A pseudo-likelihood approach to community detection in weighted networks

Andressa Cerqueira and Elizaveta Levina

Federal University of São Carlos and University of Michigan

Supplementary Material

This Supplementary Material includes the detailed proofs of Theorems 1 and 2.

S1 Proofs of Theorem 1 and 2

The main idea behind the proof of the Theorem 1 is to bound the probability of misclassification of each node under mild conditions over the parameters estimates $\widehat{a}$, $\widehat{b}$, $\widehat{\sigma}^2$ and the initial labeling e. The next lemma states how the labels assigned for each node by the pseudo-likelihood algorithm is related with the block sums defined in (2.6).

Lemma 1. *Assume that* $\pi_1 = \cdots = \pi_K = \frac{1}{K}$. *Consider the initial labeling* $e \in \mathcal{E}_\gamma$ *and let* $\hat{c}(e)$ *be the estimate of the labels obtained by* (3.14). *For any* $i \in \{1, \ldots, n\}$, $l \in \{1, \ldots, K\}$ *and initial parameter estimates* $\widehat{a}, \widehat{b}, \widehat{\sigma}^2 \in S_{a,b}$, *we have that*

1. *Given* $c_i = l$, *for* $\gamma \in (0, 1/K)$, $a < b$ *or* $\gamma \in (1/K, 1)$, $a > b$

$$\widehat{c}_i(e) = l\,, \text{ if and only if } s_{il}(e) - s_{ik}(e) > 0\,, \text{ for all } k \neq l\,.$$

2. *Given* $c_i = l$, *for* $\gamma \in (0, 1/K)$, $a > b$ *or* $\gamma \in (1/K, 1)$, $a < b$

$$\widehat{c}_i(e) = l\,, \text{ if and only if } s_{il}(e) - s_{ik}(e) < 0\,, \text{ for all } k \neq l\,.$$

Proof. Assume that $c_i = l$, for $i \in \{1, \ldots, n\}$ and $l \in \{1, \ldots, K\}$. For $e \in \mathcal{E}_\gamma$, $\gamma \in (0, 1)$, the (r,s)-th entry of the confusion matrix $R(e)$ is given by

$$R_{rs} = \begin{cases} \dfrac{\gamma}{K}, \text{ if } r = s \\ \\ \dfrac{(1-\gamma)}{K(K-1)}, \text{ if } r \neq s\,. \end{cases} \tag{S1.1}$$

Let $\widehat{a}, \widehat{b}, \widehat{\sigma}^2 \in S_{a,b}$ be the estimates of the parameters a, b, σ^2. Using the symmetry property of $\widehat{B}$, $\widehat{\Sigma}$ and R, we have that the (r,s)-th entry of $\widehat{P} = n(R(e)\widehat{B})^T$ and $\widehat{\Lambda} = n(R(e)\widehat{\Sigma})^T$ is given by

$$\widehat{P}_{rs} = \begin{cases} \dfrac{n}{K}(\,\widehat{a}\gamma + \widehat{b}(1-\gamma)\,), \text{ if } r = s \\ \\ \dfrac{n}{K(K-1)}(\,\widehat{a}(1-\gamma) + \widehat{b}(K-2+\gamma)\,), \text{ if } r \neq s \end{cases} \tag{S1.2}$$

and

$$\widehat{\Lambda}_{rs} = \frac{n}{K}\widehat{\sigma}^2\,. \tag{S1.3}$$

For any $l \in \{1, \dots, K\}$, we have that

$$\widehat{\pi}_l = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} \mathbb{1}\{e_i = l, c_i = k\} = \sum_{k=1}^{K} R_{lk} = \frac{1}{K}\,.$$

Writing $s_{im}(e) = s_{im}$, the label update of node i, given in (3.14), is obtained by

$$\widehat{c}_i(e) = \arg\min_k \left\{ \sum_{m=1}^{K} \frac{(s_{im} - \widehat{P}_{km})^2}{2\widehat{\Lambda}_{km}} + \frac{1}{2}\log\widehat{\Lambda}_{km} \right\}. \tag{S1.4}$$

