Correlated Recovery.

Recall model $G \sim SBM(n, p, q)$

$\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_n) \in \mathbb{S}^n$.

$\Pr[\text{inj}] = \begin{cases} p & \text{if } \sigma_i = \sigma_j \\ q & \text{if } \sigma_i \neq \sigma_j \end{cases}$

Correlated Recovery: An estimator $\hat{\sigma} = \hat{\sigma}(G)$ achieves correlated recovery if the overlap is strictly better than random guess $\sigma$:

$\mathbb{E} \left[ \| \hat{\sigma} - \sigma \|_1 \right] = O(n)$.

$\iff \mathbb{E} \left[ \min_{\sigma' \in \mathbb{S}^n} \| \hat{\sigma} - \sigma' \|_1 \right] \leq \frac{(1 - n(1))n}{2}$
Impossibly of correlated Reg

**Theorem 1.** Correlated Reg is possible

\[ \iff TV(P_t, P_-) = \nu(1) \]

\[ P_t \equiv \text{Law}(G \mid \sigma_1 \sigma_2 = +1) \]

\[ P_- \equiv \text{Law}(G \mid \sigma_1 \sigma_2 = -1) \]

TV and prob of error:

Suppose \( X \sim \text{Rad}(\frac{1}{2}) \) and \( Y \) (correlated)

Then

\[
\min_{X(\cdot)} P_{\text{H}_0} X \neq X(Y) = \frac{1}{2} \left( 1 - TV(P_t, P_-) \right)
\]

\[ P_t \equiv \text{Law}(Y \mid X = +1) \]

\[ P_- \equiv \text{Law}(Y \mid X = -1) \]

Why is this true?

Let \( H_0 : X = -1 \) and observe \( Y \)
\[ H_1: \ X = +1 \text{ and observe } Y. \]

\[
\min \text{ sum of Type I + II errors } = 1 - TV(P_t, P_-)
\]

\[
P(X \neq \tau(Y)) = P(X=1) \cdot P(\tau(Y) \neq 1 \mid X=1)
+ P(X=-1) \cdot P(\tau(Y) \neq -1 \mid X=-1)
\]

\[
= \frac{1}{2} \left( P(\tau(Y) \neq 1 \mid X=1) + P(\tau(Y) \neq -1 \mid X=-1) \right)
\]

\[
= \frac{1}{2} \text{ sum of Type I + II errors.}
\]

**Proof of Thm1.**

"\( \leq \)" Suppose \( TV(P_t, P_-) = \nu(1) \).

Want to show correlated recovery is possible.

By symmetry,

\[
TV(\text{Law}(G \mid \sigma_{ij} = +), \text{Law}(G \mid \sigma_{ij} = -))
\]

\[
= TV(P_t, P_-) = \nu(1) \text{ for } i \neq j.
\]
\( \Rightarrow \) for all \( i \neq j \), \( \exists \hat{G}_{ij}(G) \) such that
\[
\Pr\left( \hat{G}_{ij} = 0 \middle| \hat{G}_{j} \right) \geq \frac{1}{2} + \delta \quad \text{for } \delta \in \mathbb{R}.
\]

Then we can construct an estimate \( \hat{G}(G) \) by
\[
\hat{G}_1 = +1, \quad \hat{G}_i = \hat{G}_{ij} \quad \text{for } i = 2, \ldots, n.
\]

Then expected # of correctly classified vertices
\[
\max \sum \Pr\left( \hat{G}_i = \hat{G}_i \middle| \hat{G}_j \right) = \sum \Pr\left( \hat{G}_{ij} = 0 \middle| \hat{G}_j \right) \geq \left( \frac{1}{2} + \delta \right) n.
\]

\( \Rightarrow \) Correlated Reg is possible.

