \sum_l \sqrt{P_l Q_l} \geq 1/2\text{)} \\
&\geq -4\sum_{l<l'} \sqrt{P_l Q_l P_{l'} Q_{l'}}\left(\sqrt{T} + \frac{1}{\sqrt{T}} - 2\right) \qquad (6.5)
\end{aligned}$$

We proceed by first bounding $|T-1|$ and then taking the second order approximation of $\left(\sqrt{T} + \frac{1}{\sqrt{T}} - 2\right)$ around 1.

$$\begin{aligned}
|T-1| &= \left|\frac{\widehat{P}_l}{P_l}\frac{Q_l}{\widehat{Q}_l}\frac{P_{l'}}{\widehat{P}_{l'}}\frac{\widehat{Q}_{l'}}{Q_{l'}} - 1\right| \\
&= \left|\left(1 - \frac{P_l - \widehat{P}_l}{P_l}\right)\left(1 - \frac{\widehat{Q}_l - Q_l}{\widehat{Q}_l}\right)\left(1 - \frac{\widehat{P}_{l'} - P_{l'}}{\widehat{P}_{l'}}\right)\left(1 - \frac{Q_{l'} - \widehat{Q}_{l'}}{Q_{l'}}\right) - 1\right| \\
&\leq \left(\frac{|P_l - \widehat{P}_l|}{P_l} + \frac{|\widehat{Q}_l - Q_l|}{\widehat{Q}_l} + \frac{|\widehat{P}_{l'} - P_{l'}|}{\widehat{P}_{l'}} + \frac{|Q_{l'} - \widehat{Q}_{l'}|}{Q_{l'}}\right)
\end{aligned}$$

We use the assumption that $P_l \asymp Q_l$ and that $|\widehat{P}_l - P_l| = \eta\Delta_l$ for some $\eta = o(1)$.

We will also assume that $\frac{1}{2}P_l \leq \widehat{P}_l \leq 2P_l$.

Then, we have that

$$|T-1| \leq \eta\left(3\frac{\Delta_l}{P_l \vee Q_l} + 3\frac{\Delta_{l'}}{P_{l'} \vee Q_{l'}}\right)$$

The Taylor approximation of $\sqrt{T} + \frac{1}{\sqrt{T}} - 2$ around $T = 1$ is:

$$\begin{aligned}\sqrt{T} + \frac{1}{\sqrt{T}} - 2 &\leq \frac{1}{4}(T-1)^2 + O(T-1)^3 \\ &\leq \frac{1}{4}\eta\left(3\frac{\Delta_l}{P_l \vee Q_l} + 3\frac{\Delta_{l'}}{P_{l'} \vee Q_{l'}}\right)^2\end{aligned}$$

Continuing on from equation 6.5, we have that

$$\begin{aligned}\widehat{I} - I^* &\geq -4\sum_{l<l'}\sqrt{P_l Q_l P_{l'} Q_{l'}}\left(\sqrt{T} + \frac{1}{\sqrt{T}} - 2\right) \\ &\geq -\sum_{l<l'}\sqrt{P_l Q_l P_{l'} Q_{l'}}\eta\left(3\frac{\Delta_l}{P_l \vee Q_l} + 3\frac{\Delta_{l'}}{P_{l'} \vee Q_{l'}}\right)^2 \\ &\geq -\sum_{l<l'}\sqrt{P_l Q_l P_{l'} Q_{l'}}36\eta\left(\frac{\Delta_l}{P_l \vee Q_l}\right)^2 \quad \text{(assuming } \frac{\Delta_l}{P_l \vee Q_l} \geq \frac{\Delta_{l'}}{P_{l'} \vee Q_{l'}} \text{)} \\ &\geq -\sum_{l}\sum_{l'}36\eta\sqrt{P_{l'} Q_{l'}}\frac{\Delta_l^2}{P_l \vee Q_l} \\ &\geq -\eta\left(\sum_l \frac{\Delta_l^2}{P_l \vee Q_l}\right)\left(\sum_{l'} 36\sqrt{P_{l'} Q_{l'}}\right) \\ &\geq -o(I^*)\end{aligned}$$

The last inequality follows because $\sum_{l'}\sqrt{P_{l'} Q_{l'}} \leq 1$. This proves claim 2.
Claim 3.

$$\begin{aligned}&\left(\sum_l \sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}}Q_l\right)^{\frac{m_k - m_1}{2}}\left(\sum_l \sqrt{\frac{\widehat{Q}_l}{\widehat{P}_l}}P_l\right)^{\frac{m_1 - m_k}{2}} \\ &= \left(\sum_l \sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}}Q_l\right)^{\frac{m_k - m_1}{2}}\left(\sum_l \sqrt{\frac{\widehat{Q}_l}{\widehat{P}_l}}P_l\right)^{\frac{m_1 - m_k}{2}}\left(\frac{\sum_l \sqrt{\widehat{P}_l \widehat{Q}_l}}{\sum_l \sqrt{\widehat{P}_l \widehat{Q}_l}}\right)^{\frac{m_1 - m_k}{2}} \\ &= \left(\frac{\sum_l \sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}}Q_l}{\sum_l \sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}}\widehat{Q}_l}\right)^{\frac{m_k - m_1}{2}}\left(\frac{\sum_l \sqrt{\frac{\widehat{Q}_l}{\widehat{P}_l}}P_l}{\sum_l \sqrt{\frac{\widehat{Q}_l}{\widehat{P}_l}}\widehat{P}_l}\right)^{\frac{m_1 - m_k}{2}}\end{aligned}$$

