Computational vs. Statistical Tradeoffs

When are there fundamental gaps btwn

1. best estimator and
2. best estimator that can be computed in polynomial time?

Planted Clique [Jerrum]; [Kucera]

First consider an Erdos-Renyi random graph

$$G(n, 1/2)$$

edges included independently w/ this probability

How large is the largest clique?

$$\mathbb{E}[\#\text{ k-cliques}] = \binom{n}{k} 2^{-\binom{k}{2}}$$

$$\leq n^k 2^{-\binom{k}{2}} = 2^{k(\log n - \frac{k-1}{2})}$$

Hence if $k = (2+\delta)\log n$ then

$$\mathbb{E}[\#\text{ k-cliques}] \lessapprox 2^{-\delta \log^2 n} = n^{-\delta \log n}$$

Fact 1: whp $\omega(G) = (2 \pm o(1))\log n$

largest clique

If we plant a large clique in $G(n, 1/2)$ can we find it?

Fact 2: There is an $n^{O(\log n)}$ time algorithm to solve planted clique whenever $k \geq (2+\delta)\log n$. Conversely if $k \leq (2-\delta)\log n$, it's impossible

Proof [sketch] Brute-force search for a $(2+\delta)\log n$ sized clique and find all common neighbors

Otherwise, there are too many cliques. ∎

What about polynomial time algorithms?

Theorem [Alon, Krivelevich, Sudakov] There is a polynomial time algorithm that succeeds whp if $k \geq C\sqrt{n}$

Warm-up w/ weaker bound: A node u has degree

$$\deg(u) = \begin{cases} k-1 + \mathrm{Bin}(n-k, \frac{1}{2}) & \text{if } u \text{ in planted clique} \\ \mathrm{Bin}(n-1, \frac{1}{2}) & \text{else} \end{cases}$$

From Chernoff bounds, we have

① If $u \notin$ planted clique, whp

$$\deg(u) < \frac{n}{2} + \frac{c}{4}\sqrt{n \log n}$$

② else whp $\deg(u) > \frac{n-k}{2} - \frac{c}{4}\sqrt{n\log n} + k-1$

Thus if $k \geq C\sqrt{n \log n}$ then whp

highest degree nodes $\equiv$ planted clique

Let's get tighter bounds via random matrix theory

Recall if M is $n \times n$ and

$$M_{ij} = \begin{cases} \text{random } \pm 1 & \text{if } i < j \\ 0 & \text{if } i = j \\ M_{ji} & \text{else} \end{cases}$$

then whp $\|M\| \leq (2 + o(1))\sqrt{n}$

Now, given G, let's construct

$$A_{ij} = \begin{cases} +1 & \text{if } i = j \\ +1 & \text{if } i \neq j \text{ and } (i,j) \in E \\ -1 & \text{else} \end{cases}$$

Spectral Algorithm

- Construct A and let x be its top eigenvector
- Let T = top k coordinates of x in absolute value and set

$$H = \{u \mid u \text{ has at least } \tfrac{4}{5}k \text{ neighbors in } T\}$$

Main idea is to decompose

$$A = \begin{array}{c|c} & \begin{array}{cc} k & n-k \end{array} \\ \begin{array}{c} k \\ n-k \end{array} & \underbrace{\left[\begin{array}{c|c} J & 0 \\ \hline 0 & 0 \end{array}\right]}_{S} \end{array} + \underbrace{\left[\begin{array}{c|c} 0 & \pm 1s \\ \hline \pm 1s & \pm 1s \end{array}\right]}_{E}$$

all ones

From R.M.T. (need bounds for asymmetric matrices) we have $\|E\| \leq C\sqrt{n}$

Moreover $\|S\| = k$ since it is rank one and has top eigenvector

$$y = [\underbrace{\tfrac{1}{\sqrt{k}}, \tfrac{1}{\sqrt{k}}, \ldots \tfrac{1}{\sqrt{k}}}_{k}, \underbrace{0, \ldots 0}_{n-k}]$$

From Wedin's theorem, we have

$$\sin\theta(x,y) \leq \frac{2\|E\|}{k}$$

eigenvalue gap of S

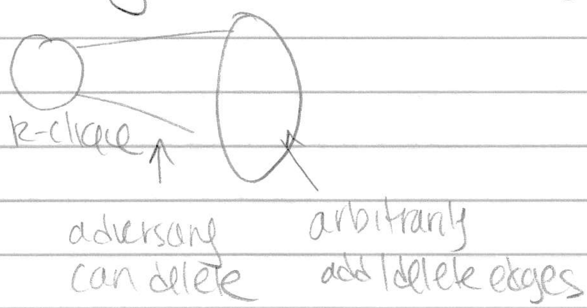
Thus we have $\langle x,y\rangle \geq 0.99$, rest of the analysis is book keeping

<u>Aside</u>: what if a monotone adversary can remove edges in the "random" portion of the graph?

