Stat260/CS294: Spectral Graph Methods **Lecture 24 - 04/21/2015**

Lecture: Some Statistical Inference Issues (3 of 3)

Lecturer: Michael Mahoney *Scribe: Michael Mahoney*

Warning: these notes are still very rough. They provide more details on what we discussed in class, but there may still be some errors, incomplete/imprecise statements, etc. in them.

24 Stochastic blockmodels

Today, we will finish up talking about statistical inference issues by discussing them in the context of stochastic blockmodels. These are different models of data generation than we discussed in the last few classes, and they illustrate somewhat different issues.

24.1 Introduction to stochastic block modeling

As opposed to working with expansion or conductance—or some other "edge counting" objective like cut value, modularity, etc.—the *stochastic block model (SBM)* is an example of a so-called probabilistic or generative model. Generative models are a popular way to encode assumptions about the way that latent/unknown parameters interact to create edges (ij) Then, they assign a probability value for each edges (ij) in a network. There are several advantages to this approach.

- It makes the assumptions about the world/data explicit. This is as opposed to encoding them into an objective and/or approximation algorithm—we saw several examples of reverse engineering the implicit properties of approximation algorithms.
- The parameters can sometimes be interpreted with respect to hypotheses about the network structure.
- It allows us to use likelihood scores, to compare different parameterizations or different models.
- It allows us to estimate missing structures based on partial observations of graph structure.

There are also several disadvantages to this approach. The most obvious is the following.

- One must fit the model to the data, and fitting the model can be complicated and/or computationally expensive.
- As a result of this, various approximation algorithms are used to fit the parameters. This in turn leads to the question of what is the effect of those approximations versus what is the effect of the original hypothesized model? (I.e., we are back in the other case of reverse engineering the implicit statistical properties underlying approximation algorithms, except here it is in the approximation algorithm to estimate the parameters of a generative model.) This problem is particularly acute for sparse and noisy data, as is common.

Like other generative models, SBMs define a probability distribution over graphs, $\mathbb{P}\left[G|\Theta\right]$, where Θ is a set of parameters that govern probabilities under the model. Given a specific Θ, we can then draw or *generate* a graph G from the distribution by flipping appropriately-biased coins. Note that *inference* is the reverse task: given a graph G, either just given to us or generated synthetically by a model, we want to recover the model, i.e., we want to find the specific values of Θ that generated it.

The simpled version of a SBM is specified by the following.

- A positive integer k, a scalar value denoting the the number of blocks.
- A vector $\vec{z} \in \mathbb{R}^n$, where z_i gives the group index of vertex i.
- A matrix $M \in \mathbb{R}^{k \times k}$, a stochastic block matrix, where M_{ij} gives the probability that a vertex of type i links to a vertex of type j.

Then, one generates edge (ij) with probability $M_{z_i z_j}$. That is, edges are not identically distributed, but they are conditionally independent, i.e., conditioned on their types, all edges are independent, and for a given pair of types (ij), edges are i.i.d.

Observe that the SBM has a relatively large number of parameters, $\binom{k}{2}$, even after we have chosen the labeling on the vertices. This has plusses and minuses.

- Plus: it allows one the flexibility to model lots of possible structures and reproduce lots of quantities of interest.
- Minus: it means that there is a lot of flexibility, thus making the possibility of overfitting more likely.

Here are some simple examples of SBMs.

- If $k = 1$ and $M_{ij} = p$, for all i, j, then we recover the vanilla ER model.
- Assortative networks, if $M_{ii} > M_{ij}$, for $i \neq j$.
- Disassortative networks, if $M_{ii} < M_{ij}$, for $i \neq j$.

24.2 Warming up with the simplest SBM

To illustrate some of the points we will make in a simple context, consider the ER model.

