Beyond Worst Case Analysis **Fall 2018**

Lecture 6: Stochastic Block Model 1

Scribe: Antares Chen 10/18/2018

6.1 Introduction

In the vein of finding planted structures in random inputs, we next look at the problem of partitioning a graph. We introduce the stochastic block model, a distribution of graphs that contain a natural partition, and provide a spectral algorithm for recovering this division. Our analysis will demonstrate that this algorithm makes few mistakes by using perturbation theory in a manner simplified from McSherry [1].

6.2 The Stochastic Block Model

So far we have seen spectral and SDP based algorithms for finding planted cliques in random $\mathcal{G}_{n,1/2}$ graphs. Continuing along the path of finding planted structures, we can ask if it is possible to find a planted partition in a random graph. The stochastic block model is a distribution of random graphs that embeds a natural partition inside the graph. Graphs sampled from this distribution are given by the following procedure.

The Stochastic Block Model
Given the number of vertices n and probabilities p, q where $p > q$, a graph $G = (V, E)$ sampled from the stochastic block model $\mathcal{G}_{n,p,q}$ is constructed via the following:

1. Choose an even partition of n vertices into two even sets V_1, V_2
2. For any $i, j \in V_1$ and any $i, j \in V_2$ where $i \neq j$, add the edge (i, j) to E with probability p independently at random.
3. For any $i \in V_1$ and $j \in V_2$, add the edge (i, j) to E with probability q independently at random.

Graphs $G \sim \mathcal{G}_{n,p,q}$ naturally have a partition between V_1 and V_2. In particular:

$$\mathbb{E}[\#\text{ of edges between } V_1, V_2] = q|V_1||V_2| = \frac{qn^2}{4}$$

$$\mathbb{E}[\#\text{ of edges within } V_1] = p\binom{|V_1|}{2} = p\binom{\frac{n}{2}}{2} = \frac{pn(n-2)}{8}$$

Assuming $p \gg q$, there tends to be more edges on expection contained within V_1 or V_2 than across V_1 and V_2.

6.2.1 The Problem with Finding Partitions

Another way of looking at this is that the balanced cut (one that evenly separates V) with the fewest number of edges maximizes the likelihood that it is the hidden cut $(S, V-S)$ that we are interested in (for example, the one separating V_1 and V_2). Let S be the event that $(S, V-S)$ is the cut we wish to discover chosen uniformly at random. Bayes rule tells us that

$$\Pr[S \mid G] = \frac{\Pr[G \mid S]\Pr[S]}{\Pr[G]}$$

Given n, p, q, the probability $\Pr[G]$ is some fixed constant while

$$\Pr[S] = \frac{1}{\binom{n}{n/2}}$$

Hence maximizing $\Pr[S \mid G]$ reduces to maximizing $\Pr[G \mid S]$ which is given by

$$\Pr[G \mid S] = p^{|E|-|E(S,V-S)|} \cdot (1-p)^{2\binom{n/2}{2}-(|E|-|E(S,V-S)|)} \cdot q^{|E(S,V-S)|} \cdot (1-q)^{\frac{n^2}{4}-|E(S,V-S)|}$$

Observe that as $p > q$, minimizing $|E(S, V-S)|$ would maximize $\Pr[G \mid S]$! This means that in cases where it is possible to find the hidden cut, calculating this maximum likelihood estimator would be sufficient. However, this in general requires us to solve a balanced cut problem which is NP-Hard. What we lack is some assumption on how well separated p and q are, because in the worst-case, the probabilities p and q could be too close. We cannot hope to recover V_1 and V_2 efficiently because random cuts of G would be nearly indistinguishable from one another. For this reason, we will develop an algorithm assuming that $p - q$ is large.

6.3 A Spectral Algorithm

Let us now develop a spectral algorithm for approximately recovering the partition in a stochastic block model. Given $G \sim \mathcal{G}_{n,p,q}$ where p and q are sufficiently separated, we wish to recover V_1 and V_2. The exact condition on p, q will be given later, but for now think of $q = p - \frac{\delta}{\sqrt{n}}$ for some large δ. Now, how did we develop the spectral algorithm for planted clique?

