A NEARLY-TIGHT SUM-OF-SQUARES LOWER BOUND FOR PLANTED CLIQUE

Boaz Barak

Sam Hopkins

Jon Kelner

Pravesh Kothari

Ankur Moitra

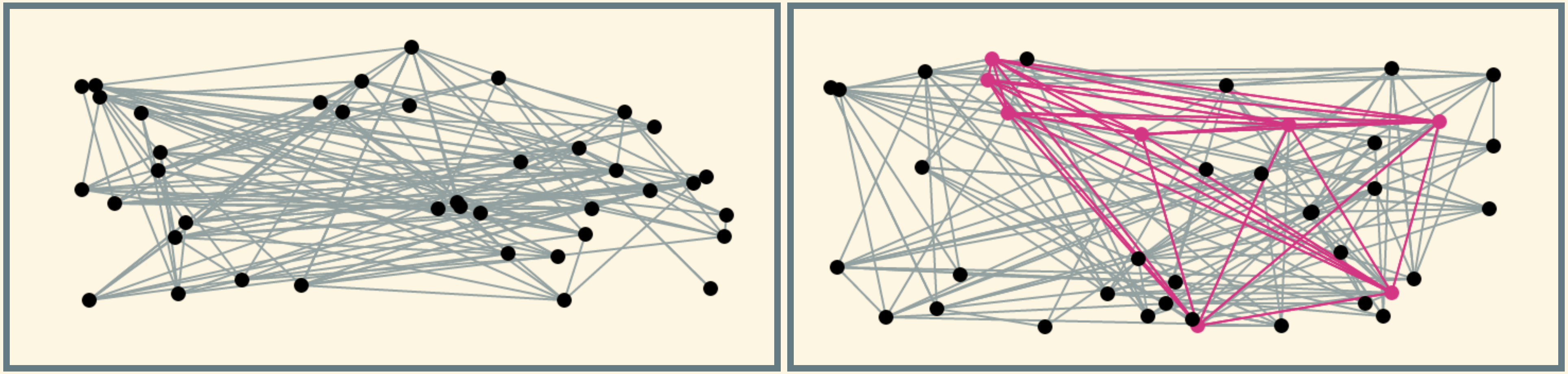
Aaron Potechin ← on the market!

PLANTED CLIQUE ("DISTINGUISHING")

Input: graph G

Goal: determine (with high probability) whether

$G \sim \mathbb{G}\left(n, \frac{1}{2}\right)$ (*"random distribution"*), or
$G \sim \mathbb{G}\left(n, \frac{1}{2}\right) + k\text{-clique}$ (*"planted distribution"*).



FOR WHAT k IS THIS POSSIBLE IN POLYNOMIAL TIME?

[J '92, K '95]

[J '92, K '95]

BASICS

$k = O(1)$: information-theoretically-impossible

BRUTE FORCE

$k >$ max clique in $\mathbb{G}\left(n, \frac{1}{2}\right) \approx 2 \log n$ [GM '75, M '76, BE '78]

requires $n^{O(\log n)}$ time ("quasipolynomial")

SPECTRAL (ADJACENCY MATRIX)

$k > \sqrt{n}$, polynomial time [AKS '98]

HYPOTHESIS: NO POLYNOMIAL-TIME ALGORITHM FOR $k = n^{0.49}$

"SPECTRAL IS BEST"

IF SPECTRAL IS BEST, HARDNESS RESULTS GALORE!

Sparse PCA [BR '13]

Compressed Sensing [KZ '14]

Property Testing [AAK+ '07]

Mathematical Finance [DBL '10]

Cryptography [JP '00, ABW '10]

Computational Biology [PS000, MSOI+

Best Nash Equilibrium [HK '11, ABC '13]

(and $P \neq NP$)

> Problems which have:
>
> a distribution on inputs
>
> $n^{O(\log n)}$-time algorithms

TO BEAT SPECTRAL SEEMS TO REQUIRE NEW ALGORITHMIC IDEAS

WHY BELIEVE SPECTRAL IS BEST?