\( \Rightarrow \) For a sake of contradiction, let's say
\[
TV\left( P_+ , P_- \right) = 0.1.
\]

\( \Rightarrow \)
\[
TV\left( \text{Law}(G \mid \hat{G}_j = +) , \text{Law}(G \mid \hat{G}_j = -) \right) = 0.1
\]
for all \( i \neq j \),
\[
\forall i \neq j, \quad \text{Pr}(\hat{r}_{ij} = 0; \delta_j) = \frac{1}{2} + o(1)
\]
This means given \( \delta = (\delta_1, \ldots, \delta_n) \),
we have
\[
\text{Pr}(\hat{G} = \hat{G}) = 0; \delta_j) = \frac{1}{2} + o(1)
\]
\[
2n^2 - \text{E}[\epsilon^2] = \text{E}[\hat{G}^T \hat{G} - \delta^T \delta] = \frac{1}{4} \sum_{i \neq j} \text{Pr}(\hat{r}_{ij} = 0; \delta_j) (\frac{1}{2} + o(1)) = 2n^2 + o(n^2)
\]
\[
\Rightarrow \text{E}[\epsilon^2] = o(n^2)
\]
\[
\Rightarrow \text{E}[\hat{G}^2] \leq \sqrt{\text{E}[\epsilon^2]} = o(n).
\]
\[
\Rightarrow \text{Correlated Reg is impossible.}
\]
Next, we show

**Thm2**: If \( \frac{(a-b)^2}{2(a+b)} < 1 \), then \( TV(P_+, P_-) = o(1) \),

where \( P_+ \equiv \text{Law}(G \mid \sigma_1 \sigma_2 = +1) \)

\( P_- \equiv \text{Law}(G \mid \sigma_1 \sigma_2 = -1) \)

which immediately implies that correlated recon is impossible by Thm1.

Before the pf of Thm2, we first introduce a variational characterization of TV.

**Lemma**: [Variational representation of total variation]:

\[
TV(P_+, P_-) = \frac{1}{2} \inf_{Q} \int \int (P_+ - P_-)^2 \sqrt{Q}
\]

**Pf.** By Cauchy-Schwarz,

\[
4TV(P_+P_-) = \left( \int |P_+ - P_-| \right)^2 = \left( \int \frac{|P_+ - P_-|}{\sqrt{Q}} \cdot \sqrt{Q} \right)^2
\]

\[
\leq \int \frac{(P_+ - P_-)^2}{Q} \cdot \int Q
\]

\[
= \int \frac{(P_+ - P_-)^2}{Q}
\]

with equality if \( Q = \frac{|P_+ - P_-|}{\int |P_+ - P_-|} \cdot 1 \).
Proof of Thm 2:

Apply Lemma 2 with \( \mathcal{Q} = \text{Law of } G(n, d/n) \).

Then it suffices to show

\[
\int \frac{(\hat{\theta} - \theta_0)^2}{\mathcal{Q}} = o(1).
\]

This is a second-moment calculation similar to what we did before for detection. The difference is that here there is no null model and we need to compute the exact asymptotics.

Write:

\[
\int \frac{(\hat{\theta} - \theta_0)^2}{\mathcal{Q}} = \int \frac{P^2}{\mathcal{Q}} + \int \frac{P_0^2}{\mathcal{Q}} - 2 \int \frac{P \cdot P_0}{\mathcal{Q}}.
\]

It remains to show

\[
\int \frac{P_z P_{\overline{z}}}{\mathcal{Q}} = o(1) \quad \text{for all } z, \overline{z} \in \mathbb{S}^1, \text{ where}
\]

\((2)\) is a constant independent of \( z \) and \( \overline{z} \).

Consider the iid labels, where \( \xi_i \overset{iid}{\sim} \text{Rad}(\mathbb{S}) \).

