Algorithms and Barriers for Random Hypergraph Partitioning

Will Perkins
Univ. of Birmingham
Average-case complexity

Take your favorite `hard' computational problem, eg. graph coloring, counting independent sets, 3-SAT, finding Nash equilibria, factoring, unique games etc.

• Can you find a concrete hard instance?

• Can you find a (natural) distribution of hard instances?

• Can you efficiently generate hard instances with known solutions?

• Harder than P vs. NP! Can we say anything rigorous at all?
Sample $m$ iid copies of a random combinatorial structure (e.g., edges or constraints) from an unknown distribution $D$.

- How large must $m$ be to (approximately) determine $D$?
- How large must $m$ be to distinguish $D$ from a reference distribution (null hypothesis)?
Statistical perspective

These are statistical questions, but what if we demand a **computationally efficient** procedure?

- How large must $m$ be to *efficiently* determine $D$?
- How large must $m$ be to *efficiently* distinguish $D$ from a reference distribution?
Statistical perspective

If the efficient procedure requires more samples, we say there is an **algorithmic gap**.

- This is a source of hard computational problems.
- Points at limits of statistical inference.
Outline

• A general model of planted hypergraph partitioning, with 3 examples.

• A general purpose algorithm.

• A geometric phenomenon that indicates a barrier to further improvement.
The Model

1. Partition a set of \( n \) vertices into two equal-sized sets. Denote the planted partition by \( \sigma \).
The Model

2. Add $m$ $k$-uniform hyperedges independently at random according to a planted distribution.
The Model

3. The planted distribution depends on a planting function

\[ Q : \{\pm 1\}^k \rightarrow [0, 1] \]

\[ \Pr[e] = \frac{Q(\sigma(e))}{\sum_{e' \in \binom{V}{k}} Q(\sigma(e'))} \]
The Model

4. **The problem**: find the planted partition $\sigma$ using as few edges as possible (and do so efficiently!).

- **Recovery**: find $\sigma$ exactly.
- **Detection**: find a partition correlated with $\sigma$.
- **Distinction**: distinguish a planted distribution from the uniformly random distribution on hyperedges.
1. **Planted k-SAT:** 

\[ Q(-1, \cdots - 1) = 0 \]

Choice over the other values of \( Q \). This can affect the difficulty of recovering the planted assignment.
2. **The Stochastic Block Model:**
Add interior edges with prob $\frac{a}{n}$ and crossing edges with prob $\frac{b}{n}$.
Examples

3. **Noisy k-XOR-SAT (parity):** $Q(\text{even})=a$, $Q(\text{odd})=b$

In fact, for $k=2$ this is the stochastic block model.
Related Problems

- Goldreich’s PRG
- Certifying sparse hypergraph quasirandomness
- Random k-SAT (or CSP) refutation
Recovery Thresholds

Information theoretically, planted partition can be recovered with $m = \tilde{O}(n)$ hyperedges/ clauses.

How many to recover/detect efficiently?
The SBM

Conjecture [Decelle, Krzakala, Moore, Zdeborova ’11]
• If \((a-b)^2 > 2(a+b)\), then the planted partition can be detected efficiently whp.
• If \((a-b)^2 \leq 2(a+b)\), then no algorithm can detect the partition whp.

Based on the analysis of belief propagation fixed points. Essentially since there are few short cycles, things *should* behave as on a tree.

Proved by Mossel, Neeman, Sly ’14 (impossibility) and Massoulie ’14 and Mossel, Neeman, Sly ’14 (detection)!

In particular, in the 2 part SBM there is no algorithmic gap.
**Impossibility**: couple the SBM to a broadcast model on Poisson Galton-Watson tree.
A Sharp Threshold

**Impossibility**: couple the SBM to a broadcast model on Poisson Galton-Watson tree.
A Sharp Threshold

**Impossibility**: couple the SBM to a broadcast model on Poisson Galton-Watson tree.
**Impossibility:** couple the SBM to a broadcast model on Poisson Galton-Watson tree.
A Sharp Threshold

**Impossibility:** couple the SBM to a broadcast model on Poisson Galton-Watson tree.

A Sharp Threshold

Detection: first attempt would be to use 2nd eigenvector of adjacency matrix and round.

This works for logarithmic average degree, but for sparser graphs, spectrum is obscured by noise from high-degree vertices.

Solution: consider the spectrum of the non-backtracking matrix. 2m x 2m matrix indexed by directed edges, with an edge between (i,j), (k,l) if j=k, i≠l
Some algorithms:
Flaxman (SODA ’03): $\tilde{O}(n)$ spectral algorithm for some distributions of planted 3-SAT.
Coja-Oghlan, Cooper, Frieze (SIAM Disc. Math ’10): $\tilde{O}(n^{3/2})$ algorithm for all planted 3-SAT distributions.
Bogdanov-Qiao (RANDOM ’09): $\tilde{O}(n)$ inversion of Goldreich’s PRG for pairwise independent predicates.
Feldman, P., Vempala (NIPS ‘15): $\tilde{O}(n^{r/2})$ algorithm via reduction to a bipartite stochastic block model, where $r$ is the distribution complexity of $Q$. 
Distribution Complexity

Let r be the smallest integer so that there is a non-empty $S \subseteq \{1, \ldots, k\}$, $|S| = r$ so that $\hat{Q}(S) \neq 0$. Distribution is (r-1)-wise independent, but not r-wise.

