CS 598: Theoretical Machine Learning

Lecturer: Pranjal Awasthi
Scribe: Aditya Potukuchi

Lecture # 15
11/03/2017

1 Sparse clustering in the stochastic block model

In the last lecture, we were looking at clustering sparse graphs in the stochastic block model with parameters $p > q$, where $p = \frac{a}{n}$, and $q = \frac{b}{n}$, we saw that we could not do perfect clustering as w.h.p, there will be a constant fraction of isolated vertices. However, we shall see that we can still do *weak* recovery, i.e., by allowing a small fraction of vertices to be misplaced, we will recover the rest of the vertices with relatively good guarantees.

Theorem 1. *If $a - b \geq \frac{c}{\epsilon^2}\sqrt{a}$, then w.h.p, we can recover S_1' and S_2' such that*

$$|S_1 \Delta S_1'| + |S_2 \Delta S_2'| \leq \epsilon n$$

Previously, we proposed the following algorithm:

1. Let A be the adjacency matrix.
2. Let d be the average degree $\frac{a+b}{n}$.
3. Remove any vertex with degree $> 4d$.
4. Run spectral clustering on $\hat{A}$, the new matrix.

To see why this is not as straightforward as the previous algorithms, let us review this analysis. Very briefly, the sketch of the proofs before, involved the following steps:

1. Set $M := \mathbf{E}[A] = \frac{a+b}{2} v_1 v_1^T + \frac{a+b}{2} v_2 v_2^T$, where $v_1 = \mathbb{1}$, and $v_2 = \mathbb{1}_{S_1} - \mathbb{1}_{S_2}$.
2. Compute the SVD $A = \sigma_1 w_1 w_1^T + \sigma_2 w_2 w_2^T + \cdots$.
3. Prove that $\|M - A\|$ is small , i.e., $\approx \sqrt{a+b}$.
4. Use Davis-Kahan theorem to say that this implies $\|v_2 - w_2\|$ is small.

However, this analysis does not easily extend, and the (only!) problem is in step 3. In particular, for sparse graphs, we have average degree $\approx \frac{a+b}{2}$, but max degree $> \sqrt{\frac{\log n}{\log \log n}}$, and so for all we know, it is possible that $\|M - A\| \approx \sqrt{\frac{\log n}{\log \log n}}$. To rectify this, we will hope that after throwing away the entries of large degrees, we are left with $\|M - A\| \approx \sqrt{a+b}$.

In order to prove the above claim, we will use the proof which goes along the lines of Le, Levina, and Vershynin, in [1].

Theorem 2. *Let A be an $n \times n$ random matrix with independent entries $\in \{0, 1\}$, let $p_{i,j} = \mathbf{E}[A_{i,j}]$, and let $d := n \max_{i,j} p_{i,j}$. Choose a set of $\frac{n}{d}$ rows and columns, and set some of the 1's to 0's. Let $\hat{A}$ be the new obtained matrix. Then, with probability at least $1 - \frac{1}{n^3}$, we have $\|\hat{A} - \mathbf{E}[A]\| \leq O\left(\sqrt{d} + \sqrt{d'}\right)$, where d' is the max degree in $\hat{A}$.*

It is clear that given the proof of the above theorem, the analysis works with the small proposed change. Indeed, there are at most vertices $4^{-4}n$ vertices of degree at least $4d$, just delete them. Now we can run the previous algorithm and the analysis carries forward verbatim. Since we are allowing an error of a constant fraction of vertices, placing the deleted vertices in partitions arbitrarily is still good enough. Hence we will now turn our attention to proving the theorem. We have the adjacency matrix A. For the sake of the proof, assume all the entries are the same. The key lemma that is proved here is the following:

Lemma 1 (Decomposition lemma)**.** *Let A be a random matrix with independent entries, and $P = \mathbf{E}[A]$ can be decomposed into 3 parts $\mathcal{N}$, $\mathcal{R}$, $\mathcal{C}$, where $\mathcal{N}$, $\mathcal{R}$, and $\mathcal{C}$ are just subsets of the entries of the matrix with the following properties:*