Assume that $m_k \geq m_1$. The reverse case can be analyzed in identical manners. Then,

$$= \left(1 + \frac{\sum_l \sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}}(Q_l - \widehat{Q}_l)}{\sum_l \sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}}\widehat{Q}_l}\right)^{\frac{m_k - m_1}{2}}\left(1 + \frac{\sum_l \sqrt{\frac{\widehat{Q}_l}{\widehat{P}_l}}(\widehat{P}_l - P_l)}{\sum_l \sqrt{\frac{\widehat{Q}_l}{\widehat{P}_l}}P_l}\right)^{\frac{m_k - m_1}{2}}$$

The term $\sum_l \sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}}\widehat{Q}_l = \sum_l \sqrt{\widehat{P}_l\widehat{Q}_l} = \Theta(\sqrt{\widehat{P}_0\widehat{Q}_0}) = \Theta(1)$ since $|\widehat{P}_0 - P_0| = o(\sum_{l\neq 0}|P_l - Q_l|) \to 0$ and $P_0 \to 1$. Likewise, $\sum_l \sqrt{\frac{\widehat{Q}_l}{\widehat{P}_l}}P_l = \Theta\left(\sqrt{\frac{\widehat{Q}_0}{\widehat{P}_0}}P_0\right) = \Theta(1)$.

To bound the numerator term, we first note that

$$\left|\sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} - 1\right| \leq \frac{\Delta_l}{Q_l}$$

Therefore, we have that

$$\begin{aligned}\left|\sum_l \sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}}(Q_l - \widehat{Q}_l)\right| &= \left|\sum_l \left(\sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} - 1\right)(Q_l - \widehat{Q}_l)\right| \\ &\leq o\left(\sum_l \frac{\Delta_l^2}{Q_l}\right) \\ &\leq o(I^*)\end{aligned}$$

$$\begin{aligned}&\left(1 + \frac{\sum_l \sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}}(Q_l - \widehat{Q}_l)}{\sum_l \sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}}\widehat{Q}_l}\right)^{\frac{m_k - m_1}{2}} \left(1 + \frac{\sum_l \sqrt{\frac{\widehat{Q}_l}{\widehat{P}_l}}(\widehat{P}_l - P_l)}{\sum_l \sqrt{\frac{\widehat{Q}_l}{\widehat{P}_l}}P_l}\right)^{\frac{m_k - m_1}{2}} \\ &\leq \exp\left((m_k - m_1)\log(1 + o(I^*))\right) \\ &\leq \exp\left(\frac{n}{k}o(I^*)\right)\end{aligned}$$

This proves claim 3.

Multiplying the bounds for term 6.3 and 6.2 completes the proof.

7 Choosing the initial clustering

Recall that the stage 1 of our algorithm is as follows. Let τ be an input parameter.

1. For each color l:
   (a) Perform spectral clustering on $\widetilde{A}_{ij} \equiv \mathbf{1}(A_{ij} = l)$ to get σ^l.
   (b) Estimate $\widehat{P}_l, \widehat{Q}_l$ via counts from σ^l.
   (c) Estimate $\sqrt{I_l}$ via $\sqrt{\widehat{I}_l} \equiv \frac{|\widehat{P}_l - \widehat{Q}_l|}{\sqrt{\widehat{P}_l \vee \widehat{Q}_l}}$.
2. Discard from L all colors l such that $\sqrt{\widehat{I}_l} \leq \tau\sqrt{\frac{1}{n}}$.
3. Output l^* for which $\sqrt{\widehat{I}}$ is maximized

Assuming that $\frac{nI_{tot}}{K} \to \infty$, we want to say that the l^* we output satisfies $\frac{nI_{l^*}}{K} \to \infty$ and that the l's we discard satisfy $\limsup_{n\to\infty} n\frac{I_l}{K} < \infty$. It must be noted that the output of the initialization algorithm depends on n; that is, l^* and the discarded colors l depend on n. We will omit the dependence in our notation.

Both of these claims follow from proposition 7.1 below. The second claim follows directly from the second statement of proposition 7.1. To see that the first claim is also true, note that there must exist an l such that $\frac{nI_l}{K}$ by proposition 3.1 and furthermore, by the first statement of proposition 7.1, $\widehat{I}_l \asymp I_l$.

Proposition 7.1. *Suppose color l satisfies*

$$\Delta_l = \omega\left(\sqrt{\frac{K(P_l \vee Q_l)}{n}}\right)$$

.

Let σ^l be a spectral clustering of the graph based on $\widetilde{A}_{ij} = \mathbf{1}(A_{ij}) = l)$ and let $\widehat{P}_l, \widehat{Q}_l$ be estimates of P_l, Q_l constructed from σ^l. Then, with probability at least $1 - 2n^{-3+\delta}$ for some $\delta > 0$, there is a sequence $\eta \to 0$ such that

$$\frac{|P_l - Q_l|}{\sqrt{P_l \vee Q_l}}(\frac{1}{\sqrt{2}} - \eta) \leq \frac{|\widehat{P}_l - \widehat{Q}_l|}{\sqrt{\widehat{P}_l \vee \widehat{Q}_l}} \leq \frac{|P_l - Q_l|}{\sqrt{P_l \vee Q_l}}(\sqrt{2} + \eta)$$

On the other hand, supposing K is fixed and consider a color l such that

$$\Delta_l = O\left(\sqrt{\frac{P_l \vee Q_l}{n}}\right)$$

(which implies that σ^l is inconsistent), then, with probability at least $1 - n^{-3+\delta}$, there exists $\eta \to 0$ such that

$$\frac{\widehat{P}_l - \widehat{Q}_l}{\sqrt{\widehat{P}_l \vee \widehat{Q}_l}} \leq C'\frac{1}{\sqrt{n}}(1 + \eta)$$

where $C' \equiv \sqrt{2}(\limsup_{n\to} \sqrt{nI_l} + 1)$ is a constant.

