Test by counting cycles

- Below, we describe a test for

\[ H_0 : G \sim G(n, \frac{p+q}{2}) \]

\[ H_1 : G \sim \text{SBM}(n, \frac{a}{n}, \frac{b}{n}) , \text{ where } \sigma_i \sim \text{Rad}(\frac{1}{2}), \text{ and } a > b. \]

Note: Here under \( H_1 \), the underlying community label \( \sigma \) is unknown.

- The average node degree is matched under \( H_0 \) and \( H_1 \) (still slowly growing with \( n \)).

- The test is based on counting "short" cycles in \( G \).

let \( X_k \equiv \# \text{ of } k-\text{cycles (not necessarily induced) in } G. \)

- High-level intuition: For appropriately chosen \( k \), and some threshold \( z \)

\[ H_0 : X_k \leq z \ \text{whp} \]

\[ H_1 : X_k > z \ \text{whp}. \]

Then under the test \( 1\Pr[X_k > 2z], \) the sum of Type I + II errors = \( o(1) \).

- Warm-up of analysis:

Let \( G \sim G(n, \frac{d}{n}) \). Then

\[ \Pr[X_k > 0] \leq \mathbb{E}[X_k] = \binom{n}{k} \cdot k! \left( \frac{d}{n} \right)^k \leq d^k. \]

Thus, there are no cycles of gray length if \( d < 1 \).

The first-moment calculation does not tell us the existence of log cycles.
Nevertheless, it is known that the longest cycle is of length $\sqrt{n}$ whp. [Bollobás Random Graph book, Section 8.1].

Let's get back to testing $G(n, \frac{\alpha n}{2})$ vs. $SBM(n, \frac{\alpha}{2}, k)$.

Assume $\alpha > b$. Define $s = \frac{\alpha - b}{2}$, $d = \frac{\alpha + b}{2}$.

The threshold is then given by $s^2 > d$.

Since $d > s \Rightarrow s > 1$ and $\alpha > 2$ and $d > 1$.

Intuition: For $k$ not too big, $X_k$ has poisson limit under $H_0$ and $H_1$:

$H_0$: $X_k \xrightarrow{d} \text{Pois} \left( \frac{d^2}{2K} \right) \Rightarrow \mathbb{E}[X_k] = \frac{d^2}{2K}$, $\text{Var}(X_k) = \frac{d^2}{2K}$

$H_1$: $X_k \xrightarrow{d} \text{Pois} \left( \frac{d^2 + s^2}{2K} \right) \Rightarrow \mathbb{E}[X_k] = \frac{d^2 + s^2}{2K}$, $\text{Var}(X_k) = \frac{d^2 + s^2}{2K}$

Under the condition $s^2 > d$, we have:

$$\sqrt{\text{Var}(X_k) + \text{Var}(X_k)}$$

as $K$ grows. Hence the test $1 \iff X_k > \frac{d^2 + s^2}{4K}$ succeeds. (follows from Chebyshev's inequality)

Rigorous analysis:

$$\text{Chebyshev: } \mathbb{P}_{H_0} \left( X_k > \frac{d^2 + s^2}{4K} \right) \leq \frac{\text{Var}(X_k)}{(\frac{d^2 + s^2}{4K})^2} = \frac{(\frac{d^2 + s^2}{2K})^2}{(\frac{d^2 + s^2}{4K})^2}$$

Step 1: First-moment calculation.

Similarly for $K = 19n$.

$$\mathbb{P}_{H_0} \left( X_k \leq \frac{d^2 + s^2}{2K} \right)$$
\[ X_k = \frac{1}{2k} \sum_{\text{all\ ordered\ k-tuple}} \frac{1}{v_{i_1}v_{i_2}v_{i_3} \ldots v_{i_k}} \]

Under Ho:
\[ E_0[X_k] = \frac{1}{2k} \binom{n}{k} \cdot k! \cdot \frac{P(n, V_k, V_k - V_{i_k})}{(d^n)^k} \]
\[ = \frac{1}{2k} \binom{n}{k} \cdot k! \cdot \left(\frac{d}{n}\right)^k \]
\[ = \frac{n(n+1)(n+2) \ldots (n-k+1)}{2k^n} \cdot d^k = \frac{1+o(n)}{2k} \cdot d^k, \]
as long as \( k = o(n) \).

Under Hi: Consider the adjacency matrix \( A \).

