Spectral algorithms are observed to work very well in practice, much better than the theoretical guarantees. Researchers have been trying to explain this phenomenon rigorously, that spectral algorithms work well in "practical" instances.

One direction is to show that spectral algorithms perform well in special instances, like the planar graph result we saw in week 2. So, if in practice the graphs are like planar graphs, these results partially explain why spectral algorithms perform better than the worst case performance guarantees.

Another direction is to show that spectral algorithms perform well in an average instance. When it comes to average case analysis we don't have a very compelling theoretical model. A commonly considered model is the random graph model, but it is not very satisfactory in capturing practical instances, e.g. a random graph is an expander graph (all sets have high conductance) with high probability but we seldom see an expander graph in practice.

The semi-random model, that we are going to study today, is proposed to (partially) address this issue.

---

**Semi-Random Graph Problems [1]**

Three problems that have been extensively studied in the semi-random model (or planted model) are:

- **Bisection**: Fix a bisection \( V = X \cup Y \). For each pair of vertices \( u \in X \) or \( v \in Y \), put an edge between \( uv \) with probability \( p \). For each pair of vertices \( u \in X \) and \( v \in Y \), put an edge between \( uv \) with probability \( q \). When \( q \geq p \), \( (X, Y) \) is the optimal bisection with high probability.

- **Clique**: Hide a clique in the graph. Add each remaining edge with probability \( p \). When \( p = \frac{1}{2} \), the maximum clique size in a random graph is \( \Theta(\log n) \) with high probability, and so if the hidden clique is large enough it will be the optimal solution with high probability.

- **k-coloring**: Fix a k-coloring. Add each edge with probability \( p \), and remove all edges within the same color classes.

For general graphs, these problems are hard problems even to give a reasonable approximation algorithm.

In these very special instances, can we solve the problems by recovering the hidden structures?

Spectral algorithms were shown to be effective in these planted instances, on a case by case analysis.

Today we study a unified framework, proposed by McSherry [1], to solve these planted problems.
Before that, we first introduce two important results: singular value decomposition and eigenvalues in random graphs.

Matrix Norm (see [2] for more details)

Before we state the singular value decomposition, we first recall some facts about matrix norms.

In general, a norm is a function \( \| \cdot \| : \mathbb{R}^n \rightarrow \mathbb{R} \) that assigns a length to each vector satisfying:

1. \( \| x \| \geq 0 \), and \( \| x \| = 0 \) only if \( x = 0 \).
2. \( \| x + y \| \leq \| x \| + \| y \| \), the triangle inequality.
3. \( \| ax \| = |a| \| x \| \) for any \( a \in \mathbb{R} \).

The most important class of norms are the \( p \)-norms, i.e., \( \| x \|_p = \left( \sum_{i=1}^{n} |x_i|^p \right)^{1/p} \) for \( 1 \leq p \leq \infty \).

For today we are going to focus on the 2-norm; check that it satisfies the three conditions above.

The matrix 2-norm \( \| A \|_2 \) (or just \( \| A \| \)) is defined as \( \| A \|_2 = \sup_{\| x \|_2 = 1} \| Ax \|_2 = \sup_{\| x \| = 1} \| Ax \|_2 \).

It measures the maximum factor \( A \) can "stretch" a vector \( x \). Here are some basic facts about matrix norms that we will use (leaving the proofs as exercises).

1. \( \| A + B \| \leq \| A \| + \| B \| \)
2. \( \| AB \| \leq \| A \| \| B \| \)
3. \( \| A \| = \| A_i \| \) for all \( i \), where \( A_i \) is the \( i \)-th column of \( A \).
4. \( \| A \| = \| A^T \| \) for all \( i \), where \( A^T \) is the \( i \)-th row of \( A \).
5. \( \| QA \| = \| A \| \) for any orthogonal matrix \( Q \). A matrix is orthogonal if its columns (and hence rows) are orthogonal, i.e., \( QQ^T = Q^TQ = I \). It is equivalent to saying that \( \| Qx \| = \| x \| \) for all \( x \in \mathbb{R}^n \).

Actually, it is also equivalent to saying that \( \langle Qx, Qy \rangle = \langle x, y \rangle \) for all \( x, y \in \mathbb{R}^n \). So, an orthogonal transformation preserves lengths and angles. It is called a rigid motion.