- If, say, $p = \frac{1}{2}$ and the graph G has more than a handful of nodes, then it will be very easy to estimate p, i.e., to estimate the parameter vector Θ of this simple SBM, basically since measure concentration will occur very quickly and the empirical estimate of p we obtain by counting the number of edges will be very close to its expected value, i.e., to p. More generally, if n is large and $p \gtrsim \frac{\log(n)}{n}$, then measure will still concentrate, i.e., the empirical and expected values of p will be close, and we will be able to estimate p well. (This is related to the well-known observation that if $p \gtrsim \frac{\log(n)}{n}$, then G_{np} and G_{nm} are very similar, for appropriately chosen values of p and m.)

- If, on the other hand, say, $p = \frac{3}{n}$, then this is not true. In this regime, measure has *not* concentrated for most statistics of interest: the graph is not even fully connected; the giant component has nodes of degree almost $O\left(\log(n)\right)$; and the giant component has small sets of nodes of size $\Theta\left(\log(n)\right)$ that have conductance $O\left(\frac{1}{\log(n)}\right)$. (Contrast all of these the a 3-regular random graph, which: is fully connected, is degree-homogeneous, and is a very good expander.)

In these cases when measure concentration fails to occur, e.g., due to exogenously-specified degree heterogeneity or due to extreme sparsity, then one will have difficulty with recovering parameters of hypothesized models. More generally, similar problems arise, and the challenge will be to show that one can reconstruct the model under as broad a range of parameters as possible.

24.3 A result for a spectral algorithm for the simplest nontrivial SBM

Let's go into detail on the following simple SBM (which is the simplest aside from ER).

- Choose a partition of the vertics, call them V^1 and V^2, and WLOG let $V^1 = \{1, \ldots, \frac{n}{2}\}$ and $V^2 = \{\frac{n}{2}+1, \ldots, n\}$.
- Then, choose probabilities $p > q$ and place edges between vertices i and j with probability

$$\mathbb{P}\left[(ij) \in E\right] = \begin{cases} q & \text{if } i \in V^1 \text{ and } j \in V^2 \text{ of } i \in V^2 \text{ and } j \in V^1 \\ p & \text{otherwise} \end{cases},$$

In addition to being the "second simplest" SBM, this is also a simple example of a *planted partition model*, which is commonly studied in TCS and related areas.

Here is a fact:

$$\mathbb{E}\left[\text{number of edges crossing bw } V^1 \text{ and } V^2\right] = q|V^1||V^2|.$$

In addition, if p is sufficiently larger than q, then every other partition has more edges. This is the basis of recovering the model. Of course, if p is only slightly but not sufficiently larger than q, then there might be fluctuational effects such that it is difficult to find this from the empirical graph. This is analogous to having difficulty with recovering p from very sparse ER, as we discussed.

Within the SBM framework, the most important inferential task is recovering cluster membership of nodes from a single observation of a graph (i.e., the two clusters in this simple planted partition form of the SBM). There are a variety of procedures to do this, and here we will describe spectral methods.

In particular, we will follow a simple analysis motivated by McSherry's analysis, as described by Spielman, that will provide a "positive" result for sufficiently dense matrices where p and q are sufficiently far apart. Then, we will discuss this model more generally, with an emphasis on how to deal with very low-degree nodes that lead to measure concentration problems. In particular, we will focus on a form of regularized spectral clustering, as done by Qin and Rohe in their paper "Regularized spectral clustering under the degree-corrected stochastic blockmodel." This has connections with what we have done with the Laplacian over the last few weeks.

To start, let M be the *population adjacency matrix*, i.e., the hypothesized matrix, as described above. That is,

$$M = \begin{pmatrix} p\vec{1}\vec{1}^T & q\vec{1}\vec{1}^T \\ q\vec{1}\vec{1}^T & p\vec{1}\vec{1}^T \end{pmatrix}$$

Then, let A be the *empirical adjacency matrix*, i.e., the actual matrix that is generated by flipping coins and on which we will perform computations. This is generated as follows: let $A_{ij} = 1$ w.p. M_{ij} and s.t. $A_{ij} = A_{ji}$. So, the basic goal is going to be to recover clusters in M by looking at information in A.

