Structure in noisy networks - WASP Community Building 2025

Please choose some questions below amounting to at least (3) points. Either find me in person on 26th August (or, if you miss these, email to me `fiona.skerman@math.uu.se`), photos or scans of handwritten solutions are ok. Feel free to work in groups of up to six, and only one submission per group is required.

I: Modularity-based questions

M1. (3) (`leiden`) The `leiden` algorithm is an alternative algorithm which also works by iteratively refining partitions based on modularity score, and has been proposed by some of the same authors as `louvain`. Read up on it and write an explanation of how it works, including (possibly hand-drawn) diagrams.

M2. (3) (`louvain`) It has been claimed the `louvain` output can have disconnected clusters. Read the appendix of the paper introducing `louvain` (file on kanvas). There are calculations shown for a different quality function on partitions ('CPM quality function' not modularity), check if the same example works for modularity.

M3. (3) (`louvain`) In Louvain, What is the tie-breaking rule used in the original implementation by Blondel et al.? Find original documentation or original code, show relevant snippet of text or code and explain.

M4. (theory) Let G be a graph with $m > 1$ edges and no isolated vertices[1]. Suppose also that $\mathcal{A}^*$ is a modularity-optimal partition of G, i.e. $q_{\mathcal{A}^*}(G) = q^*(G)$

(a) (1) A *leaf vertex* (or *pendant vertex*) is a vertex which has exactly one edge connected to it. Show that if v is a leaf vertex joined by edge uv to u (which may join other vertices) then v and u are in the same part in $\mathcal{A}^*$.

(b) (2) Show that all parts/communities A in $\mathcal{A}^*$ have $\geqslant 2$ vertices.

M5. (resolution limit - the $\sqrt{2m}$ threshold) We re-prove the threshold claimed in the lecture.

Suppose H is a connected component with h edges, and it sits inside a larger graph G which has m edges in total (including the h edges in H). In an optimal partition $\mathcal{A}^*$ of G the component H is partitioned separately to the rest of G, let $\mathcal{A}_H$ denote partition $\mathcal{A}$ restricted to H. Define partial modularity

$$q_{\mathcal{A}_H}(H, m) = \sum_{A \in \mathcal{A}_H} \frac{e(A)}{m} - \frac{\mathrm{vol}(A)^2}{4m^2}.$$

(a) (1) Show that for the partition $\mathcal{A}_{H\text{ together}}$ of H which places all of the vertices in the same part

$$q_{\mathcal{A}_{H\text{ together}}}(H, m) = \frac{h}{m} - \frac{h^2}{m^2}$$

(b) (1) Show that if H is connected then for any partition $\mathcal{A}_{H\text{ split}}$ that splits H into two pieces

$$q_{\mathcal{A}_{H\text{ split}}}(H, m) \leqslant \frac{h}{m} - \frac{1}{m} - \frac{h^2}{2m^2} \tag{1.1}$$

[1] isolated vertices are those which have no edges connected to them

(c) (bonus) Show that if H is connected then for any partition $\mathcal{A}_{H \text{ split}}$ that splits H into $k \geqslant 2$ pieces

$$q_{\mathcal{A}_{H \text{ split}}}(H, m) \leqslant \frac{h}{m} - \frac{k-1}{m} - \frac{h^2}{km^2}$$

(d) (1) Conclude for any H the optimal partiton of G will places vertices of H together in same part if $h < \sqrt{2m}$

(e) (1) Show that if H is two equal sized cliques connected by a single edge (call this dumbbell graph) then the 'natural' split achieves the RHS of the expression (1.1) above.

(f) (1) Conclude that for a dumbbell graph there is an optimal partition of G will split the dumbbell graph into the two natural halves.