Using the estimator (S1.4), we conclude that $\widehat{c}_i(e) = l$, if and only if, for all $k \neq l$

$$\sum_{m=1}^{K} \frac{(s_{im} - \widehat{P}_{lm})^2}{2\widehat{\Lambda}_{lm}} - \sum_{m=1}^{K} \frac{(s_{im} - \widehat{P}_{km})^2}{2\widehat{\Lambda}_{km}} < \frac{1}{2}\sum_{m=1}^{K}(\log\widehat{\Lambda}_{km} - \log\widehat{\Lambda}_{lm})\,. \tag{S1.5}$$

Notice that the RHS of (S1.5) is equal to 0 by (S1.3). By symmetry of $\widehat{P}$, the LRS of (S1.5) can be written as

$$
\begin{aligned}
&\frac{(s_{il}-\widehat{P}_{ll})^2}{2\widehat{\Lambda}_{ll}}-\frac{(s_{ik}-\widehat{P}_{kk})^2}{2\widehat{\Lambda}_{kk}}+\frac{(s_{ik}-\widehat{P}_{lk})^2}{2\widehat{\Lambda}_{lk}}-\frac{(s_{il}-\widehat{P}_{kl})^2}{2\widehat{\Lambda}_{kl}}+\sum_{\substack{m=1\\ m\neq l,k}}^{K}\frac{(s_{im}-\widehat{P}_{lm})^2}{2\widehat{\Lambda}_{lm}}-\frac{(s_{im}-\widehat{P}_{km})^2}{2\widehat{\Lambda}_{km}}\\
&\quad=\frac{(s_{il}-\widehat{P}_{ll})^2}{2\widehat{\Lambda}_{ll}}-\frac{(s_{ik}-\widehat{P}_{ll})^2}{2\widehat{\Lambda}_{ll}}+\frac{(s_{ik}-\widehat{P}_{lk})^2}{2\widehat{\Lambda}_{lk}}-\frac{(s_{il}-\widehat{P}_{lk})^2}{2\widehat{\Lambda}_{lk}}\\
&\quad=\frac{1}{2\widehat{\Lambda}_{ll}}(s_{il}-s_{ik})(s_{il}+s_{ik}-2\widehat{P}_{ll})+\frac{1}{2\widehat{\Lambda}_{lk}}(s_{ik}-s_{il})(s_{ik}+s_{il}-2\widehat{P}_{lk})\\
&\quad=\frac{K}{2n\widehat{\sigma}^2}(s_{il}-s_{ik})\left(-2\widehat{P}_{ll}+2\widehat{P}_{lk}\right).
\end{aligned}
\tag{S1.6}
$$

By (S1.5) and (S1.6) we conclude that $\widehat{c}_i(e)=l$, if and only if, for all $k\neq l$

$$
\frac{K}{2n\widehat{\sigma}^2}(s_{il}-s_{ik})\left(-2\widehat{P}_{ll}+2\widehat{P}_{lk}\right)<0\,. \tag{S1.7}
$$

Inequality (S1.7) holds in the following cases:

(C_1) $\widehat{P}_{lk}<\widehat{P}_{ll}$ and $s_{il}(e)-s_{ik}(e)>0$

(C_2) $\widehat{P}_{lk}>\widehat{P}_{ll}$ and $s_{il}(e)-s_{ik}(e)<0$.

By (S1.2) and (S1.3), we can write

$$
\begin{aligned}
\widehat{P}_{lk}-\widehat{P}_{ll}&=\frac{n}{K}\left(\widehat{a}\left(\frac{1-\gamma K}{K-1}\right)+\widehat{b}\left(\frac{-1+\gamma K}{K-1}\right)\right)\\
&=\frac{n}{K(K-1)}(\widehat{a}-\widehat{b})(1-\gamma K)\,.
\end{aligned}
\tag{S1.8}
$$

In this case, we have that

(C_3) $\widehat{P}_{lk}<\widehat{P}_{ll}$ if $\widehat{a}<\widehat{b}$, $\gamma\in(0,1/K)$ or $\widehat{a}>\widehat{b}$, $\gamma\in(1/K,1)$.

(C_4) $\widehat{P}_{lk}>\widehat{P}_{ll}$ if $\widehat{a}>\widehat{b}$, $\gamma\in(0,1/K)$ or $\widehat{a}<\widehat{b}$, $\gamma\in(1/K,1)$.

By the definition of the set $S_{a,b}$ in (3.15) we have that if $a<b$ then $\widehat{a}<\widehat{b}$. Thus, combining (C_1), (C_2), (C_3) and (C_4) the result follows. □

The next proposition gives an upper bound on the probability of misclassification of each node in the network. This proposition shows that the upper bound does not depend on the initial parameter estimates $\widehat{a},\widehat{b}$ and $\widehat{\sigma}^2$.