Let's look at the eigenvectors. First, since $M\vec{1} = \frac{n}{2}(p+q)\vec{1}$, we have

$$\begin{aligned} \mu_1 &= \frac{n}{2}(p+q) \\ w_1 &= \vec{1}, \end{aligned}$$

where μ_1 and w_1 are the leading eigenvalue and eigenvector, respectively. Then, since the second eigenvector (of M) is constant on each cluster, we have that $Mw_2 = \mu_2 w_2$, where

$$\begin{aligned} \mu_2 &= \frac{n}{2}(p-q) \\ w_2 &= \begin{cases} \frac{1}{\sqrt{n}} \text{ if } i \in V^1 \\ -\frac{1}{\sqrt{n}} \text{ if } i \in V^2 \end{cases} . \end{aligned}$$

In that case, here is a simple algorithm for finding the planted bisection.

1. Compute v_2, the eigenvector of second largest eigenvalue of A.
2. Set $S = \{i : v_2(i) \geq 0\}$
3. Guess that S is one side of the bisection and that $\bar{S}$ is the other side.

We will show that under not unreasonable assumptions on p, q, and S, then by running this algorithm one gets the hypothesized cluster mostly right.

Why is this?

The basic idea is that A is a perturbed version of M, and so by perturbation theory the eigenvectors of A should look like the eigenvectors of M.

Let's define $R = A - M$. We are going to view R as a random matrix that depends on the noise/randomness in the coin flipping process. Since matrix perturbation theory bounds depend on (among other things) the norm of the perturbation, the goal is to bound the probability that $\|R\|_2$ is large. There are several methods from random matrix theory that give results of this general form, and one or the other is appropriate, depending on the exact statement that one wants to prove. For example, if you are familiar with Wigner's semi-circle law, it is of this general form. More recently, Furedi-Komlos got another version; as did Krivelevich and Vu; and Vu. Here we state a result due to Vu.

Theorem 1. *With probability tending to one, if $p \geq c\frac{\log^4(n)}{n}$, for a constant c, then*

$$\|R\|_2 \leq 3\sqrt{pn}.$$

The key question in theorems like this is the value of p. Here, one has that $p \gtrapprox \frac{\log(n)}{n}$, meaning that one can get pretty sparse (relative to $p = 1$) but not extremely sparse (relative to $p = \frac{1}{n}$ or $p = \frac{3}{n}$). If one wants stronger results (e.g., not just mis-classifying only a constant fraction of the vertices, which we will do below, but instead that one predicts correctly for all but a small fraction of the vertices), then one needs p to be larger and the graph to be denser. As with the ER example, the reason for this is that we need to establish concentration of appropriate estimators.

Let's go onto perturbation theory for eigenvectors. Let $\alpha_1 \geq \alpha_2 \geq \cdots \alpha_n$ be the eigenvalues of A, and let $\mu_1 > \mu_2 > \mu_3 = \cdots \mu_n = 0$ be the eigenvalues of M.

Here is a fact from matrix perturbation theory that we mentioned before: for all i,

$$|\alpha_i - \mu_i| \leq \|A - M\|_2 = \|R\|_2.$$

The following two claims are easy to establish.

Claim 1. *If $\|R\|_2 < \frac{n}{4}(p-q)$, then*

$$\frac{n}{4}(p-q) < \alpha_2 < \frac{3n}{4}(p-q)$$

Claim 2. *If, in addition, $q > \frac{p}{3}$, then $\frac{3n}{4}(p-q) < \alpha_1$.*

From these results, we have a separation, and so we can view α_2 as a perturbation of μ_2. The question is: can we view v_2 as a perturbation of w_2? The answer is Yes. Here is a statement of this result.