1. We found an ideal matrix whose top eigenvector encoded the location of the planted clique.
2. Using a matrix concentration result, we demonstrated that the adjacency matrix (after some manipulation) of the randomly sampled graph G had a spectrum that was close to the ideal matrix.
3. Once we knew that the eigenvalues were close, the Davis-Kahan theorem implied that the eigenvectors of the sampled matrix were close to that of the ideal matrix whose only eigenvector was the clique indicator.
4. Finally, we applied a rounding step that takes the top eigenvector of the sampled matrix to calculate the algorithm's solution. We were able to show, with high probability, that the solution fully recovers the clique because the eigenvectors were close.

Can we hope to apply a similar process to find the partition? Yes, in fact this technique has found success in not only finding planted cliques and bisections, but has also lead to the development of nice algorithms for image processing problems [2, 3]. This is sometimes called a "matrix perturbative" analysis because it uses matrix perturbation theory (i.e. tools such as Davis-Kahan and various matrix concentration bounds that we have seen so far) to argue that the sampled matrix is a *perturbation* of the ideal matrix.

Let's start by finding the ideal matrix. Assume that the vertices are ordered such that $V_1 = \{1, \dots, \frac{n}{2}\}$ and $V_2 = \{\frac{n}{2} + 1, \dots, n\}$ and consider the expectation of $\boldsymbol{A}$ the adjacency matrix of $G \sim \mathcal{G}_{n,p,q}$.

$$\mathbb{E}[\boldsymbol{A}] = \begin{pmatrix} 0 & \dots & p & q & \dots & q \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ p & \dots & 0 & q & \dots & q \\ p & \dots & p & 0 & \dots & q \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ p & \dots & p & q & \dots & 0 \end{pmatrix} = \left(\begin{array}{c:c} p\boldsymbol{J}_{\frac{n}{2}} & q\boldsymbol{J}_{\frac{n}{2}} \\ \hdashline q\boldsymbol{J}_{\frac{n}{2}} & p\boldsymbol{J}_{\frac{n}{2}} \end{array}\right) - p\boldsymbol{I}$$

where $\boldsymbol{J}_{\frac{n}{2}}$ denotes the $\frac{n}{2} \times \frac{n}{2}$ all-ones matrix and $\boldsymbol{I}$ denotes the $n \times n$ identity matrix. Define $\hat{\boldsymbol{A}} = \mathbb{E}[\boldsymbol{A}] + p\boldsymbol{I}$. This should not effect our analysis too much since $\hat{\boldsymbol{A}}$ and $\mathbb{E}[\boldsymbol{A}]$ have the same eigenvectors. As an added bonus, the partitions are also encoded as separates blocks in $\hat{\boldsymbol{A}}$! Let's now find the eigenvectors of $\hat{\boldsymbol{A}}$. It is always useful to check the all-ones vector.

$$\hat{\boldsymbol{A}}\mathbf{1} = \frac{n}{2}(p+q)\mathbf{1}$$

However, the second eigenvector $\boldsymbol{w}_1$ of $\hat{\boldsymbol{A}}$ is constructed via

$$[\boldsymbol{w}_1]_i = \begin{cases} +1 & \text{if } i \in V_1 \\ -1 & \text{if } i \in V_2 \end{cases}$$

which is exactly the ± 1 indicator for the partition! Specifically

$$\hat{\boldsymbol{A}}\boldsymbol{w}_1 = \frac{n}{2}(p-q)\boldsymbol{w}_1$$

and because $\hat{\boldsymbol{A}}$ is rank 2, all other eigenvalues of $\hat{\boldsymbol{A}}$ are 0. We can rewrite $\hat{\boldsymbol{A}}$ as the following.

$$\hat{\boldsymbol{A}} = \left(\frac{p+q}{2}\right)\boldsymbol{J} + \left(\frac{p-q}{2}\right)\boldsymbol{w}_1\boldsymbol{w}_1^\top \qquad \iff \qquad \left(\frac{p-q}{2}\right)\boldsymbol{w}_1\boldsymbol{w}_1^\top = \hat{\boldsymbol{A}} - \left(\frac{p+q}{2}\right)\boldsymbol{J}$$

Our ideal matrix should thus be $\hat{\boldsymbol{M}} = \mathbb{E}[\boldsymbol{A}] - \left(\frac{p+q}{2}\right)\boldsymbol{J} + p\boldsymbol{I}$ because its top eigenvector is the ± 1 indicator of the partition! It is now clear what our spectral algorithm should look like.