Proposition 1. *Assume that* $\pi_1 = \cdots = \pi_K = \frac{1}{K}$. *Consider the initial labeling* $e \in \mathcal{E}_\gamma$ *and let* $\hat{c}(e)$ *be the estimate of the labels obtained by* (3.14). *For any node* $i \in \{1, \ldots, n\}$, $a, b \in \mathbb{R}$, $a \neq b$, $\sigma^2 > 0$, $\gamma \in (0,1)$, $\gamma \neq \frac{1}{K}$, *we have that*

$$\mathbb{P}(\,\widehat{c}_i(e) \neq c_i\,) \leq (K-1)\exp\left\{-\frac{1}{4}\frac{(\gamma K - 1)^2}{K(K-1)^2}\frac{n(a-b)^2}{\sigma^2}\right\}. \tag{S1.9}$$

Proof. Without loss of generality we assume that $c_i = l$, for $i \in \{1, \ldots, n\}$ and $l \in \{1, \ldots, K\}$. Conditioning on $c_i = l$, we have that

$$S_{il}(e) - S_{ik}(e) \sim \mathscr{N}(P_{ll} - P_{lk}\,,\, \Lambda_{ll} + \Lambda_{lk})\,, \tag{S1.10}$$

for all $k \in \{1, \ldots, K\}$ and $k \neq l$.

The matrix $\Lambda = n(R(e)\Sigma)^T$ can be obtained as in (S1.3) using the entries σ^2 of the matrix Σ. Thus,

$$\Lambda_{ll} + \Lambda_{lk} \;=\; \frac{2n\sigma^2}{K}\,. \tag{S1.11}$$

In the same way, the entries of the matrix $P = n(R(e)B)^T$ can be obtained as in (S1.8) and we get

$$P_{ll} - P_{lk} = \frac{n}{K(K-1)}(a-b)(\gamma K - 1)\,. \tag{S1.12}$$

Using the Chernoff tail bound for Gaussian random variables (Boucheron et al., 2013, p. 22) with mean given by (S1.12) and variance given by (S1.11), we obtain

$$\mathbb{P}\left(S_{il}(e) - S_{ik}(e) \geq n(a-b)\left(\frac{(\gamma K-1)}{K(K-1)}\right) + t\right) \leq \exp\left\{-\frac{Kt^2}{4n\sigma^2}\right\}, \quad \text{if} \quad t \geq 0 \tag{S1.13}$$

and

$$\mathbb{P}\left(S_{il}(e) - S_{ik}(e) \leq n(a-b)\left(\frac{(\gamma K-1)}{K(K-1)}\right) - t\right) \leq \exp\left\{-\frac{Kt^2}{4n\sigma^2}\right\}, \quad \text{if} \quad t > 0\,. \tag{S1.14}$$

Notice that the upper bounds (S1.13) and (S1.14) do not depend on node i and labels l and k. Thus, for $\gamma \in (0, 1/K)$, $a < b$ or $\gamma \in (1/K, 1)$, $a > b$, applying Lemma 1, the union bound inequality and the concentration inequality (S1.14) with $t = n(a-b)\left(\frac{(\gamma K - 1)}{K(K-1)}\right) > 0$, we get

$$\mathbb{P}(\,\widehat{c}_i(e) \neq l\,) = \mathbb{P}\left(\bigcup_{\substack{k=1\\k\neq l}}^{K}\{S_{il} - S_{ik} \leq 0\}\right) \leq (K-1)\exp\left\{-\frac{1}{4}\frac{(\gamma K-1)^2}{K(K-1)^2}\frac{n(a-b)^2}{\sigma^2}\right\}. \tag{S1.15}$$

In the same way, for $\gamma \in (0, 1/k)$, $a > b$ and $c > d$ or $\gamma \in (1/K, 1)$, $a < b$ and $c < d$, setting $t = -n(a-b)\left(\frac{(\gamma K - 1)}{K(K-1)}\right) > 0$, we have that

$$\mathbb{P}(\,\widehat{c}_i(e) \neq l\,) = \mathbb{P}\left(\bigcup_{\substack{k=1\\k\neq l}}^{K}\{S_{il} - S_{ik} > 0\}\right) \leq (K-1)\exp\left\{-\frac{1}{4}\frac{(\gamma K-1)^2}{K(K-1)^2}\frac{n(a-b)^2}{\sigma^2}\right\}. \tag{S1.16}$$