Given any two vertices \( V_i, V_i+1 \), we have
\[ A_{V_iV_i+1} \sim \begin{cases} \text{Ber}(p) & \text{if } \sigma_i = \sigma_{i+1} \\ \text{Ber}(q) & \text{if } \sigma_i \neq \sigma_{i+1} \end{cases} \]

Given any \( k \)-tuple \( \{V_i, V_{i+1}, \ldots, V_k\} \) of vertices,
let \( N = \sum_{i=1}^{k-1} 1_{\sigma_i = \sigma_{i+1}} + 1_{\sigma_k = \sigma_{i+1}} \)
where \( k+1 \) is understood as 1.

\[ N = \sum_{i=1}^{k-1} 1_{\sigma_i = \sigma_{i+1}} + 1_{\sigma_k = \sigma_{\text{next}}} \]

We have \( T \sim \text{Bin}(k, \frac{1}{2}) \) and
\[ S = \begin{cases} 0 & \text{if } T \text{ is even} \\ 1 & \text{o.w.} \end{cases} \]
So that \( N = S + T \) is always even.

\[ \text{Ex: } \]

\[ v_1, v_2, v_3, \ldots, v_k \] \( T = 4 \)

\[ S = 0 \]

\[ N = 4 \]

Now, conditional on \( N = m \),

\[ \Pr \{ V_1 \sim V_2, V_3 \sim V_k \mid N = m \} = q^m \cdot p^{k-m} \]

and

\[ \Pr \{ N = m \} = \sum_{k \in \text{even}} \Pr \{ \text{Bin}(k-1, \frac{1}{2}) = m \} \cdot q^m \cdot p^{k-m} \]

\[ = \sum_{k \in \text{even}} \binom{k-1}{m-1} \left( \frac{1}{2} \right)^{k-1} \left( \frac{1}{2} \right)^{k-1} \]

\[ = \sum_{k \in \text{even}} \binom{k}{m} \left( \frac{1}{2} \right)^{k-1} \]

\[ \text{Pascal's identity} \]

Thus

\[ \Pr \{ V_1 \sim V_2, V_3 \sim V_k \mid N = m \} \]

\[ = \sum_{k \in \text{even}} \Pr \{ N = m \} \cdot q^m \cdot p^{k-m} \]

\[ = \sum_{m \text{ even}} \binom{m}{k} \left( \frac{1}{2} \right)^k + \frac{q^m \cdot p^{k-m}}{2} \cdot \binom{k}{m} \left( \frac{1}{2} \right)^{k-1} \]

\[ = \left( \frac{p+q}{2} \right)^k + \left( \frac{p-q}{2} \right)^k \]

\[ = \tilde{p}^{-k} \left( s^k + d^k \right) \quad \text{if} \quad p = \frac{\alpha}{n} \quad \text{and} \quad q = \frac{\beta}{n} \]
Thus under $H_1$:

$$E[X_k] = \frac{n(n-1)\ldots(n-k+1)}{2K^n} (s^k + d^k)$$

for $k = o(n^{1/2})$.

\[= \frac{1+o(1)}{2K} (s^k + d^k) \]

**Step 2:** Variance analysis.

We only consider the variance under the null; the calculation under $H_1$ is similar, but more tedious.

Given ordered $k$-tuple of vertices $T = (V_1, V_2, \ldots, V_k)$, define $b_T = \{V_1, V_2, \ldots, V_k\}$.

Under $H_0$:

$$\text{Var}(X_k) = \frac{1}{4K^2} \sum_{T, T'} \text{cov}(b_T, b_{T'})$$

\[= \frac{1}{4K^2} \left( \frac{1}{T \cap T'} \text{Var}(b_T) + \sum_{T \cap T' \neq \emptyset} \text{cov}(b_T, b_{T'}) \right) \]

\[+ \sum_{T \cap T' = \emptyset} \text{cov}(b_T, b_{T'}) \]

\[= 0 \quad \text{no shared pairs of nodes}\]

Now, consider two distinct $k$-cycles $T$ and $T'$ that are overlapping.

Let $I = \#$ of common edges

$V = \#$ of common vertices.
Note that:

1. \( \text{Cov}(b_T, b_{T'}) \leq \mathbb{E}[b_T b_{T'}] = p^{2k-2} \).

2. Crucially, \( V \geq k+1 \). This is because 

   \( T \cap T' \) is a forest (each connected component is a path).

   So that

   \[
   V = k + \# \text{ of connected components of } T \cap T'
   \]

   \( = k + 1 \).

   e.g.