Proof. Let color l be fixed and let γ be the error rate of the spectral clustering σ^l.

From the proof of estimation error of $\widehat{P}_l$ (5.1)

$$|\widehat{P}_l - P_l| \leq \beta^2 \gamma \Delta_l (1 + o(1)) + 2\sqrt{\frac{P_l \vee Q_l}{n}}\sqrt{CK\gamma \log \frac{1}{\gamma}} + \frac{K\gamma \log \frac{1}{\gamma}}{n}$$

with probability at least $1 - n^{-(3+\delta)}$ and likewise for $\widehat{Q}_l - Q_l$.

First, suppose $\Delta = \omega\left(\sqrt{\frac{K(P_l \vee Q_l)}{n}}\right)$. Then, by Theorem 10.1, we know that, with probability at least $1 - n^{-4}$, $\gamma K \log K \to 0$. We also assume without loss of generality that $P_l \geq Q_l$.

In this case, $\sqrt{\frac{P_l \vee Q_l}{n}} = o(\Delta_l)$. Also, $\Delta_l \geq \frac{1}{n}$ and so we have that $|\widehat{P}_l - P_l| = o(\Delta_l)$. Since $\Delta_l \leq P_l$, we also have that $\frac{1}{2}P_l \leq \widehat{P}_l \leq 2P_l$. Likewise, we have that $|\widehat{Q}_l - Q_l| = o(\Delta_l)$ and $\widehat{Q}_l \leq 2P_l$.

$$\begin{aligned}
\frac{|\widehat{P}_l - \widehat{Q}_l|}{\sqrt{\widehat{P}_l \vee \widehat{Q}_l}} &\geq \frac{|P_l - Q_l| - o(\Delta_l)}{\sqrt{2P_l}} \\
&\geq \frac{\Delta_l}{\sqrt{2P_l}}(1 - o(1))
\end{aligned}$$

$$\begin{aligned}
\frac{|\widehat{P}_l - \widehat{Q}_l|}{\sqrt{\widehat{P}_l \vee \widehat{Q}_l}} &\leq \frac{|P_l - Q_l| + o(\Delta_l)}{\sqrt{\frac{1}{2}P_l}} \\
&\leq \frac{\sqrt{2}\Delta_l}{\sqrt{P_l}}(1 + o(1))
\end{aligned}$$

Now, we turn to the second statement of the proposition and suppose $\Delta_l = O\left(\sqrt{\frac{P_l \vee Q_l}{n}}\right)$. With no assumption on γ, we must carefully examine the proof of proposition 5.1.

From the proof, it is clear that the bias $|\mathbb{E}\widehat{P}_l - P_l| \leq \Delta_l = O\left(\sqrt{\frac{P_l \vee Q_l}{n}}\right)$ and that the variance satisfies

$$\begin{aligned}
|\widehat{P}_l - \mathbb{E}\widehat{P}_l| &\leq \frac{\sqrt{2(P_l \vee Q_l)C_\delta \gamma n \log \frac{1}{\gamma}}}{\sqrt{\sum_k \widehat{n}_k(\widehat{n}_k - 1)}} + \frac{C_\delta \gamma n \log \frac{1}{\gamma}}{\sum_k \widehat{n}_k(\widehat{n}_k - 1)} \\
&\leq \frac{\sqrt{2(P_l \vee Q_l)C_\delta \gamma K \log \frac{1}{\gamma}}}{\sqrt{\max_k \widehat{n}_k}} + \frac{C_\delta \gamma K \log \frac{1}{\gamma}}{\max_k \widehat{n}_k} \\
&\leq \sqrt{\frac{P_l \vee Q_l}{n}}\sqrt{C_\delta \gamma K^2 \log \frac{1}{\gamma}} + \frac{C_\delta \gamma K^2 \log \frac{1}{\gamma}}{n}
\end{aligned}$$

Since $\frac{1}{n} \leq \sqrt{\frac{P_l \vee Q_l}{n}}$ by our assumption that $P_l \vee Q_l = \omega\left(\frac{1}{n}\right)$, we have that

$$|\widehat{P}_l - P_l| = O\left(\sqrt{\frac{P_l \vee Q_l}{n}}\right)$$

Since $\frac{P_l \vee Q_l}{n} \to 0$, we also have that $\widehat{P}_l \geq \frac{1}{2}P_l$. Therefore,

$$\begin{aligned}\frac{|\widehat{P}_l - \widehat{Q}_l|}{\sqrt{\widehat{P}_l \vee \widehat{Q}_l}} &\leq \frac{|P_l - Q_l| + O\left(\sqrt{\frac{P_l \vee Q_l}{n}}\right)}{\sqrt{\frac{1}{2}P_l}} \\ &\leq \sqrt{2}\frac{|P_l - Q_l|}{\sqrt{P_l}} + \sqrt{2}\sqrt{\frac{1}{n}}(1+\eta) \\ &\leq C'\sqrt{\frac{1}{n}}(1+\eta)\end{aligned}$$

□

Theorem 7.1. *Suppose* $\frac{nI_{tot}}{K} \to \infty$ *and let* L' *be the set of colors such that* $\limsup_{n\to\infty} \sqrt{nI_l} < \infty$.

