The hidden clique problem and graphical models

Andrea Montanari

Stanford University

July 15, 2013

Outline

Finding a clique in a haystack

General Problem

$G = (V, E)$ a graph.

$S \subseteq V$ supports a clique (i.e. $(i,j) \in E$ for all $i,j \in S$)

Problem : Find S.

General Problem

$G = (V, E)$ a graph.

$S \subseteq V$ supports a clique (i.e. $(i, j) \in E$ for all $i, j \in S$)

Problem : Find S.

General Problem

$G = (V, E)$ a graph.

$S \subseteq V$ supports a clique (i.e. $(i,j) \in E$ for all $i, j \in S$)

Problem : Find S.

Example 1: Zachary's karate club

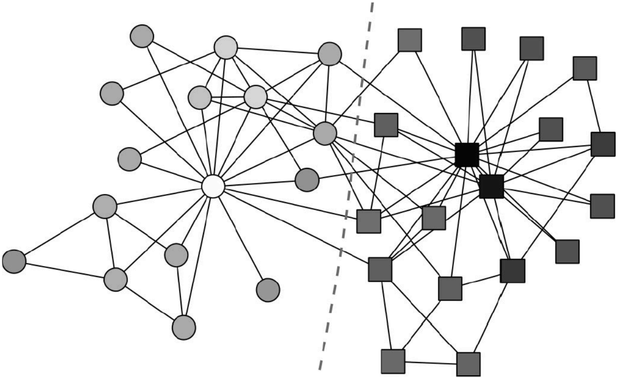
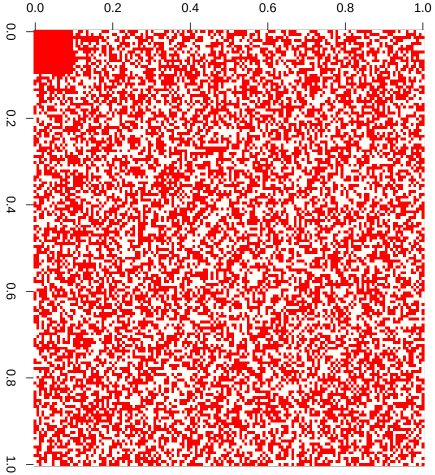


Fig. 2. Application of the eigenvector-based method to the karate club

A catchier name

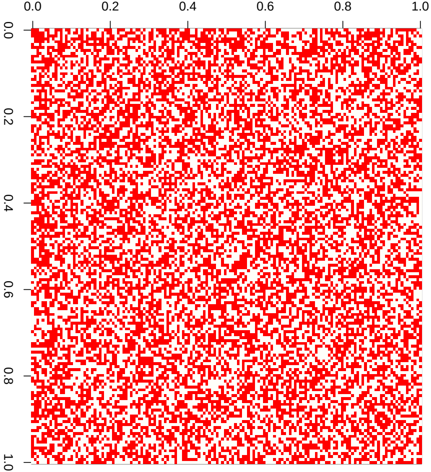
Finding a terrorist cell in a social network

Toy example: 150 nodes, 15 highly connected



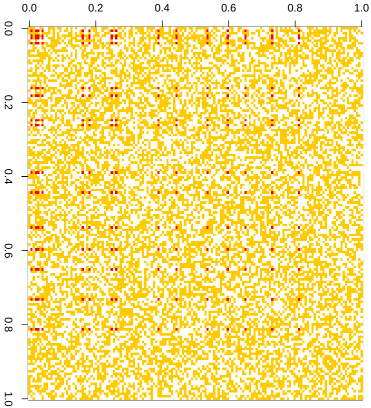
Here binary data: Can generalize...

Of course not the first 15...



Where are the highly connected nodes? 10^{21} possilities.

An efficient algorithm



The model [Alon,Krivelevich,Sudakov 1998]

- Choose $S \subseteq V$ with $|S| = k$ uniformly at random.
- Add an edge (i, j) for each pair s.t. $i, j \in S$.
- Add an edge for each other pair (i, j) independently with prob p.

$$G \sim \mathbb{G}(n, p, k)$$

Will assume $p = 1/2$

The model [Alon,Krivelevich,Sudakov 1998]

- Choose $S \subseteq V$ with $|S| = k$ uniformly at random.
- Add an edge (i, j) for each pair s.t. $i, j \in S$.
- Add an edge for each other pair (i, j) independently with prob p.

$$G \sim \mathbb{G}(n, p, k)$$

Will assume $p = 1/2$

The model [Alon,Krivelevich,Sudakov 1998]

- Choose $S \subseteq V$ with $|S| = k$ uniformly at random.
- Add an edge (i,j) for each pair s.t. $i,j \in S$.
- Add an edge for each other pair (i,j) independently with prob p.