1. $\|(A - \mathbf{E}[A])_{\mathcal{N}}\| = O\left(\sqrt{d}\right)$.
2. *$\mathcal{C}$ will intersect at most $\frac{n}{d}$ rows.*
3. *Every column has only $O(1)$ many nonzero points in $\mathcal{C}$.*
4. *$\mathcal{R}$ intersects at most $\frac{n}{d}$ columns.*
5. *Ever row has at most $O(1)$ many nonzero points in $\mathcal{R}$.*

A brief sketch of the proof is given below, we visit the proof of this in (slightly more) detail in Section 2:

First, we establish that a random matrix, there is a big block of $\frac{n}{d}$ where the entries concentrate well, Let I be the remaining rows and S be the remaining matrix. We have that $\|(A - \mathbf{E}[A])_{S_1}\|_{\infty\to 2} \leq O\left(\sqrt{dn}\right)$ (easy). Then establish the *Grothendick-Pietsch* factorization: There is a $J \subseteq [n]$, such that $|J| \geq \frac{3n}{4}$. such that $\|(A - \mathbf{E}[A])_{i\times J}\| \leq O\left(\sqrt{d}\right)$.

We then do the same procedure on the transpose to eliminate some columns, and get a large chunk of matrix which can be put into $\mathcal{N}$, and a remaining $n/2 \times n/2$ matrix where we don't know what's happening, however, we recursively to the same procedure, and get a bound.

On the way to proving Theorem 2, we will also need the following elementary fact:

Fact 1. *If A is a matrix with maximum ℓ^1 norm of any a, and maximum ℓ^1 norm of any column at most b, then $\|A\| \leq \sqrt{ab}$.*

Now we are ready to prove the theorem.

Proof of Theorem 2. Assume that we are given the partition of the entries of A into three subset of coordinates $\mathcal{N}$, $\mathcal{R}$, and $\mathcal{C}$ as in Lemma 1, since this happens with high probability at least $1 - \frac{1}{n^3}$. This is all we need, and the rest of the proof is deterministic. We have $\|\hat{A} - \mathbf{E}[A]\| \leq \|(\hat{A} - \mathbf{E}[A])_{\mathcal{N}}\| + \|(\hat{A} - \mathbf{E}[A])_{\mathcal{R}}\| + \|(\hat{A} - \mathbf{E}[A])_{\mathcal{C}}\|$. We handle each summand separately, in a relatively straightforward way.

For the first summand, let I_1 be the subset of rows of A that changed to obtain $\hat{A}$, and I_2 be the subset of columns of A that are changed to obtain $\hat{A}$. Denote the $I_1 \times [n]$ submatrix of A to be E_1, and the $[n] \times I_2$ submatrix of A to be E_2. We have

$$\begin{aligned}
\|(\hat{A} - \mathbf{E}[A])_{\mathcal{N}}\| &\leq \|(A - \hat{A})_{\mathcal{N}}\| + \|(\mathbf{E}[A] - A)_{\mathcal{N}}\| \\
&\leq \|A - \hat{A}\| + O\left(\sqrt{d}\right) \\
&\leq \|(A - \hat{A})_{\mathcal{N}\cap E_1}\| + \|(A - \hat{A})_{\mathcal{N}\cap E_2}\| + O\left(\sqrt{d}\right) \\
&\leq 2\|A_{\mathcal{N}\cap E_1}\| + 2\|A_{\mathcal{N}\cap E_2}\| + O\left(\sqrt{d}\right) \\
&\leq 2\|(A - \mathbf{E}[A])_{\mathcal{N}\cap E_1}\| + 2\|\mathbf{E}[A]_{\mathcal{N}\cap E_1}\| + 2\|A_{\mathcal{N}\cap E_2}\| + O\left(\sqrt{d}\right) \\
&\leq 2\|A_{\mathcal{N}\cap E_2}\| + O\left(\sqrt{d}\right) \\
&\leq O\left(\sqrt{d}\right)
\end{aligned}$$

Every inequality above is proved using standard techniques: we mostly used the triangle inequality, and the fact that since the entries of A and $A - \hat{A}$ and are nonnegative, reducing some (nonzero) entries, and in particular, setting them to 0 may only decrease the norm. The bound on $\|A_{\mathcal{N}\cap E_2}\|$ is obtained in the same was as on $\|A_{\mathcal{N}\cap E_1}\|$. Also, Fact 1 was used to bound $\mathbf{E}[A]_{\mathcal{N}\cap E_1}$.