Theorem 2. *Let A, M be symmetric matrices, and let $R = M - A$. Let $\alpha_1 \geq \cdots \geq \alpha_n$ be the eigenvectors of A, with $v_1, \cdots, v_n$ the corresponding eigenvectors. Let $\mu_1 \geq \cdots \geq \mu_n$ be the eigenvectors of M, with $w_1, \cdots, w_n$ the corresponding eigenvectors. Let θ_i be the angle between v_i and w_i. Then,*

$$\sin\theta_i \leq \frac{2\|R\|_2}{\min_{j\neq i}|\alpha_i - \alpha_j|}$$
$$\sin\theta_i \leq \frac{2\|R\|_2}{\min_{j\neq i}|\mu_i - \mu_j|}$$

Proof. WLOG, we can assume $\mu_i = 0$, since the matrices $M - \mu_i I$ and $A - \alpha_i I$ have the same eigenvectors as M and A, and $M - \mu_i I$ has the ith eigenvalue being 0. Since the theorem is vacuous if μ_i has multiplicities, we can assume unit multiplicity, and that w_i is a unit vector in the null space of M. Due to the assumption that $\mu_i = 0$, we have that $|\alpha_i| \leq \|R\|_2$.

Then, expand v_i in an eigenbasis of M: $v_i = \sum_j c_j w_j$, where $c_j = w_j^T v_i$. Let $\delta = \min_j |\mu_j|$. Then observe that

$$\|Mv_i\|_2^2 = \sum_j c_j^2 \mu_j^2 \geq \sum_{j\neq i} c_j^2 \delta^2 = \delta^2 \sum_{j\neq i} c_j^2 = \delta^2\left(1 - c_i^2\right) = \delta^2 \sin^2\theta_i$$

and also that

$$\|Mv_i\| \leq \|Av_i\| + \|Rv_i\| = \alpha_i + \|Rv_i\| \leq 2\|R\|_2.$$

So, from this it follows that $\sin\theta_i \leq \frac{2\|R\|_2}{\delta}$. □

This is essentially a version of the Davis-Kahan result we saw before. Note that it says that the amount by which eigenvectors are perturbed depends on how close are other eigenvalues, which is what we would expect.

Next, we use this for partitioning the simple SBM. We want to show that not too many vertices are mis-classified.

Theorem 3. *Given the two-class SBM defined above, assume that* $p \geq c\frac{\log^4(n)}{n}$ *and that* $q > p/3$. *If one runs the spectral algorithm described above, then at most a constant fraction of the vertices are misclassified.*

Proof. Consider the vector $\vec{\delta} = v_2 - w_2$. For all $i \in V$ that are misclassified by v_2, we have that $|\delta(i)| \geq \frac{1}{\sqrt{n}}$. So, if v_2 misclassified k vertices, then $\|\delta\| \geq \sqrt{k/n}$. Since u and v are unit vectors, we have the crude bound that $\|\delta\| \leq \sqrt{2}\sin\theta_2$.

Next, we can combine this with the perturbation theory result above. Since $q > p/3$, we have that $\min_{j \neq 2} |\mu_2 - \mu_i| = \frac{n}{2}(p-q)$; and since $p \geq c\frac{\log^4(n)}{n}$, we have that $\|R\| \leq 3\sqrt{pn}$. Then,

$$\sin\theta_2 \leq \frac{3\sqrt{pn}}{\frac{n}{2}(p-q)} = \frac{6\sqrt{p}}{\sqrt{n}(p-q)}.$$

So, the number k of mis-classified vertices satisfies $\sqrt{\frac{k}{n}} \leq \frac{6\sqrt{p}}{\sqrt{n}(p-q)}$, and thus $k \leq \frac{36p}{(p-q)^2}$. □

So, in particular, if p and q are both constant, then we expect to misclassify at most a constant fraction of the vertices. E.g., if $p = \frac{1}{2}$ and $q = p - \frac{12}{\sqrt{n}}$, then $\frac{36p}{(p-q)^2} = \frac{n}{8}$, and so only a constant fraction of the vertices are misclassified.