$$G \sim \mathbb{G}(n,p,k)$$

Will assume $p = 1/2$

The model [Alon,Krivelevich,Sudakov 1998]

- Choose $S \subseteq V$ with $|S| = k$ uniformly at random.
- Add an edge (i,j) for each pair s.t. $i, j \in S$.
- Add an edge for each other pair (i,j) independently with prob p.

$$G \sim \mathbb{G}(n,p,k)$$

Will assume $p = 1/2$

The model [Alon,Krivelevich,Sudakov 1998]

- Choose $S \subseteq V$ with $|S| = k$ uniformly at random.
- Add an edge (i, j) for each pair s.t. $i, j \in S$.
- Add an edge for each other pair (i, j) independently with prob p.

$$G \sim \mathbb{G}(n, p, k)$$

Will assume $p = 1/2$

Question

How big k has to be for us to find the clique?

If you could wait forever

EXHAUSTIVE SEARCH

Input : Graph $G = (V, E)$, Clique size k
Output : Clique of size k
1: For all $S \subset V$, $|S| = k$;
3: Check if G_S is a clique;
4: Output all cliques found;

Works if $k > k_*$, typical size of largest clique in $G \sim \mathbb{G}(n, 1/2, k)$ that is not supported on S.

If you could wait forever

Exhaustive search

Input : Graph $G = (V, E)$, Clique size k
Output : Clique of size k
1: For all $S \subset V$, $|S| = k$;
3: Check if G_S is a clique;
4: Output all cliques found;

Works if $k > k_*$, typical size of largest clique in $G \sim \mathbb{G}(n, 1/2, k)$ that is not supported on S.

Largest random clique

Largest clique that does not share any vertex with S

Equivalently $G \sim \mathbb{G}(n-k, 1/2, 0) \approx \mathbb{G}(n, 1/2, 0)$.

Idea: compute expected number of cliques of size k

Largest random clique

Expected nb of cliques of size k

$$
\begin{aligned}
\mathbb{E}\, N(k) &= \binom{n}{k} \mathbb{P}\{(1,\dots,k) \text{ form a clique}\} \\
&= \binom{n}{k} 2^{-k(k-1)/2} \\
&\le \left(\frac{ne}{k}\right)^k 2^{-k(k-1)/2} \\
&\le \left(\frac{ne}{k} 2^{-(k-1)/2}\right)^k
\end{aligned}
$$

$$k_*(n) \approx 2 \log_2 n$$

Largest random clique

Expected nb of cliques of size k

$$\begin{aligned}
\mathbb{E}\, N(k) &= \binom{n}{k} \mathbb{P}\big\{(1,\dots,k) \text{ form a clique}\big\}\\
&= \binom{n}{k} 2^{-k(k-1)/2}\\
&\le \left(\frac{ne}{k}\right)^k 2^{-k(k-1)/2}\\
&\le \left(\frac{ne}{k} 2^{-(k-1)/2}\right)^k
\end{aligned}$$

$$k_*(n) \approx 2\log_2 n$$

Largest random clique

Expected nb of cliques of size k

$$\begin{aligned}
\mathbb{E}\, N(k) &= \binom{n}{k} \mathbb{P}\{(1,\dots,k) \text{ form a clique}\} \\
&= \binom{n}{k} 2^{-k(k-1)/2} \\
&\le \left(\frac{ne}{k}\right)^k 2^{-k(k-1)/2} \\
&\le \left(\frac{ne}{k} 2^{-(k-1)/2}\right)^k
\end{aligned}$$

$$k_*(n) \approx 2 \log_2 n$$

Largest random clique

Expected nb of cliques of size k

$$
\begin{aligned}
\mathbb{E}\, N(k) &= \binom{n}{k} \mathbb{P}\big\{(1,\dots,k) \text{ form a clique}\big\} \\
&= \binom{n}{k} 2^{-k(k-1)/2} \\
&\le \Big(\frac{ne}{k}\Big)^k 2^{-k(k-1)/2} \\
&\le \Big(\frac{ne}{k} 2^{-(k-1)/2}\Big)^k
\end{aligned}
$$

$$k_*(n) \approx 2 \log_2 n$$

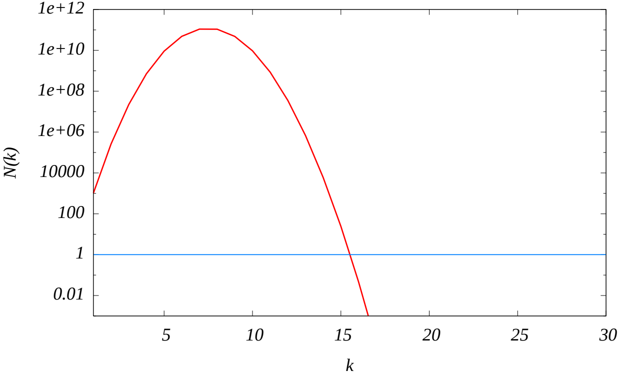
Largest random clique

Expected nb of cliques of size k

$$
\begin{aligned}
\mathbb{E}\, N(k) &= \binom{n}{k} \mathbb{P}\big\{(1,\dots,k) \text{ form a clique}\big\} \\
&= \binom{n}{k} 2^{-k(k-1)/2} \\
&\le \left(\frac{ne}{k}\right)^k 2^{-k(k-1)/2} \\
&\le \left(\frac{ne}{k} 2^{-(k-1)/2}\right)^k
\end{aligned}
$$