Suppose the initialization algorithm is set with $\tau = \max_{l\in L'} \sqrt{2}(\limsup_{n\to\infty} nI_l + 2)$. *Then, the following holds with probability at least* $1 - 2Ln^{-(3+\delta)}$.

1. *Let* l^* *be the color outputted by the initialization algorithm. Then, for all large enough* n, *we have that* l^* *satisfies*
$$\frac{nI_{l^*}}{K} \to \infty$$
2. *A color* l *is discarded by the initialization algorithm iff* $l \in L'$.

Proof. By taking a union bound, we reason that conclusions of Proposition 7.1 hold uniformly for all l with probability at least $1 - 2Ln^{-(3+\delta)}$.

By proposition 3.1, there must exists a color l such that $\frac{nI_l}{K} \to \infty$. By proposition 7.1, with probability at least $1 - 2Ln^{-(3+\delta)}$,

$$\begin{aligned}\frac{|P_{l^*} - Q_{l^*}|}{\sqrt{P_{l^*} \vee Q_{l^*}}} &\geq \frac{1}{\sqrt{2}+\eta}\frac{|\widehat{P}_{l^*} - \widehat{Q}_{l^*}|}{\sqrt{\widehat{P}_{l^*} \vee \widehat{Q}_{l^*}}} \\ &\geq \frac{1}{\sqrt{2}+\eta}\frac{|\widehat{P}_l - \widehat{Q}_l|}{\sqrt{\widehat{P}_l \vee \widehat{Q}_l}} \\ &\geq \frac{|P_l - Q_l|}{\sqrt{P_l \vee Q_l}}\frac{1/\sqrt{2} - \eta}{\sqrt{2}+\eta}\end{aligned}$$

where we let n be large enough such that $|\eta| < 1/\sqrt{2}$. Therefore, we see that $\frac{nI_{l^*}}{K} \to \infty$.

Let us turn to the second claim of the theorem. If $l \in L'$, then it is clear by proposition 7.1 that l will be discarded.

If $l \notin L'$, then, by proposition 7.1, we have that

$$\frac{|\widehat{P}_l - \widehat{Q}_l|}{\sqrt{\widehat{P}_l \vee \widehat{Q}_l}} \geq (\frac{1}{\sqrt{2}} - \eta)\frac{|P_l - Q_l|}{\sqrt{P_l \vee Q_l}}$$

For large enough n, $\frac{|P_l - Q_l|}{\sqrt{P_l \vee Q_l}} > \tau\sqrt{\frac{1}{n}}$ and so l would not be discarded.

□

8 Continuous Distributions

In this section, we suppose that the weights of within-cluster edges are drawn from a density p and that of between-cluster edges are drawn from a density q.

$$A_{ij} \sim p \quad \text{if } \sigma_0(i) = \sigma_0(j)$$
$$A_{ij} \sim q \quad \text{if } \sigma_0(i) = \sigma_0(j)$$

In this case, the continuous Renyi divergence is

$$I = -2 \log \int \sqrt{p(x)q(x)} dx$$

Under the oracle setting, the rate of recovery is similar to that of proposition 2.1.

Proposition 8.1. *(Oracle Setting Upper Bound for Continuous Distributions) Assume* $\frac{nI}{K \log K} \to \infty$. *The maximum likelihood estimator* $\widehat{\sigma}$ *in the oracle setting achieves:*

$$\sup_{\Theta(n,K,\beta,P,Q)} \mathbb{E} r(\widehat{\sigma}, \sigma) \leq \begin{cases} \exp\left(-(1+o(1))\frac{nI}{2}\right), & K = 2, \\ \exp\left(-(1+o(1))\frac{nI}{\beta K}\right), & K \geq 3 \end{cases}$$

The proof is identical to that of proposition 2.1, replacing sums with integrals where needed.

8.1 Rate Optimal Recovery

Let us go to the general setting. Our goal is to show that under certain conditions, rate-optimal recovery is possible for continuous distributions.

8.1.1 Assumptions

A1 $p(x), q(x)$ are supported on $[0,1]$, and $p(x), q(x) \geq c > 0$.

A2 $p(x) - q(x) = \gamma(x)\alpha$ where $|\gamma(x)| \leq M$ for some constant M. $\alpha \to 0$.

A3 $|p'(x)|, |q'(x)|, |\gamma'(x)| \leq M'$ for some constant M'.

8.1.2 Continuous and Discretized Renyi Divergence

First, we show that, under the assumptions we have listed, the continuous Renyi divergence scales as α^2.

Proposition 8.2. *Let* $I = -2 \log \int \sqrt{p(x)q(x)} dx$ *be the continuous Renyi divergence. Suppose assumptions A1 and A2 are satisfied with constants* M, c. *We have that, with* $d = \int p(x) \left(\frac{1}{2} \frac{\gamma(x)}{p(x)}\right)^2 dx$,

$$I = d\alpha^2(1+\eta)$$

where, for any $\alpha < \frac{c}{4M}$, $|\eta| \leq \frac{12M}{c}\alpha$.

