

Lectures 2: Random Graphs

Radu Balan

Department of Mathematics, AMSC and NWC
University of Maryland, College Park, MD

March 28, 2024

The Erdős-Rényi class $\mathcal{G}_{n,p}$

Definition

Today we discuss about random graphs. The *Erdős-Rényi class* $\mathcal{G}_{n,p}$ of random graphs is defined as follows.

The Erdős-Rényi class $\mathcal{G}_{n,p}$

Definition

Today we discuss about random graphs. The *Erdős-Rényi class* $\mathcal{G}_{n,p}$ of random graphs is defined as follows.

Let \mathcal{V} denote the set of n vertices, $\mathcal{V} = \{1, 2, \dots, n\}$, and let \mathcal{G} denote the

set of all graphs with vertices \mathcal{V} . There are exactly $2^{\binom{n}{2}}$ such graphs.

The probability mass function on \mathcal{G} , $P : \mathcal{G} \rightarrow [0, 1]$, is obtained by assuming that, as random variables, edges are independent from one another, and each edge occurs with probability $p \in [0, 1]$. Thus a graph $G \in \mathcal{G}$ with m edges will have probability $P(G)$ given by

$$P(G) = p^m (1 - p)^{\binom{n}{2} - m}.$$

(explain why)

The Erdős-Rényi class $\mathcal{G}_{n,p}$

Probability space

Formally, $\mathcal{G}_{n,p}$ stands for the the probability space (\mathcal{G}, P) composed of the set \mathcal{G} of all graphs with n vertices, and the probability mass function P defined above.

The Erdős-Rényi class $\mathcal{G}_{n,p}$

Probability space

Formally, $\mathcal{G}_{n,p}$ stands for the the probability space (\mathcal{G}, P) composed of the set \mathcal{G} of all graphs with n vertices, and the probability mass function P defined above.

A reformulation of P : Let $G = (\mathcal{V}, \mathcal{E})$ be a graph with n vertices and m edges and let A be its adjacency matrix. Then:

$$\begin{aligned} P(G) &= \prod_{(i,j) \in \mathcal{E}} \text{Prob}((i,j) \text{ is an edge}) \prod_{(i,j) \notin \mathcal{E}} \text{Prob}((i,j) \text{ is not an edge}) = \\ &= \prod_{1 \leq i < j \leq n} p^{A_{i,j}} (1-p)^{1-A_{i,j}} \end{aligned}$$

where the product is over all ordered pairs (i,j) with $1 \leq i < j \leq n$. Note:

$$|\{(i,j), 1 \leq i < j \leq n\}| = \binom{n}{2} \quad \& \quad |\{(i,j) \in \mathcal{E}\}| = |\mathcal{E}| = m = \sum_{1 \leq i < j \leq n} A_{i,j}.$$

The Erdős-Rényi class $\mathcal{G}_{n,p}$

Computations in $\mathcal{G}_{n,p}$

How to compute the expected number of edges of a graph in $\mathcal{G}_{n,p}$?

The Erdős-Rényi class $\mathcal{G}_{n,p}$

Computations in $\mathcal{G}_{n,p}$

How to compute the expected number of edges of a graph in $\mathcal{G}_{n,p}$?

Let $X_2 : \mathcal{G} \rightarrow \{0, 1, \dots, \binom{n}{2}\}$ be the random variable of *number of edges of a graph G* .

$$X_2 = \sum_{1 \leq i < j \leq n} 1_{(i,j)} \quad , \quad 1_{(i,j)}(G) = \begin{cases} 1 & \text{if } (i,j) \text{ is edge in } G \\ 0 & \text{if otherwise} \end{cases}$$

Use linearity and the fact that $\mathbb{E}[1_{(i,j)}] = \text{Prob}((i,j) \in \mathcal{E}) = p$ to obtain:

$$\mathbb{E}[\text{Number of Edges}] = \binom{n}{2} p = \frac{n(n-1)}{2} p$$

The Erdős-Rényi class $\mathcal{G}_{n,p}$

MLE of p

Given a realization G of a graph with n vertices and m edges, how to estimate the most likely p that explains the graph.

Concept: The Maximum Likelihood Estimator (MLE).

In statistics: The MLE of a parameter θ given an observation x of a random variable $X \sim p_X(x; \theta)$ is the value θ that maximizes the probability $P_X(x; \theta)$:

$$\theta_{MLE} = \operatorname{argmax}_{\theta} P_X(x; \theta).$$

In our case: our observation G has m edges. We know

$$P(G; p) = p^m (1 - p)^{\binom{n}{2} - m}.$$

The Erdős-Rényi class $\mathcal{G}_{n,p}$

MLE of p

Lemma

Given a random graph with n vertices and m edges, the MLE estimator of p is

$$p_{MLE} = \frac{m}{\binom{n}{2}} = \frac{2m}{n(n-1)}.$$

The Erdős-Rényi class $\mathcal{G}_{n,p}$

MLE of p

Lemma

Given a random graph with n vertices and m edges, the MLE estimator of p is

$$p_{MLE} = \frac{m}{\binom{n}{2}} = \frac{2m}{n(n-1)}.$$

Why

Note $\log(P(G; p)) = m \log(p) + \left(\binom{n}{2} - m \right) \log(1-p)$ and solve for p :

$$\frac{d \log(P)}{dp} = \frac{m}{p} - \frac{\binom{n}{2} - m}{1-p} = 0.$$

The Erdős-Rényi class $\mathcal{G}_{n,p}$

Method of Moments Estimator for p

An alternative parameter estimation method is the moment matching method. Given a likelihood function for observed data $p(x; \theta)$ and a sequence of observations (x_1, x_2, \dots, x_N) , the moment matching method computes the parameters $\theta \in \mathbb{R}^d$ by solving the system of equations:

$$\mathbb{E}[X] = \frac{1}{N} \sum_{t=1}^N x_t \quad \dots \quad \mathbb{E}[X^d] = \frac{1}{N} \sum_{t=1}^N x_t^d$$