$$k_*(n) \approx 2 \log_2 n$$

Largest random clique



Can we do it in reasonable time?

NAIVE ALGORITHM

Input : Graph $G = (V, E)$, Clique size k
Output : Clique of size k
1: Sort vertices by degree;
2: Check if the k vertices with largest degree form a clique;
3: If yes, output them;

When does naive work?

For $i \notin S$

$$d_i \sim \mathrm{Binom}(n-1, 1/2) \approx \mathrm{Normal}(n/2, n/4)$$

$$\mathbb{P}\Big\{d_i \geq \frac{n}{2} + \frac{n^{1/2}}{2}\, t\Big\} \leq e^{-t^2/2} \leq \frac{1}{n^2} \quad \text{for } t \geq \sqrt{4 \log n}$$

Proposition

With high probability

$$\max_{i \notin S} d_i \leq \frac{n}{2} + \sqrt{n \log n}\,.$$

$$\min_{i \in S} d_i \geq \frac{n}{2} + k - 1 - \sqrt{n \log n}\,.$$

Works for $k \geq 2\sqrt{n \log n}$

When does naive work?

For $i \notin S$

$$d_i \sim \mathrm{Binom}(n-1, 1/2) \approx \mathrm{Normal}(n/2, n/4)$$

$$\mathbb{P}\Big\{ d_i \geq \frac{n}{2} + \frac{n^{1/2}}{2}\, t \Big\} \leq e^{-t^2/2} \leq \frac{1}{n^2} \quad \text{for } t \geq \sqrt{4 \log n}$$

Proposition

With high probability

$$\max_{i \notin S} d_i \leq \frac{n}{2} + \sqrt{n \log n}\,.$$

$$\min_{i \in S} d_i \geq \frac{n}{2} + k - 1 - \sqrt{n \log n}\,.$$

Works for $k \geq 2\sqrt{n \log n}$

When does naive work?

For $i \notin S$

$$d_i \sim \mathrm{Binom}(n-1, 1/2) \approx \mathrm{Normal}(n/2, n/4)$$

$$\mathbb{P}\Big\{ d_i \geq \frac{n}{2} + \frac{n^{1/2}}{2}\, t \Big\} \leq e^{-t^2/2} \leq \frac{1}{n^2} \quad \text{for } t \geq \sqrt{4 \log n}$$

Proposition

With high probability

$$\max_{i \notin S} d_i \leq \frac{n}{2} + \sqrt{n \log n}\,.$$

$$\min_{i \in S} d_i \geq \frac{n}{2} + k - 1 - \sqrt{n \log n}\,.$$

Works for $k \geq 2\sqrt{n \log n}$

When does naive work?

For $i \not\in S$

$$d_i \sim \mathrm{Binom}(n-1, 1/2) \approx \mathrm{Normal}(n/2, n/4)$$

$$\mathbb{P}\Big\{ d_i \geq \frac{n}{2} + \frac{n^{1/2}}{2}\, t \Big\} \leq e^{-t^2/2} \leq \frac{1}{n^2} \quad \text{for } t \geq \sqrt{4 \log n}$$

Proposition

With high probability

$$\max_{i \not\in S} d_i \leq \frac{n}{2} + \sqrt{n \log n}\,.$$

$$\min_{i \in S} d_i \geq \frac{n}{2} + k - 1 - \sqrt{n \log n}\,.$$

Works for $k \geq 2\sqrt{n \log n}$

A spectral algorithm

Idea

$$W_{ij} = \begin{cases} +1 & \text{if } (i,j) \in E, \\ -1 & \text{otherwise.} \end{cases}$$

$$(u_S)_i = \begin{cases} +1 & \text{if } i \in S, \\ 0 & \text{otherwise.} \end{cases}$$

Want to find u_S from W.

Idea

$$W_{ij} = \begin{cases} +1 & \text{if } (i,j) \in E, \\ -1 & \text{otherwise.} \end{cases}$$

$$(u_S)_i = \begin{cases} +1 & \text{if } i \in S, \\ 0 & \text{otherwise.} \end{cases}$$

Want to find u_S from W.