The result follows by observing that the upper bounds (S1.15) and (S1.16) do not depend on l. □

Proof of Theorem 1. The proportion of misclassified nodes given in (3.12) can be upper bounded using ϕ as the identity permutation function. By Lemma 1, for any initial estimates $\widehat{a}, \widehat{b}, \widehat{\sigma}^2 \in S_{a,b}$ the estimated labels $\widehat{c}(e)$ depend only on the block sums using the partition e. Using the fact that the upper bound obtained in Proposition 1 does not depend on the node i, $i = 1, \ldots, n$, we have that

$$\mathbb{E}\left[\,L(\widehat{c}\,(e), c)\,\right] \leq (K-1)\exp\left\{-\frac{1}{4}\frac{(\gamma K-1)^2}{K(K-1)^2}\frac{n(a-b)^2}{\sigma^2}\right\}. \tag{S1.17}$$

Using the Markov inequality for any $\epsilon > 0$ we conclude that

$$\mathbb{P}\left(\,L(\widehat{c}\,(e), c) > \epsilon\,\right) \leq \frac{(K-1)}{\epsilon}\exp\left\{-\frac{1}{4}\frac{(\gamma K-1)^2}{K(K-1)^2}\frac{n(a-b)^2}{\sigma^2}\right\}. \tag{S1.18}$$

By the union bound inequality we have that

$$\mathbb{P}\left(\sup_{e \in \mathcal{E}_\gamma} L(\widehat{c}\,(e), c) > \epsilon \right) \leq |\mathcal{E}_\gamma| \, \frac{(K-1)}{\epsilon} \exp\left\{ -\frac{1}{4} \frac{(\gamma K - 1)^2}{K(K-1)^2} \frac{n(a-b)^2}{\sigma^2} \right\}. \qquad \text{(S1.19)}$$

The cardinality of $\mathcal{E}_\gamma$ is upper bounded by the number of configurations $e \in \{1, \ldots, K\}^n$ such that, for each community $k = 1, \ldots, K$, we have that $\sum_{i=1}^{n} \mathbb{1}\{e_i = k, c_i = k\} = \gamma m$. In this way, since $m = n/K$, we have

$$\begin{aligned} |\mathcal{E}_\gamma| &\leq \binom{m}{m\gamma}^K (K-1)^{(1-\gamma)mK} \\ &\leq \exp\left[\, n(\, h(\gamma) + \kappa_\gamma(2n/K) + (1-\gamma)\log(K-1\,))\,\right], \end{aligned} \qquad \text{(S1.20)}$$

where the last inequality follows from Lemma 6 in the supplementary material of Amini et al. (2013). The result follows taking $\epsilon = \exp\left\{ -\frac{1}{8} \frac{(\gamma K-1)^2}{K(K-1)^2} \frac{n(a-b)^2}{\sigma^2} \right\}$. □

The proof of an upper bound on the probability of misclassification of each node in the case of unbalanced networks is more complex because we can not eliminate the dependence on the initial parameters estimates $\widehat{a}, \widehat{b}$ and $\widehat{\sigma}^2$.

Proof of Theorem 2. Let $\hat{\sigma} > 0$ and $\hat{a}, \hat{b} \in \mathbb{R}$ be fixed values. Using the estimated matrices $\widehat{B}$ and $\widehat{\Sigma}$, we have that

$$\widehat{P} = n(R(e)\widehat{B})^T = n \begin{pmatrix} \hat{a}\gamma_1\pi_1 + \hat{b}(1-\gamma_2)\pi_2 & \hat{a}(1-\gamma_1)\pi_1 + \hat{b}\gamma_2\pi_2 \\ \hat{b}\gamma_1\pi_1 + \hat{a}(1-\gamma_2)\pi_2 & \hat{b}(1-\gamma_1)\pi_1 + \hat{a}\gamma_2\pi_2 \end{pmatrix} \qquad \text{(S1.21)}$$

and

$$\widehat{\Lambda} = n(R(e)\widehat{\Sigma})^T = n\widehat{\sigma}^2 \begin{pmatrix} \widetilde{\pi}_1 & \widetilde{\pi}_2 \\ \widetilde{\pi}_1 & \widetilde{\pi}_2 \end{pmatrix}. \qquad \text{(S1.22)}$$