Proof. First, denote $H = \int(\sqrt{p(x)} - \sqrt{q(x)})^2 dx$ as the continuous Hellinger distance and we will first show that $c_1\alpha^2 \leq H \leq c_2\alpha^2$ for some constants c_1, c_2.

$$\begin{aligned}
&\int(\sqrt{p(x)} - \sqrt{q(x)})^2 dx \\
=&\int(\sqrt{p(x)} - \sqrt{p(x) - \gamma(x)\alpha})^2 dx \\
=&\int p(x)\left(1 - \sqrt{1 - \frac{\gamma(x)\alpha}{p(x)}}\right)^2 dx
\end{aligned}$$

Since $|\gamma(x)| \leq M$ and $p(x) \geq c$, we have that, for any $|\alpha| \leq \frac{c}{2M}$, $\left|\frac{\gamma(x)}{p(x)}\alpha\right| \leq \frac{1}{2}$. Thus, we can take Taylor series expansion of $f(z) = \sqrt{1-z}$ around 0 and plug in $\frac{\gamma(x)\alpha}{p(x)}$.

Therefore, there exists some function $\xi(\alpha, x)$ satisfying $|\xi(\alpha, x)| \leq \frac{4M}{c}\alpha$ such that

$$\begin{aligned}
=&\int p(x)\left(1 - (1 - \frac{1}{2}\frac{\gamma(x)\alpha}{p(x)}(1 + \xi(\alpha, x)))\right)^2 dx \\
=&\int p(x)\left(\frac{1}{2}\frac{\gamma(x)\alpha}{p(x)}(1 + \xi(\alpha, x))\right)^2 dx \\
=&\alpha^2\int p(x)\left(\frac{1}{2}\frac{\gamma(x)}{p(x)}\right)^2(1 + \xi(\alpha, x))^2 dx \\
=&\alpha^2\left(\int p(x)\left(\frac{1}{2}\frac{\gamma(x)}{p(x)}\right)^2 dx\right)(1 + \eta)
\end{aligned}$$

where $\eta = \frac{\int p(x)\left(\frac{1}{2}\frac{\gamma(x)}{p(x)}\right)^2(2\xi(x,\alpha) + \xi(x,\alpha)^2)dx}{\int p(x)\left(\frac{1}{2}\frac{\gamma(x)}{p(x)}\right)^2 dx}$. Since $|\xi(\alpha, x)| \leq \frac{4M}{c}\alpha \leq 1$,

$$\begin{aligned}
|\eta| &\leq \frac{1}{\int p(x)\left(\frac{1}{2}\frac{\gamma(x)}{p(x)}\right)^2 dx}\int p(x)\left(\frac{1}{2}\frac{\gamma(x)}{p(x)}\right)^2\frac{12M}{c}\alpha dx \\
&\leq \frac{12M}{c}\alpha
\end{aligned}$$

Now, it remains to bound I in terms of H. Since

$$\begin{aligned}
I =& -2\log\int\sqrt{p(x)q(x)}dx \\
=& -2\log\left(1 - \frac{1}{2}\int(\sqrt{p(x)} - \sqrt{q(x)})^2 dx\right) \\
=& -2\log(1 - \frac{1}{2}H)
\end{aligned}$$

If $\alpha \leq \frac{c}{4M}$, then $H \leq \frac{1}{16}$, thus, we have that

$$(1 + \eta')H \geq I \geq H$$

where $|\eta'| \leq \frac{1}{4}H \leq \frac{1}{4}\alpha^2 d(1+\eta)$.

This completes the proof with $d = \int p(x) \left(\frac{1}{2}\frac{\gamma(x)}{p(x)}\right)^2 dx$.

□

Next, we define the **discretized Renyi divergence**. Let the interval $[0,1]$ be divided into L equally spaced sub-intervals. Let l index each of these sub-intervals and let $B = 1/L$ be the length of each sub-interval.

Let $P_l = \int_{\text{Bin}_l} p(x)dx$ and $Q_l = \int_{\text{Bin}_l} q(x)dx$. We define the *discretized Renyi divergence* as $\widetilde{I} = -2\log\sum_{l=1}^{L}\sqrt{P_l Q_l}$ and $\widetilde{H} = \sum_{l=1}^{L}(\sqrt{P_l} - \sqrt{Q_l})^2$.

Proposition 8.3. *Let $L \geq 1$ be arbitrary and let $\widetilde{I}_L = -2\log\sum_{l=1}^{L}\sqrt{P_l Q_l}$ be the discretized Renyi divergence.*

Suppose assumptions A1 and A2 hold.

Let $\gamma_l = \int_{Bin_l}\gamma(x)dx$ and $d_L = \sum_l P_l\left(\frac{1}{2}\frac{\gamma_l}{P_l}\right)^2$

Then, we have that,

$$\widetilde{I}_L = d_L\alpha^2(1+\eta_L)$$

where, for any $\alpha < \frac{c}{2M}$, $|\eta_L| \leq \frac{8M}{c}\alpha$.