For the second term $\|(\hat{A} - \mathbf{E}[A])_{\mathcal{R}}\|$, we have:

$$\begin{aligned}
\|(\hat{A} - \mathbf{E}[A])_{\mathcal{R}}\| &\leq \|\hat{A}_{\mathcal{R}}\| + \|\mathbf{E}[A]_{\mathcal{R}}\| \\
&\leq O\left(\sqrt{d'} + \sqrt{d}\right)
\end{aligned}$$

Where we have again used Fact 1 to bound both summands.

And the bound $\|(\hat{A} - \mathbf{E}[A])_{\mathcal{C}}\| \leq O\left(\sqrt{d} + \sqrt{d'}\right)$ is obtained similarly, completing the proof.

□

2 Proof of Lemma 1

In this section, we will give a slightly more detailed sketch of Lemma 1. We will be slightly loose with the constants here, since these do not affect the stated asymptotics of the performance. First, we define the $\ell^{\infty\to 2}$ norm, for an $n \times n$ matrix M, let us denote:

$$\|M\|_{\infty\to 2} := \max_{\|x\|_\infty = 1} \|Mx\|_2 = \max_{x\in\{\pm 1\}^n} \|Mx\|_2$$

Clearly, we have

$$\frac{\|M\|_{\infty\to 2}}{\sqrt{n}} \leq \|M\| \leq \|M\|_{\infty\to 2}$$

The key tool used in the proof is the *Grothendeick-Pietsch* factorization, which gives us a better handle on going from the $\ell^{\infty\to 2}$ norm to the ℓ^2 norm. We will state the exact version that we are going to use:

Theorem 3 (Grothendeick-Pietsch factorization for submatrices)**.** *Let M be an $m \times n$ matrix, then there is a subset $J \subseteq [n]$ of size $(1 - \delta)n$ such that*

$$\|M_{[m]\times J}\| \leq \frac{2\|M\|_{\infty\to 2}}{\sqrt{\delta n}}$$

Let $I' \subseteq [n]$ be the set of rows of A that have at most αd ones for some $\alpha \geq 1$ (think of $\alpha \sim 10$). We start off by getting bounds on $\left\|A_{I' \times [n]}\right\|_{\infty \to 2}$. The following lemma may be taken as an exercise. Intuitively, it gets moderate bounds on some statistics of all cuts in the underlying graph:

Lemma 2. *For A sampled as per the stochastic block model with parameters $p = \frac{a}{n}$, and $q = \frac{b}{n}$, where the expected average degree is at most $d := a$. Let $I,\ J \subseteq [n]$ where $|I| = |J| = m$ and $I' \subseteq I$ are all the rows with at most αd ones where $\alpha \geq \frac{m}{n}$, we have that with probability at least $1 - n^{-r}$,*

$$\left\|(A - \mathbf{E}A)_{I' \times [n]}\right\|_{\infty \to 2} = O\left(\sqrt{\alpha d n r \ln\left(\frac{n}{m}\right)}\right)$$

Let us prove Lemma 1 assuming that we have Lemma 2 and Theorem 3. We will need a couple more simple lemmas. Intuitively, the first one says that in every subgraph, the degrees are essentially what they ought to be:

Lemma 3. *With the notation as above, consider subsets $I, J \subseteq [n]$ with $|I| = |J| = m$, and a parameter $\alpha \geq \sqrt{\frac{m}{n}}$. The probability that the number of rows of this submatrix with greater than $8\alpha d$ ones is more than $\frac{m}{\alpha d}$ at most n^{-r}.*

Proof sketch. Fix sets $I, J \subset [n]$ with $|I| = |J| = m$, and a vertex $i \in I$. Denote $d_J(i)$ to be its degree in J. We have $\mathbf{E}[d_J(i)] = \frac{dm}{n}$. We have that (by approximating using Poisson tail):

$$\Pr(d_J(i) > 8r\alpha d)) \leq \left(\frac{8r\alpha n}{m}\right)^{-8r\alpha d} =: p$$