M6. (3)(resolution limit - affect of the resolution parameter) Now considered the general modularity formula with resolution parameter λ, i.e.

$$q^{\lambda}_{\mathcal{A}}(G) = \sum_{A \in \mathcal{A}} \frac{e(A)}{m} - \lambda \frac{\text{vol}(A)^2}{4m^2}$$

which has partial modularity for the partition restricted to $\mathcal{A}_H$

$$q_{\mathcal{A}_H}(H, m) = \sum_{A \in \mathcal{A}_H} \frac{e(A)}{m} - \lambda \frac{\text{vol}(A)^2}{4m^2}.$$

Suppose you have a 'dumbbell' connected component H consisting of two equal sized cliques connected by a single edge, h edges total, and that H is a connected component in a larger graph G. Compare two partitions of G - where either H as a single part or H split into two parts. Find the threshold for h in terms of λ and m such that splitting is better (higher partial modularity score) if $h > f(\lambda, m)$ and keeping together is better if $h < f(\lambda, m)$.

M7. (3)(simulations, modularity and stochastic block model) Generate random graphs according to the stohcastic block model (exact model of your choice). Find the partition output by `louvain`, `leiden` or another modularity-based community recovery algorithm of your choice[2]. Investigate the closeness of the ouputs of these algorithms to the planted communities in the stochastic block model. (Use any 'closeness' measure of the two partitions, see also 'classification agreement indices', or just check whether the algorithm outputs the correct number of parts).

M8. (3) (degree tax – investigating the penalty term) In the usual modularity formula, for vertices i and j, approximate the probability there is an edge ij by $d_i d_j/(2m)$, where m is the number of edges in the graph.

For a given graph G, we may choose a random graph G' uniformly at random from the set of graphs with the same degree sequence as G, and calculate the precise probability that ij is an edge in G'. Call this the 'exact penalty'.

Can you design a function which is more accurate than $d_i d_j/(2m)$?[3]

See an example of this in the paper by Chang and Van Mieghem Figure 8 at this link.

[2] python package `networkx` contains an implementation of `louvain`: link and there are also packages `leidenalg`, `louvain` and `igraph` which may be helpful.

[3] Alternately, try the following: for $i \neq j$ set $d_i d_j/(2m-1)$ and 0 for loops, and see how the error compares to the standard modularity.

IIa: Planted network models - one hidden community

C1. (3) (finding a planted clique using spectral methods). Let $G \sim G'(n, k, 1/2)$ be a random graph with planted clique on k vertices, and all other edges appear with probability $1/2$. Let A denote the adjacency matrix of G and let M be matrix defined by subtracting a half from each entry $M_{ij} = A_{ij} - 1/2$. Let $\underline{x}$ be an eigenvector of largest eigenvalue of M. Let I be the k vertices with largest $|x_v|$ value.[4] Simulate this for different values of n and k and display results on the success of this algorithm at recovering the planted clique (e.g. with a heatmap).

C2. (3) Consider the planted dense subgraph model $\mathrm{PDS}'(n, k, p, q)$. It is important that we consider the uniform model, i.e. where the set of community vertices S^* has size exactly k and is chosen uniformly at random from all $\binom{n}{k}$ subsets of size k. We want to simulate this graph, then use 2-3 different algorithms to attempt recovery and plot rate of success of each for different parameter values. A good measure of success is the overlap, if $\hat{S}$ is the k-vertex set returned by the algorithm then the overlap $o(S^*, \hat{S}) = |S^* \cap \hat{S}|/k$.

A possible set of three algorithms to compare are $\hat{S}_1$ the k vertices of highest degree, $\hat{S}_2$ an iteration of this, where we pick the k vertices which have the largest number of neighbours in $\hat{S}_1$, and $\hat{S}_3$ (the spectral output from the previous question). A possible set of parameters would be ($n = 100$, $k = 10$, $q = 0.4$, $p = 0.5, 0.6, 0.7, 0.8$) but you may have to play with these a little to get some interesting behaviour.

C3 (maximum likelihood estimator MLE) Consider the planted dense subgraph model $\mathrm{PDS}'(n, k, p, q)$ where the planted clique has exactly k vertices. Denote by S^* the planted clique.