(or unbiased estimates of the moments). In particular, for the Erdős-Rényi class, we match the first moment with the observation: $\frac{n(n-1)}{2} p = m$.

Hence

$$p_{MM} = \frac{2m}{n(n-1)},$$

same as the MLE estimator.

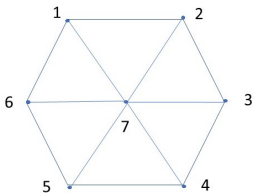
Cliques

q -cliques

Definition

Given a graph $G = (\mathcal{V}, \mathcal{E})$, a subset of q vertices $S \subset \mathcal{V}$ is called a q -clique if the subgraph $(S, \mathcal{E}|_{S \times S})$ is complete.

In other words, S is a q -clique if for every $i \neq j \in S$, $(i, j) \in \mathcal{E}$ (or $(j, i) \in \mathcal{E}$), that is, (i, j) is an edge in G .



- Each edge is a 2-clique.

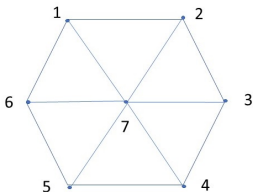
Cliques

q -cliques

Definition

Given a graph $G = (\mathcal{V}, \mathcal{E})$, a subset of q vertices $S \subset \mathcal{V}$ is called a q -clique if the subgraph $(S, \mathcal{E}|_{S \times S})$ is complete.

In other words, S is a q -clique if for every $i \neq j \in S$, $(i, j) \in \mathcal{E}$ (or $(j, i) \in \mathcal{E}$), that is, (i, j) is an edge in G .



- Each edge is a 2-clique.
- $\{1, 2, 7\}$ is a 3-clique. And so are $\{2, 3, 7\}$, $\{3, 4, 7\}$, $\{4, 5, 7\}$, $\{5, 6, 7\}$, $\{1, 6, 7\}$

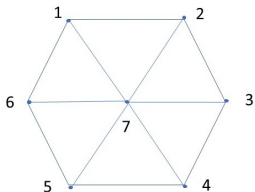
Cliques

q -cliques

Definition

Given a graph $G = (\mathcal{V}, \mathcal{E})$, a subset of q vertices $S \subset \mathcal{V}$ is called a q -clique if the subgraph $(S, \mathcal{E}|_{S \times S})$ is complete.

In other words, S is a q -clique if for every $i \neq j \in S$, $(i, j) \in \mathcal{E}$ (or $(j, i) \in \mathcal{E}$), that is, (i, j) is an edge in G .



- Each edge is a 2-clique.
- $\{1, 2, 7\}$ is a 3-clique. And so are $\{2, 3, 7\}$, $\{3, 4, 7\}$, $\{4, 5, 7\}$, $\{5, 6, 7\}$, $\{1, 6, 7\}$
- There is no k -clique, with $k \geq 4$.

Finding the largest clique is a NP-hard problem, see for instance:

https://en.wikipedia.org/wiki/Clique_problem

The Erdős-Rényi class $\mathcal{G}_{n,p}$

Computations in $\mathcal{G}_{n,p}$: q -cliques

How to compute the expected number of q -cliques?

The Erdős-Rényi class $\mathcal{G}_{n,p}$

Computations in $\mathcal{G}_{n,p}$: q -cliques

How to compute the expected number of q -cliques?

For $k = 2$ we computed earlier the number of edges, which is also the number of 2-cliques.

We shall compute now the number of 3-cliques: triangles, or 3-cycles.

Let $X_3 : \mathcal{G} \rightarrow \mathbb{N}$ be the random variable of number of 3-cliques. Note the

maximum number of 3-cliques is $\binom{n}{3}$.

Let S_3 denote the set of all distinct 3-cliques of the complete graph with n vertices, $S_3 = \{(i, j, k), 1 \leq i < j < k \leq n\}$.

Let

$$1_{(i,j,k)}(G) = \begin{cases} 1 & \text{if } (i, j, k) \text{ is a 3-clique in } G \\ 0 & \text{if otherwise} \end{cases}$$

The Erdős-Rényi class $\mathcal{G}_{n,p}$

Expectation of the number of 3-cliques

Note: $X_3 = \sum_{(i,j,k) \in \mathcal{S}_3} 1_{(i,j,k)}$. Thus

$$\mathbb{E}[X_3] = \sum_{(i,j,k) \in \mathcal{S}_3} \mathbb{E}[1_{(i,j,k)}] = \sum_{(i,j,k) \in \mathcal{S}_3} \text{Prob}((i,j,k) \text{ is a clique}).$$

Since $\text{Prob}((i,j,k) \text{ is a clique}) = p^3$ we obtain:

$$\mathbb{E}[\text{Number of 3-cliques}] = \binom{n}{3} p^3 = \frac{n(n-1)(n-2)}{6} p^3.$$

The Erdős-Rényi class $\mathcal{G}_{n,p}$

Number of 3 cliques

Assume we observe a graph G with n vertices and m edges. What would be the expected number N_3 of 3-cliques?

$$\mathbb{E}[X_3 | X_2 = m] = \frac{1}{L} \sum_{k=1}^L X_3(G_k)$$

where L denotes the number of graphs with m edges and n vertices, and G_1, \dots, G_L is an enumeration of these graphs.