Proof. Again, we first analyze the discretized Hellinger distance. The discretized Renyi divergence can be bounded in terms of the discretized Hellinger distance in the same fashion as the continuous case.

$$\begin{aligned}
\widetilde{H} &= \sum_{l=1}^{L}(\sqrt{P_l} - \sqrt{Q_l})^2 \\
&= \sum_{l=1}^{L} P_l\left(1 - \sqrt{Q_l}P_l\right)^2 \\
&= \sum_{l=1}^{L} P_l\left(1 - \sqrt{1 - \frac{P_l - Q_l}{P_l}}\right)^2
\end{aligned}$$

We simplify the term $P_l - Q_l$.

$$\begin{aligned}
P_l - Q_l &= \int_{\text{Bin}_l} p(x) - q(x)dx \\
&= \int_{\text{Bin}_l}\gamma(x)\alpha dx \\
&= \gamma_l\alpha
\end{aligned}$$

where $|\gamma_l| \leq MB$.

$$P_l = \int_{\text{Bin}_l} p(x)dx \leq cB$$

For $\alpha < \frac{c}{2M}$, we have that $\frac{\gamma_l}{P_l}\alpha < \frac{1}{2}$ and thus, there exists η_l satisfying $|\eta_l| \leq \frac{4M}{c}\alpha$ such that

$$\begin{aligned}
\widetilde{H} &= \sum_{l=1}^{L} P_l \left(1 - \sqrt{1 - \frac{\gamma_l \alpha}{P_l}}\right)^2 \\
&= \sum_{l=1}^{L} P_l \left(\frac{1}{2}\frac{\gamma_l \alpha}{P_l}(1+\eta_l)\right)^2 \\
&= \alpha^2 \sum_{l=1}^{L} P_l \left(\frac{1}{2}\frac{\gamma_l}{P_l}\right)^2 (1+\eta_l)^2 \\
&= \alpha^2 \left(\sum_{l=1}^{L} P_l \left(\frac{1}{2}\frac{\gamma_l}{P_l}\right)^2\right)(1+\eta_L)
\end{aligned}$$

where $\eta_L = \frac{\sum_{l=1}^{L} P_l \left(\frac{1}{2}\frac{\gamma_l}{P_l}\right)^2 (2\eta_l + \eta_l^2)}{\sum_{l=1}^{L} P_l \left(\frac{1}{2}\frac{\gamma_l}{P_l}\right)^2}$. And since $|\eta_l| \leq \alpha \frac{4M}{c} < 1$, we have that

$$|\eta_L| \leq \frac{12M}{c}\alpha$$

The claim thus follows. □

Next, we will show that $\lim_{L\to\infty} d_L = d$.

Proposition 8.4. *Let* $d = \int p(x) \left(\frac{1}{2}\frac{\gamma(x)}{p(x)}\right)^2 dx$ *and* $d_L = \sum_l P_l \left(\frac{1}{2}\frac{\gamma_l}{P_l}\right)^2$. *Suppose assumptions A1, A3 hold. Then, we have that*

$$\lim_{L\to\infty} d_L = d$$

Note that $B \to 0$ is equivalent to $L \to \infty$ since $B = 1/L$.

Proof. Let $\mathrm{Bin}_l = [a_l, b_l]$.

$$\begin{aligned}
P_l &= \int_{\mathrm{Bin}_l} p(x)dx \\
&= \int_{a_l}^{b_l} p(x)dx \\
&= \int_{a_l}^{b_l} p(a_l) + p'(c_x)(x - a_l)dx \quad \text{for some } c_x \in [a_l, b_l] \\
&= Bp(a_l) + B^2\xi_l \quad \text{where } |\xi_l| \leq M'/2
\end{aligned}$$

Likewise, we have that $\gamma_l = B\gamma(a_l) + B^2\xi_l'$ for some ξ_l'.

$$\begin{aligned}
d_L &= \sum_{l=1}^{L} P_l \left(\frac{1}{2}\frac{\gamma_l}{P_l}\right)^2 \\
&= \sum_{l=1}^{L} B \left((p(a_l) + B\xi_l)\left(\frac{1}{2}\frac{\gamma(a_l) + B\xi_l'}{p(a_l) + B\xi_l}\right)^2\right)
\end{aligned}$$

Since $p(a_l) \geq c > 0$ is bounded away from 0 and $|\xi_l|, |\xi_l'| \leq M'$ is bounded away from ∞, we have that

$$= \sum_{l=1}^{L} Bp(a_l) \left(\frac{1}{2} \frac{\gamma(a_l)}{p(a_l)} \right)^2 (1 + O(B))$$

Thus, nothing that $\sum_{l=1}^{L} B = 1$, we have that

$$\begin{aligned} \lim_{B \to 0} d_L &= \lim_{B \to 0} \sum_{l=1}^{L} Bp(a_l) \left(\frac{1}{2} \frac{\gamma(a_l)}{p(a_l)} \right)^2 \\ &= \int p(x) \left(\frac{1}{2} \frac{\gamma(x)}{p(x)} \right)^2 dx \end{aligned}$$

Since $\frac{\gamma^2(x)}{p(x)}$ is Riemann integrable.

□

These propositions together imply the following:

Theorem 8.1. *Suppose assumptions A1-A3 are satisfied. Let $\widetilde{I}_L$ be the discretized Renyi divergence discretized at level L.*

*Let $n \to 0$, then, we have that, for **any sequences** $L_n \to \infty, \alpha_n \to 0$,*

$$\left| \frac{\widetilde{I}_L}{I} - 1 \right| \to 0$$

Proof. By proposition 8.2 and proposition 8.3, we have that, for all $\alpha < \frac{c}{4M}$,

$$\begin{aligned} |\widetilde{I}_L - I| &\leq \left| d_L \alpha^2 (1 + \eta_L) - d\alpha^2 (1 + \eta) \right| \\ &\leq d\alpha^2 \left| \frac{d_L}{d} (1 + \eta_L) - (1 + \eta) \right| \\ &\leq I \left| \frac{d_L}{d} \frac{(1 + \eta_L)}{(1 + \eta)} - 1 \right| \\ &\Rightarrow \\ \left| \frac{\widetilde{I}_L}{I} - 1 \right| &\leq \left| \frac{d_L}{d} \frac{(1 + \eta_L)}{1 + \eta} - 1 \right| \end{aligned}$$

where $|\eta|, |\eta_L| \leq \frac{12M}{c} \alpha$ for all L and d_L, d do not depend on α.

