Lecture 4

In which we explore the Stochastic Block Model.

1 The $G_{n,p,q}$ problem

The Stochastic Block Model is a generic model for graphs generated by some parameters. The simplest model and one we will consider today is the $G_{n,p,q}$ problem.

Definition 1 ($G_{n,p,q}$ graph distribution) The $G_{n,p,q}$ distribution is a distribution on graphs of $n$ vertices where $V$ is partitioned into two subsets of equal size: $V = V_1 \sqcup V_2$. Then for each $\{i,j\}$ pair of vertices in the same subset, $\Pr((i,j) \in E) = p$ and otherwise $\Pr((i,j) \in E) = q$.

We will only consider the regime under which $p > q$. If we want to find the partition $V = V_1 \sqcup V_2$, it is intuitive to look at the problem of finding the minimum balanced cut. The cut $(V_1, V_2)$ has expected size $qn^2/4$ and any other cut will have greater expected size.

Our intuition should be that as $p \to q$, the problem only gets harder. And for fixed ratio $p/q$, as $p, q \to 1$, the problem only gets easier. This can be stated rigorously as follows: If we can solve the problem for $p, q$ then we can also solve it for $cp, cq$ where $c > 1$, by keeping only $1/c$ edges and reducing to the case we can solve.

Recall that for the $k$-planted clique problem, we found the eigenvector $x$ corresponding to the largest eigenvalue of $A - \frac{1}{2}J$. We then defined $S$ as the vertices $i$ with the $k$ largest values of $|x_i|$ and cleaned up $S$ a little to get our guess for the planted clique.

In the Stochastic Block Model we are going to follow a similar approach, but we are instead going to find the largest eigenvalue of $A - \left(\frac{p+q}{2}\right)J$. Note this is intuitive as the average degree of the graph is $p(n/2 - 1) + q(n/2) \approx \frac{n}{2}(p + q)$. The idea is simple: Solve $x$ the largest eigenvector corresponding to the largest eigenvalue and define

$$V_1 = \{i : x_i > 0\}, \quad V_2 = \{i : x_i \leq 0\}$$

As we proceed to the analysis of this procedure, we fix $V_1, V_2$. Prior to fixing, the adjacency matrix $A$ was $\left(\frac{p+q}{2}\right)J$.\footnote{The diagonal should be zeroes, but this is close enough.} Upon fixing $V_1, V_2$, the average adjacency matrix $R$ looks different.
For ease of notation, if we write a bold constant $c$ for a matrix, we mean the matrix $cJ$. It will be clear from context.

$$R = \begin{pmatrix} p & q \\ q & p \end{pmatrix}$$  \hspace{1cm} (2)

Here we have broken up $R$ into blocks according to the partition $V_1, V_2$.

**Theorem 2** If $p, q > \log n/n$ then with high probability, $\|A - R\| < O(\sqrt{n(p + q)})$.

**Proof:** Define the graph $G_1$ as the union of a $G_{n/2, p}$ graph on $V_1$ and $G_{n/2, p}$ graph on $V_2$. Define the graph $G_2$ as a $G_{n,q}$ graph. Note that the graph $G$ is distributed according to picking a $G_1$ and $G_2$ graph and adding the partition crossing edges of $G_2$ to $G_1$. Let $A_1$ and $A_2$ be the respective adjacency matrices and define the follow submatrices:

$$A_1 = \begin{pmatrix} A_1' & A_1'' \\ A_1'' & A_1''' \end{pmatrix}, \quad A_2 = \begin{pmatrix} A_2' & A_2'' \\ A_2'' & A_2''' \end{pmatrix}. \hspace{1cm} (3)$$

Then the adjacency matrix $A$ is defined by

$$A = A_1 + A_2 - \begin{pmatrix} A_2' & A_2'' \\ A_2'' & A_2''' \end{pmatrix} \hspace{1cm} (4)$$

Similarly, we can generate a decomposition for $R$:

$$R = \begin{pmatrix} p & q \\ q & p \end{pmatrix} + \begin{pmatrix} q & 0 \\ 0 & q \end{pmatrix} - \begin{pmatrix} q & 0 \\ 0 & q \end{pmatrix} \hspace{1cm} (5)$$

Then using triangle inequality we can bound $\|A - R\|$ by bounding the difference in the various terms.