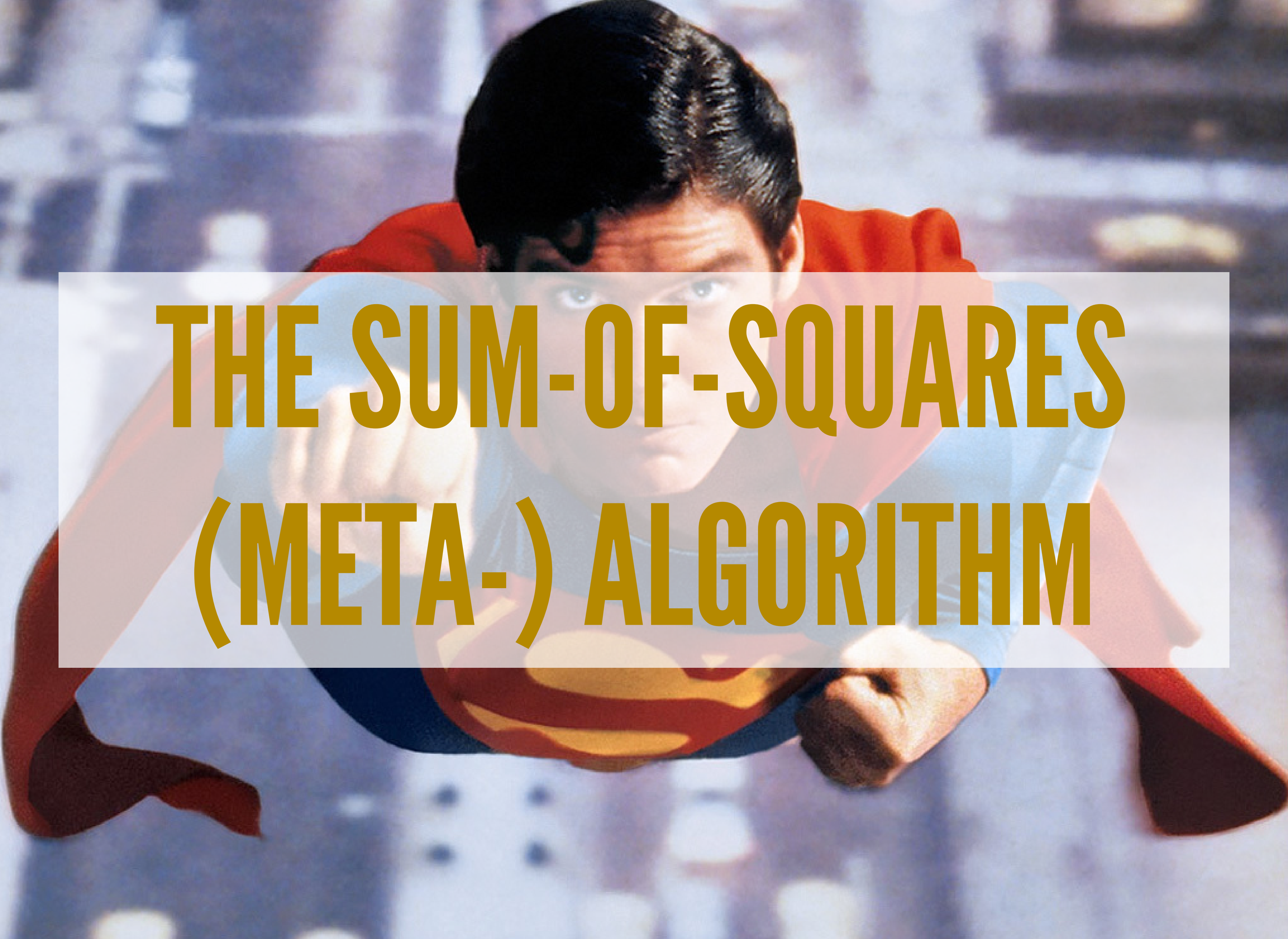
no algorithmic progress in 20 years? — bad science

reductions? (3SAT $\leq$ planted clique) — distribution on inputs

Rule out large classes of algorithms

> Markov-Chain Monte Carlo [Jerrum '93]
>
> Lovasz-Schrijver+ Convex Hierarchy [Feige-Krauthgamer '04]
>
> Statistical Algorithms [Feldman et al '12]
>
> **All do not beat spectral**

WHY BELIEVE SPECTRAL COULD BE BEATEN?