Singular Value Decomposition (see [3] and [4])

The singular value decomposition says that any linear transformation \( A : \mathbb{R}^m \rightarrow \mathbb{R}^n \) can be described as follows:

1. Identify an orthonormal basis \( v_1, v_2, \ldots, v_n \) of \( \mathbb{R}^n \).
2. Map \( v_1, v_2, \ldots, v_n \) to the standard basis \( e_1, e_2, \ldots, e_n \). Scale each vector \( e_i \) by a nonnegative factor \( \sigma_i \).

If \( m \leq n \), at most \( n \) of them will be nonzero.

3. Map the standard basis to an orthonormal basis \( u_1, u_2, \ldots, u_m \) of \( \mathbb{R}^m \).
Geometrically, \( A \rightarrow D \).

We assume \( x_1, x_2, \ldots, x_n \) and call them the singular values of \( A \).

We call \( u_1, \ldots, u_m \) the right singular vectors and \( v_1, \ldots, v_n \) the left singular vectors.

Algebraically, let \( V \) be the \( n \times n \) matrix where the \( i \)-th column is \( v_i \), and let \( U \) be the \( m \times m \) matrix where the \( i \)-th column is \( u_i \), and let \( \Sigma \) be the \( m \times n \) diagonal matrix where the \( i \)-th entry is \( \sigma_i \).

Then \( AV = UV \). Since \( V \) is orthogonal, we have \( A = UV \Sigma V^T \) (that's why \( U \) left, \( V \) right).

**Theorem.** Every matrix \( A \in \mathbb{R}^{m \times n} \) has a singular value decomposition. Furthermore, all the singular values are uniquely determined.

**Proof.** First we prove the existence of bases and scalars.

Let the rank of \( A \) be \( r \). Consider \( A^T A \).

Note that \( A^T A \) is symmetric, positive semidefinite and of rank \( r \).

Therefore, there is an orthonormal basis of eigenvectors \( \{v_1, v_2, \ldots, v_n\} \) of \( A^T A \) with eigenvalues \( \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r > 0 \) and \( \lambda_i = 0 \) for \( i > r \).

Set \( \sigma_i = \sqrt{\lambda_i} \) and \( u_i = \frac{1}{\sigma_i} Av_i \). Then it is easy to see that \( AV = UV \).

We check that \( \{u_1, u_2, \ldots, u_m\} \) form an orthonormal subset.

We have \( \langle u_i, u_j \rangle = \langle \frac{1}{\sigma_i} Av_i, \frac{1}{\sigma_j} Av_j \rangle = \frac{1}{\sigma_i \sigma_j} \langle Av_i, Av_j \rangle = \frac{1}{\sigma_i \sigma_j} \lambda_i v_i^T v_j = \frac{\sigma^2}{\sigma_i \sigma_j} \langle v_i, v_j \rangle \).

Since \( \{v_1, v_2, \ldots, v_n\} \) are orthonormal, it follows that \( \{u_1, u_2, \ldots, u_m\} \) are orthonormal.

To show that the singular values are uniquely determined, suppose that we have a SVD,

we show that each \( v_i \) is an eigenvector of \( A^T A \) with eigenvalue \( \sigma_i^2 \) if \( i \leq r \) and 0 if \( i > r \).

We have \( A^T A v_i = \sigma_i^2 Av_i = \sigma_i^2 \sum_{j=1}^{n} \langle Av_j, v_i \rangle v_j = \sigma_i^2 \sum_{j=1}^{n} (u_i^T A v_j) v_j = \sigma_i^2 v_j \).

Since the eigenvalues are uniquely determined, the singular values are thus uniquely determined.

Here are some remarks comparing singular value decomposition and eigenvalue decomposition.

- SVD was two bases, eigenvalue decomposition just was one.
- SVD was orthonormal bases, eigenvalue decomposition in general (except symmetric) does not
have an orthonormal basis.
- every matrix (even rectangular) has a SVD, but not every (square) matrix has an eigenvalue decomposition.

An important property of SVD is that it gives the best low rank approximation of a matrix.

**Theorem** Let \( A_k = \sum_{i=1}^{k} \lambda_i u_i v_i^T \). Then \( \|A - A_k\|_2 = \|A - B\|_2 \) for any rank \( k \) matrix \( B \).

**Proof** The proof uses similar ideas as in the proof of the Courant-Fischer theorem in week 2.

It is not difficult to check that \( \|A - A_k\|_2 = \|B\|_2 \), attained by \( u_k, v_k \).