   \( f = 3 \)

   \( V = 6 \)

\( k = 7 \)

Combining all these, we get that

\[
\sum_{\substack{T \not= T' \mid \text{Cov}(b_T, b_{T'}) \leq \mathbb{E}[b_T b_{T'}] \leq p^{2k-2} \\leq k \leq k-1}} [n]_{2k-v} [k]_v \left( \frac{d}{n} \right)^{2k-2}
\]

\[
= \sum_{\substack{T \not= T' \mid \text{Cov}(b_T, b_{T'}) \leq \mathbb{E}[b_T b_{T'}] \leq p^{2k-2} \\leq k \leq k-1}} n^{2k-v} k! \left( \frac{d}{n} \right)^{2k-2}
\]

\[
= \sum_{\substack{T \not= T' \mid \text{Cov}(b_T, b_{T'}) \leq \mathbb{E}[b_T b_{T'}] \leq p^{2k-2} \\leq k \leq k-1}} n^{2k-v} k! \left( \frac{d}{n} \right)^{2k-2}
\]

\[
= \frac{1}{n} \sum_{\substack{T \not= T' \mid \text{Cov}(b_T, b_{T'}) \leq \mathbb{E}[b_T b_{T'}] \leq p^{2k-2} \\leq k \leq k-1}} n^{2k-v} k! \left( \frac{d}{n} \right)^{2k-2}
\]

\[
\leq \frac{1}{n} \sum_{\substack{T \not= T' \mid \text{Cov}(b_T, b_{T'}) \leq \mathbb{E}[b_T b_{T'}] \leq p^{2k-2} \\leq k \leq k-1}} n^{2k-v} k! \left( \frac{d}{n} \right)^{2k-2}
\]

\[
\leq \frac{1}{n} \sum_{\substack{T \not= T' \mid \text{Cov}(b_T, b_{T'}) \leq \mathbb{E}[b_T b_{T'}] \leq p^{2k-2} \\leq k \leq k-1}} n^{2k-v} k! \left( \frac{d}{n} \right)^{2k-2}
\]

\[
\leq \frac{1}{n} \sum_{\substack{T \not= T' \mid \text{Cov}(b_T, b_{T'}) \leq \mathbb{E}[b_T b_{T'}] \leq p^{2k-2} \\leq k \leq k-1}} n^{2k-v} k! \left( \frac{d}{n} \right)^{2k-2}
\]

\[
\leq \frac{1}{n} \sum_{\substack{T \not= T' \mid \text{Cov}(b_T, b_{T'}) \leq \mathbb{E}[b_T b_{T'}] \leq p^{2k-2} \\leq k \leq k-1}} n^{2k-v} k! \left( \frac{d}{n} \right)^{2k-2}
\]
\[ = o(1), \text{ provided that } k = o\left(\frac{\log n}{\log\log n}\right) \]

Thus we get that
\[ \text{Var}(X_k) = \frac{1}{4k^2} \sum_{T, T' : T \neq T'} \text{Var}(b\mathbf{T}) + o(1)\]
\[ = \frac{1}{4k^2} \sum_{T, T' : T \neq T'} (\frac{\mathbf{d}}{n})^k \left(1 - (\frac{\mathbf{d}}{n})^k\right) + o(1) \]
\[ = \frac{1}{4k^2} \ln \mathbf{d} \cdot (\frac{\mathbf{d}}{n})^k \left(1 - (\frac{\mathbf{d}}{n})^k\right) + o(1) \]
\[ = \frac{1}{2k} \mathbf{d}^k (1 + o(1)). \]

\textbf{Remark.} In fact, we can show that
\[ \mathbb{E} \left[ [X_k]^m \right] = \mathbb{E} \left[ (X_k)(X_k-1) \cdots (X_k-m+1) \right] \]
\[ \sim \left(\frac{\mathbf{d}^k}{2k}\right)^m \]
\[ \Rightarrow X_k \Rightarrow \text{pois} \left(\frac{\mathbf{d}^k}{2k}\right) \]
follows from method of moments.

\textbf{Thm.:} Let \( X_k \) be a sequence of r.v.s. for every fixed \( m \geq 1 \).
\[ \mathbb{E} \left[ [X_k]^m \right] \Rightarrow \lambda^m. \]
Then \( X_k \overset{d}{\rightarrow} \text{pois}(\lambda) \).