THE SUM-OF-SQUARES
(META-) ALGORITHM

Generalization of linear programming, basic semidefinite programming, spectral algorithms

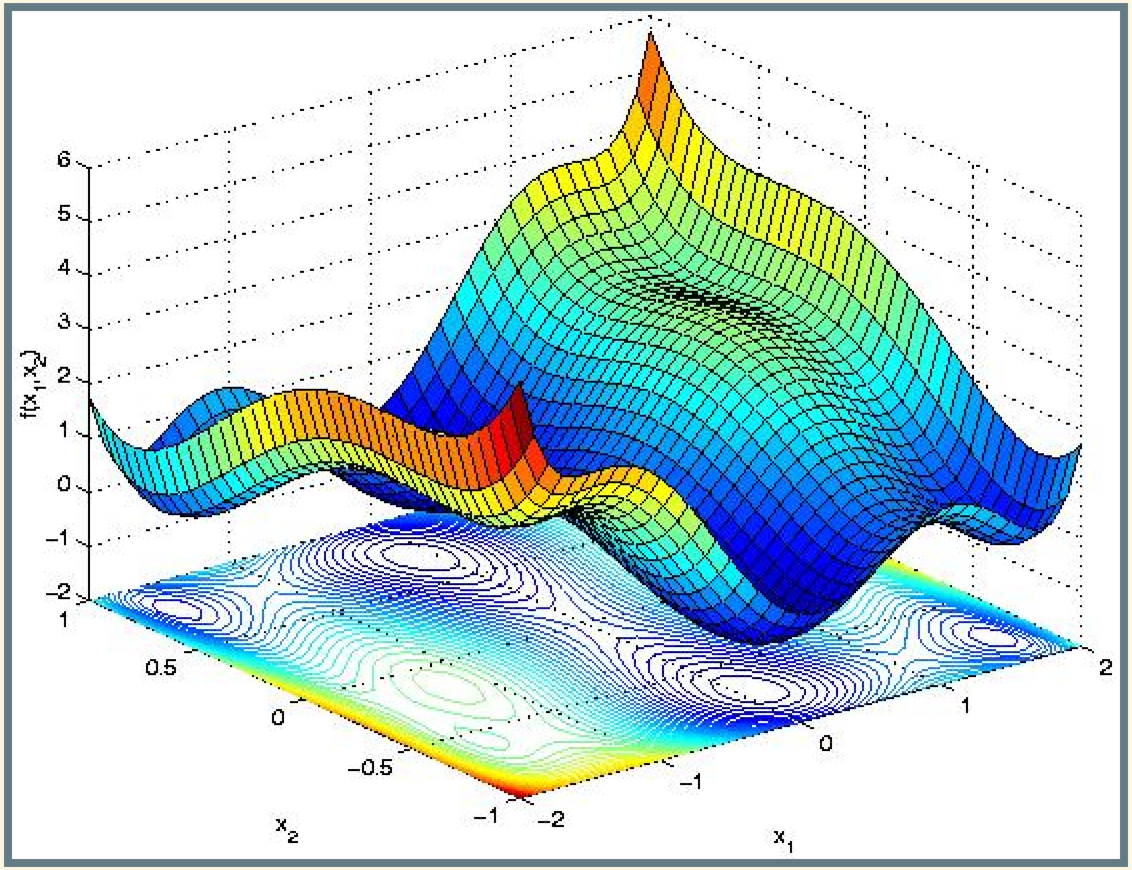
> Optimal among all (poly-sized) SDPs for constraint satisfaction [LRS '15]
>
> Solves all known hard instances of *unique games*, *max cut* in polynomial time [BBHKSZ '12, OZ '13, DMN '12]
>
> Best known algorithm for many *planted problems*, **beating corresponding spectral algorithms!**
> (dictionary learning, planted CSPs, tensor PCA, tensor decomposition, …) [BKS '15ab, AOW '15, RRS '16, **HSS** '15, **HSSS** '16, GM '15, MSS '16, …]

QUESTION: DOES SPECTRAL-IS-BEST WITHSTAND THE SUM-OF-SQUARES ALGORITHM?

