1. Simulations in Experiment 4

In Experiment 4, we compare the power of the power-enhanced test with the likelihood-based model selection criterion of [3] and the eigenvalue-based goodness-of-fit test of [2]. We consider networks with n actors and adjacency matrix $A \in \{0,1\}^{K\times K}$, sampled from a Stochastic Block Model with $K = 2$ communities, community mixing matrix $P \in [0,1]^{K\times K}$ and community assignment multinomial probabilities $a \in [0,1]^K$. We choose a fixed level γ for the PET and the test of [2]. We do not need to define the level for the method of [3], which is not a test but a model selection criterion.

1.1. Likelihood-based model selection criterion of [3]

The model selection criterion of [3] leverages the complete data likelihood to consistently estimate the number of communities in a Stochastic Block Model under the regime of growing degrees. For $K' \in \{1,2\}$, we optimize the following mean field approximation of the criterion:

$$\beta(K';A) = \sup_{(P,a)\in\mathcal{M}_{K'}} \sup_{q\in\mathcal{D}_{K'}} J(q,P,a;A) - \lambda\frac{K'(1+K')}{2} n\log(n),$$

where $\mathcal{D}_{K'}$ is the set of product distributions over $\{0,1\}^n$, $\mathcal{M}_{K'}$ is the set of proper SBM parameters and λ is a tuning parameter chosen using the procedure described in section 3.2. of [3]. $J(q,P,a;A)$ corresponds to the mean field approximation of the log-likelihood, which can be expressed as

$$\begin{aligned} J(q,P,a;A) = &\sum_{i=1}^{n}\sum_{k=1}^{K'} \tau_{ik}(-\log(\tau_{ik}) + \log(a_k)) \\ &+ \frac{1}{2}\sum_{i\neq j}\sum_{k,l=1}^{K'} \tau_{ik}\tau_{jl}(A_{ij}\log(P_{kl}) + (1-A_{ij})\log(1-P_{kl})), \end{aligned}$$

where $\{\tau_{ik}\}_{i,k}$ are the multinomial parameters of the mean field distribution q, with the constraint that $\sum_k \tau_{ik} = 1$ for all $i \in [\![1,n]\!]$. [1] showed that for fixed (P,a), maximizing $J(q,P,a;A)$ with respect to q can be done by solving the fixed point equation

$$\widehat{\tau}_{ik} \propto a_k \prod_{\substack{j=1\\ j\neq i}}^{n}\prod_{l=1}^{K'} \left(P_{kl}^{A_{ij}}(1-P_{kl})^{1-A_{ij}}\right)^{\widehat{\tau}_{jl}}.$$

Then, given the variational parameters $\{\tau_{ik}\}_{i,k}$, the values of the parameters P and a that maximize $J(q,P,a;A)$ are

$$\widehat{a}_k = \frac{1}{n}\sum_{i=1}^{n}\widehat{\tau}_{ik}, \qquad \widehat{P}_{kl} = \frac{\sum_{i\neq j}\widehat{\tau}_{ik}\widehat{\tau}_{jl}A_{ij}}{\sum_{i\neq j}\widehat{\tau}_{ik}\widehat{\tau}_{jl}}.$$

Following the estimation algorithm of [1], we start from an initial guess for $\{\widehat{\tau}_{ik}^{(0)}\}_{i,k}$ (which could be random or using spectral clustering of the regularized Laplacian matrix) and iteratively update $\{\widehat{a}_k\}_k$, $\widehat{P}$ and $\{\widehat{\tau}_{ik}\}_{i,k}$ until convergence of $J(q,P,a;A)$. For a network A drawn from our model, we use this procedure to compute $\beta(1;A)$ and $\beta(2;A)$. We call a true positive discovery when $\beta(2;A) > \beta(1;A)$, and a false negative discovery when $\beta(2;A) \leq \beta(1;A)$.

1.2. Eigenvalue-based test of [2]

We also compare the PET to the goodness-of-fit test of [2], which uses the largest singular value of a residual matrix $\widetilde{A}$ obtained by subtracting the estimated block mean effect from the adjacency matrix. We choose an Erdös-Rényi null model. In this setting, we reject the null hypothesis if

$$T_{n,1} := n^{2/3}[\sigma_1(\widetilde{A}) - 2] \geq t(\gamma/2),$$

where γ is the level of the test, $\sigma_1(\widetilde{A})$ is the largest singular value of $\widetilde{A}$ and $t(\gamma/2)$ is the $\gamma/2$ upper quantile of the Tracy-Widom TW_1 distribution. In the test statistic $T_{n,1}$, the residual matrix $\widetilde{A}$ is computed using an estimated community membership vector $(\widehat{z}_i)_{i=1}^n$. The vector $(\widehat{z}_i)_{i=1}^n$ is obtained from spectral clustering, where we apply k-means clustering to the rows of the matrix formed by the 2 leading singular vectors of A. Then, $\widetilde{A}$ can be expressed as

$$\widetilde{A}_{ij} = \begin{cases} \frac{A_{ij} - \widehat{\Omega}_{ij}}{\sqrt{(n-1)\widehat{\Omega}_{ij}(1-\widehat{\Omega}_{ij})}}, & i \neq j, \\ 0, & i = j, \end{cases}$$

where $\widehat{\Omega}_{ij} = \widehat{P}_{\widehat{z}_i \widehat{z}_j}$ and

$$\widehat{\mathcal{N}}_k = \{i : 1 \leq i \leq n, \widehat{z}_i = k\}, \qquad \widehat{n}_k = |\widehat{\mathcal{N}}_k|, \qquad \widehat{P}_{kl} = \begin{cases} \frac{\sum_{i \in \widehat{\mathcal{N}}_k, j \in \widehat{\mathcal{N}}_l} A_{ij}}{\widehat{n}_k \widehat{n}_l}, & k \neq l, \\ \frac{\sum_{i,j \in \widehat{\mathcal{N}}_k, i<j} A_{ij}}{\widehat{n}_k(\widehat{n}_k - 1)/2}, & k = l. \end{cases}$$

References

[1] DAUDIN, J.-J., PICARD, F. and ROBIN, S. (2008). A mixture model for random graphs. *Statistics and computing* **18** 173–183.

[2] LEI, J. (2016). A goodness-of-fit test for stochastic block models. *Ann. Statist.* **44** 401–424.

[3] WANG, Y. R. and BICKEL, P. J. (2017). Likelihood-based model selection for stochastic block models. *Ann. Statist.* **45** 500–528.