Suppose \( \|A - B\|_2 < \|B\|_2 \) for a rank \( k \) matrix \( B \).

Let \( W \) be the \( n-k \) dimensional nullspace of \( B \), i.e. \( Bw = 0 \) for \( w \in W \).

Then \( \|A_w\|_2 = \|(A-B)w\|_2 \leq \|A-B\|_2 \|w\|_2 < \|B\|_2 \|w\|_2 \).

Thus \( W \) is an \( n-k \) dimensional subspace such that \( \|A_w\|_2 < \|B\|_2 \|w\|_2 \) for \( w \in W \).

On the other hand \( V' = \text{span}(v_1, \ldots, v_k) \) is a \( k \) dimensional subspace such that \( \|A_v\|_2 \geq \|B\|_2 \|v\|_2 \) for \( v \in V' \).

Since the sum of the dimensions are larger than \( n \), there exists a vector in their intersection, a contradiction.

This result has many applications in compression and data analysis. We will use it later.

---

**Eigenvalues of Random Graphs** [5]

To solve the semi-random graph problems, we first need to understand the eigenvalues of random graphs.

Consider a random graph where each edge \( ij \) for \( i \neq j \) appears with probability \( 1/2 \).

What do you expect its eigenvalue distribution to be?

Let's consider the "expectation" matrix \( M \) where each entry is \( 1/2 \) except the diagonal entries are 0. Then \( M = \frac{1}{2} (I+J) \), where \( J \) is the all-one matrix. In week 1, we have seen that \( M \) has one eigenvalue of \( (n-1)/2 \) and \( n-1 \) eigenvalues of \(-1/2\).

We expect that the adjacency matrix \( A \) of a random graph should have a similar spectrum as that of \( M \) with high probability.

**Theorem** \( \|A-M\|_2 \leq \frac{5\sqrt{n}}{3} \) with high probability.
Theorem \( \|A-M\| \leq 5\sqrt{n}/3 \) with high probability.

Since \( A-M \) is symmetric, there are \( n \) orthonormal eigenvectors \( v_1, \ldots, v_n \) of eigenvalues \( \lambda_1, \ldots, \lambda_n \).

We have \( \|A-M\| = \max \frac{\| (A-M)v \|}{\|v\|} = \lambda_1 \).

On the other hand, write \( x = c_1v_1 + \ldots + c_nv_n \), then \( \frac{\| (A-M)x \|}{\|x\|} \leq \frac{\| (A-M)c_1v_1 \|}{\|c_1v_1\|} \leq \lambda_1 \). So \( \|A-M\| \leq \lambda_1 \).

Recall that \( \lambda_1 = \max_x \frac{x^T(A-M)x}{x^Tx} \).

So, \( \|A-M\| \leq 5\sqrt{n}/3 \) implies that the Rayleigh quotient of any vector changes by at most \( 5\sqrt{n}/3 \).

Let \( \alpha_1 \geq \beta_1 \geq \ldots \geq \beta_n \) be the eigenvalues of \( A \) and \( \beta_1 \geq \beta_2 \geq \ldots \geq \beta_n \) be the eigenvalues of \( M \).

Therefore, \( \alpha_1 \) is within \( 5\sqrt{n}/3 \) of \( \beta_1 \). Otherwise, say \( \alpha_1 > \beta_1 + 5\sqrt{n}/3 \), then the first \( \beta \) eigenvectors of \( A \) form an \( \beta \)-dimensional subspace in \( M \) with Rayleigh quotient \( > \beta \), a contradiction (e.g., to the Courant-Fischer theorem).

So, if the theorem is true, then with high probability a random graph has one eigenvalue within \( 5\sqrt{n}/3 \) of \( (n-1)/2 \) and \( n-1 \) eigenvalues within \( 5\sqrt{n}/3 \) of \( -1/2 \).

Consider \( R = A-M \). Each non-diagonal entry is an independent random variable with probability \( 1/2 \) to be \( 1/2 \), and probability to be \( -1/2 \).

Our goal is to bound \( \frac{x^TRx}{x^Tx} \).

We want to show that for a random \( R \), \( \frac{x^TRx}{x^Tx} \) is small for all \( x \).

The proof is quite interesting.

First it considers \( x^TRx \) for a fixed unit vector \( y \).

Claim 1: For every unit vector \( x \), \( \Pr_R \{ x^TRx \geq t \} \leq 2e^{-ct^2/2} \).