Theorem (informal): The Sum-of-Squares hierarchy requires $n^{\Omega(\log n)}$ time to distinguish planted from random when $k = n^{0.49}$.

Spectral-is-best withstands the Sum-of-Squares algorithm.

WHAT IS THE SUM-OF-SQUARES ALGORITHM?



nasty optimization problem

WHAT IS THE SUM-OF-SQUARES ALGORITHM?

$$\max_{S \text{ a clique in G}} |S|$$

WHAT IS THE SUM-OF-SQUARES ALGORITHM?

A *hierarchy* of increasing-strength semidefinite programming (SDP) relaxations of an underlying (nonconvex) problem.

Generalizing linear programming, basic SDP, spectral methods.

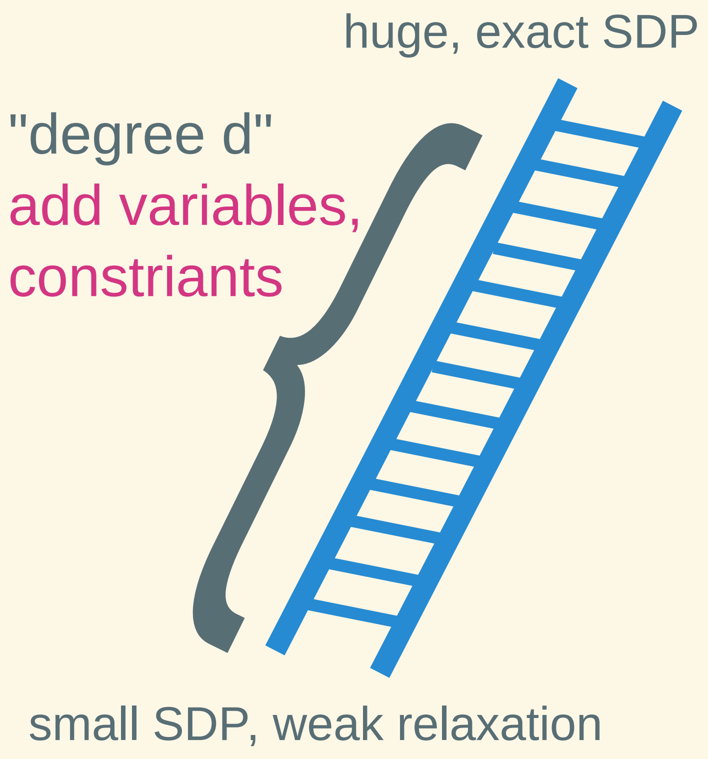
WHAT IS THE SUM-OF-SQUARES ALGORITHM?