**Question:** Can we count \# of \( k \)-cycles in poly-time for grows \( k \)?

**Answer:** Exhaustive search takes \( n^k \) time, not poly-time when \( k \geq 80 \).

- We can approximately count \# of \( k \)-cycles when the graph is sparse.
- Approximately count cycles in poly-time

**Def** \(( \ell \text{-tangle free})\). An \( \ell \)-tangle is a connected subgraph of diameter at most \( 2 \ell \) that contains at least \( 2 \) cycles.

A graph \( G \) is called \( \ell \)-tangle free if no subgraph of \( G \) is an \( \ell \)-tangle. In other words, for all \( v \in V(G) \), its \( \ell \)-hop neighborhood \( N_{\ell}(v) \) contains at most one cycle.

**Example:**

![Diagram](image)

\( l = 3 \) without the edge, then \( N_{\ell}(v) \) is \( \ell \)-tangle free.

**Lemma:** If \( G \in G(n, \frac{\ell}{n}) \) and \( \ell \) is a constant, then \( G \) is \( \ell \)-tangle free if \( \ell = o \left( \log n \right) \). (In general, \( \ell \log \ell = c \cdot \log n \) for suit constant \( c \) suffices.)
Suppose $G$ contains an $l$-tangle. Then $G$ must contain a subgraph $H$ of the following form:

With $m$ edges and $v$ vertices, such that $m \leq 4l$ and $m \geq v + 1$.

There are $O(l^3)$ such graph $H$ up to isomorphism, as there are $O(e)$ choices of length for each cycle and the connecting path.

For each such graph $H$,

By union bound, $H$ is a subgraph of $G$ w.p.

$$n^v \cdot \left( \frac{d}{n} \right)^m \leq n^{m-1} \cdot \left( \frac{d}{n} \right)^m = \frac{d^4}{n}$$

Thus $\Pr \{ G \text{ contains an } l \text{-tangle} \} \leq O(l^3) \cdot \frac{d^4}{n} \to 0$, when $\log d \ll \log n$.

Next, we discuss the connection between County and linear algebra.

Let's start with $k=3$.

**Ex**: (County Triangles). Suppose $A$ is the adjacency matrix of $G$.

Given any vertex $v$ in $G$,

$$(A^3)_{vv} = \sum_{a,b} A_{va} A_{ab} A_{bv} = 2 \times \# \text{ of triangles incident to } v.$$
Thus $\text{Tr}(A^3) = 6 \times \text{# of triangles in } G$.

- To count $K$-cycles, one can compute $\text{Tr}(A^K)$, which can be done in $O(n^3)$ of EVD, as $\text{Tr}(A^K) = \sum_{i=1}^{n} \lambda_i^K$.

But $\text{Tr}(A^K) = \# \text{ of closed walks of length } K \Rightarrow \# \text{ of } K\text{-cycles}.

- Strategy: use the tangle-free structure and count $\# \text{ of non-backtracking (NB) walks}$.

**Def [Non-back-tracking walk]:** We say

- $(V_1, V_2, \ldots, V_k)$ is a NB walk, if $V_t \neq V_{t+1}$ and $V_t \neq V_{t-2}$ for all $t$.

- $(V_1, V_2, \ldots, V_k)$ is a closed NB walk, if $(V_1, \ldots, V_k)$ is a NB walk and $V_1 = V_k$.

**Ex:**

Not a NB walk.

Closed NB walk.

Consequence: Conditioned on $G$ being 2k-tangle-free, any closed NB walk of $k$ steps is either a $K$-cycle or an $M$-cycle traversed for $\frac{k}{m}$ times. Otherwise, we have a 2k-tangle such as two short cycles connected by a path.
This reduces the problem to counting \# of closed NB walks of length \( m \) for all \( m=1,2,\ldots k \).

Count closed NB walk of length \( m \) recursively:

Let \( N_{uv}^m = \# of NB walks from \( u \) to \( v \) of length \( m \)

Let \( N_m = \sum_{u\in V} N_{uu}^m \)

Then
\[
2kX_k = N_k - \sum_{m:\text{index}} \sum_{2mXm} \]

\[
6X_3 = N_3.
\]

Let's compute \( N_{uv}^m \):

\[
N_{uv}^{m+1} = \sum_{u\in V} N_{uw}^m - (\text{all})N_{uv}^m
\]

In matrix notation, let \( N^m = (N_{uv}^m) \) and \( D = \text{diag}(d_{uv}) \). Then

\[
N^{m+1} = N^m \cdot A - N^{m-1} (D-I)
\]

\[
N^0 = A, \quad N^{(2)} = A^2 - D
\]

zero-diagonal