It is clear then that $\lim_{\alpha \to 0} \left| \frac{\widetilde{I}_L}{I} - 1 \right| = \left| \frac{d_L}{d} - 1 \right|$ and that $\lim_{L \to \infty} \lim_{\alpha \to 0} \left| \frac{\widetilde{I}_L}{I} - 1 \right| = 0$. We need to show additionally that the the convergence is uniform, that is, $\lim_{\alpha \to 0} \sup_L \left| \left| \frac{\widetilde{I}_L}{I} - 1 \right| - \left| \frac{d_L}{d} - 1 \right| \right| = 0$.

This is true since $\lim_{\alpha \to 0} \sup_L \eta_L = 0$ by proposition 8.3. The claim follows immediately. □

8.1.3 Recovery Procedure

1. For each edge (i,j), set $\widetilde{A}_{ij} = \mathbf{I}(A_{ij} < \tau)$ for some $\tau \in (0,1)$. Use the $\widetilde{A}_{ij}$ labels for the initial rough clustering.
2. Bin the interval $[0,1]$ into L bins and estimate P_l, Q_l.
3. Refine clustering with MLE based on $\widehat{P}_l, \widehat{Q}_l$.

The initial clustering is consistent if we assume that $\left|\int_0^\tau \gamma(x)dx\right| = c_1 > 0$. Let $\bar{P} = \int_0^\tau p(x)dx, \bar{Q} = \int_0^\tau q(x)dx$, $\bar{P} - \bar{Q} = c_1\alpha$. Furthermore, $\bar{P}, \bar{Q} \geq c$. Thus,

$$\bar{I} = -2\log\left(\sqrt{\bar{P}\bar{Q}} + \sqrt{(1-\bar{P})(1-\bar{Q})}\right) = \Theta(\alpha^2)$$

The discretization step is fine so long as $L \to \infty$.

8.2 Generalization

The assumptions listed in section 8.1.1 can be generalized in several ways. In the first way, γ can be a function of α under weak assumptions.

B1 $p(x), q(x)$ are supported on $[0,1]$, and $p(x), q(x) \geq c > 0$.

B2 $p(x) - q(x) = \gamma(x,\alpha)\alpha$ where $|\gamma(x,\alpha)| \leq M$ for some constant M. $\alpha \to 0$.

B3 For all α, $0 < c_1 \leq \int \frac{\gamma(x,\alpha)^2}{p(x)}dx \leq c_2 < \infty$.

B4 $|p'(x)|, |q'(x)|, |\partial_x\gamma(x,\alpha)| \leq M'$ for some constant M'.

Let us see a specific example in which these generalizations are relevant. Suppose that $f(x)$ is a density supported on $[0, 1-\alpha]$ and $g(x)$ is a density, bounded away from 0, and supported on $[0,1]$. We define $p(x), q(x)$ as mixtures.

$$\begin{aligned} p(x) &= \lambda f(x) + (1-\lambda)g(x) \\ q(x) &= \lambda f(x-\alpha) + (1-\lambda)g(x) \end{aligned}$$

Suppose that $|f'(x)|, |g'(x)| \leq M$, then, we have that

$$\begin{aligned} p(x) - q(x) &= \lambda(f(x) - f(x-\alpha)) \\ &= \lambda f'(c_{x,\alpha})\alpha \quad \text{for some } c_{x,\alpha} \text{ in between } x, x-\alpha \end{aligned}$$

Since $|f'(c_{x,\alpha})| \leq M$, condition $B2$ is satisfied. Condition $B3$ amounts to requiring that $\int \frac{(f'(c_{x,\alpha}))^2}{p(x)}dx$ be bounded away from 0 and ∞ for all α.

9 Technical Lemmas

Lemma 9.1. *Let $P = \{P_l\}_{l=0,\ldots,\infty}$ and $Q = \{Q_l\}_{l=0,\ldots,\infty}$ be two discrete distributions and suppose $P_0, Q_0 \to 1$. Let $I = -2\log\sum_l \sqrt{P_l Q_l}$.*

Then, we have that $I \to 0$ and

$$I = (1+o(1))\sum_{l=1}^{\infty}(\sqrt{P_l}-\sqrt{Q_l})^2$$

Proof. First, it is clear that if $P_0, Q_0 \to 1$, then

$$\begin{aligned}\sum_{l=0}^{\infty}(\sqrt{P_l}-\sqrt{Q_l})^2 &= (\sqrt{P_0}-\sqrt{Q_0})^2 + \sum_{l=1}^{\infty}(\sqrt{P_l}-\sqrt{Q_l})^2 \\ &= (\sqrt{P_0}-\sqrt{Q_0})^2 + \sum_{l=1}^{\infty}P_l + \sum_{l=1}^{\infty}Q_l - 2\sum_{l=1}^{\infty}\sqrt{P_l Q_l} \\ &\le (\sqrt{P_0}-\sqrt{Q_0})^2 + \sum_{l=1}^{\infty}P_l + \sum_{l=1}^{\infty}Q_l\end{aligned}$$