By a suitable argument,

\[
\int \frac{P_z P_{\overline{z}}}{\mathcal{Q}} = \int \mathbb{E}_0 \left[ P_0(G) \mid \xi \sim z \right] \cdot \mathbb{E}_\mathcal{Q} \left[ P_0(G) \mid \xi \sim \overline{z} \right].
\]
\[
\begin{align*}
\text{Fabini:} & = \mathbb{E}_{\mu \mid \sigma^2} \left[ \int \frac{P_0^{(a)} \, P_0^{(b)}}{Q} \left| \begin{array}{c}
\sigma_1 \sigma_2 = \tilde{z} \\
\tilde{\sigma}_1 \tilde{\sigma}_2 = \tilde{z}
\end{array} \right. \right] \\
& = \mathbb{E}_{\mu \mid \sigma^2} \left[ \left( 1 + \rho \sigma_1 \sigma_2 \right) \left| \begin{array}{c}
\sigma_1 \sigma_2 = \tilde{z} \\
\tilde{\sigma}_1 \tilde{\sigma}_2 = \tilde{z}
\end{array} \right. \right] \\
\text{where} & \rho = \frac{\int (p_0 - a)^2}{2(p_0^2 + a^2)}, \\
& = \mathbb{E}_{\mu \mid \sigma^2} \left[ e^{\sum_{i \neq j} \rho \sigma_i \sigma_j (1 + \rho (\sigma_i \sigma_j))^2} \left| \begin{array}{c}
\sigma_1 \sigma_2 = \tilde{z} \\
\tilde{\sigma}_1 \tilde{\sigma}_2 = \tilde{z}
\end{array} \right. \right] \\
& = \mathbb{E}_{\mu \mid \sigma^2} \left[ e^{\frac{\sum_{i \neq j} \rho \sigma_i \sigma_j (1 + \rho (\sigma_i \sigma_j))^2}{2} + O(n^3)} \left| \begin{array}{c}
\sigma_1 \sigma_2 = \tilde{z} \\
\tilde{\sigma}_1 \tilde{\sigma}_2 = \tilde{z}
\end{array} \right. \right] \\
& = e^{-\frac{\rho}{2} + \frac{\rho^2}{4} + O(n^3)} \mathbb{E}_{\mu \mid \sigma^2} \left[ e^{\frac{\rho}{2} \sigma_1 \sigma_2} \left| \begin{array}{c}
\sigma_1 \sigma_2 = \tilde{z} \\
\tilde{\sigma}_1 \tilde{\sigma}_2 = \tilde{z}
\end{array} \right. \right]
\end{align*}
\]

Now, recall

\[
\rho = \frac{\int (p_0 - a)^2}{2(p_0^2 + a^2)} = \frac{1}{n} + \frac{(a-b)^2}{4n^2} + O\left(\frac{1}{n^3}\right), \quad \text{where}
\]

\[
\tilde{z} = \frac{(a-b)^2}{2(a+b)}.
\]
\[
\frac{v}{4} = e^{-\frac{v}{2} - \frac{v^2}{4} + o(1)} \left( e^{\frac{v^2}{2N} \left( \frac{\sigma_1^2}{\sigma_2^2} \right)} \right) \]

This step follows from CLT as before.

\[
\frac{1}{\sqrt{n}} \left< 0, \sigma \right> = \frac{1}{\sqrt{n}} \sum_{j=3}^{n} \sigma_j \sigma_j + \frac{1}{\sqrt{n}} (o_1 \sigma_1 + o_2 \sigma_2)
\]

negligible

\[
\rightarrow N(0,1)
\]

Remark: For exact bisector, the see statement holds true, but one should be more careful with conditioning.

\[
\text{If } v < 1, \Rightarrow \text{Correlated Reg } is \text{ possible}
\]

In fact, one can directly prove this (by split splitting)

imposible of detect, \Rightarrow \text{imposible of correlated Reg.}
Correlated Recovery by spectral methods

Spiked Wigner model:

\[ W = \sqrt{\frac{\lambda}{n}} \sigma \sigma^T + Z \]

where \( Z \) is symmetric with \( Z_{ij} \overset{iid}{\sim} \mathcal{N}(0,2) \) and \( Z_{ij} = Z_{ji} \overset{iid}{\sim} \mathcal{N}(0,1) \).

Recall \( |Z| = O(n^{1/2}) \). So it is natural to study the limit of eigenvalues of \( \frac{1}{\sqrt{n}} Z \).

The empirical eigenvalue distribution of \( \frac{1}{\sqrt{n}} Z \) converges to the semi-circular law \( \mathcal{N}(x) \) supported on \([-2, 2]\):

\[ \frac{1}{n} \sum_{i=1}^{n} \delta_{\lambda_i(Z)} \quad \text{a.s.} \quad \mathcal{N}(x) = \frac{1}{2\pi} \sqrt{4-x^2} \quad \text{for} \quad x \in [-2, 2] \]

Figure 1.9: Panel (a): Histogram of the eigenvalues of a 1000 x 1000 symmetric matrix with independent \( \mathcal{N}(0,1) \) entries. Refer to Benedek Valko’s course on random matrices [http://www.math.wisc.edu/~valko/courses/833/833.html]. Panel (b): Semi-circular law distribution. Image by Alan Edelman, MIT open courseware 18.996 / 16.399 Random Matrix Theory and Its Applications.
**Proof.** Beyond the scope. One approach is via moment method.