(a) (1) (theory MLE) The first step is to calculate the probability of generating a particular graph G given that the planted clique is on particular vertices. Label the vertices $1, \ldots, n$ and suppose the clique is on vertices 1 and 2, and label an edge between vertices i and j as ij. Show

$$\begin{aligned}&\mathbb{P}(G|S^* = \{1,2\})\\ &\quad = \mathbb{E}\Big[\prod_{i,j\leqslant 2} (\mathbf{1}[ij \in E]p + \mathbf{1}[ij \notin E](1-p)) \prod_{\max\{i,j\}>2} (\mathbf{1}[ij \notin E]q + \mathbf{1}[ij \in E](1-q))\Big]\end{aligned}$$

(b) (2) The maximum likelihood estimator for the planted community $\hat{S}$ planted set to be such that the probability of generating the observed graph is maximal (with ties broken by choosing a random set with maximal probability). That is,

$$\hat{S}(G) = \max_{S \subseteq V(G)} \mathbb{P}(G|S^* = S)$$

Simulate the planted dense subgraph (or planted clique) model for various values of n, k, p, q and record the proportion of times that the maximal likelihood estimator returns the right vertex set.

(c) (2) Implement another method for guessing the position of the planted subgraph S^*, e.g. picking the k vertices of highest degree, or a spectral method as in other questions. Can you find some values for n, k, p, q such that the maximum likelihood method performs better than your chosen fast method?

[4] One can also implement a 'clean up' step: let C be the set of vertices in the graph which have $\geqslant 3k/4$ neighbours in I.

IIb: Planted network models - multiple hidden communities

Several questions will relate to the stochastic block model, SBM so we define it here, and some graph theory notation for Q3.

Definition - Stochastic Block Model - vanilla model (For Q3)
Let SBM(n, p, q) be the model constructed as follows. For each vertex $v \in [n]$ independently let $v \in S^*$ with probability 1/2. Let $\sigma_v = 1$ if $v \in S^*$ and $\sigma_v = -1$ if $v \notin S^*$. Construct G by choosing each edge to be present independently with probability

$$\mathbb{P}(uv \in E \mid \sigma_u, \sigma_v) = \begin{cases} p & \text{if } \sigma_u \sigma_v = 1 \\ q & \text{otherwise.} \end{cases}$$

We also consider fixed size version SBM$'(n, p, q)$ which is as above except we take $S^* \in \binom{[n]}{n/2}$, i.e. let S^* be a set of $n/2$ vertices chosen uniformly from all sets of that size in $[n]$. For this model we assume n is even.

Definition - Stochastic Block Model many unequal size parts (For Q2) - see Figure 1.
Let SBM$(n, q, s, (x_1, x_2, \ldots, x_\ell))$ be the model constructed as follows. For each vertex $v \in [n]$, $\sigma(v) \in \{1, \ldots, k\}$, we independently choose $\sigma(v) = i$ with probability x_i. Construct G by choosing each edge to be present independently with probability

$$\mathbb{P}(uv \in E \mid \sigma_u, \sigma_v) = \begin{cases} q + \frac{s}{x_i} & \text{if } \sigma_u = \sigma_v = i \\ q & \text{otherwise.} \end{cases}$$

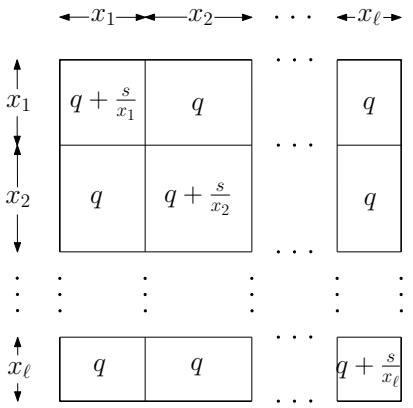


Figure 1: Stochastic Block Model (SBM). General model for many communities of unequal sizes.