This analysis is a very simple result, and it has been extended in various ways.

- The Ng et al. algorithm we discussed before computes k vectors and then does k means, making similar gap assumptions.
- Extensions to have more than two blocks, blocks that are not the same size, etc.
- Extensions to include degree variability, as well as homophily and other empirically-observed properties of networks.

The general form of the analysis we have described goes through to these cases, under the following types of assumptions.

- The matrix is dense enough. Depending on the types of recovery guarantees that are hoped for, this could mean that $\Omega(n)$ of the edges are present for each node, or perhaps $\Omega(\text{polylog}(n))$ edges for each node.
- The degree heterogeneity is not too severe. Depending on the precise algorithm that is run, this can manifest itself by placing an upper bound on the degree of the highest degree node and/or placing a lower bound on the degree of the lowest degree node.
- The number of clusters is fixed, say as a function of n, and each of the clusters is not too small, say a constant fraction of the nodes.

Importantly, *none* of these simplifying assumptions are true for most "real world" graphs. As such, there has been a lot of recent work focusing on dealing with these issues and making algorithms for SBMs work under broader assumptions. Next, we will consider one such extension.

24.4 Regularized spectral clustering for SBMs

Here, we will consider a version of the degree-corrected SBM, and we will consider doing a form of *regularized spectral clustering (RSC)* for it.

Recall the definition of the basic SBM.

Definition 1. *Given nodes $V = [n]$, let $z : [n] \to [k]$ be a partition of the n nodes into k blocks, i.e., z_i is the block membership of the i^{th} node. Let $B \in [0,1]^{k \times k}$. Then, under the SBD, we have that the probability of an edge between i and j is*

$$P_{ij} = B_{z_i z_j}, \text{ for all } i, j \in \{1, \ldots, n\}.$$

In particular, this means that, given z, the edges are independent.

Many real-world graphs have substantial degree heterogeneity, and thus it is common to in corporate this into generative models. Here is the extension of the SBM to the *Degree-corrected stochastic block model (DC-SBM)*, which introduces additional parameters θ_i, for $i \in [n]$, to control the node degree.

Definition 2. *Given the same setup as for the SBM, specify also additional parameters θ_i, for $i \in [n]$. Then, under the DC-SBM, the probability of an edge between i and j is*

$$P_{ij} = \theta_i \theta_j B_{z_i z_j},$$

where $\theta_i \theta_j B_{z_i z_j} \in [0,1]$, for all $i, j \in [n]$.

Note: to make the DC-SBM identifiable (i.e., so that it is possible in principle to learn the true model parameters, say given an infinite number of observations, which is clearly a condition that is needed for inference), one can impose the constraint that $\sum_i \theta_i \delta_{z_i,r} = 1$, for each block r. (This condition says that $\sum_i \theta_i = 1$ within each block.) In this case B_{st}, for $s \neq t$, is the expected number of links between block s and block t; and B_{st}, for $s = t$, is the expected number of links within block s.

Let's say that $A \in \{0,1\}^{n \times n}$ is the adjacency matrix; $L = D^{-1/2} A D^{-1/2}$. In addition, let $\mathcal{A} = \mathbb{E}[A]$ be the population matrix, under the DC-SBM. Then, one can express $\mathcal{A}$ as $\mathcal{A} = \Theta Z B Z^T \Theta$, where $\Theta \in \mathbb{R}^{n \times n} = \mathrm{diag}(\theta_i)$, and where $Z \in \{0,1\}^{n \times k}$ is a membership matrix with $Z_{it} = 1$ iff node i is in block t, i.e., if $z_i = t$.

We are going to be interested in very sparse matrices, for which the minimum node degree is very small, in which case a vanilla algorithm will fail to recover the SBM blocks. Thus, we will need to introduce a regularized version of the Laplacian. Here is the definition.