The Erdős-Rényi class $\mathcal{G}_{n,p}$

Number of 3 cliques

Assume we observe a graph G with n vertices and m edges. What would be the expected number N_3 of 3-cliques?

$$\mathbb{E}[X_3 | X_2 = m] = \frac{1}{L} \sum_{k=1}^L X_3(G_k)$$

where L denotes the number of graphs with m edges and n vertices, and G_1, \dots, G_L is an enumeration of these graphs.

We approximate:

$$\mathbb{E}[X_3 | X_2 = m] \approx \mathbb{E}[X_3; p = p_{MLE}(m)]$$

and obtain:

$$E[X_3 | X_2 = m] \approx \frac{4(n-2)}{3n^2(n-1)^2} m^3.$$

The Erdős-Rényi class $\mathcal{G}_{n,p}$

Expectation of the number of q -cliques

Let $X_q : \mathcal{G} \rightarrow \mathbb{N}$ be the random variable of number of q -cliques. Note the maximum number of q -cliques is $\binom{n}{q}$.

Let S_q denote the set of all distinct q -cliques of the complete graph with n vertices, $S_q = \{(i_1, i_2, \dots, i_q), 1 \leq i_1 < i_2 < \dots < i_q \leq n\}$. Note

$$|S_q| = \binom{n}{q}.$$

Let

$$1_{(i_1, i_2, \dots, i_q)}(G) = \begin{cases} 1 & \text{if } (i_1, i_2, \dots, i_q) \text{ is a } q\text{-clique in } G \\ 0 & \text{if otherwise} \end{cases}$$

The Erdős-Rényi class $\mathcal{G}_{n,p}$

Expectation of the number of q -cliques

Since $X_q = \sum_{(i_1, \dots, i_q) \in S_q} \mathbf{1}_{i_1, \dots, i_q}$ and

$Prob((i_1, \dots, i_q) \text{ is a clique}) = p \binom{q}{2}$ we obtain:

$$\mathbb{E}[\text{Number of } q\text{-cliques}] = \binom{n}{q} p^{q(q-1)/2}.$$

The Erdős-Rényi class $\mathcal{G}_{n,p}$

Expectation of the number of q -cliques

Since $X_q = \sum_{(i_1, \dots, i_q) \in S_q} \mathbf{1}_{i_1, \dots, i_q}$ and

$Prob((i_1, \dots, i_q) \text{ is a clique}) = p \binom{q}{2}$ we obtain:

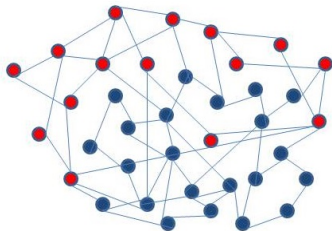
$$\mathbb{E}[\text{Number of } q\text{-cliques}] = \binom{n}{q} p^{q(q-1)/2}.$$

Using a similar argument as before, if G has m edges, then

$$\mathbb{E}[X_q | X_2 = m] \approx \binom{n}{q} \left(\frac{2m}{n(n-1)} \right)^{q(q-1)/2}.$$

The Stochastic Block Model

The *Stochastic Block Model* (SBM), a.k.a. *Planted Partition Model*, was introduced in mathematical sociology by Holland, Laskey and Leinhardt in 1983 and by Wang and Wong in 1987. Here we follow Abbe (2017).



A Stochastic Block Model with $k = 2$ classes ('red' and 'blue') over $n = 15 + 22 = 37$ nodes. Number of edges: $m_{rr} = 21$, $m_{rb} = 6$, $m_{bb} = 35$.

Figure: Example of a SBM

The Stochastic Block Model

The general SBM

Data. Let n be a positive integer (the number of vertices), k be a positive integer (the number of communities), $\mathbf{p} = (p_1, p_2, \dots, p_k)$ be a probability vector on $[k] := \{1, 2, \dots, k\}$ (the prior on the k communities), and Q be a $k \times k$ symmetric matrix with entries in $[0, 1]$ (the connectivity probabilities).

Definition

The pair (Z, G) is drawn under $SBM(n, \mathbf{p}, Q)$ if Z is an n -dimensional random vector with i.i.d. components distributed under \mathbf{p} , and G is an n -vertex graph where vertices i and j are connected with probability Q_{Z_i, Z_j} , independently of other pairs of vertices.

The *community sets* are defined by $\Omega_i = \Omega_i(Z) = \{v \in [n] , Z_v = i\}$, $1 \leq i \leq k$.