In particular

\[
\mathbb{E}\left[\frac{1}{n} \text{Tr} \left( \frac{1}{n} Z \right)^k \right] \to \int x^k \text{d}u(x)
\]

see section 2 of Terence Tao's random matrix book.

**Remark.** The semi-circle law holds beyond the Gaussian orthogonal ensemble, as long as

- \( Z_{ij} = Z_{ji} \)
- \( Z_{ij}'s \) are iid for \( i < j \) and \( Z_{ii}'s \) are iid for \( i = j \)

\[ \mathbb{E}[Z_{ij}] = 0, \quad \text{Var}(Z_{ij}) = 1, \]

- All moments of \( Z_{ij} \) exist and independent of \( n \).

Now back to spiked wigner model.

Let's renormalize \( W \) as

\[
\frac{W}{\sqrt{n}} = \frac{\sqrt{\lambda}}{n} \sigma \sigma^T + \frac{Z}{\sqrt{n}}.
\]

Question: what does the empirical eigenvalue of \( W/\sqrt{n} \) converge to?

Answer: BBP phase transition.
1. $\lambda_i \left( \frac{W}{n} \right) \xrightarrow{a.s.} \sqrt{\lambda_i} + \sqrt{n} \quad \text{if } \lambda > 1$

2. Let $\hat{u}$ denote the eigenvector of $\frac{W}{n}$ corresponding to $\lambda_i \left( \frac{W}{n} \right)$

   Let $u = \frac{1}{\sqrt{n}} \sigma$

   Then $| \langle \hat{u}, u \rangle | \xrightarrow{a.s.} \sqrt{1 - \frac{1}{\lambda_i}} \quad \text{if } \lambda > 1$

This suggests the fully specfied method achieves correlated reg of $\sigma$: $\sigma = \sqrt{n} \hat{u}$ if $\lambda > 1$:

$$\mathbb{E} [k_{\hat{u}, u^2}] = n \mathbb{E} [\langle \hat{u}, u \rangle] = n o(n) \quad \text{if } \lambda > 1.$$
SBM \((n, p, q)\) model: \(p = \frac{a}{n}\) and \(q = \frac{b}{n}\).

Recall that \(A\) denotes the adjacency matrix of graph \(G\). Mimicking the spiked Wigner model:

\[
A = \mathbb{E}[A] + (A - \mathbb{E}[A])
\]

where \(\mathbb{E}[A] = \begin{pmatrix} p & q \\ q & p \end{pmatrix}\) except for zero diagonal.

\[
\begin{align*}
= & \frac{p+q}{2} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \frac{p-q}{2} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \\
\triangleq & \mathbb{E}^T
\end{align*}
\]

So consider:

\[
A - \frac{p+q}{2} \mathbb{J} = \frac{p-q}{2} \mathbb{E}^T + (A - \mathbb{E}[A]).
\]

\[
= \frac{a-b}{2n} \mathbb{E}^T + (A - \mathbb{E}[A]).
\]

\[
\Rightarrow A - \frac{p+q}{2} \mathbb{J} = \frac{a-b}{\sqrt{2(ab)}} \frac{\mathbb{E}^T}{n} + \left( \frac{A - \mathbb{E}[A]}{\sqrt{2}} \right)
\]

\[
\begin{align*}
\triangleq & \sqrt{Z} \\
\approx & \frac{1}{\sqrt{n}} Z.
\end{align*}
\]

Note that:

\[
\mathbb{E}[Z_{ij}] = 0 \quad \text{if } i \neq j
\]

\[
\mathbb{E}[Z_{ij}^2] = \begin{cases} \frac{a}{n} \left(1 - \frac{a}{n}\right) & \text{if } i = j \\ \frac{a+b}{2n} & \text{if } i \neq j \end{cases}
\]
\[ \sum_j \text{Var}(Z_{ij}) = n \left( \frac{a^2 + b^2}{2} \right) \left( \frac{\text{atb}}{\text{atb}} \right) \sim n. \]

- If we pretend that \( Z_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, 1) \) and use the spiked Wigner result, we get the sharp threshold \( \tau \rightarrow 1 \), which is exactly the correlated Decoy threshold we want to show.