Proof: Note that \( x^TRx = \sum_{i,j} x_iR_{ij}x_j \). And \( E_R[x^TRx] = \sum_{i,j} 2x_ix_j E[R_{ij}] = 0 \).

This is a sum of independent random variables, each has small contribution to the sum. So we expect that the sum is "close" to the expected value with high probability.

We can use the Hoeffding’s inequality to prove it formally.

(Hoeffding’s inequality) Let \( X_i \) be independent random variable with \( a_i \leq X_i \leq b_i \). Let \( \mu = \mathbb{E}[X_i] \).

Then \( \Pr \{ \sum X_i \geq \mu + t \} \leq \exp \left( -\frac{t^2}{\sum (b_i-a_i)^2} \right) \).
In our situation, set \( X_{ij} = 2r_{ij}x_{ij} \), then \( a_{ij} = -x_{ij} \) and \( b_{ij} = x_{ij} \).

So, \( \sum_{i,j} (b_{ij} - a_{ij})^2 = 4 \sum_{i,j} x_{ij}^2 x_{ij}^2 = 2 \sum_{i,j} x_{ij}^4 \leq 2 \left( \sum_{i,j} x_{ij}^2 \right) \left( \sum_{i,j} x_{ij}^2 \right) = 2 \), where the inequality is Cauchy-Schwarz.

Note \( \sum_{i,j} x_{ij} = x^T R x \) and \( \mu = 0 \).

Thus, \( \Pr( |x^T R x| > t) \leq \Pr( |x^T R x > t|) + \Pr( |x^T R x < -t|) \), since \( R \) and \( -R \) have the same distribution.

\[
= \Pr( x^T R x > t) + \Pr( x^T R x < -t) = 2 \Pr( x^T R x > t) \leq 2e^{-t}.
\]

The next step is to show that vectors close to \( v \) also have a large Rayleigh quotient.

**Claim 2** Let \( v \) be a unit eigenvector of \( R \) whose eigenvalue has absolute value \( |\lambda| \).

If \( x \) is another unit vector such that \( \sqrt{x^T x} \geq 1/2 \), then \( x^T R x \geq 1/2 |\lambda| \).

**Proof** Assume \( \lambda \geq 1/2 \) and \( v = v \).

Write \( x = \frac{1}{\sqrt{x^T v}} v + \frac{1}{\sqrt{x^T v}} \xi \).

Since \( \sqrt{x^T x} \geq 1/2 \), we have \( c_1 \geq 1/2 \). Also, we have \( 1 = x^T x = \frac{1}{2} c_1 \).

Hence, \( x^T R x = \frac{1}{2} c_1 \lambda c_1 + \frac{1}{2} c_1 \xi^T \lambda \xi = c_1 \lambda + \frac{1}{2} c_1 (\lambda - \xi^T \lambda) \geq c_1 (\lambda^2 - \frac{1}{4} c_1^2) = \lambda (2c_1 - 1) \geq \lambda / 2 \).

We now compute the probability that a random vector is close to \( v \).

**Claim 3** Let \( v \) be a unit vector, and \( x \) be a random unit vector. Then,

\[
\Pr \left( \sqrt{x^T x} \geq \frac{1}{2} \right) \geq \frac{1}{\sqrt{\pi n} 2^{-n}}.
\]

**Proof** Let \( B^n \) denote the unit ball in \( \mathbb{R}^n \), and let \( C \) denote the cap on the surface of \( B^n \) containing all vectors such that \( \sqrt{x^T x} \geq \frac{1}{2} \).

The surface area of \( B^n \) is \( \frac{n \pi^{n/2}}{\Gamma \left( \frac{n}{2} + 1 \right)} \),

where \( \Gamma(n) = (n-1)! \) for positive integer \( n \) and \( \Gamma(x) \) is increasing for \( x > 1 \).

Now, consider the \( (n-1) \)-dimensional sphere whose boundary is the boundary of the cap \( C \).

As the cap lies above this sphere, the \( (n-1) \)-dimensional volume of this sphere is a lower bound on the surface area of the cap.

The radius of this sphere is \( r = \sin \left( \frac{1}{2} \right) = \frac{1}{2} \).

The volume of a \( (n-1) \)-dimensional sphere is \( \frac{n \pi^{(n-1)/2}}{\Gamma \left( \frac{n-1}{2} + 1 \right)} = \frac{\pi^{(n-1)/2}}{2^{n-1} \Gamma \left( \frac{n}{2} + 1 \right)} \).