Starting the EM algorithm with labels $e \in \mathcal{E}_{\gamma_1,\gamma_2}$ and estimated matrices $\widehat{B}$ and $\widehat{\Sigma}$, the algorithm returns after one iteration the label of node i such that

$$\widehat{c}_i(e) = \arg\max_{k=1,2} \left\{ \log \widetilde{\pi}_k - \left(\sum_{m=1}^{K} \frac{(s_{im}(e) - \widehat{P}_{km})^2}{2\widehat{\Lambda}_{km}} + \frac{1}{2}\log \widehat{\Lambda}_{km} \right) \right\}. \qquad \text{(S1.23)}$$

Using the estimator defined above and the fact that $\widehat{\Lambda}_{12} = \widehat{\Lambda}_{22}$ and $\widehat{\Lambda}_{11} = \widehat{\Lambda}_{21}$, we conclude that $\widehat{c}_i(e) = 1$, if and only if,

$$\sum_{k=1}^{2} \frac{(s_{ik}(e) - \widehat{P}_{1k})^2}{\widehat{\Lambda}_{1k}} - \sum_{k=1}^{2} \frac{(s_{ik}(e) - \widehat{P}_{2k})^2}{\widehat{\Lambda}_{2k}} < 2\log\left(\frac{\widehat{\pi}_1}{\widehat{\pi}_2}\right). \tag{S1.24}$$

Computing the difference of squares, we have that

$$\begin{aligned}\frac{(s_{i1}(e) - \widehat{P}_{11})^2}{\widehat{\Lambda}_{11}} - \frac{(s_{i1}(e) - \widehat{P}_{21})^2}{\widehat{\Lambda}_{21}} &= \frac{1}{n\widehat{\sigma}^2\widehat{\pi}_1}\left[(s_{i1}(e) - \widehat{P}_{11})^2 - (s_{i1}(e) - \widehat{P}_{21})^2\right] \\ &= \frac{1}{n\widehat{\sigma}^2\widehat{\pi}_1}\left(2s_{i1}(e) - (\widehat{P}_{11} + \widehat{P}_{21})\right)\left(\widehat{P}_{21} - \widehat{P}_{11}\right)\end{aligned} \tag{S1.25}$$

and

$$\frac{(s_{i2}(e) - \widehat{P}_{12})^2}{\widehat{\Lambda}_{12}} - \frac{(s_{i2}(e) - \widehat{P}_{22})^2}{\widehat{\Lambda}_{22}} = \frac{1}{n\widehat{\sigma}^2\widehat{\pi}_2}\left(2s_{i2}(e) - (\widehat{P}_{12} + \widehat{P}_{22})\right)\left(\widehat{P}_{22} - \widehat{P}_{12}\right). \tag{S1.26}$$

After some calculations of the entries of $\widehat{P}$ given in (S1.21), we conclude that

$$\begin{aligned}\widehat{P}_{21} - \widehat{P}_{11} &= n(\hat{a} - \hat{b})\left((1-\gamma_2)\pi_2 - \gamma_1\pi_1\right) \\ \widehat{P}_{22} - \widehat{P}_{12} &= n(\hat{a} - \hat{b})\left(\gamma_2\pi_2 - (1-\gamma_1)\pi_1\right)\end{aligned} \tag{S1.27}$$

By the definition of β_1 and β_2 given in (3.24), the LRS of (S1.24) is given by

$$\begin{aligned}&\frac{(\widehat{a} - \widehat{b})}{\widehat{\sigma}^2\widetilde{\pi}_1\widetilde{\pi}_2}\left[\beta_1\left(2s_{i1}(e) - (\widehat{P}_{11} + \widehat{P}_{21})\right) - \beta_2\left(2s_{i2}(e) - (\widehat{P}_{12} + \widehat{P}_{22})\right)\right] \\ &= \frac{(\widehat{a} - \widehat{b})}{\widehat{\sigma}^2\widetilde{\pi}_1\widetilde{\pi}_2}\left[2\beta_1 s_{i1}(e) - 2\beta_2 s_{i2}(e) - \beta_1\left(\widehat{P}_{11} + \widehat{P}_{21}\right) + \beta_2\left(\widehat{P}_{12} + \widehat{P}_{22}\right)\right].\end{aligned} \tag{S1.28}$$