$$\|A - R\| \leq \left\| A_1 - \begin{pmatrix} p & 0 \\ 0 & p \end{pmatrix} \right\| + \left\| A_2 - \begin{pmatrix} q & 0 \\ 0 & q \end{pmatrix} \right\| + \left\| \begin{pmatrix} A_2' & A_2'' \\ A_2'' & A_2''' \end{pmatrix} - \begin{pmatrix} q & 0 \\ 0 & q \end{pmatrix} \right\|$$

$$\leq O(\sqrt{np}) + O(\sqrt{nq}) + O(\sqrt{nq}) \hspace{1cm} (6)$$

The last line follows as the submatrices are adjacency matrices of $G_{n,p}$ graphs and we can apply the results we proved in that regime for $p, q > \log n/n$. \quad \square

But the difficulty is that we don’t know $R$ as $R = R(V_1, V_2)$. If we knew $R$, then we would know the partition. What we can compute is $\|A - \left( \frac{p + q}{2} \right) J\|^2$. We can rewrite $R$ as

$$R = \left( \frac{p + q}{2} \right) J + \frac{p - q}{2} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \hspace{1cm} (7)$$

\[^2\]The rest of this proof actually doesn’t even rely on knowing $p$ or $q$. We can estimate $p + q$ by calculating the average vertex degree.
Call the matrix on the right \( C \). It is clearly rank-one as it has decomposition \( n \chi \chi^T \) where \( \chi = \frac{1}{\sqrt{n}} \begin{pmatrix} 1 \\ -1 \end{pmatrix} \). Therefore

\[
\left\| \left( A - \left( \frac{p+q}{2} \right) J \right) - \left( \frac{p-q}{2} \right) C \right\| = \| A - R \| \leq O \left( \sqrt{n(p+q)} \right).
\] (8)

Then \( A - \left( \frac{p+q}{2} \right) J \) is close (in operator norm) to the rank 1 matrix \( \left( \frac{p-q}{2} \right) C \). Then their largest eigenvalues are close. But since \( \left( \frac{p-q}{2} \right) C \) has only one non-zero eigenvalue \( \chi \), finding the corresponding eigenvector to the largest eigenvalue of \( A - \left( \frac{p+q}{2} \right) J \) will be close to the ideal partition as \( C \) describes the ideal partition. This can be formalized with the Davis-Kahan Theorem.

**Theorem 3 (Davis-Kahan)** Given matrices \( M, M' \) with \( \| M - M' \| \leq \varepsilon \) where \( M \) has eigenvalues \( \lambda_1 \leq \ldots \leq \lambda_n \) and corresponding eigenvectors \( v_1, \ldots, v_n \) and \( M' \) has eigenvalues \( \lambda'_1 \leq \ldots \leq \lambda'_n \) and corresponding eigenvectors \( v'_1, \ldots, v'_n \), then

\[
\sin \left( \text{angle} \left( \text{span}(v_1), \text{span}(v'_1) \right) \right) \leq \frac{\varepsilon}{|\lambda'_1 - \lambda_2|} \leq \frac{\varepsilon}{|\lambda_1 - \lambda_2 - \varepsilon|}.
\] (9)

Equivalently,

\[
\min \left\{ \| v_1 \pm v'_1 \| \right\} \leq \frac{\sqrt{2\varepsilon}}{\lambda_1 - \lambda_2 - \varepsilon}.
\] (10)

The Davis Kahan Theorem with \( M' = A - \left( \frac{p+q}{2} \right) J, M = \left( \frac{p-q}{2} \right) C \), and \( \varepsilon = O \left( \sqrt{n(p+q)} \right) \) states that

\[
\min \left\{ \| v' \pm \chi \| \right\} \leq O \left( \frac{\sqrt{a+b}}{a-b - O \left( \sqrt{a+b} \right)} \right)
\] (11)

where \( v' \), the eigenvector associated with the largest eigenvalue of \( A - \left( \frac{p+q}{2} \right) J \) and \( a = pm/2, b = qn/2 \), the expected degrees of the two parts of the graph. Choose between \( \pm v' \) for the one closer to \( \chi \). Then

\[
\| v' - \chi \|^2 \leq O \left( \left( \frac{\sqrt{a+b}}{a-b - O \left( \sqrt{a+b} \right)} \right)^2 \right).
\] (12)

Recall that \( \sum_i (v'_i - \chi_i)^2 = \| v' - \chi \|^2 \). If \( v'_i \) and \( \chi_i \) disagree in sign, then this contributes at least \( 1/n \) to the value of \( \| v' - \chi \|^2 \). Equivalently, \( n \cdot \| v' - \chi \|^2 \) is at least the number of misclassified vertices. It is simple to see from here that if \( a-b \geq c\varepsilon \sqrt{a+b} \) then we can bound the number of misclassified vertices by \( \varepsilon n \). This completes the proof that the proposed algorithm does well in calculating the partition of the Stochastic Block Model.