- However, this Gaussian analogy does not work. In particular, consider

\[ \mathbb{E} \left[ Z_{ij}^4 \right] = \mathbb{P}(\mathbf{A}_{ij} = 1) \cdot \left( 1 - o(1) \right) \frac{(1 + o(1))}{\sqrt{\frac{\text{atb}}{2}}} \left( \frac{n}{\text{atb}} \right)^2 \left( \frac{n}{\text{atb}} \right)^4 \]

\[ = 0 \left( \frac{1}{n^2} \times n^2 \right) = O(n). \]

The fourth moment is not bounded independent of \( n \).

- Even worse, we can show that

\[ \lambda_i \left( \frac{A - \frac{\text{atb}}{2} J}{\sqrt{\frac{\text{atb}}{2}}} \right) \text{ is not bounded} \]

and its corresponding leading eigenvector \( U \) is not
Correlated with $\Omega$ at all.

- This is because of existence of high-degree vertices in sparse graphs.
  
  In particular, consider $G(n, \frac{d}{n})$ with constant $d$.

  It is known that

  \[ \lambda_i(A) = \|A\|_1 = \sqrt{d_{\text{max}}} \left(1+o(1)\right) \]

  where the max degree $d_{\text{max}} = \Theta\left(\frac{\log n}{\log \log n}\right)$.

  In fact, not only $\lambda_i(A)$ is unbounded,

  \[ \lambda_i(A) = \lambda_i(A) \left(1-o(1)\right) \text{ for unbounded many of } i. \]

- To see why $\|A\|_1 \geq \sqrt{d_{\text{max}}}$, consider $d_i = d_{\text{max}}$ for some node $i$. Then

  \[ \|A\|_1 \geq \frac{\|Ae_i\|_1}{\|e_i\|_1} = \sqrt{d_{\text{max}}}. \]

- Alternatively, let's consider
\[(A^k)_{i j} = \sum_{i_2, \ldots, i_{2k}} A_{i i_2} A_{i_2 i_3} \cdots A_{i_{2k} i}.
\]

The number of closed walks from \(i\) to \(j\) of length \(2k\).

\[
\text{\# of closed walks from } i \text{ to } j \text{ of length } 2k
\]

Thus, \(\| A^{2k} \| = \| A^k \|^2 \geq \| A^k e_i \|^2.
\]

\[
e_i^T A^k A^k e_i
\]

\[
e_i^T A^{2k} e_i
\]

\[
= (A^{2k})_{i i}
\]

\[
\geq d_{\text{max}}^k
\]

Thus, \(\| A \| \geq \sqrt{k} d_{\text{max}}\).

\[
\text{\# of backtracking walks that goes from } i \text{ to } 1 \text{ neighbor and immediately goes back to } i.
\]

- why \(d_{\text{max}} = \Theta \left( \frac{\log n}{\log \log n} \right) \) in \(G(n, \frac{d}{n})\) even when \(d\) is bounded?

Note that \(d_i \sim \text{Bin} \left( n-1, \frac{d}{n} \right) \rightarrow \text{pois}(d)\).

Pretend \(d_i\)'s are independent \(\text{pois}(d)\).

The maximum of \(n \) \(i\) \(\text{pois}(d)\) is given by the \(\frac{1}{n}\)th quantile, that is
\[
\frac{\epsilon^d dk}{K^j} \sim \frac{1}{n} \implies K \sim \frac{\log n}{(\log \log n)}.
\]

- In summary, the adjacency matrix of sparse graphs is plagued by high-degree vertices, and the leading eigenvectors is localized on these vertices and are not informative. (see HW for one simulation experiment)

**Solution:**

1. **Regularize:** remove high-degree vertices and apply spectral methods.

   However, this destroys some information and it's hard to show this achieves the correlated decy threshold.

2. **Resolve the pathological behavior due to backtracking:**

   ⇒ Belief propagation and non-backtracking matrices.

3. **Turn to SDP:** can resolve the high-degree issue, but difficult to achieve the sharp threshold

   (will study SDP later).