By (S1.24) and (S1.28) we conclude that $\widehat{c}_i(e) \neq 1$ if

$$\frac{(\widehat{a} - \widehat{b})}{\widehat{\sigma}^2\widetilde{\pi}_1\widetilde{\pi}_2}\left[2\beta_1 s_{i1}(e) - 2\beta_2 s_{i2}(e) - \beta_1\left(\widehat{P}_{11} + \widehat{P}_{21}\right) + \beta_2\left(\widehat{P}_{12} + \widehat{P}_{22}\right)\right] \geq 2\log\left(\frac{\widetilde{\pi}_1}{\widetilde{\pi}_2}\right), \tag{S1.29}$$

and $\widehat{c}_i(e) \neq 2$ if

$$\frac{(\widehat{a} - \widehat{b})}{\widehat{\sigma}^2\widetilde{\pi}_1\widetilde{\pi}_2}\left[2\beta_1 s_{i1}(e) - 2\beta_2 s_{i2}(e) - \beta_1\left(\widehat{P}_{11} + \widehat{P}_{21}\right) + \beta_2\left(\widehat{P}_{12} + \widehat{P}_{22}\right)\right] \leq 2\log\left(\frac{\widetilde{\pi}_1}{\widetilde{\pi}_2}\right). \tag{S1.30}$$

Given $c_i = 1$, we have that

$$2\beta_1 s_{i1}(e) - 2\beta_2 s_{i2}(e) \sim \mathcal{N}\left(2\beta_1 P_{11} - 2\beta_2 P_{12}\,,\ (2\beta_1)^2\Lambda_{11} + (2\beta_2)^2\Lambda_{12}\right). \tag{S1.31}$$

In the same way, given $c_i = 2$, we have that

$$2\beta_1 s_{i1}(e) - 2\beta_2 s_{i2}(e) \sim \mathscr{N}(2\beta_1 P_{21} - 2\beta_2 P_{22}\,,\ (2\beta_1)^2\Lambda_{21} + (2\beta_2)^2\Lambda_{22})\,. \tag{S1.32}$$

Calculating the matrix Λ as computed in (S1.22) using the matrix Σ instead of $\widehat{\Sigma}$, we have

$$\begin{aligned}(2\beta_1)^2\Lambda_{11} + (2\beta_2)^2\Lambda_{12} &= (2\beta_1)^2\Lambda_{21} + (2\beta_2)^2\Lambda_{22} \\ &= 4n\sigma^2(\beta_1^2\pi_1 + \beta_2^2\pi_2) \\ &= 4n\sigma^2\tau^2\end{aligned} \tag{S1.33}$$

Using the Chernoff tail bound for Gaussian random variables we obtain

$$\mathbb{P}\left(2\beta_1 s_{i1}(e) - 2\beta_2 s_{i2}(e) > 2\beta_1 P_{11} - 2\beta_2 P_{12} + t \,|\, c_i = 1\right) \le \exp\left\{-\frac{t^2}{8n\sigma^2\tau^2}\right\}\,,\ t \ge 0 \tag{S1.34}$$

and

$$\mathbb{P}\left(2\beta_1 s_{i1}(e) - 2\beta_2 s_{i2}(e) \le 2\beta_1 P_{21} - 2\beta_2 P_{22} - t \,|\, c_i = 2\right) \le \exp\left\{-\frac{t^2}{8n\sigma^2\tau^2}\right\}\,,\ t > 0 \tag{S1.35}$$

If $\widehat{a} > \widehat{b}$ and $t_1 = \frac{2\sigma^2\widetilde{\pi}_1\widetilde{\pi}_2}{\widehat{a}-\widehat{b}} \log\left(\frac{\widetilde{\pi}_1}{\widetilde{\pi}_2}\right) + \beta_1\left(\widehat{P}_{11} + \widehat{P}_{21}\right) - \beta_2\left(\widehat{P}_{12} + \widehat{P}_{22}\right) - 2\beta_1 P_{11} + 2\beta_2 P_{12} \ge 0$ we have

$$\mathbb{P}(\widehat{c}_i(e) \ne 1 \,|\, c_i = 1) \le \exp\left\{-\frac{t_1^2}{8n\sigma^2\tau^2}\right\}\,. \tag{S1.36}$$

