Supplemental Materials: A generalized hypothesis test for community structure in networks

Technical Proofs

Upper bound on E2D2 parameter

We want to show that $\{\bar{p}_{in}(\boldsymbol{c})-\bar{p}_{out}(\boldsymbol{c})\}/(K\bar{p}) \leq 1$. Notice that for any $\boldsymbol{c}$, $\bar{p} = r\bar{p}_{in}+(1-r)\bar{p}_{out}$ for $r = m_{in}/\binom{n}{2}$ where $0 \leq r \leq 1$ and $r = r(K)$ depends on K, the number of communities. Thus, we equivalently want to maximize

$$f(x,y,r) = \frac{x-y}{rx+(1-r)y} \tag{1}$$

where $0 \leq r, x, y \leq 1$. First, let's consider a fixed r. Then $f(x,y,r)$ will clearly be maximized when $y = 0$ which yields

$$f(x,0,r) = \frac{1}{r}. \tag{2}$$

Thus, $f(x,y,r)$ is maximized when r is minimized, or, equivalently, when m_{in} is minimized for a fixed K.

Let m_k be the number of nodes in community $k \in \{1,\dots,K\}$ where $m_1+\cdots+m_K = n$. Then we want to minimize $m_{in} = \frac{1}{2}\sum_{j=1}^{K} m_j(m_j-1)$ subject to $\sum_{j=1}^{K} m_j = n$. We can use Lagrange multipliers:

$$\mathcal{L}(m_1,\dots,m_K,\lambda) = \frac{1}{2}\sum_{j=1}^{K} m_j(m_j-1) - \lambda\left(\sum_{j=1}^{K} m_j - n\right) \tag{3}$$

Take the gradient:

$$\nabla\mathcal{L}(m_j,\lambda) = \left(m_1 - \tfrac{1}{2} - \lambda, \dots, m_K - \tfrac{1}{2} - \lambda, n - \sum_{j=1}^{K} m_j\right) \tag{4}$$

Setting equal to 0 means that for all j, $m_j = \lambda + \frac{1}{2}$ so

$$0 = n - \sum_{j=1}^{K}(\lambda + \tfrac{1}{2}) \implies \lambda = \frac{n}{K} - \frac{1}{2}. \tag{5}$$

Thus, m_{in} is minimized at $m_1 = \cdots = m_k = \frac{n}{K}$ so

$$m_{in} \geq \frac{1}{2}\sum_{j=1}^{K} \tfrac{n}{K}(\tfrac{n}{K}-1) = \frac{n(n-K)}{2K}. \tag{6}$$

Thus,

$$f(x,y,r) \leq \frac{1}{r} \leq \frac{\binom{n}{2}}{n(n-K)/2K} = K\frac{n-1}{n-K}. \tag{7}$$

For large n, $(n-1)/(n-K) \approx 1$ so we have the desired result.

Theorem 2.1

First, note that since we assume K_n is known, we can ignore it during the proof and simply divide the final cutoff by K_n. Now, let $\gamma_0 = \xi_0/\bar{p}$. Assume a rejection region of the form $R = \{T_*(n) > c(n)\}$ where $c(n) = \frac{\xi_0 + k(n)}{\bar{p}(n)/(1+\epsilon)}$ and

$$T_*(n) = \frac{U_*(n)}{S(n)} \tag{8}$$

where $U_*(n) = \max_{\boldsymbol{c}}\{\hat{p}_{in}(\boldsymbol{c}) - \hat{p}_{out}(\boldsymbol{c})\}$ with the max taken over all possible community assignments $\boldsymbol{c}_i$ for $i = 1, \ldots, N_{n,K}$; $S(n) = \hat{p}(n)$ and

$$\bar{p}(n) = \frac{1}{\binom{n}{2}} \sum_{i>j} P_{ij}(n).$$

From this point, we suppress the dependence on n. Using DeMorgan's Law, we can show that

$$P(T_* > c) \leq P(U_* > \xi_0 + k) + P(S < \bar{p}/(1+\epsilon)). \tag{9}$$

where

$$\bar{p} = \frac{1}{\binom{n}{2}} \sum_{i<j} P_{ij}. \tag{10}$$

Under H_0, we show that each term on the right-hand side goes to 0. Assume the null model P_0 and consider a fixed community assignment with K_n communities, $\boldsymbol{c}_i$, for $i \in \{1, \ldots, N_{n,k}\}$ where $N_{n,K_n} \leq K_n^n$ and let $U_i = \hat{p}_{in}(\boldsymbol{c}_i) - \hat{p}_{out}(\boldsymbol{c}_i)$. Then