Graph theory notation.(For Q1) For graph $g = ([n], e)$ write $e(g)$ for the number of edges in g. For vertex subset $S \subset [n]$, write $\bar{S} = [n] \backslash S$, write $e_g(S)$ for the number of edges in g with both end points in set S and write $e_g(S, \bar{S})$ for the number of edges in g with one endpoint in S and the other endpoint in $\bar{S}$.

Q1 (3) (theory) Let G be a random sample of $\mathrm{SBM}'(n,p,q)$. Then show[5]

$$\mathbb{P}(G=g)=\binom{n}{n/2}^{-1}\sum_{|S|=n/2}\left(\frac{p(1-q)}{q(1-p)}\right)^{e(g)-e_g(S,\bar{S})}((1-p)(1-q))^{n^2/4}.$$

Q2 (theory) We want to show that counts of a small subgraph will distinguish the stochastic block model with equal size parts from the stochastic blockmodel with non-equal sized parts.

Let $x\neq 1/2$. Distinguishing $H_1:\mathrm{SBM}(n,p,q,(x,1-x))$ and $H_0:\mathrm{SBM}(n,p,q,(1/2,1/2))$, see Figure 2

Denote the adjacency matrix of the observed graph by A, it may be easier to count triangles, $\#\triangle=\sum_{i,j,k}A_{ij}A_{ik}A_{jk}$ or signed triangles $\#\triangle_s=\sum_{i,j,k}(A_{ij}-q)(A_{ik}-q)(A_{jk}-q)$.

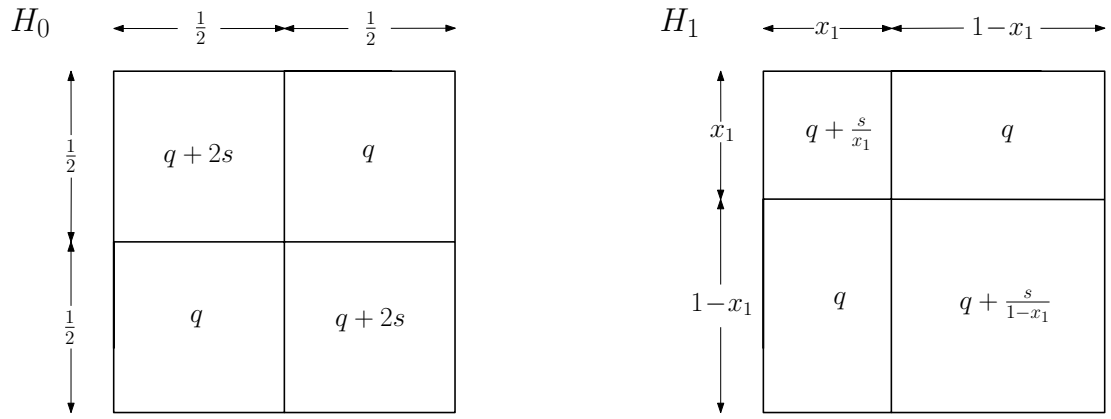
(a) (1) Show that triangles (or signed triangles) will not work. i.e. show that

$$\mathbb{E}_0[\#\triangle]=\mathbb{E}_1[\#\triangle].$$

(b) (1) Find a small subgraph H (or the signed version) such that $\mathbb{E}_0[\#H]\neq\mathbb{E}_1[\#H]$.

(c) (1) (Bonus) For a subgraph H satisfying (b) characterise which distributions it can not distinguish.

(d) (1) (Bonus) For a subgraph H satisfying (b) find the variance of $\#H$ under H_0 and under H_1.



(a) $H_0=\mathrm{SBM}(n,p,s,(\frac{1}{2},\frac{1}{2}))$ (b) $H_1=\mathrm{SBM}(n,p,s,(x_1,1-x_1))$

Figure 2: The distinguishing problem in question 1.

Q3 (theory) (1) **Prove, disprove or salvage if possible.** In the SBM for any two distinct nodes the probability that they have common neighbours is independent of whether they share an edge or not.

Feel free to consider either SBM or SBM' and to change the wording slightly, e.g. to consider expected number of common neighbours etc.

[5] or similar, typos expected.