$$d = n : \quad \underbrace{\left(\quad \right)}_{2^n \times 2^n} \succeq 0$$

$$\cdots$$

$$d = 2 : \quad \underbrace{\left(\quad \right)}_{n^2 \times n^2} \succeq 0$$

$$d = 1 : \quad \underbrace{\left(\quad \right)}_{n \times n} \succeq 0$$



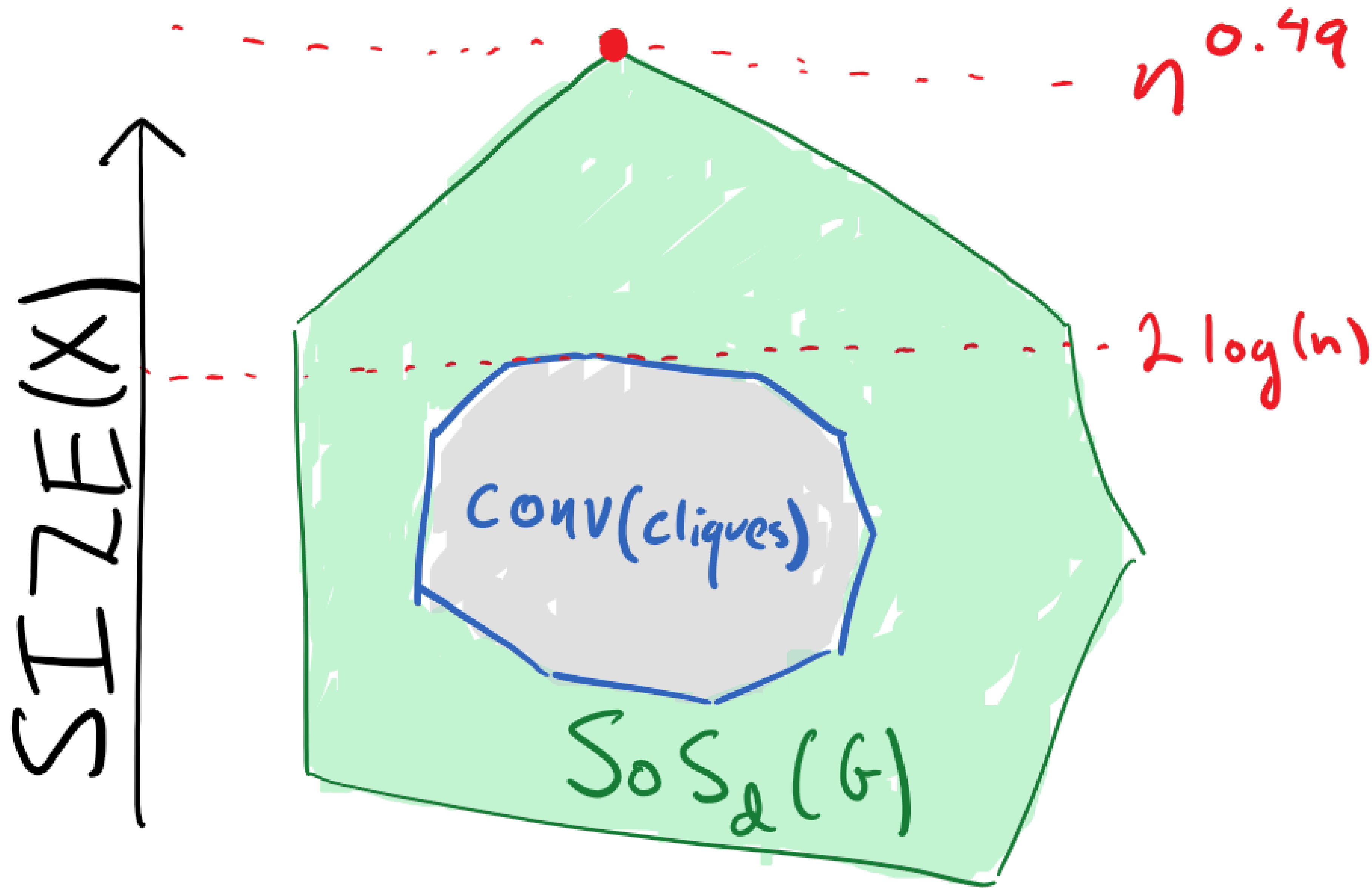
Regime of interest: $d < o(\log n)$.

Convex Relaxations for Planted Clique

If $\max_{X \in SoS_d(G)} SIZE(X) \geq n^{0.49}$, output PLANTED else RANDOM.

Question: How big is $\max_{X \in SoS_d(G)} SIZE(X)$ in the *random* case?

(If $\ll \sqrt{n}$, we can beat the spectral algorithm!)