Definition 3. *Let $\tau > 0$. The* regularized graph Laplacian *is $L_\tau = D_\tau^{-1/2} A D_\tau^{-1/2} \in \mathbb{R}^{n \times n}$, with $D_\tau = D + \tau I$, for $\tau > 0$.*

This is defined for the empirical data; but given this, we can define the corresponding population quantities:

$$\begin{aligned}
\mathcal{D}_{ii} &= \sum_j \mathcal{A}_{ij} \\
\mathcal{D}_\tau &= \mathcal{D} + \tau I \\
\mathcal{L} &= \mathcal{D}^{-1/2}\mathcal{A}\mathcal{D}^{-1/2} \\
\mathcal{L}_\tau &= \mathcal{D}_\tau^{-1/2}\mathcal{A}\mathcal{D}_\tau^{-1/2}
\end{aligned}$$

Two things to note.

- Under the DC-SBM, if the model is identifiable, then one should be able to determine the partition from $\mathcal{A}$ (which we don't have direct access to, given the empirical data).
- One also wants to determine the partition from the empirical data A, under broader assumptions than before, in particular under smaller minimum degree.

Here is a description of the basic algorithm of Qin and Rohe. Basically, it is the Ng et al. algorithm that we described before, except that we apply it to the regularized graph Laplacian, i.e., it involves finding the leading eigenvectors of L_τ and then clustering in the low dimensional space.

Given as input an Adjacency Matrix A, the number of clusters k, and the regularizer $\tau \geq 0$.

1. Compute L_τ.
2. Compute the matrix $X_\tau = [X_1^\tau, \ldots, X_k^\tau] \in \mathbb{R}^{n \times k}$, the orthogonal matrix consisting of the k largest eigenvectors of L_τ.
3. Compute the matrix $X_\tau^* \in \mathbb{R}^{n \times k}$ by normalizing each row of X_τ to have unit length, i.e., project each row of X_τ onto the unit sphere in $\mathbb{R}^k$, i.e., $X_{ij}^{*,\tau} = X_{ij}^\tau / \sum_j X_{ij}^{\tau,2}$.
4. Run k means on the rows of X_τ^* to create k non-overlapping clusters $V_1, \ldots, V_k$.
5. Output $V_1, \ldots, V_k$; node i is assigned to cluster r if the i^{th} tow of X_τ^* is assigned to V.

There are a number of empirical/theoretical tradeoffs in determining the best value for τ, but one can think of τ as being the average node degree.

There are several things one can show here.

First, one can show that L_τ is close to $\mathcal{L}_\tau$.

Theorem 4. *Let G be the random graph with $\mathbb{P}\left[edge\ bw\ ij\right] = P_{ij}$. Let $\delta = \min_i \mathcal{D}_{ii}$ be the minimum expected degree of G. If $\delta + \tau > O\left(\log(n)\right)$, then with constant probability*

$$\|L_\tau - \mathcal{L}_\tau\| \leq O(1)\sqrt{\frac{\log(n)}{\delta + \tau}}.$$

Remark. Previous results required that the minimum degree $\delta \geq O(\log(n))$, so this result generalizes these to allow δ to be much smaller, assuming the regularization parameter τ is large enough.

Importantly, typical real networks do *not* satisfy the condition that $\delta \geq O(\log(n))$, and RSC is most interesting when this condition fails. So, we can apply this result in here to graph with small node degrees.

Remark. The form of L_τ is similar to many of the results we have discussed, and one can imagine implementing RSC (and obtaining this theorem as well as those given below) by computing approximations such as what we have discussed. So far as I know, that has not been done.

Second, one can bound the difference between the empirical and population eigenvectors. For this, one needs an additional concept.

- Given an $n \times k$ matrix A, the *statistical leverage scores* of A are the diagonal elements of the projection matrix onto the span of A.