The Stochastic Block Model

The Symmetric SBM (SSBM)

Definition

The pair (Z, G) is drawn under $SSBM(n, k, a, b)$ if Z is an n -dimensional random vector with i.i.d. components uniformly distributed over $[k] = \{1, 2, \dots, k\}$, and G is an n -vertex graph where distinct vertices i and j are connected with probability a if $Z_i = Z_j$ and probability b if $Z_i \neq Z_j$, independently of other pairs of vertices.

Data:

- the number of vertices: n ;
- the number of communities: k ;
- prior on k communities: $\mathbf{p} = (\frac{1}{k}, \frac{1}{k}, \dots, \frac{1}{k})$ on $[k] := \{1, 2, \dots, k\}$;
- connectivity probabilities: Q

$$Q = \begin{bmatrix} a & b & \dots & b \\ b & a & \dots & b \\ \vdots & \vdots & \ddots & \vdots \\ b & b & \dots & a \end{bmatrix}.$$

The Erdős-Rényi random graph is obtained when $a = b = p$.

The Binary Symmetric Stochastic Block Model

Distributions (1)

Consider a realization (Z, G) drawn randomly under $SSBM(n, 2, a, b)$ that models two communities. This means every node belongs with equal probability to either community, 1 or 2: $Z = (Z_1, Z_2, \dots, z_n)$, where $Z_i \in \{1, 2\}$, $P(Z_i = 1) = P(Z_i = 2) = \frac{1}{2}$. The graph G of n nodes has adjacency matrix A . The conditional probability of realization A given the vector Z :

$$\begin{aligned} P(A|Z) &= \prod_{1 \leq u < v \leq n} Q_{Z_u, Z_v}^{A_{u,v}} (1 - Q_{Z_u, Z_v})^{1 - A_{u,v}} = \\ &= a^{m_{11} + m_{22}} b^{m_{12}} (1 - a)^{m_{11}^c + m_{22}^c} (1 - b)^{m_{12}^c} \end{aligned}$$

where m_{11}, m_{22} are the number of edges inside community 1, respectively 2, m_{12} is the number of edges between the two communities, and $m_{11}^c, m_{22}^c, m_{12}^c$ are the number of missing edges inside each community/between the two communities.

The Binary Symmetric Stochastic Block Model

Distributions (2)

Explicitly these numbers are given by:

$$m_{11} = \# \text{Edges inside community 1} = \sum_{\substack{i < j \\ i, j \in \Omega_1}} A_{i,j}$$

$$m_{11}^c = \binom{n_1}{2} - m_{11} \quad n_1 = |\Omega_1|$$

$$m_{22} = \# \text{Edges inside community 2} = \sum_{\substack{i < j \\ i, j \in \Omega_2}} A_{i,j}$$

$$m_{22}^c = \binom{n_2}{2} - m_{22} \quad n_2 = |\Omega_2|$$

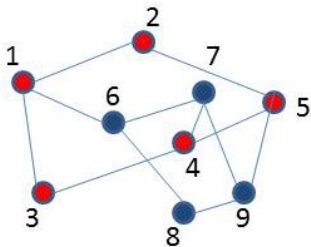
The Binary Symmetric Stochastic Block Model

Distributions (3)

$$m_{12} = \# \text{Edges between community 1 and 2} = \sum_{\substack{i \in \Omega_1 \\ j \in \Omega_2}} A_{i,j}$$

$$m_{12}^c = n_1 n_2 - m_{12}$$

Example:



$$n = 9, \quad \Omega_1 = \{1, 2, 3, 4, 5\}, \quad \Omega_2 = \{6, 7, 8, 9\}.$$

$$m_{11} = 5, \quad m_{11}^c = 5$$

$$m_{22} = 4, \quad m_{22}^c = 2$$

$$m_{12} = 3, \quad m_{12}^c = 17$$

The Stochastic Block Model

Community Detection

The main problem: Community Detection.

This means a partition of the set of vertices $\mathcal{V} = \{1, 2, \dots, n\}$ compatible with the observed graph G for a given connectivity probability matrix W . To formulate mathematically we need to define the *agreement* between two community vectors.

Definition

The *agreement* between two community vectors $x, y \in [k]^n$ is obtained by maximizing the number of common components of these two vectors over all possible relabelling (i.e., permutations):

$$\text{Agr}(x, y) = \max_{\pi \in S_k} \frac{1}{n} \sum_{i=1}^n \mathbf{1}(x_i = \pi(y_i))$$

where S_k denotes the group of permutations.

The Binary Symmetric Stochastic Block Model

Model Calibration: Supervised Learning

How to estimate parameters a, b in the 2-community symmetric stochastic block model $SSBM(n, 2, a, b)$. Use the Maximum Likelihood Estimator (MLE):

$$(a_{MLE}, b_{MLE}) = \operatorname{argmax}_{a,b} \operatorname{Prob}(G|Z, a, b)$$