Note that $1 \leq r \leq k$ for any non-uniform planted distribution.

**Examples**: uniform planted k-SAT has $r=1$; SBM has $r=2$; noisy k-XOR-SAT has $r=k$. 
An Algorithm

1. Restrict each \( k \)-edge to an \( r \)-edge indicated by \( S \).

This induces the noisy \( r \)-XOR-SAT distribution.
An Algorithm

2. Split each r-edge into a singleton and an (r-1)-tuple. Then form an unbalanced bipartite incidence graph.
An Algorithm

3. Use the bipartite graph to partition the vertex set!
An Algorithm

3. Use the bipartite graph to partition the vertex set!
An Algorithm

3. Use the bipartite graph to partition the vertex set!

L
R
vertices

(r-1)-tuples

even
odd
An Algorithm

4. Paths of length 2 induce a SBM on the small side. Then use one of the optimal SBM detection algorithms.

This is optimal for the bipartite SBM, by coupling with a broadcast process on a two-type Galton-Watson tree.
An Algorithm

What about a spectral approach to the bipartite SBM?

For \( m < n^{r-1} \), \( \| M - \mathbb{E}M \| \gg \| \mathbb{E}M \| \)
An Algorithm

But now look at $M M^T$. No longer has independent entries, but close, and its eigenvectors are singular vectors of $M$.

Much of the noise in the spectrum comes from the diagonal. When we remove it, the spectral algorithm recovers the partition at near optimal threshold $m = n^{r/2} \log n$. 

An Algorithm

In fact there are 2 thresholds:

- $n^{r/2}$: impossible
- $m$: Diagonal Deletion works, singular vectors localized
- $n^{2r/3-1/3}$: SVD works, sing. vectors delocalized

This is a simple toy model of a localization/delocalization phase transition.
Need to consider restricted models of computation: Integrality gaps for convex programs, slow mixing results for Markov Chains etc.

Is \( m = n^{r/2} \) an algorithmic barrier?

\[ Q(1, 1) = Q(1, 1) = (1) \]

\[ Q : \{\pm 1\}^k \rightarrow [0, 1]^m = \tilde{O}(n) \]

\[ O(n) \quad O(n^{r/2}) \]
Statistical Algorithms

- Recall the statistical formulation: determine unknown distribution $D$ from $m$ samples.

- A *statistical algorithm* interacts with data indirectly, through expectations of arbitrary functions with respect to $D$.

- Introduced by Kearns in machine learning to capture noise-tolerant learning.

- Extended to problems over distributions by Feldman, Grigorescu, Reyzin, Vempala, Xiao, STOC ’13.
Statistical Algorithms

- Query a function $h : \mathcal{X} \to \{0, 1\}$, get $\hat{h} \in [\mathbb{E}_D h - \tau, \mathbb{E}_D h + \tau]$ where $\tau = 1/\sqrt{m}$

- Can be simulated using $O(m)$ samples (edges).

- The number of queried functions is a proxy for computation cost; interested in tradeoff between accuracy and query complexity.
Statistical Algorithms

• Can implement many known algorithmic approaches in the statistical framework: gradient descent, convex programming, spectral methods etc.

• Turns the computational problem into an information theoretic one.
Statistical Algorithms

**Theorem** (Feldman, P., Vempala, STOC ’15)
At least $n^{c \log n}$ queries to the statistical oracle with $m = n^{r/2}/\log n$ are required to distinguish a planted distribution with complexity $r$ from the uniformly random distribution.
Statistical Algorithms

Proof is based on a geometric phenomenon. Define:

\[ d_h(D_\sigma, D_\tau) = \frac{|\mathbb{E}_{D_\sigma} h - \mathbb{E}_{D_\tau} h|}{\|h\|} \]

\[ D_h^*(m) = \{\sigma : d_h(D_\sigma, U) > \frac{1}{\sqrt{m}}\} \]

\[ \beta(m) = \sup_{h} \frac{|D_h^*(m)|}{2^n} \]
Statistical Algorithms

Theorem

- For $m < n^{r/2}/\log n$, $\beta < n^{-c \log n}$
- For $m \geq Cn^{r/2}$, $\beta > n^{-c}$
Statistical Algorithms

A challenge:
Find a new algorithmic approach, that is not statistical and not Gaussian elimination for any of the related problems: partitioning, planted k-SAT, k-SAT refutation, certifying quasirandomness.

Some hope:
Feige, Kim, Ofek FOCS ’06: a non-deterministic polynomial time algorithm for 3-SAT refutation which succeeds with $m=n^{1.4}$. 
Open Questions

• Find an algorithm to beat the $n^{k/2}$ barrier for noisy k-XOR-SAT.

• Alternatively, come up with some prediction of a sharp threshold for detection. Not even the physicists have a guess!

• Find a planted version of your favorite combinatorial problem: can you design a distribution that produces hard instances?
Thank You!