$$U_i = \sum_{j<k} X_{jk} \tag{11}$$

where $X_{jk} = m_{in,i}^{-1}$ if $(\boldsymbol{c}_i)_j = (\boldsymbol{c}_i)_k$ and $-m_{out,i}^{-1}$ otherwise. From the proof of the upper bound on the E2D2 parameter, we have that $m_{in,i} = O(n^2)$ and $m_{out,i} = O(n^2)$. Thus, letting $k_i' = \mathsf{E}(U_i) + k$ and using Hoeffding's inequality,

$$\begin{aligned}
\frac{\eta}{N_{n,K}} &= P(U_i \geq k_i') && (12)\\
&= P(U_i \geq \mathsf{E}(U_i) + k) && (13)\\
&\leq \exp\left(\frac{-2k^2}{\binom{n}{2}(\frac{1}{m_{in}} + \frac{1}{m_{out}})^2}\right) && (14)\\
&\leq \exp\left(-n^2k^2\right) && (15)\\
\implies k &\leq \left(\frac{\log N_{n,K} - \log \eta}{n^2}\right)^{1/2} \sim \left(\frac{\log K_n}{n}\right)^{1/2} && (16)
\end{aligned}$$

Now, under the null hypothesis, $\mathsf{E}(U_i) \leq \xi_0$. Then we have

$$P\{U_* > \xi_0 + k\} = P\left(\bigcup_{i=1}^{N_{n,K}} \{U_i > \xi_0 + k\}\right)$$
$$\leq P\left(\bigcup_{i=1}^{N_{n,K}} \{U_i > k_i'\}\right) \leq \sum_{i=1}^{N_{n,K}} P\{U_i > k_i'\} \leq \sum_{i=1}^{N_{n,K}} \frac{\eta}{N_{n,K}} \leq \eta. \quad (17)$$

We also have

$$P(S < \bar{p}/(1+\epsilon)) = P(S < \bar{p} - \tfrac{\epsilon}{1+\epsilon}\bar{p}) \leq e^{-\epsilon^2\bar{p}^2 n(n-1)/(1+\epsilon)^2} \to 0 \quad (18)$$

since $n^{1/2}\bar{p} \to \infty$. Combining these two results we have that

$$P(T_* > c) \leq P(U_* > \xi_0 + k) + P(S < \bar{p}/(1+\epsilon)) \leq \eta \quad (19)$$

as we hoped to show.

Under H_1, let $\gamma_1 = \xi_1/\bar{p}$ and let $T_{oracle} = T(\boldsymbol{c}_\gamma, A) = U_{oracle}/S$ where $\boldsymbol{c}_\gamma = \arg\max_{\boldsymbol{c}}\{\gamma(\boldsymbol{c}, P)\}$, i.e., $\boldsymbol{c}_\gamma$ is the community assignment which maximizes the E2D2 parameter. This is reasonable because we assume that the algorithm finds the global maximum $\tilde{T}(A)$ so $T_{oracle} \leq \tilde{T}(A)$. We will use a similar approach to the proof of H_0 noting that

$$\{U_{oracle} > (\xi_0 + k)\tfrac{1+\epsilon}{1-\epsilon}\} \cap \{S \leq \tfrac{\bar{p}}{1-\epsilon}\} \subseteq \{T_{oracle} > c\} \quad (20)$$

so

$$P(T_{oracle} > c) \geq P\{U_{oracle} > (\xi_0 + k)\tfrac{1+\epsilon}{1-\epsilon}\} \cap \{S \leq \tfrac{\bar{p}}{1-\epsilon}\}) \quad (21)$$
$$\geq P\{U_{oracle} > (\xi_0 + k)\tfrac{1+\epsilon}{1-\epsilon}\} + P\{S \leq \tfrac{\bar{p}}{1-\epsilon}\} - 1. \quad (22)$$

Thus, we want to show that the first two terms on the right-side go to 1. For the first term, we note that U_{oracle} is the sum of $O(n^2)$ independent random variables, each of which takes values between $[-m_{out}^{-1}, m_{in}^{-1}]$. Moreover, $\mathsf{E}(U_{oracle}) = \xi_1 > \xi_0$. Let $1_\epsilon := (1+\epsilon)/(1-\epsilon)$. Then,

$$P\{U_{oracle} \leq (\xi_0 + k)1_\epsilon\} = P\{U_{oracle} \leq \xi_1 1_\epsilon - (\xi_1 - \xi_0 - k)1_\epsilon\} \quad (23)$$
$$= P\{U_{oracle} \leq \xi_1 - \underbrace{(\xi_1 - \xi_0 - \tfrac{2\epsilon}{1+\epsilon}\xi_1 - k)}_{z}\}. \quad (24)$$