Setup: Assume we have access to a training (i.e., labelled) data set (Z, G) . Then for parameters a, b maximize:

$$a^{m_{11}+m_{22}}(1-a)^{m_{11}^c+m_{22}^c} b^{m_{12}}(1-b)^{m_{12}^c}$$

Take the logarithm and obtain:

$$a_{MLE} = \frac{m_{11} + m_{22}}{\binom{n_1}{2} + \binom{n_2}{2}} = \frac{2(m_{11} + m_{22})}{n_1(n_1 - 1) + n_2(n_2 - 1)}$$

$$b_{MLE} = \frac{m_{12}}{n_1 n_2}$$

The Binary Symmetric Stochastic Block Model

Model Calibration: Unsupervised Learning

Assume we have access to only one realization $G = (\mathcal{V}, A)$ of the random graph drawn from a binary symmetric SBM $SSBM(n, 2, a, b)$. The MLE is hard to solve. Instead we use the Method of Moment Matching. Since there are two parameters to estimate, a and b , we need two equations. We choose to match the numbers of 2-cliques (edges) and the number of 3-cliques. The expectations are computed by conditioning first on $n_1 = |\Omega_1|$ the size of partition, with $n_2 = n - n_1$:

$$\mathbb{E}[X_2 | n_1] = \binom{n_1}{2} a + n_1 n_2 b + \binom{n_2}{2} a$$

$$\mathbb{E}[X_3 | n_1] = \binom{n_1}{3} a^3 + \left[\binom{n_1}{2} n_2 + n_1 \binom{n_2}{2} \right] ab^2 + \binom{n_2}{3} a^3$$

$$\begin{aligned}
 \mathbb{E}[X_2|n_1] &= \binom{n_1}{2} a + n_1 n_2 b + \binom{n_2}{2} a = \\
 &= \frac{n_1(n_1 - 1) + (n - n_1)(n - n_1 - 1)}{2} a + n_1(n - n_1)b \\
 &= \frac{n_1^2 - n_1 + n^2 - 2nn_1 + n_1^2 - n + n_1}{2} a + (nn_1 - n_1^2)b \\
 &= \left(n_1^2 - nn_1 + \frac{n(n-1)}{2} \right) a + (nn_1 - n_1^2)b
 \end{aligned}$$

Next compute the expectation of the number of edges by double expectation. To do so we need

$$\begin{aligned}
 \mathbb{E}[n_1] &= \mathbb{E} \left[\sum_{v=1}^n 1_{Z_v=1} \right] = \frac{n}{2} \\
 \mathbb{E}[n_1^2] &= \mathbb{E} \left[\left(\sum_{v=1}^n 1_{Z_v=1} \right)^2 \right] = n \frac{1}{2} + 2 \frac{n(n-1)}{2} \frac{1}{4} = \frac{n(n+1)}{4}
 \end{aligned}$$

Thus

$$\begin{aligned}\mathbb{E}[X_2] &= \mathbb{E}[\mathbb{E}[X_2|n_1]] = \left(\frac{n^2+n}{4} - \frac{n^2}{2} + \frac{n^2-n}{2}\right)a + \left(\frac{n^2}{2} - \frac{n^2+n}{4}\right)b = \\ &= \frac{n^2-n}{4}(a+b)\end{aligned}$$

Similarly,

$$\begin{aligned}\mathbb{E}[X_3|n_1] &= \binom{n_1}{3}a^3 + \left[\binom{n_1}{2}n_2 + n_1\binom{n_2}{2}\right]ab^2 + \binom{n_2}{3}a^3 \\ &= \frac{n_1(n_1-1)(n_1-2) + n_2(n_2-1)(n_2-2)}{6}a^3 + \frac{n_1n_2(n_1-1+n_2-1)}{2}ab^2 \\ &= \frac{n_1^3 + n_2^3 - 3(n_1^2 + n_2^2) + 2(n_1 + n_2)}{6}a^3 + \frac{(nn_1 - n_1^2)(n-2)}{2}ab^2 \\ &= \frac{(n_1 + n_2)(n_1^2 - n_1n_2 + n_2^2) - 3(n_1^2 + n_2^2) + 2n}{6}a^3 + \frac{(nn_1 - n_1^2)(n-2)}{2}ab^2\end{aligned}$$

$$= \frac{(n-3)(n^2 - 2nn_1 + 2n_1^2) - nn_1(n - n_1) + 2n}{6} a^3 + \frac{(nn_1 - n_1^2)(n-2)}{2} ab^2$$

$$= \frac{n^3 - 3n^2 + 2n + (3n-6)n_1^2 - (3n^2 - 6n)n_1}{6} a^3 + \frac{(nn_1 - n_1^2)(n-2)}{2} ab^2$$

Substitute $\mathbb{E}[n_1] = \frac{n}{2}$ and $\mathbb{E}[n_1^2] = \frac{n^2+n}{4}$:

$$\mathbb{E}[X_3] = \frac{n(n-2)}{6} \left(n - 1 + \frac{3}{4}(n+1) - \frac{3}{2}n \right) a^3 + \frac{n(n-2) \left(\frac{n}{2} - \frac{n+1}{4} \right)}{2} ab^2$$

$$= \frac{n(n-1)(n-2)}{24} a^3 + \frac{n(n-1)(n-2)}{8} ab^2 = \frac{n(n-1)(n-2)}{24} (a^3 + 3ab^2)$$

The Binary Symmetric Stochastic Block Model

Model Calibration: Unsupervised Learning (2)

Assuming the graph has m 2-cliques (=edges) and t 3-cliques (=triangles) then by the moment matching method:

$$m = \frac{n(n-1)}{4}(a+b) \quad , \quad t = \frac{n(n-1)(n-2)}{24}(a^3 + 3ab^2)$$

Note: the $SSBM(n, 2, a, b)$ class reduces to the Erdős-Rényi class $\mathcal{G}_{n,p}$ if $a = b = p$.