Spectral Algorithm
Given $G \sim \mathcal{G}_{n,p,q}$, do the following:

1. Compute the top eigenvector $\boldsymbol{v}_1$ of $\boldsymbol{M} = \boldsymbol{A} - \left(\frac{p+q}{2}\right)\boldsymbol{J} + p\boldsymbol{I}$.
2. Return $V_1 = \{i : [\boldsymbol{v}_1]_i < 0\}$ and $V_2 = \{i : [\boldsymbol{v}_1]_i \geq 0\}$.

6.4 Analysis

Similar to the AKS algorithm for planted clique, our spectral algorithm for finding the planted partition splits nicely into two parts: (1) computing the top eigenvector of $\boldsymbol{M}$ and (2) rounding the top eigenvector to recover V_1 and V_2. We will divide our analysis of this algorithm into these two parts accordingly.

6.4.1 Analysis of the Retrieved Eigenvector

We want to show that $\boldsymbol{v}_1$ is close to the top eigenvector $\boldsymbol{w}_1$ of $\hat{\boldsymbol{M}}$. To begin, let's show that $\boldsymbol{M}$ and $\hat{\boldsymbol{M}}$ are close in terms of the spectral norm. Our analysis will require theorem 1.4 of Vu [4] which implies the following.

Theorem 6.1. *Let $\boldsymbol{R}$ be a matrix satisfying $|\boldsymbol{R}_{ij}| \leq 1$, $\mathbb{E}[\boldsymbol{R}_{ij}] = 0$, and $\mathbf{Var}[\boldsymbol{R}_{ij}] \leq \sigma^2$ for any $1 \leq i, j \leq n$. Then there exists absolute constants c_1, c_2 such that*

$$\|\boldsymbol{R}\| \leq 2\sigma\sqrt{n} + c_1\sigma^{1/2}n^{1/4}\ln n$$

Provided that,

$$\sigma^2 \geq c_2\frac{\ln^4 n}{n}$$

This inequality is stated with high probability; Vu proves that the probability $\|\boldsymbol{R}\|$ exceeds the above value by more than t is exponentially small in t. Theorem 6.1 allows us to demonstrate this lemma

Lemma 6.2. *Given $G \sim \mathcal{G}_{n,p,q}$, there exists an absolute constant c such that with high probability*

$$\|\boldsymbol{M} - \hat{\boldsymbol{M}}\| \leq c\sqrt{n(p+q)}$$

Provided that,

$$p + q \geq c\frac{\ln^4 n}{n}$$

Proof. Observe that

$$\begin{aligned}
\|\boldsymbol{M} - \hat{\boldsymbol{M}}\| &= \left\| \left(\boldsymbol{A} - \left(\frac{p+q}{2}\right)\boldsymbol{J} + p\boldsymbol{I} \right) - \left(\frac{p-q}{2}\right)\boldsymbol{w}_1\boldsymbol{w}_1^\top \right\| \\
&= \left\| \boldsymbol{A} - \left(\left(\frac{p+q}{2}\right)\boldsymbol{J} + \left(\frac{p-q}{2}\right)\boldsymbol{w}_1\boldsymbol{w}_1^\top \right) + p\boldsymbol{I} \right\| \\
&= \|\boldsymbol{A} - \hat{\boldsymbol{A}} + p\boldsymbol{I}\| \\
&= \|\boldsymbol{A} - \mathbb{E}[\boldsymbol{A}]\|
\end{aligned}$$

Let $\boldsymbol{R} = \boldsymbol{A} - \mathbb{E}[\boldsymbol{A}]$ and notice that $|\boldsymbol{R}_{ij}| \leq 1$ and $\mathbb{E}[\boldsymbol{R}_{ij}] = 0$ for any $1 \leq i, j \leq n$. For the variance, observe that if $(i, j) \notin E$ then $\mathbf{Var}[\mathbf{R}_{ij}] = 0$. If $(i, j) \in E$ then there are two cases:

1. Suppose $i, j \in V_1$ or $i, j \in V_2$. Then $\mathbb{E}[\boldsymbol{R}_{ij}^2] = p$ and the variance is

$$\mathbf{Var}[\boldsymbol{R}_{ij}] = \mathbb{E}[\boldsymbol{R}_{ij}^2] - \mathbb{E}[\boldsymbol{R}_{ij}]^2 = p$$

2. Suppose without loss of generality $i \in V_1$ and $j \in V_2$. Then $\mathbb{E}[\boldsymbol{R}_{ij}^2] = q$ and the variance is

$$\mathbf{Var}[\boldsymbol{R}_{ij}] = \mathbb{E}[\boldsymbol{R}_{ij}^2] - \mathbb{E}[\boldsymbol{R}_{ij}]^2 = q$$

In either case, we can say that $\mathbf{Var}[\boldsymbol{R}_{ij}] \leq p + q$. Invoking theorem 6.1 with $\sigma = \sqrt{p+q}$, we have for constants c_1 and c_2

$$\|\boldsymbol{R}\| \leq 2\sqrt{n(p+q)} + c_1\big(n(p+q)\big)^{1/4} \ln n$$

which for sufficiently large c admits

$$\|\boldsymbol{R}\| \leq c\sqrt{n(p+q)} \qquad \square$$

Now that we have a bound on $\|\boldsymbol{M} - \hat{\boldsymbol{M}}\|$, we can use the Davis-Kahan theorem to argue their top eigenvectors are close. We use a more general version of the theorem that was provided in lecture 3.

Theorem 6.3 (Davis and Kahan)**.** *Let $\boldsymbol{M}$ and $\hat{\boldsymbol{M}}$ be symmetric matrices with eigenvalues $\lambda_1 \geq \ldots \geq \lambda_n$ and $\mu_1 \geq \ldots \geq \mu_n$ as well as corresponding eigenvectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ and $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n$ respectively. Let θ_i denote the angle between $\boldsymbol{v}_i$ and $\boldsymbol{w}_i$. Then the following holds*

$$\sin(\theta_i) \leq \frac{\|\boldsymbol{M} - \hat{\boldsymbol{M}}\|}{\min_{j \neq i} |\mu_i - \mu_j|}$$

We can use the Davis-Kahan theorem to provide a very crude bound on the ℓ_2-distance between the computed top eigenvector $\boldsymbol{v}_1$ and the indicator vector of the partition $\boldsymbol{w}_1$.

Lemma 6.4. *Let $\boldsymbol{v}_1$ be the computed top eigenvector of $\boldsymbol{M}$ and $\boldsymbol{w}_1$ be the normalized ± 1 indicator vector of the hidden partition. Then for sufficiently large constant c', if $p + q \geq c' \frac{\ln^4 n}{n}$*

$$\min\left\{\|\boldsymbol{v}_1 - \boldsymbol{w}_1\|_2^2\,,\, \|-\boldsymbol{v}_1 - \boldsymbol{w}_1\|_2^2\right\} \leq \frac{c'}{p-q}\sqrt{n(p+q)}$$

Proof. Recall from lecture 3 that

$$\min\left\{\|\boldsymbol{v}_1 - \boldsymbol{w}_1\|_2^2\,,\, \|-\boldsymbol{v}_1 - \boldsymbol{w}_1\|_2^2\right\} \leq \|\boldsymbol{v}_1\|_2^2 + \|\boldsymbol{w}_1\|_2^2 - 2|\langle \boldsymbol{v}_1, \boldsymbol{w}_1\rangle| = 2(1 - |\cos(\theta_1)|)$$

which is at most $2\sin(\theta_1)$ from $0 \leq \theta_1 \leq \pi$. By theorem 6.3, we have then that

$$\min\left\{\|\boldsymbol{v}_1 - \boldsymbol{w}_1\|_2^2\,,\, \|-\boldsymbol{v}_1 - \boldsymbol{w}_1\|_2^2\right\} \leq \frac{2\|\boldsymbol{M} - \hat{\boldsymbol{M}}\|}{\min_{j \neq i} |\mu_i - \mu_j|}$$