In particular, if the $n \times k$ matrix U is an orthogonal matrix for the column span of A, then the leverage scores of A are the Euclidean norms of the *rows* of U. For a "tall" matrix A, the i^{th} leverage score has an interpretation in terms of the leverage or influence that the i^{th} row of an A has on the least-squares fit problem defined by A. In the following, we will use an extension of the leverage scores, defined relative to the best rank-k approximation the the matrix.

Theorem 5. *Let X_τ and $\mathcal{X}_\tau$ be in $\mathbb{R}^{n \times k}$ contain the top k eigenvectors of L_τ and $\mathcal{L}_\tau$, respectively. Let*

$$\xi = \min_i \{ \min\{ \|X_\tau^i\|_2, \|\mathcal{X}_\tau^i\|_2 \} \}.$$

Let X_τ^* and $\mathcal{X}_\tau^*$ be the row normalized versions of X_τ and $\mathcal{X}_\tau$. Assume that $\sqrt{\frac{k\log(n)}{\delta+\tau}} \leq O(\lambda_k)$ and $\delta + \tau > O(\log(n))$. Then, with constant probability,

$$\begin{aligned} \|X_\tau - \mathcal{X}_\tau O\|_F &\leq O\left(\frac{1}{\lambda_k} \sqrt{k\log(n)}\delta + \tau \right) \\ \|X_\tau^* - \mathcal{X}_\tau^* O\|_F &\leq O\left(\frac{1}{\xi\lambda_k} \sqrt{k\log(n)}\delta + \tau \right), \end{aligned}$$

where O is a rotation matrix.

Note that the smallest leverage score enters the second expression but not the first expression. That is, it does not enter the bounds on the empirical quantities, but it does enter into the bounds for the population quantities.

We can use these results to derive misclassification rate for RSC. The basic idea for the misclassification rate is to run k-means on the rows of X_τ^* and also on the rows of $\mathcal{X}_\tau^*$. Then, one can say that a node on the empirical data is clustered correctly if it is closer to the centroid of the corresponding cluster on the population data. This basic idea needs to be modified to take into account the fact that if any λ_i are equal, then only the subspace spanned by the eigenvectors is identifiable, so we consider this up to a rotation O.

Definition 4. *If $C_i O$ is closer to $\mathcal{C}_i$ than any other $\mathcal{C}_j$, then we say that the node is correctly clustered; and we define the misclassified nodes to be*

$$\mathcal{M} = \left\{ i : \exists j \neq i \ s.t. \ \|C_i O^T - \mathcal{C}_i\|_2 > \|\mathcal{C}_i O^T - C_j \right\}.$$

Third, one can bound the misclassification rate of the RCS classifier with the following theorem.

Theorem 6. *With constant probability, the misclassification rate is*

$$\frac{|\mathcal{M}|}{n} \leq c\frac{k\log(n)}{n\xi^2(\delta+\tau)\lambda_k^2}.$$

Here too the smallest leverage score determines the overall quality.

Remark. This is the first result that explicitly relates leverage scores to the statistical performance of a spectral clustering algorithm. This is a large topic, but to get a slightly better sense of it, recall that the leverage scores of $\mathcal{L}_\tau$ are $\|\mathcal{X}_\tau^i\|_2^2 = \frac{\theta_i^\tau}{\sum_j \theta_j^\tau \delta_{z_j z_i}}$. So, in particular, if a node i has a small expected degree, then θ_i^τ is small and $\|\mathcal{X}_\tau^i\|_2$ is small. Since ξ appears in the denominator of the above theorems, this leads to a worse bound for the statistical claims in these theorems. In particular, the problem arises due to projecting X_τ^i onto the unit sphere, i.e., while large-leverage nodes don't cause a problem, errors for small-leverage rows can be amplified—this didn't arise when we were just making claims about the empirical data, e.g., the first claim of Theorem **??**, but when considering statistical performance, e.g., the second claim of Theorem **??** or the claim of Theorem **??**, for nodes with small leverage score it amplifies noisy measurements.