SIZE(X)
$n^{0.49}$
2 log(n)
conv(cliques)
$SoS_d(G)$

$$G \sim G(n, \tfrac{1}{2})$$

Goal: $X = X(G) \in SoS_d(G)$ so that

$$\sqrt{n} \geq \mathbb{E}_{G \sim \mathbb{G}(n,1/2)} SIZE(X(G)) \geq n^{0.49}$$

Goal: $X = X(G) \in SoS_d(G)$ so that

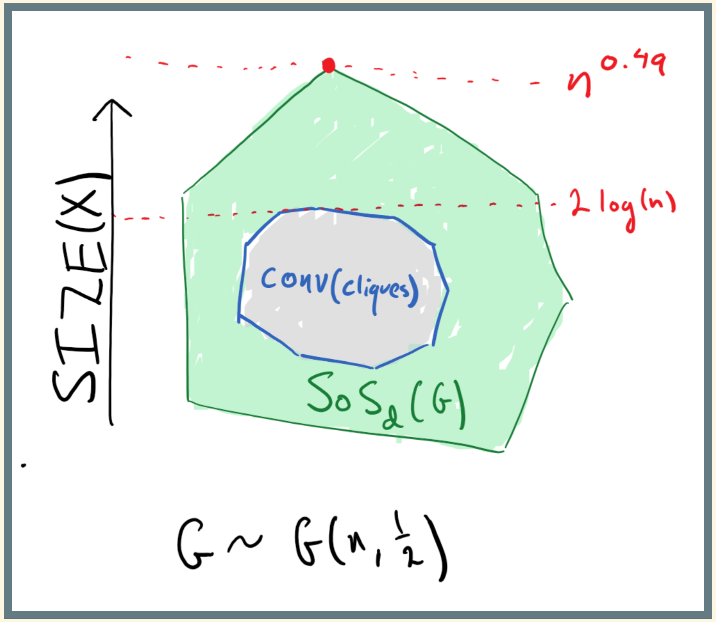
$$\mathbb{E}_{G \sim \mathbb{G}(n,1/2)} SIZE(X(G)) \geq n^{0.49}$$

Prior work [FK '04, folklore]:
$X(G)$ feasible for LP, basic SDP, spectral, Sherali-Adams, Lovasz-Schrijver+ with $SIZE(X(G)) \geq n^{0.49}$ w.h.p.

Related [MPW '15, DM '15, HKPRS '16]:
$X(G) \in SoS_d(G)$

$$\mathbb{E}_{G \sim \mathbb{G}(n,1/2)} SIZE(X(G)) \geq n^{1/\mathrm{poly}(d)} \approx n^{0.001}$$

Same $X(G)$ cannot work for tight SoS bound [Kelner]



Original Goal: $X = X(G) \in SoS_d(G)$ so that $\mathbb{E}_{G \sim \mathbb{G}(n,1/2)} SIZE(X(G)) \geq n^{0.49}$

Difficult to construct $X(G) \in SoS_d(G)$ (☹ ☹ high-dimensional positive-semidefinite matrices)

New (harder) Goal ***(pseudocalibration)*** : Construct $X = X(G) \in SoS_d(G)$ which shares more properties of a planted clique than just $SIZE(X(G)) \geq n^{0.49}$.

New (harder) Goal ***(pseudocalibration)*** :

$X = X(G) \in SoS_d(G)$ so that

$$\mathbb{E}_{G \sim \mathbb{G}(n,1/2)} T_G(X(G)) = \mathbb{E}_{\text{planted}} T_G(1_{\text{clique}})$$

for some family $\{T_G : \mathbb{R}^n \to \mathbb{R}\}$ of G-dependent linear functions ("tests"), including $SIZE$.

Example: $T_G(1_{\text{clique}}) =$ number of 4-cliques containing a typical clique vertex.

Lemma: If $\{T_G\}$ = linear functions whose coefficients are low-degree polynomials in entries of A_G and $X(G)$ satisfies

1. $\mathbb{E}_{G \sim \mathbb{G}(n,1/2)} T_G(X(G)) = \mathbb{E}_{\text{planted}} T_G(1_{\text{clique}})$
2. $\mathbb{E}_{G \sim \mathbb{G}(n,1/2)} \|X(G)\|_2^2$ is "small"

then with high probability $X(G) \in SoS_d(G)$ for $d = o(\log n)$ and $SIZE(X(G)) \geq n^{0.49}$.

Lemma: $X(G)$ satisfying these conditions exists.

$SoS_d(G)$ thinks there is an $n^{0.49}$-size clique in $G \sim \mathbb{G}(n, \frac{1}{2})$

$$G \sim \mathbb{G}\left(n, \frac{1}{2}\right).$$

WRAPPING UP

(Nearly) resolved "SoS versus planted clique" using new *pseudocalibration* approach.

Future work/open problems: Pseudocalibration suggests SoS lower bound construction for other planted problems.

THANKS! QUESTIONS?