So, the probability is at least \( \frac{\pi^{n/2}}{2^{n-1} \Gamma \left( \frac{n}{2} + 1 \right)} \approx \frac{1}{\sqrt{\pi n} 2^{-n}} \).

Now we are ready to prove the theorem.
Think of an infinite matrix where the rows are indexed by the unit vectors of $\mathbb{R}^n$, the columns are indexed by the matrix $R$. We put a "tick" in the $(e,R)$-th entry if $|x^TRx| \geq t/2$.

We want to argue that when $t$ is large, most of the columns have no ticks at all.

Consider the event that $\|R\| > t$.

Claim 1 showed that a row has very few ticks, with probability $\leq 2e^{-t/4}$.

Claim 2 says that if $\|R\| > t$, then all vectors $x$ with $x^TRx \geq \sqrt{3}t$ will satisfy $x^TRx \geq t/2$.

Claim 3 says that each such column has many ticks, with probability $\geq \frac{1}{\sqrt{n}n2^n}$.

Therefore, $\frac{\Pr(\|R\| > t)}{\sqrt{n}n2^n} \leq \Pr(\text{ticks}) \leq 2e^{-t/4}$, hence $\Pr(\|R\| > t) \leq e^{-t/4} \sqrt{n}n2^n$.

So, when $e^{-t/4} \geq \sqrt{n}n2^n$, the probability becomes small.

This happens when $t > 2\sqrt{n}n \approx 5.3t/3$. $\blacksquare$

In general the eigenvalues of a random symmetric matrix satisfies the semicircle law; see wiki.
Semi-Random Graph Problems

Let $\Psi: \{1, \ldots, n\} \rightarrow \{1, \ldots, k\}$ be a partition of $n$ nodes into $k$ classes. Let $P$ be a $k \times k$ matrix where $P_{ij} \in [0, 1]$ for all $i, j$. Include edge $(u, v)$ with probability $P_{\Psi(u), \Psi(v)}$.

Let $S_i$ be the size of class $i$.

This model captures many problems:

Planted multipartition $P$ is $p$ everywhere, except the diagonal is $q$. $\Psi$ is the multisection.

Planted clique $P$ is $p$ everywhere, except $P_{ii}$ is $1$. $\Psi$ is the coloring.

Planted $k$-coloring $P$ is $p$ everywhere, except the diagonal is $0$. $\Psi(v)=i$ if $v$ is in the clique.

Now, from a graph $\hat{G}$ generated by a known $P$ but an unknown $\Psi$, our task is to produce a partition $\hat{\Psi}$ such that $\hat{\Psi}(u)=\hat{\Psi}(v)$ if and only if $\Psi(u)=\Psi(v)$.

We assume that we also know $S_i$ for $1 \leq i \leq k$.

Approach

Let $\hat{G}$ be the expectation matrix, i.e. $\hat{G}_{uv} = P_{\Psi(u), \Psi(v)}$.

Suppose we know $\hat{G}$. Then we can easily cluster the columns and find the hidden partition.

But of course we do not know $\hat{G}$, and the columns $\hat{G}_u$ could be very different from $G_u$.

The idea is to cluster the columns, by using some appropriate projection.

Suppose we are given $P_G$, the projection on the column space of $G$.

Then we know that $\|P_G(G_u - \hat{G}_u)\|$ is zero

\[ \|P_G(G_u - \hat{G}_u)\| \text{ is small, as we expect that } G_u \text{ has a large projection on } G_u. \]

Then, by the triangle inequality, we have $P_G(G_u)$ is very close to $G_u$, and so we can find the hidden partition by clustering based on $P_G(G_u)$.

Of course, we do not have $P_G$ either, but we know that $\|\hat{G} - G\|$ is small, which implies that $\|\hat{G}_u - G_u\|$ is small.

Our goal is to find a projection $P_x$ so that:

1. Data is preserved: $\|P_x(G_u) - G_u\|$ is small
2. Noise is removed: $\|P_x(G_u - \hat{G}_u)\|$ is small

If so, then $P_x(G_u)$ is equal to $G_u$, up to a small error. If $\|G_u - \hat{G}_u\|$ is much larger
than this error, then we can apply a simple greedy clustering algorithm to recover the hidden partition.

The second requirement is kind of easy, by using some concentration of expectation result.

The first requirement is the interesting part, where we use the results on random matrices and also the singular value decomposition.