Now, $z > 0$ since $\xi_1 - \xi_0 > 0$ and we can choose ϵ small enough such that $\xi_1 - \xi_0 - \frac{2\epsilon}{1+\epsilon}\xi_1 > 0$. Additionally, $k \to 0$ by A3 so there exists an N such that for all $n \geq N$, $\xi_1 - \xi_0 - \frac{2\epsilon}{1+\epsilon}\xi_1 > k$. Thus, we can use Hoeffding's inequality to show

$$P\{U_{oracle} \leq (\xi_0 + k)1_\epsilon\} = P\{U_{oracle} \leq \xi_1 - z\} \quad (25)$$
$$\leq \exp\left(-\frac{2z^2}{\sum_{i=1}^{n^2}\frac{1}{n^4}}\right) \quad (26)$$
$$= \exp\left(-2n^2z^2\right) \quad (27)$$
$$\to 0, \quad (28)$$

or equivalently,

$$P\{U_{oracle} > (\xi_0 + k)1_\epsilon\} \to 1. \tag{29}$$

Next, consider S. First, notice that

$$\frac{\bar{p}}{1-\varepsilon} = \bar{p} + \frac{\varepsilon}{1-\varepsilon}\bar{p} \tag{30}$$

Then, by Hoeffding's inequality, we can show

$$\begin{aligned} P(S \geq \bar{p}/(1-\varepsilon)) &= P(S \geq \bar{p} + \tfrac{\varepsilon}{1+\varepsilon}\bar{p}) && (31)\\ &\leq e^{-\varepsilon^2(\bar{p})^2/(1+\varepsilon)^2 n(n-1)} && (32)\\ &\to 0 && (33)\end{aligned}$$

since $n^{1/2}\bar{p} \to \infty$. Then

$$\lim_{n\to\infty} P(\tilde{T}(A) > C) \geq \lim_{n\to\infty} P(T_{oracle} > C) \geq 1 + 1 - 1 \geq 1.\square \tag{34}$$

Proposition in Section 2.4

Claim: $\tilde{\gamma}(P) = 0$ *if and only if* P *is from an ER model.*

Proof. The only if direction of the claim is immediate. To prove the forward direction, we first show that $\gamma(\boldsymbol{c}, P) \leq 0$ for all $\boldsymbol{c}$ implies that $\gamma(\boldsymbol{c}, P) = 0$ for all $\boldsymbol{c}$. Then we show that if $\gamma(\boldsymbol{c}, P) = 0$ for all $\boldsymbol{c}$, then P is from an ER model which is equivalent to showing $\tilde{\gamma}(P) = 0$.

For the first part, this is equivalent to showing that if $\gamma(\boldsymbol{c}, P) < 0$ for some $\boldsymbol{c}$, then there exists some $\boldsymbol{c}'$ such that $\gamma(\boldsymbol{c}', P) > 0$. If there exists some $\boldsymbol{c}$ such that $\gamma(\boldsymbol{c}, P) < 0$, then

$$\frac{1}{\sum_{k=1}^K \binom{n_k}{2}} \sum_{i<j}^K \delta_{c_i,c_j} P_{ij} < \frac{1}{\sum_{k>l} n_k n_l} \sum_{i<j} (1 - \delta_{c_i,c_j}) P_{ij}.$$

But this means that there is some P_{ij} such that $P_{ij} \geq P_{kl}$ for all $i \neq k$ or $j \neq l$ and is strictly greater for at least one P_{kl}. Thus, if we consider the community assignment $\boldsymbol{c}'$ where nodes i and j are in one community and all other nodes are in the other, then $\bar{p}_{in}(\boldsymbol{c}') > \bar{p}_{out}(\boldsymbol{c}')$ and thus $\gamma(\boldsymbol{c}', P) > 0$.

We will prove the second part by induction. Let $n = 3$ and we are given that $\gamma(\boldsymbol{c}, P) = 0$ for all $\boldsymbol{c}$. We start by writing out the probability matrix.

$$P = \begin{pmatrix} - & P_{12} & P_{13} \\ & - & P_{23} \\ & & - \end{pmatrix}.$$

There are three possible community assignments: $\boldsymbol{c}_1 = \{1, 1, 2\}, \boldsymbol{c}_2 = \{1, 2, 1\}$ and $\boldsymbol{c}_3 = \{2, 1, 1\}$. From each of these assignments, we have a corresponding statement relating the probabilities:

$$\begin{aligned}\bar{p}_{in} = P_{12} = \bar{p}_{out} = \frac{1}{2}(P_{13} + P_{23})\\ \bar{p}_{in} = P_{13} = \bar{p}_{out} = \frac{1}{2}(P_{12} + P_{23})\\ \bar{p}_{in} = P_{23} = \bar{p}_{out} = \frac{1}{2}(P_{12} + P_{13}).\end{aligned}$$

Plugging the first equation into the second equation we find:

$$P_{13} = \frac{1}{2}(\tfrac{1}{2}(P_{13} + P_{23}) + P_{23}) \implies P_{13} = P_{23}.$$

Plugging this into the first equation we have $P_{12} = P_{13} = P_{23} := p$ which means that this must be an ER model.