Analogously, if $\widehat{a} > \widehat{b}$ and

$$t_2 = -\frac{2\sigma^2\widetilde{\pi}_1\widetilde{\pi}_2}{\widehat{a}-\widehat{b}} \log\left(\frac{\widetilde{\pi}_1}{\widetilde{\pi}_2}\right) - \beta_1\left(\widehat{P}_{11} + \widehat{P}_{21}\right) + \beta_2\left(\widehat{P}_{12} + \widehat{P}_{22}\right) + 2\beta_1 P_{21} - 2\beta_2 P_{22} > 0$$

we have

$$\mathbb{P}(\widehat{c}_i(e) \ne 2 \,|\, c_i = 2) \le \exp\left\{-\frac{t_2^2}{8n\sigma^2\tau^2}\right\}\,. \tag{S1.37}$$

After some calculations using (S1.21) and (3.24) to compute t_1 and t_2 we obtain that $t_1 = \frac{2\sigma^2\widetilde{\pi}_1\widetilde{\pi}_2}{\widehat{a}-\widehat{b}} \log\left(\frac{\widetilde{\pi}_1}{\widetilde{\pi}_2}\right) + nF(a,b)$ and $t_2 = -\frac{2\sigma^2\widetilde{\pi}_1\widetilde{\pi}_2}{\widehat{a}-\widehat{b}} \log\left(\frac{\widetilde{\pi}_1}{\widetilde{\pi}_2}\right) - nF(b,a)$. The result follows analogously when $\widehat{a} < \widehat{b}$. □

S2 A general result for balanced communities

In this section we consider a general case for balanced communities where the initial labeling $c_0 \in \{1, \dots, K\}^n$ matches a potentially different proportion of labels in each community. Define

$$\mathcal{E}_\alpha = \Big\{ e \in \{1, \dots, K\}^n : \text{ for all } k, l = 1, \dots, K$$
$$\sum_{i=1}^n \mathbb{1}\{e_i = k, c_i = l\} = \alpha_{kl} \frac{n}{K} \quad \text{with} \quad \sum_{k=1}^K \alpha_{kl} = 1 \text{ and } \sum_{l=1}^K \alpha_{kl} = 1 \Big\}.$$

We need to enforce $\sum_{k=1}^K \alpha_{kl} = 1$, to guarantee that the sum of each row of R is equal to n/K, that is, the number of nodes in each community, and $\sum_{l=1}^K \alpha_{kl} = 1$ to guarantee that the number of nodes in each community of e is equal to n/K, since all of this is for equal community sizes. Otherwise the proportion of labels matching can be general. Under this assumption, Proposition 2 gives an upper bound for the misclassification probability of each node; unlike Theorem 1, this bound holds conditionally on the initial values $\hat{a}, \hat{b}$ and $\hat{\sigma}^2$.

Proposition 2. *Assume that* $\pi_1 = \cdots = \pi_K = 1/K$*. Let the initial labeling* $e \in \mathcal{E}_\alpha$ *and let* $\hat{c}(e)$ *be the estimate of the labels obtained from* (3.14)*. Conditionally on* $\hat{a}, \hat{b}$ *and* $\hat{\sigma}^2$*, we have that*

$$\mathbb{P}(\hat{c}_i(e) \neq l \mid c, c_i = l) \leq \sum_{k \neq l} \exp\left\{ -\frac{n t_k^2}{2K\hat{\sigma}^2} \right\} \tag{S2.38}$$

where $t_k = \dfrac{\sum\limits_{m=1}^K (\alpha_{mk} - \alpha_{ml}) \left(b \sum_{j \neq r} \alpha_{mj} + a \alpha_{ml} \right) + \frac{(\hat{a} - \hat{b})}{2} \sum\limits_{m=1}^K (\alpha_{mk}^2 - \alpha_{ml}^2)}{\left(\sum\limits_{m=1}^K (\alpha_{ml} - \alpha_{mk})^2 \right)^{1/2}}$.