[Feige, Krauthgamer] show Lovasz theta function still works

[Feige, Kilian]

Another interesting model [Steinhardt]



No longer possible to find <u>the</u> planted clique, but can we find <u>any</u> large clique?

<u>Thm</u> [Buhai, Kothari, Steurer] There is an $n^{O(1/\varepsilon)}$ time algorithm in FK model above that finds a clique of size k for $k \geq n^{\frac{1}{2}+\varepsilon}$

Actually outputs a list of size $\lesssim \frac{n}{k}$ that contains the planted clique

---

Are there computational vs. statistical tradeoffs for related problems?

Consider the $k>2$ community detection case

$$W = \begin{bmatrix} \frac{a}{n} & \frac{b}{n} & \cdots & \\ \frac{b}{n} & \ddots & & \frac{b}{n} \\ \vdots & & \frac{b}{n} & \frac{a}{n} \end{bmatrix}$$

Let $d = \dfrac{a+(k-1)b}{k}$ be the average degree and let $\lambda = \frac{a-b}{kd}$ = second eigenvalue of "transmission" matrix

It satisfies

$$-\frac{1}{k-1} \le \lambda \le 1$$

planted coloring, i.e. $a=0$

Conjecture [Decelle et al] There are computationally efficient algorithms for partial recovery iff

$$\lambda^2 d > 1 \quad \text{(Kesten-Stigum for higher } k\text{)}$$

Moreover for $k \ge 5$, there is a computationally hard but detectable regime

Consider coloring: K-S bound becomes

$$d > (k-1)^2$$

[Abbe, Sandon] gave matching algorithms

Thm [Achlioptas, Naor] The k-colorability threshold for E-R random graphs with average degree d grows as $2k\ln k$

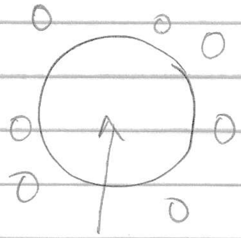
Thus we can solve the "distinguishing problem" whp

Note: There are computationally efficient distinguishers that beat random guessing for

SBM vs ER

but they do not work whp, e.g. cycle counts

the information theoretic threshold is called condensation b/c the posterior looks like



planted clustering

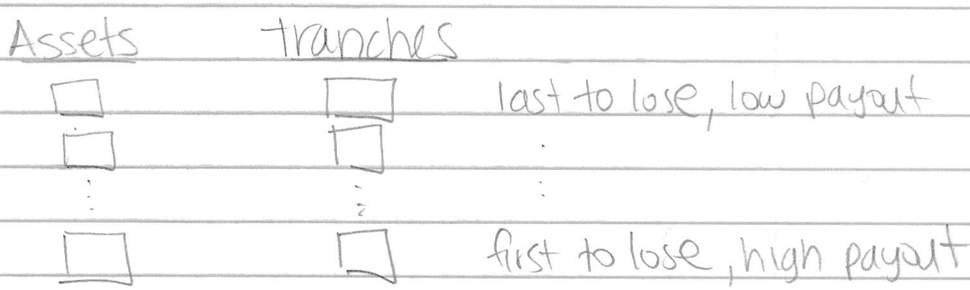
i.e. it's dominated by clusters correlated with planted one.

[Banks, Moore, Neeman, Netrapalli] showed tight bounds for condensation in the planted coloring case

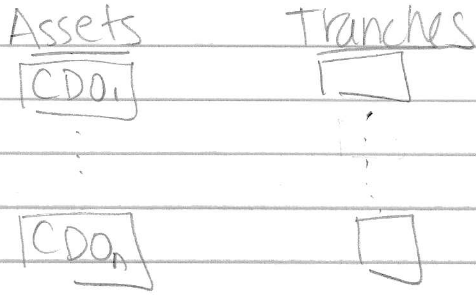
---

Some applications of these sorts of gaps

① Financial derivatives, e.g. CDOs



CDO^2:



Traditional view: CDOs help resolve information asymmetry b/c higher tranches less info-sensitive

[Arora, Barak, Brunnermeier, Ge]: With computational complexity, can amplify information asymmetry, if densest subgraph is hard

"Is there a dense subgraph — small set of assets — that make highest tranch more likely to fail?"

The party that creates the derivative has an advantage!

---

(2) Nash Equilibrium

Find a pair of strategies x, y so that no player can do ε-better

Thm [Lipton, Markakis, Mehta] There is a $n^{O(\frac{\log n}{\varepsilon^2})}$ time algorithm to find an ε-approximate Nash equilibrium

Thm [Hazan, Krauthgamer] If planted clique is hard, so is finding the best

(i.e. maximize social welfare) approximate
Nash equilibrium

[Rubinstein] later showed hardness for finding
any approximate Nash