Therefore, $\lim_{n\to\infty}\sum_{l=0}^{\infty}(\sqrt{P_l}-\sqrt{Q_l})^2 = 0$.

$$\begin{aligned}I &= -2\log\sum_{l=0}^{\infty}\sqrt{P_l Q_l} \\ &= -2\log\left(1-\frac{1}{2}\sum_{l=0}^{\infty}(\sqrt{P_l}-\sqrt{Q_l})^2\right) \\ &= (1+o(1))\sum_{l=0}^{\infty}(\sqrt{P_l}-\sqrt{Q_l})^2 \quad \text{(since the sum tends to 0)}\end{aligned}$$

We will show that $(\sqrt{P_0}-\sqrt{Q_0})^2 = o\left(\sum_{l=1}^{\infty}(\sqrt{P_l}-\sqrt{Q_l})^2\right)$ and the result follows immediately.

Let $P' = 1-P_0$ and $Q' = 1-Q_0$.

$$\begin{aligned}(\sqrt{P_0}-\sqrt{Q_0})^2 &= (\sqrt{1-P'}+\sqrt{1-Q'})^2 \\ &= (1-P')\left(1-\sqrt{\frac{1-Q'}{1-P'}}\right)^2 \\ &= (1-P')\left(1-\sqrt{1-\frac{Q'-P'}{1-P'}}\right)^2 \\ &\le (1-P')\left(1-(1-\frac{1}{2}\left(\frac{Q'-P'}{1-P'}\right)(1+o(1)))\right)^2 \\ &\le (1-P')\left(\frac{1}{2}\left(\frac{Q'-P'}{1-P'}\right)(1+o(1))\right)^2 \\ &\le \frac{1}{4}\left(\frac{Q'-P'}{1-P'}\right)^2(1+o(1)) \le \frac{1}{4}(Q'-P')^2(1+o(1))\end{aligned}$$

$$\begin{aligned}
\sum_{l=1}^{\infty}(\sqrt{P_l}-\sqrt{Q_l})^2 &= \sum_{l=1}^{\infty} P_l + Q_l - 2\sqrt{P_l Q_l} \\
&\geq P' + Q' - 2\sqrt{\left(\sum_{l=1}^{\infty} P_l\right)\left(\sum_{l=1}^{\infty} Q_l\right)} \\
&= P' + Q' - 2\sqrt{P'Q'} \\
&= (\sqrt{P'}-\sqrt{Q'})^2 \\
&= P'\left(1-\sqrt{\frac{Q'}{P'}}\right)^2 \\
&= P'\left(1-\sqrt{1-\frac{P'-Q'}{P'}}\right)^2 \\
&\geq P'\left(1-(1-\frac{1}{2}\frac{P'-Q'}{P'}(1+o(1)))\right)^2 \\
&\geq P'\left(\frac{1}{2}\frac{P'-Q'}{P'}(1+o(1))\right)^2 \\
&\geq \frac{1}{4}\left(\frac{(P'-Q')^2}{P'}\right)(1+o(1))
\end{aligned}$$

Thus, we have shown that

$$\begin{aligned}
(\sqrt{P_0}-\sqrt{Q_0})^2 &\leq \frac{1}{4}(Q'-P')^2(1+o(1)) \\
\sum_{l=1}^{\infty}(\sqrt{P_l}-\sqrt{Q_l})^2 &\geq \frac{1}{4}\frac{(P'-Q')^2}{P'}(1+o(1))
\end{aligned}$$

Since $P' \to 0$, the proof is complete.

□

10 Reference Results

10.1 Existing Results from Literature

Let $\Theta_0(n,k,p,q,\beta)$ be the parameter space of homogeneous stochastic block model with p as the within-cluster probability and q as the between-cluster probability. The following theorem follows from Theorem 3 and Proposition 1 of [1].

Theorem 10.1. *Assume $p \leq C_1 q$ and that $p, q = \Omega(\frac{1}{n})$ and suppose that $n \geq 2\beta k$. Suppose there is some $c \in (0,1)$ such that*

$$\frac{k^3 p}{(p-q)^2 n^2} \leq c$$

.

Suppose we apply Unnormalized-Spectral-Clustering with trim constant $\tau = C_2\bar{d}$ and a sufficiently small post-processing constant $\mu > 0$. Then, for any constant C', there exists some $C > 0$ dependent only on C', C_1, C_2, μ such that

$$l(\widehat{\sigma}, \sigma_0) \leq C \frac{\beta^2 k^2 p}{(p-q)^2 n}$$

with probability at least $1 - n^{-C'}$.

We note that $\frac{(p-q)^2}{p} = I$. Restated, this theorem says that if $p \asymp q$ and if $\frac{k}{nI} \to 0$, then, the error rate γ of spectral clustering goes to zero with probability $1 - n^{-C'}$ for any constant $C' > 0$.

References

[1] Chao Gao, Zongming Ma, Anderson Y Zhang, and Harrison H Zhou. Achieving optimal misclassification proportion in stochastic block model. *arXiv preprint arXiv:1505.03772*, 2015.

[2] Varun Jog and Po-Ling Loh. Information-theoretic bounds for exact recovery in weighted stochastic block models using the renyi divergence. *arXiv preprint arXiv:1509.06418*, 2015.

[3] Anderson Y Zhang and Harrison H Zhou. Minimax rates of community detection in stochastic block model.