---

**Projection**

Given a subspace $S$ of $\mathbb{R}^n$, $P_S(x)$ is the vector in $S$ that is closest to $x$.

Let $\{q_i\}$ be a set of orthonormal vectors that spans the subspace $S$.

Then $x = r + \sum \frac{\langle q_i, x \rangle}{\langle q_i, q_i \rangle} q_i$, where $r$ is a vector orthogonal to $S$.

So $x = r + \sum \frac{\langle q_i, x \rangle}{\langle q_i, q_i \rangle} q_i$ and $x = \sum \frac{\langle q_i, x \rangle}{\langle q_i, q_i \rangle} q_i$ is the projection to $S$.

Let $Q$ be the matrix whose $i$-th column is $q_i$. Then $P_S(x) = QQ^T(x)$.

**Good projector**

Let $A$ be an $m \times n$ matrix, and let $A = U \Sigma V^T$ be its singular value decomposition.

Let $U_k$ be the first $k$ columns of $U$.

Let $P_k$ be the projection that projects any vector into $U_k$.

Then $P_k(A) = U_kU_k^T U \Sigma V^T = \sum_{i=1}^{k} \theta_i u_i v_i^T$, the best rank $k$ approximation to $A$.

So $\|A - P_k(A)\|$ is small.

We will use $P_k$ as our projector.

---

**Algorithm**

1. Randomly divide $[1, ..., n]$ into two parts. Let this division split the columns of $G$ as $[A | B]$.
2. Let $H = [P_A(A) | P_B(B)]$.
3. While there are unpartitioned nodes
   - Choose an unpartitioned node $i$ arbitrarily.
   - For each $v$, set $\hat{y}(v) = i$ if $\|H_u - H_v\| = 2(r_i + r_v)$.
4. Return the partition $\hat{y}$.

**Analysis**

To show that the algorithm can correctly recover the hidden partition, it suffices to prove:
- \| P_\tilde{G} (A_u) - A_u \| \leq r_1 \quad \text{and} \quad \| P_\tilde{G} (A_u - \tilde{A}_u) \| \leq r_2 \\
- \| P_\tilde{G} (B_u - \tilde{B}_u) \| \leq r_1 \\
- \| G_w - \tilde{G}_w \| \leq 4(\epsilon + r_2) \quad \text{when} \quad \psi(w) = \psi(u)

Note that we divide the columns into two groups to avoid the dependency in analyzing \( P_\tilde{G}(G-G) \); now to analyze \( P_\tilde{G}(B-B) \) we can assume independence of the random variables.

**Proofs**

By doing a random division of the columns into \( A \) and \( B \), we expect that each class has close to half the nodes in \( A \) and in \( B \). To simplify the analysis, we assume (incorrectly) that exactly half the nodes in each class is in \( A \) (and in \( B \)).

Let \( s_m = \min \{ n, \frac{n}{2} \} \) to be the minimum size of a class. Think of \( s_m \) as quite big, e.g. \( s_m = \sqrt{n} \), and so the simplifying assumption is quite close to the truth.

To bound \( \| G - \hat{G} \| \), we use the following classical result by Furedi and Komlos.

**Theorem.** Let \( \hat{M} \) be a matrix generated by randomly rounding the entries of a matrix of probabilities \( M \) preserving symmetry. Let \( b \) be the largest deviation of an entry in \( \hat{M} \).

If \( b^2 \gg \log^2 n / n \), then \( \| \hat{M} - M \| \leq 4\sqrt{n} \) with probability at least \( 1 - 2 e^{-b^2 / 8} \).

Now we are ready to prove the main theorem.

**Theorem.** With probability at least \( 1 - 3\delta \), we have

\[
\| P_\tilde{G}(B_u) - B_u \| \leq 8\sqrt{n / s_m} \quad \text{and} \quad \| P_\tilde{G}(B_u - \tilde{B}_u) \| \leq \sqrt{2k \log(\kappa n / \delta)}
\]

**Proof.** \( \| P_\tilde{G}(B) - B \| = \| P_\tilde{G}(A) - A \| \) (because we assume each class has half the size in \( A, B \))

\[
= \| (P_\tilde{G} - I)A \|
\leq \| (P_\tilde{G} - I)A \| + \| (P_\tilde{G} - I)(A - \tilde{A}) \|
\leq \| P_\tilde{G}A - A \| + \| (P_\tilde{G} - I)(A - \tilde{A}) \|
\leq \| A - \tilde{A} \| + \| A - \tilde{A} \| \quad \text{(since \( P_\tilde{G}A \) is the best rank \( k \) approximation)}
\leq 2\| A - \tilde{A} \| = 8\sqrt{n}.
\]