For the numerator, observe that lemma 6.2

$$\|\boldsymbol{M} - \hat{\boldsymbol{M}}\| \leq c\sqrt{n(p+q)}$$

while for the denominator, the top eigenvalue of $\hat{\boldsymbol{M}}$ is $\frac{p-q}{2}$ with all other eigenvalues being 0, hence

$$\min_{j\neq 1}|\mu_1 - \mu_j| = \frac{p-q}{2}$$

Putting it all together:

$$\min\left\{\|\boldsymbol{v}_1 - \boldsymbol{w}_1\|_2^2\,,\, \|-\boldsymbol{v}_1 - \boldsymbol{w}_1\|_2^2\right\} \leq \frac{4c\sqrt{n(p+q)}}{p-q} = \frac{c'}{p-q}\sqrt{n(p+q)} \qquad \square$$

We have finally shown that the computed top eigenvector of $\boldsymbol{M}$ is close to the top eigenvector of $\hat{\boldsymbol{M}}$ or the normalized ± 1 indicator of the hidden partitions V_1, V_2. This gives us very good reason to believe that the rounding stage will succeed in approximately recovering the hidden partitions with high probability which we will now proceed to demonstrate!

6.4.2 Analysis of Rounding Stage

Culminating our analysis of the spectral algorithm, we have the following theorem.

Theorem 6.5. *Suppose $G \sim \mathcal{G}_{n,p,q}$ where $p - q \geq \frac{c'}{\epsilon}\sqrt{\frac{p+q}{n}}$ for c' a sufficiently large constant and $0 < \epsilon < 1$. Then with high probability, the spectral algorithm classifies at most ϵn vertices incorrectly provided that $p + q \geq c'\frac{\ln^4 n}{n}$*

Proof. Let $\boldsymbol{\chi} = \pm\boldsymbol{v}_1$ as to pick the minimum of $\left\{\|\boldsymbol{v}_1 - \boldsymbol{w}_1\|_2^2\,,\, \|-\boldsymbol{v}_1 - \boldsymbol{w}_1\|_2^2\right\}$. Observe that if $i \in V$ is classified incorrectly then $\boldsymbol{\chi}_i$ and $[\boldsymbol{w}_1]_i$ must have different signs. As $[\boldsymbol{w}_1]_i = \pm 1$, it must be that every i that is incorrectly classified contributes at least $+1$ to $\|\boldsymbol{v}_1 - \boldsymbol{w}_1\|_2^2$. Supposing there are k incorrectly classified vertices, we have that

$$k \leq \min\left\{\|\boldsymbol{v}_1 - \boldsymbol{w}_1\|_2^2\,,\, \|-\boldsymbol{v}_1 - \boldsymbol{w}_1\|_2^2\right\}$$

which by lemma 6.2 is at most

$$\left\{\|\boldsymbol{v}_1 - \boldsymbol{w}_1\|_2^2\,,\, \|-\boldsymbol{v}_1 - \boldsymbol{w}_1\|_2^2\right\} \leq \frac{c'}{p-q}\sqrt{n(p+q)}$$

Given $p - q \geq \frac{c'}{\epsilon}\sqrt{\frac{p+q}{n}}$, we have

$$\frac{c'}{p-q}\sqrt{n(p+q)} \leq \epsilon n$$

hence $k \leq \epsilon n$ as required. $\square$

Theorem 6.5 states that as p and q become more separated, the algorithm recovers the planted partition with greater accuracy matching our intuition of the problem exactly.

References

[1] McSherry, F. (2001, October). Spectral partitioning of random graphs. In *FOCS '01: Proceedings of the 42nd IEEE symposium on Foundations of Computer Science* (p. 529). IEEE.

[2] Singer, A., & Shkolnisky, Y. (2011). Three-dimensional structure determination from common lines in cryo-EM by eigenvectors and semidefinite programming. *SIAM journal on imaging sciences*, 4(2), 543-572.

[3] Shkolnisky, Y., & Singer, A. (2012). Viewing direction estimation in cryo-EM using synchronization. *SIAM journal on imaging sciences*, 5(3), 1088-1110.

[4] Vu, V. H. (2007). Spectral norm of random matrices. *Combinatorica*, 27(6), 721-736.