Proof. Consider one node with $c_i = l$, for some $i \in \{1, \dots, n\}$ and $l \in \{1, \dots, K\}$. For $e \in \mathcal{E}_\alpha$, the entry (r, s) of the confusion matrix is given by

$$R_{rs} = \frac{\alpha_{rs}}{K}.$$

Given the values $\hat{a}, \hat{b}$ and $\hat{\sigma}^2$ and using that $\hat{P} = n(R(e)\hat{B})^T$ and $\hat{\Lambda} = n(R(e)\hat{\Sigma})^T$, we have

$$\hat{P}_{rs} = \frac{n}{K}\left(\hat{b}\sum_{j\neq r}\alpha_{sj} + \hat{a}\alpha_{sr}\right), \tag{S2.39}$$

$$\hat{\Lambda}_{rs} = \frac{n}{K}\hat{\sigma}^2 . \tag{S2.40}$$

By (3.14), $\hat{c}_i(e) = l$ if and only if for all $k \neq l$

$$\sum_{m=1}^{K}(s_{im} - \hat{P}_{lm})^2 - \sum_{m=1}^{K}(s_{im} - \hat{P}_{km})^2 < 0 ,$$

which is straightforward to rewrite as

$$Y_i^{lk} = \sum_{m=1}^{K} 2s_{im}(\hat{P}_{km} - \hat{P}_{lm}) + (\hat{P}_{lm}^2 - \hat{P}_{km}^2) < 0 . \tag{S2.41}$$

Conditionally on c with $c_i = l$, $\{S_{i1}(e), \ldots, S_{iK}(e)\}$ are mutually independent random variables, with S_{im} following the normal distribution with mean P_{lm} and variance Λ_{lm}. Thus,

$$Y_i^{lk} \sim \mathscr{N}\left(\sum_{m=1}^{K} 2P_{lm}(\hat{P}_{km} - \hat{P}_{lm}) + (\hat{P}_{lm}^2 - \hat{P}_{km}^2), 4\sum_{m=1}^{K}(\hat{P}_{km} - \hat{P}_{lm})^2\Lambda_{lm}\right) . \tag{S2.42}$$

Using (S2.40), we get

$$\begin{aligned}\hat{P}_{km} - \hat{P}_{lm} &= \frac{n}{K}(\hat{a} - \hat{b})(\alpha_{mk} - \alpha_{ml}) ,\\ \hat{P}_{lm}^2 - \hat{P}_{km}^2 &= \frac{n^2}{K^2}(\hat{a} - \hat{b})(\alpha_{ml} - \alpha_{mk})(2\hat{b} + (\hat{a} - \hat{b})(\alpha_{mk} + \alpha_{ml})) .\end{aligned}$$

Summing over m and using that $\sum_{k=1}^{K}\alpha_{kl} = 1$ we get

$$\begin{aligned}\sum_{m=1}^{K}(\hat{P}_{lm}^2 - \hat{P}_{km}^2) &= \frac{n^2}{K^2}(\hat{a} - \hat{b})^2\sum_{m=1}^{K}(\alpha_{ml}^2 - \alpha_{mk}^2)\\ \sum_{m=1}^{K} P_{lm}(\hat{P}_{km} - \hat{P}_{lm}) &= \frac{n^2}{K^2}(\hat{a} - \hat{b})\sum_{m=1}^{K}(\alpha_{mk} - \alpha_{ml})\left(b\sum_{j\neq r}\alpha_{mj} + a\alpha_{ml}\right) .\end{aligned}$$

By (S2.41) and the union bound, we have

$$\mathbb{P}(\hat{c}_i(e) \neq l \mid c, c_i = l) = \sum_{k\neq l}\mathbb{P}\left(Y_i^{lk} > 0\right) .$$

Using the Chernoff tail bound for Gaussian random variables, we obtain

$$\mathbb{P}\left(Y_i^{lk} > 0\right) \leq \exp\left\{ -\frac{n\left(\sum_{m=1}^{K}(\alpha_{mk}-\alpha_{ml})\left(b\sum_{j\neq r}\alpha_{mj}+a\alpha_{ml}\right)+\frac{(\hat{a}-\hat{b})}{2}\sum_{m=1}^{K}(\alpha_{ml}^2-\alpha_{mk}^2)\right)^2}{2K\hat{\sigma}^2\sum_{m=1}^{K}(\alpha_{mk}-\alpha_{ml})^2}\right\}$$

which concludes the proof. □

References

Amini, A. A., A. Chen, P. J. Bickel, and E. Levina (2013). Pseudo-likelihood methods for community detection in large sparse networks. *The Annals of Statistics 41*(4), 2097–2122.

Boucheron, S., G. Lugosi, and P. Massart (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press.