From where we solve for a and b in terms of n , m and t : Let $c_1 = \frac{4m}{n(n-1)}$ and $c_2 = \frac{24t}{n(n-1)(n-2)}$. Thus $b = c_1 - a$ and

$$4a^3 - 6c_1a^2 + 3c_1^2a - c_2 = 0 \Rightarrow (2a - c_1)^3 + c_1^3 - 2c_2 = 0$$

Thus:

$$a_{MM} = \frac{1}{2} \left(c_1 + \sqrt[3]{2c_2 - c_1^3} \right) \quad , \quad b_{MM} = \frac{1}{2} \left(c_1 - \sqrt[3]{2c_2 - c_1^3} \right)$$

The Binary Symmetric Stochastic Block Model

Model Calibration: Unsupervised Learning. Modified Estimator

The closed form expression deduced earlier using the moment matching method may produce un-feasible solutions. Specifically, the estimates a_{MM}, b_{MM} may not remain in the range $[0, 1]$. Now we derive a modified estimator that satisfies the feasibility constraints $a, b \in [0, 1]$.

Our designing principle was to satisfy *exactly*:

$$m = \mathbb{E}[X_2] \quad , \quad t = \mathbb{E}[X_3]$$

Instead the modified estimator will satisfy the first constraint exactly, but will strive to satisfy the second constraint as much as possible, Specifically, it solves the following optimization problem:

$$\begin{aligned} & \text{minimize} && |\mathbb{E}[X_3] - t| \\ & \text{subject to :} && \\ & && m = \mathbb{E}[X_2] \\ & && 0 \leq a, b \leq 1 \end{aligned}$$

The Binary Symmetric Stochastic Block Model

Model Calibration: Unsupervised Learning. Modified Estimator (2)

Substituting $a + b = 2p = \frac{4m}{n(n-1)}$ into the objective function, after a bit of algebra we obtain:

$$\frac{6}{n(n-1)(n-2)} |t - \mathbb{E}[X_3]| = |(a-p)^3 - \delta|$$

where $p = \frac{2m}{n(n-1)}$, $\delta = \frac{6t}{n(n-1)(n-2)} - p^3$. Let $P(x) = (x-p)^3 - \delta$. Note $P'(x) = 3(x-p)^2 \geq 0$. Hence $x \mapsto P(x)$ is monotone increasing (in fact, strictly increasing).

On the other hand, $b = 2p - a$ and the constraint $b \in [0, 1]$ imply $0 \leq 2p - a \leq 1$. Since $a \in [0, 1]$ we obtain:

$$\max(0, 2p - 1) \leq a \leq \min(1, 2p)$$

With $A_1 = \max(0, 2p - 1)$ and $A_2 = \min(1, 2p)$ we obtain:

$$\begin{aligned} & \text{minimize } |P(a)| \\ & A_1 \leq a \leq A_2 \end{aligned}$$

The Binary Symmetric Stochastic Block Model Calibration

Algorithm 1

The last optimization problem can be solved exactly.
The solutions is as follows:

Algorithm (1)

Input: n, m, t .

① *Compute:*

$$p = \frac{2m}{n(n-1)}, \delta = \frac{6t}{n(n-1)(n-2)} - p^3$$

$$A_1 = \max(0, 2p - 1), A_2 = \min(1, 2p)$$

$$P(A_1) = (A_1 - p)^3 - \delta, P(A_2) = (A_2 - p)^3 - \delta.$$

The Binary Symmetric Stochastic Block Model Calibration

Algorithm 1 - cont'ed

Algorithm (1 continued)

- ② *Test and compute the Constrained Moment Matching estimates:*
- *If $P(A_1) \leq 0 \leq P(A_2)$ then*

$$a_{CMM} = p + \sqrt[3]{\delta} \quad , \quad b_{CMM} = p - \sqrt[3]{\delta}$$

- *If $P(A_2) < 0$ then*

$$a_{CMM} = A_2 \quad , \quad b_{CMM} = 2p - A_2$$

- *If $P(A_1) > 0$ then*

$$a_{CMM} = A_1 \quad , \quad b_{CMM} = 2p - A_1$$

Output: a_{CMM} and b_{CMM} .

The Binary Symmetric Stochastic Block Model Calibration

Algorithm 2

While the Algorithm 1 produces estimates $a_{CMM}, b_{CMM} \in [0, 1]$ it is often the case that one would like to obtain $a, b > 0$. The following algorithm provides such an “engineering fix”:

Algorithm (2)

Input: n, m, t .

① *Compute:*

$$p = \frac{2m}{n(n-1)}, \delta = \frac{6t}{n(n-1)(n-2)} - p^3$$

$$A_1 = \max(0, 2p - 1), A_2 = \min(1, 2p)$$

$$P(A_1) = (A_1 - p)^3 - \delta, P(A_2) = (A_2 - p)^3 - \delta.$$

The Binary Symmetric Stochastic Block Model Calibration

Algorithm 2 - cont'ed

Algorithm (2 continued)

② *Test and compute:*

- *If $P(A_1) \leq 0 \leq P(A_2)$ then*

$$a_{CMM} = p + \sqrt[3]{\delta} \quad , \quad b_{CMM} = p - \sqrt[3]{\delta}$$

- *If $P(A_2) < 0$ then*

$$a_{CMM} = A_2 \quad , \quad b_{CMM} = 2p - A_2$$

- *If $P(A_1) > 0$ then*

$$a_{CMM} = A_1 \quad , \quad b_{CMM} = 2p - A_1$$

The Binary Symmetric Stochastic Block Model Calibration

Algorithm 2 - cont'ed

Algorithm (2 continued)