Since there are at least \( s_m \) identical columns in \( \| (P_\tilde{G} - I)A \| \), each column \( u \) cannot have

\[
\| (P_\tilde{G} - I)A_u \| > 8\sqrt{n / s_m}, \quad \text{otherwise by setting} \quad x = (\frac{1}{s_m}, \frac{1}{s_m}, ..., \frac{1}{s_m}, 0, ..., 0), \quad \text{then}
\|

\[
\| (P_\tilde{G} - I)A_x \| > s_m \cdot 8\sqrt{n / s_m} \cdot \frac{1}{s_m} = 8\sqrt{n}, \quad \text{contradicting the above bound}.
\]

L05 Page 10
\[ \| (P_R - I) A x \| > 8 \varepsilon m \cdot 86 \sqrt{n/\varepsilon} \cdot \frac{1}{\sqrt{m}} = 86 \sqrt{n}, \text{ contradicting the above bound.} \]

So we have \[ \| P_R (B_u - \hat{B}_u) \| < 86 \sqrt{n/\varepsilon} \], proving the first part.

Next we bound \[ \| P_R (B_u - \hat{B}_u) \| \].

Let \( \{ q_i \} \) be the left singular vectors of \( \hat{A} \), and \( Q \) be the matrix whose \( i \)-th column is \( q_i \).

Then \[ \| P_R (B_u - \hat{B}_u) \|^2 = \| Q Q^T (B_u - \hat{B}_u) \|^2 \]
\[ = \sum_i < q_i, B_u - \hat{B}_u >^2 \]

Note that \[ < B_u - \hat{B}_u, q_i > \overset{\Delta}{=} \sum_{i} (B_u - \hat{B}_u) q_i v_i \]

Fix any \( q_i \), each term in the sum is an independent zero mean random variable.

Furthermore, \[ \sum_{i} | (B_u - \hat{B}_u) q_i v_i |^2 \overset{\Delta}{=} \sum_{i} | q_i v_i |^2 = 1 \]

We can apply Azuma's inequality to bound \[ | \langle (B_u - \hat{B}_u), q_i \rangle | \]. Then we can just use a union bound to bound \[ \| P_R (B_u - \hat{B}_u) \| \] for all \( i \).

**Azuma's inequality**

Let \( \{ X_i \} \) be independent zero mean variables such that \( |X_i| \leq c_i \) and \( \mathbb{E} X_i^2 \leq 1 \). Then, for any \( \lambda > 0 \), \( \Pr[ \sum_{i} X_i > \lambda ] \leq 2 e^{-\lambda^2/2} \).

By setting \( \lambda = \sqrt{2 \log(kn/\delta)} \), we have the probability is at most \( \frac{2 \delta}{kn} \).

This implies that \( \| P_R (B_u - \hat{B}_u) \|^2 \leq k \lambda^2 = 2 k \log(kn/\delta) \) is at most \( \frac{2 \delta}{kn} \).

This implies that \( \| P_R (B_u - \hat{B}_u) \|^2 \) for some \( i \) is at most \( 2 \delta \).

**Hidden Clique**

Let's apply the main result to the hidden clique problem, other problems are similar.

Suppose there is a hidden clique of size \( c \sqrt{n} \).

Then \( \| G_u - G_v \|_2 = \sqrt{c \sqrt{n} (\frac{1}{2})} = c' n^{\frac{1}{4}} \), where \( c' \) is a constant.

By the main result, \( \| P_R (B_u) - B_n \| \leq 86 \sqrt{n k/\varepsilon} \cdot 86 \sqrt{n/\varepsilon} \cdot \sqrt{c \sqrt{n}} = c'' n^{\frac{1}{4}} \), and

\[ \| P_R (B_u - \hat{B}_u) \| \leq O(\log n) \]

So, setting \( c' \) to be a large enough constant, we can recover a hidden clique of size \( c \sqrt{n} \).

In a random graph, we expect that the largest clique is of size \( \Theta(\log n) \).

It is a wide open problem whether there is an efficient algorithm to recover a hidden clique of size \( O(\sqrt{n}) \).

Recently the hidden clique problem is even used a hardness assumption in computing equilibrium and
References