Now assume that the claim holds for $n-1$ and show it holds for n. For convenience, assume n is even but the proof can easily be extended if n is odd. Consider a network with n nodes such that $\gamma(\boldsymbol{c}, P) = 0$. Remove an arbitrary node such that we have a network with $n-1$ nodes and apply the induction hypothesis, i.e. $P_{ij} = p$ for all i, j. We now add the removed node back to the network such that the node has probability $P_{i,n}$ of an edge between itself and node i for $i = 1, \ldots, n-1$. Thus, the probability matrix is:

$$P = \begin{pmatrix} - & p & p & \cdots & p & P_{1n} \\ & - & p & \cdots & p & P_{2n} \\ & & & \ddots & & \vdots \\ & & & & & P_{n-1,n} \\ & & \ddots & & & - \end{pmatrix}.$$

Since $\gamma(\boldsymbol{c}, P) = 0$ for all $\boldsymbol{c}$, then we want to show that $P_{i,n} = p$ for $i = 1, \ldots, n$. Assume for contradiction that P is not ER and we will show that $\gamma(\boldsymbol{c}, P) \neq 0$ for some $\boldsymbol{c}$. Without loss of generality, let $\{P_{1n}, \ldots, P_{n/2,n}\}$ be the smaller values of the last column and $\{P_{n/2+1,n}, \ldots, P_{n-1,n}\}$ be the larger values and consider the community assignment where nodes $\{1, \ldots, n/2\}$ are in one community and nodes $\{n/2+1, \ldots, n\}$ are in the other community. Then

$$\bar{p}_{in} = \frac{1}{2\binom{n/2}{2}} \left(p \cdot (\tfrac{n}{2} - 1)^2 + \sum_{i=n/2+1}^{n-1} P_{i,n} \right) > \bar{p}_{out} = \frac{1}{n^2/4} \left(p \cdot (\tfrac{n^2}{4} - \tfrac{n}{2}) + \sum_{i=1}^{n/2} P_{i,n} \right)$$

since

$$\sum_{i=n/2+1}^{n-1} P_{i,n} > \sum_{i=1}^{n/2} P_{i,n}.$$

Thus $\gamma(\boldsymbol{c}, P) \neq 0$ for this particular choice of $\boldsymbol{c}$ and we have completed the proof. $\square$

Lemma 3.1

We follow closely the ideas of the proof of Theorem 5 in Levin and Levina (2019). Assume that $p \sim F(\cdot)$ and $A, H|p \sim ER(p)$, $\hat{A}^*|\hat{p} \sim ER(\hat{p})$ where $\hat{p} = \sum_{i,j} A_{ij}/\{n(n-1)\}$. We will use the well-known property of Bernoulli random variables that if $X \sim \mathsf{Bernoulli}(p_1)$ and $Y \sim \mathsf{Bernoulli}(p_2)$, then $d_1(X, Y) \leq |p_1 - p_2|$. Thus,

$$P(\hat{A}^*_{ij} \neq H_{ij}|p, \hat{p}) \leq |\hat{p} - p|.$$

Let ν be the coupling such that A and H are independent. Then

$$W_p^p(\hat{A}^*, H) \leq \int d^p_{GM}(\hat{A}^*, H) d\nu(\hat{A}^*, H).$$

Using Jensen's inequality,

$$d^p_{GM}(\hat{A}^*, H) \leq \left(\frac{1}{2}\binom{n}{2}^{-1} ||\hat{A}^* - H||_1\right)^p \leq \binom{n}{2}^{-1} \sum_{i<j} |\hat{A}^*_{ij} - H_{ij}|^p = \binom{n}{2}^{-1} \sum_{i<j} |\hat{A}^*_{ij} - H_{ij}|.$$