③ *Adjust to produce the Modified Constrained Moment Matching estimates*

- *If $0 < a_{CMM}, b_{CMM}$ then*

$$a_{MCMM} = a_{CMM} \quad , \quad b_{MCMM} = b_{CMM}$$

- *If $b_{CMM} = 0$ then*

$$a_{MCMM} = 0.99a_{CMM} \quad , \quad b_{MCMM} = 0.01a_{CMM}$$

- *If $a_{CMM} = 0$ then*

$$a_{MCMM} = 0.01b_{CMM} \quad , \quad b_{MCMM} = 0.99b_{CMM}$$

Output: a_{MCMM} and b_{MCMM} .

The Stochastic Block Model

Types of Community Detection Algorithms

Types of *algorithm*:

Let $(Z, G) \sim SBM(n, \mathfrak{p}, Q)$. Then the following recovery requirements are solved if there exists an algorithm that takes G as input and outputs $\hat{Z} = \hat{Z}(G)$ such that:

- **Exact recovery:** $P\{Agr(Z, \hat{Z}) = 1\} = 1 - o(1)$
- **Almost exact recovery:** $P\{Agr(Z, \hat{Z}) = 1 - o(1)\} = 1 - o(1)$
- **Partial recovery:** $P\{Agr(Z, \hat{Z}) \geq \alpha\} = 1 - o(1), \alpha \in (0, 1)$.

Note these definitions apply to an algorithm, where probabilities are computed over all realizations of $SBM(n, \mathfrak{p}, Q)$ model.

The Symmetric Stochastic Block Model $SSBM(n, 2, a, b)$

Expectation of number of 4-cliques (1)

Under $SSBM(n, 2, a, b)$ the conditional expectation of X_4 given the size n_1 of the first community, is given by the following formula:

$$\begin{aligned} \mathbb{E}[X_4 | n_1] = & \binom{n_1}{4} a^6 + \binom{n_1}{3} n_2 a^3 b^3 + \binom{n_1}{2} \binom{n_2}{2} a^2 b^4 + \\ & + n_1 \binom{n_2}{3} a^3 b^3 + \binom{n_2}{4} a^6 \end{aligned}$$

where the terms represent the cases when all four vertices are in community 1, three vertices in community 1 and one vertex in community 2, two vertices in each community, one vertex in community 1 and three in community 2, and finally, all four vertices are in community 2.

Next, the expectation of the number of 4-cliques given parameters a, b is obtained by iterating the expectation operator over n_1 :

$$\mathbb{E}[X_4; a, b] = \mathbb{E}[\mathbb{E}[X_4 | n_1]]$$

The Symmetric Stochastic Block Model $SSBM(n, 2, a, b)$

Expectation of number of 4-cliques (2)

Since n_1 follows the binomial distribution $B(n, \frac{1}{2})$,

$$\mathbb{E}[n_1] = \frac{n}{2}, \quad \mathbb{E}[n_1^2] = \frac{n^2 + n}{4}$$

$$\mathbb{E}[n_1^3] = \frac{n^2(n+3)}{8}, \quad \mathbb{E}[n_1^4] = \frac{n(n+1)(n^2+5n-2)}{16}$$

These expressions come from the moment generating function of the binomial distribution $M_X(t) = (1 - p + pe^t)^n$ which for $p = \frac{1}{2}$ becomes $M_{n_1}(t) = \frac{1}{2^n}(1 + e^t)^n$. Then the k^{th} moment is given by

$$\mathbb{E}[n_1^k] = \frac{d^k}{dt^k} M_{n_1}(t)|_{t=0}$$

See: <http://mathworld.wolfram.com/BinomialDistribution.html> for details.

The expectation over n_1 is obtained by substituting $n_2 = n - n_1$, expanding the expression of $\mathbb{E}[X_4|n_1]$ and then using the moments of n_1, n_1^2, n_1^3, n_1^4 .

The Symmetric Stochastic Block Model $SSBM(n, 2, a, b)$

Expectation of number of 4-cliques (3)

Expanding, making the substitution $n_2 = n - n_1$ and combining the terms we get:

$$\begin{aligned} \mathbb{E}[X_4 | n_1] = & \frac{a^6}{24} \left(2n_1^4 - 4nn_1^3 + (6n^2 - 18n + 22)n_1^2 + (-4n^3 + 18n^2 - 22n)n_1 \right. \\ & \left. + n^4 - 6n^3 + 11n^2 - 6n \right) + \\ & + \frac{a^3b^3}{6} \left(-2n_1^4 + 4nn_1^3 + (-3n^2 + 3n - 4)n_1^2 + (n^3 - 3n^2 + 4n)n_1 \right) \\ & + \frac{a^2b^4}{4} \left(n_1^4 - 2nn_1^3 + (n^2 + n - 1)n_1^2 + (-n^2 + n)n_1 \right) \end{aligned}$$