Now we bound the probability that too many vertices have high degree in J. Let $X := |\{i \in I \mid d_J(i) \geq 8r\alpha d\}|$. We have $\mathbf{E}[X] \leq m \cdot \left(\frac{8r\alpha n}{m}\right)^{-8r\alpha d} = pm$

$$\begin{aligned}
\Pr(X \geq m/(\alpha d)) &\leq (p\alpha d)^{\frac{m}{\alpha d}} \\
&\lesssim \left(\frac{r\alpha n}{m}\right)^{-4rm} \\
&\ll \binom{n}{m}^{-2}
\end{aligned}$$

for, say, $r > e$, and $\alpha \geq \sqrt{\frac{m}{n}}$. This allows us to union bound over all sets I, and J of sizes at most m. The error probability bound is dominated by the largest summand the union bound, where m is small, say, $m = 1$ for example. □

The next one says that even if a few rows have many ones, many columns have only $O(1)$ ones in those rows.

Lemma 4. *Let $I, J \subseteq [n]$ with $|I| = k$ and $|J| = m$, where $k \leq m/\alpha d$, where $\alpha \geq \sqrt{\frac{m}{n}}$ is a parameter. Then, the probability that more than $m/4$ columns have more than $32r$ ones is at most n^{-r}.*

The proof is exactly the same as the one above

Proof sketch. Fix an I and J satisfying the conditions above. For $j \in J$, let $d_I(j)$ denote the degree in J. We have that $\mathbf{E}[d_I(j)] = k\frac{d}{n}$. Therefore, we have:

$$\Pr(d_I(j) \geq 32r) \leq \left(\frac{32nr}{dk}\right)^{-32r} =: p$$

And similarly, as before let $X := |\{j \in J \mid d_I(j) \geq 32r\}|$. We have:

$$\begin{aligned}\Pr(X \geq m/4) &\leq (4p)^{m/4} \\ &\leq \left(\frac{dk}{32nr}\right)^{8mr} \\ &\leq \left(\frac{m}{32n\alpha d}\right)^{8mr} \\ &\ll \binom{n}{m}^{-1}\binom{n}{k}^{-1}\end{aligned}$$

And we use the union bound again, as above.

□

The idea for the rest of the proof is as follows: We start off first find a set of rows $I' \subset [n]$ containing more than $8d$ ones. By Lemma 3, we have w.h.p, $|I'| \leq \frac{n}{d}$, and by Lemma 4, we have that at least $\frac{3n}{4}$ columns have at most $32r$ ones in I', let the set of these 'good' columns be called J_1. Also, from Theorem 3, we know that there is a set of at least $\frac{3m}{4}$ columns J_2 such that

$$\|(A - \mathbf{E}[A])_{I' \times J_2}\| \leq \frac{4\|(A - \mathbf{E}[A])_{I' \times [n]}\|_{\infty \to 2}}{\sqrt{n}} \leq C\sqrt{dr},$$

by Lemma 2. Setting $J := J_1 \cap J_2$, we have that $|J| \geq \frac{n}{2}$, and, using, standard inequalities, we have $\|A_{([n]\setminus I') \times J}\| \leq C\sqrt{dr}$.

Applying the same procedure to A^t, let J' be all the columns that have more than $8d$ ones, we have a set of rows I such that $|I| \geq \frac{n}{2}$ and $\|(A - \mathbf{E}[A])_{I \times ([n]\setminus J')}\| \leq C\sqrt{dr}$. So we add all entries of $I' \times J$ to $\mathcal{R}$, all entries of $I \cap J'$ to $\mathcal{C}$, and all the entries of $(([n] \setminus I') \times J) \cup (I \times ([n] \setminus J'))$ to $\mathcal{N}$. So we have an $n/2 \times n/2$ submatrix of entries to account for. We recursively apply the same procedure and use standard inequalities to bound the total sizes of $\mathcal{R}$, $\mathcal{C}$, and the norm of $(A - \mathbf{E}[A])_{\mathcal{N}}$.

References

[1] Can M. Le, Elizaveta Levina, Roman Vershynin, *Concentration and regularization of random graphs*, (Preliminary version), `https://arxiv.org/abs/1506.00669`