Average-case complexity and statistical inference

Exercise Sheet 2

Please choose some questions below amounting to at least (4) points. Deadline 30th May, email to me `fiona.skerman@math.uu.se` or put a physical copy in my pigeon-hole.

Several questions will relate to the stochastic block model, *stochastic block model* so we define it here.

Definition - Stochastic Block Model - vanilla model (For Q2)
Let $\mathrm{SBM}(n, p, q)$ 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 $\mathrm{SBM}'(n, p, q)$ 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 Q1) - see Figure 1.
Let $\mathrm{SBM}(n, q, s, (x_1, x_2, \ldots, x_\ell))$ 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

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

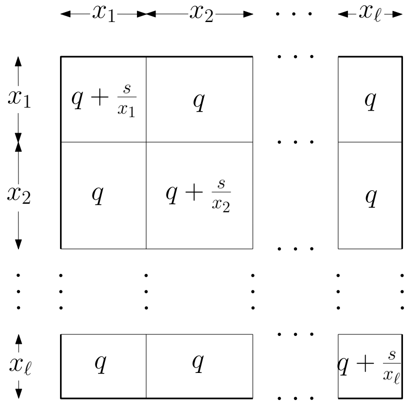


Figure 1: Stochastic Block Model (SBM). General model for many communities of unequal sizes.

Q1. 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 $\#\triangle_s = \sum_{i,j,k}(A_{ij} - q)(A_{ik} - q)(A_{jk} - q)$.

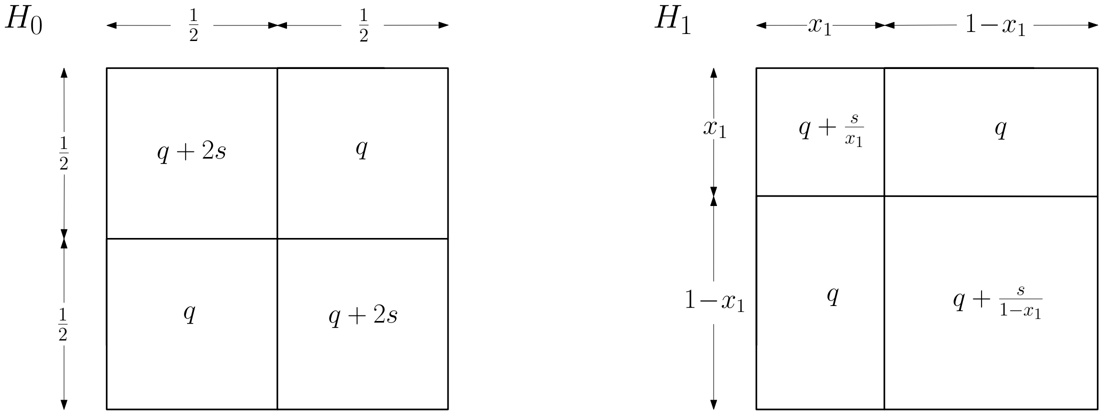
(a) (1) Show that triangles (or signed triangles) will not work. i.e. show that

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

(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.

Q2. (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.

Q3. (1) Prove Lemma 5.3 in the notes, i.e. prove the following.

If P, P_1 and P_2 be three probability spaces, and $\mathcal{A}_1$ and $\mathcal{A}_2$ algorithms such that

$$P \xrightarrow{\mathcal{A}_1}_{\varepsilon_1} P_1 \quad \text{and} \quad P_1 \xrightarrow{\mathcal{A}_2}_{\varepsilon_2} P_2.$$

Then

$$P \xrightarrow{\mathcal{A}_2 \circ \mathcal{A}_1}_{\varepsilon_1 + \varepsilon_2} P_2.$$

Q4. In the reduction in lectures, we reduced from the planted clique detection to a symmetrised version of planted submatrix with zeros on the diagonal. (Good references for this are the video of Guy Bresler linked from the homepage, as well as the original paper of Brennan, Bresler and Huleihel cited in the lecture notes.)

(a) (1) Show the reduction from the symmetrised to non-symmetrised version

(b) (1) Show we may 'fill in' the diagonal.

Q5. (1) Find another example of a worst-case to average-case reduction and write down a clear explanation of the reduction and why it works.

Q6. (1) Read then write in your own words a proof of impossibility of distinguishing with vanishing risk in the 'impossible region' for distinguishing the planted submatrix model from a matrix of independent Gaussians – see Figure 6 (detection) in the lecture notes. A good reference is the lecture notes by Wu and Xu. Feel free to use some facts without proof - just make it clear which facts you assume.

Q7. (1) Find a gap in your knowledge and write about it. It could be a detail skipped in the lecture or reading a section of the lecture notes of Lugosi or Wu and Xu and explaining it in your own words (and equations), or something else.