Thus,

$$\begin{aligned}
\int d^p_{GM}(\hat{A}^*, H) d\nu(\hat{A}^*, H) &\leq \binom{n}{2}^{-1} \sum_{i<j} \int |\hat{A}^*_{ij} - H_{ij}| d\nu \\
&= \binom{n}{2}^{-1} \sum_{i<j} \nu(\{\hat{A}^*_{ij} \neq H_{ij}\}) \\
&\leq \binom{n}{2}^{-1} \sum_{i<j} |\hat{p} - p| \\
&= |\hat{p} - p| \\
&= O(n^{-1}). \ \square
\end{aligned}$$

Lemma 3.2

It's easy to see that the CL model falls into the Random Dot Product Graph framework where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ correspond to the latent positions and the dimension $d = 1$. Then by Theorem 5 of Levin and Levina (2019), we have that

$$W_p^p(\hat{A}^*, H) = O((n^{-1/2} + n^{-1/1}) \log n) = O(n^{-1/2} \log n)$$

since $\hat{\boldsymbol{\theta}}$ is estimated using the ASE.

Lemma 3.3

Let $t(H, \boldsymbol{c}) = \sum_{i<j} C_{ij} H_{ij}$ where $C_{ij} = m_{in}^{-1}$ if $c_i = c_j$ and m_{out}^{-1} otherwise and $H_{ij} \sim \mathsf{Bernoulli}(p)$. Define $\mathsf{E}\{t(H, \boldsymbol{c})\} = \xi(H, \boldsymbol{c})$ and

$$s_n^2 = \sum_{i<j} \mathsf{Var}(C_{ij} H_{ij}) = p(1-p) \sum_{i<j} C_{ij}^2.$$

We want to invoke Lyapunov's CLT so we must check the follow condition: for some $\delta > 0$,

$$\lim_{n\to\infty} \frac{1}{s_n^{2+\delta}} \sum_{i=1}^{n} \mathsf{E}(|X_i - \mathsf{E}(X_i)|^{2+\delta}) \to 0.$$

Let $\delta = 1$ and recall that $C_{ij} = O(n^{-2})$. Then, ignoring constants,

$$\begin{aligned}
\frac{1}{s_n^3} \sum_{i<j} \mathsf{E}(|C_{ij}H_{ij} - C_{ij}p|^3) &= \frac{1}{s_n^3} \sum_{i<j} C_{ij}^3 \mathsf{E}(|H_{ij} - p|^3) \\
&= \frac{1}{s_n^3} \sum_{i<j} C_{ij}^3 \\
&= O(n^3) \sum_{i<j} O(n^{-6}) \\
&= O(n^{-1}) \checkmark .
\end{aligned}$$

Thus, by Lyapunov's CLT,

$$\frac{1}{s_n} \sum_{i<j} (C_{ij}H_{ij} - C_{ij}p) = \frac{1}{s_n} \{t(H, \boldsymbol{c}) - \xi(H, \boldsymbol{c})\} \xrightarrow{d} \mathsf{N}(0, 1). \tag{35}$$

Finally, note that $T(H, \boldsymbol{c}) = t(H, \boldsymbol{c})/(K\hat{p})$ and $\gamma(H, \boldsymbol{c}) = \xi(H, \boldsymbol{c})/(Kp)$. Since $\hat{p} \xrightarrow{P} p$, by Slutsky's theorem,

$$\frac{1}{s_n} \{T(H, \boldsymbol{c}) - \gamma(H, \boldsymbol{c})\} \xrightarrow{d} \mathsf{N}(0, K^2p^2). \tag{36}$$

The results for $\tilde{T}(\hat{A}^*, \boldsymbol{c})$ are the same noting that:

$$\mathsf{E}(\hat{A}_{ij}^*) = \mathsf{E}(\mathsf{E}(\hat{A}_{ij}^*|\hat{p})) = \mathsf{E}(\hat{p}) = p = \mathsf{E}(H_{ij})$$

so $\mathsf{E}(t(\hat{A}^*, \boldsymbol{c}) = \mathsf{E}(t(H, \boldsymbol{c}))$; and

$$\mathsf{Var}(\hat{A}_{ij}^*) = \mathsf{Var}(\mathsf{E}(\hat{A}_{ij}^*|\hat{p})) + \mathsf{E}(\mathsf{Var}(\hat{A}_{ij}^*|\hat{p})) = \mathsf{Var}(\hat{p}) + \mathsf{E}(\hat{p}(1 - \hat{p})) = p(1 - p) = \mathsf{Var}(H_{ij})$$

so $\mathsf{Var}(t(\hat{A}^*, \boldsymbol{c}) = \mathsf{Var}(t(H, \boldsymbol{c}))$. $\square$

References

Levin, K. and Levina, E. (2019). Bootstrapping networks with latent space structure. *arXiv preprint arXiv:1907.10821*.