The Symmetric Stochastic Block Model $SSBM(n, 2, a, b)$

Expectation of number of 4-cliques (4)

$$\begin{aligned} \mathbb{E}[X_4] &= \frac{a^6}{24} \left(2\mathbb{E}[n_1^4] - 4n\mathbb{E}[n_1^3] + (6n^2 - 18n + 22)\mathbb{E}[n_1^2] \right. \\ &\quad \left. + (-4n^3 + 18n^2 - 22n)\mathbb{E}[n_1] + n^4 - 6n^3 + 11n^2 - 6n \right) + \\ &+ \frac{a^3b^3}{6} \left(-2\mathbb{E}[n_1^4] + 4n\mathbb{E}[n_1^3] + (-3n^2 + 3n - 4)\mathbb{E}[n_1^2] + (n^3 - 3n^2 + 4n)\mathbb{E}[n_1] \right) \\ &+ \frac{a^2b^4}{4} \left(\mathbb{E}[n_1^4] - 2n\mathbb{E}[n_1^3] + (n^2 + n - 1)\mathbb{E}[n_1^2] + (-n^2 + n)\mathbb{E}[n_1] \right) \end{aligned}$$

where the expectations $\mathbb{E}[n_1]$, $\mathbb{E}[n_1^2]$, $\mathbb{E}[n_1^3]$ and $\mathbb{E}[n_1^4]$ have been computed before.

Numerical Computation of Number of Cliques

An Iterative Algorithm

We discuss two algorithms to compute X_q : iterative, and adjacency matrix based algorithm.

Framework: we are given a sequence $(G_t)_{t \geq 0}$ of graphs on n vertices, where G_{t+1} is obtained from G_t by adding one additional edge:

$G_t = (\mathcal{V}, \mathcal{E}_t)$, $\emptyset = \mathcal{E}_0 \subset \mathcal{E}_1 \subset \dots$ and $|\mathcal{E}_t| = t$.

Iterative Algorithm: Assume we know $X_q(G_t)$, the number of q -cliques of graph G_t . Then $X_q(G_{t+1}) = X_q(G_t) + D_q(e; G_t)$ where $D_q(e; G_t)$ denotes the number of q -cliques in G_{t+1} formed by the additional edge $e \in \mathcal{E}_{t+1} \setminus \mathcal{E}_t$.

Computation of Number of Cliques

An Analytic Formula

Laplace Matrix $\Delta = D - A$ contains all connectivity information.

Idea: Note the (i, j) element of A^2 is

$$(A^2)_{i,j} = \sum_{k=1}^n A_{i,k}A_{k,j} = |\{k : i \sim k \sim j\}|.$$

This means $(A^2)_{i,j}$ is the number of paths of length 2 that connect i to j . Hence $m = \frac{1}{2} \text{trace}(A^2)$.

Remark: The diagonal elements of $A(A^2 - D)$ represent twice the number of 3-cycles (= 3-cliques) that contain that particular vertex.

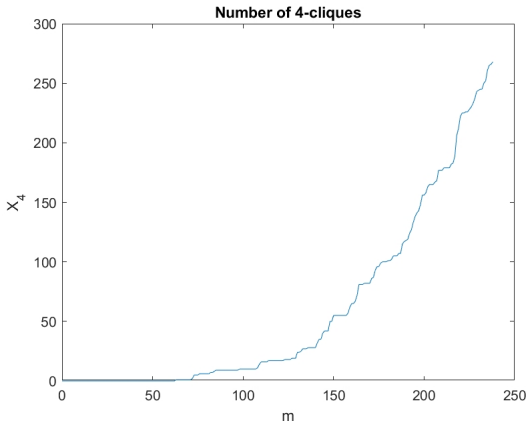
Conclusion:

$$X_3 = \frac{1}{6} \text{trace}\{A(A^2 - D)\} = \frac{1}{6} \text{trace}(A^3).$$








Numerical results





Plot of X_4 for the BKOFF dataset

Weighted graph: Ordered $m = 238$ edges for $n = 40$ nodes. The plot of X_4 the number of 4-cliques:



References

-  E. Abbe, Community detection and stochastic block models: recent developments, arXiv:1703.10146 [math.PR] 29 Mar. 2017.
-  B. Bollobás, **Graph Theory. An Introductory Course**, Springer-Verlag 1979. **99**(25), 15879–15882 (2002).
-  F. Chung, **Spectral Graph Theory**, AMS 1997.
-  F. Chung, L. Lu, The average distances in random graphs with given expected degrees, Proc. Nat.Acad.Sci. 2002.
-  R. Diestel, **Graph Theory**, 3rd Edition, Springer-Verlag 2005.
-  P. Erdős, A. Rényi, On The Evolution of Random Graphs
-  G. Grimmett, **Probability on Graphs. Random Processes on Graphs and Lattices**, Cambridge Press 2010.

-  P.W. Holland, K.B. Laskey, S. Leinhardt, Stochastic Blockmodels: First Steps, *Social Networks* **5**(1983), 109–137.
-  J. Leskovec, J. Kleinberg, C. Faloutsos, Graph Evolution: Densification and Shrinking Diameters, *ACM Trans. on Knowl.Disc.Data*, **1**(1) 2007.
-  M. Penrose, **Random Geometric Graphs**, Oxford University Press, 2003.
-  Y.J. Wang, G.Y. Wong, Stochastic Blockmodels for Directed Graphs, *J. Amer.Stat.Assoc.* **82**(397), pp.8–19, 1987.