Idea

$$W_{ij} = \begin{cases} +1 & \text{if } (i,j) \in E, \\ -1 & \text{otherwise.} \end{cases}$$

$$(u_S)_i = \begin{cases} +1 & \text{if } i \in S, \\ 0 & \text{otherwise.} \end{cases}$$

Want to find u_S from W.

Idea

$$W = u_S u_S^{\mathsf{T}} + Z - Z_{S,S}$$

$(Z_{ij})_{i<j}$ i.i.d.

$$Z_{ij} = \begin{cases} +1 & \text{with probability } 1/2, \\ -1 & \text{with probability } 1/2. \end{cases}$$

$(Z_{S,S})_{ij} = Z_{ij}$ if $i, j \in S$ and $= 0$ otherwise

Idea

$$W = u_S u_S^{\mathsf{T}} + Z - Z_{S,S}$$

$(Z_{ij})_{i<j}$ i.i.d.

$$Z_{ij} = \begin{cases} +1 & \text{with probability } 1/2, \\ -1 & \text{with probability } 1/2. \end{cases}$$

$(Z_{S,S})_{ij} = Z_{ij}$ if $i, j \in S$ and $= 0$ otherwise

Idea

$$W = u_S u_S^{\mathsf{T}} + Z - Z_{S,S}$$

With overwhelming probability

$$\|u_S u_S\|_2 = k\,,$$
$$\|Z\|_2 \approx 2\sqrt{n}\,,$$
$$\|Z_{S,S}\|_2 \approx 2\sqrt{k} \ll \|Z\|_2\,.$$

Idea

$$W = u_S u_S^{\mathsf{T}} + Z - Z_{S,S}$$

With overwhelming probability

$$\begin{aligned} \|u_S u_S\|_2 &= k\,, \\ \|Z\|_2 &\approx 2\sqrt{n}\,, \\ \|Z_{S,S}\|_2 &\approx 2\sqrt{k} \ll \|Z\|_2\,. \end{aligned}$$

Use matrix perturbation theory

Unperturbed matrix

$$W_0 = u_S u_S^{\mathsf{T}} ,$$

$$\lambda_1(W_0) = k,\ \lambda_2(W_0) = \cdots = \lambda_n(W_0) = 0$$

The sin theta theorem

$\widehat{u}_S = u_S/\sqrt{k}$ principal eigenvector of W_0
v principal eigenvector of W

$$
\begin{aligned}
\|v - \widehat{u}_S\|_2 &\le \sqrt{2}\sin\theta(v, \widehat{u}_S) \le \frac{\sqrt{2}\|Z + Z_{S,S}\|_2}{\lambda_1(W_0) - \lambda_2(W)} \\
&\le \frac{3\sqrt{n}}{k - 3\sqrt{n}}
\end{aligned}
$$

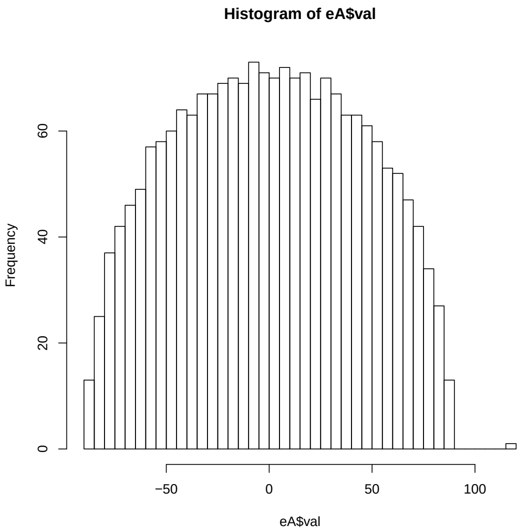
Summarizing

Proposition

For $k \geq 100\sqrt{n}$, whp

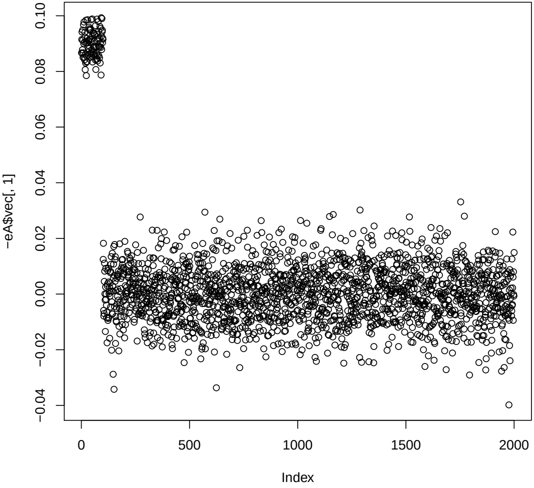
$$\|v - \widehat{u}_S\|_2 \leq \frac{1}{10}$$

Let's check how does it work. . .



$n = 2000, \ k = 100$

Let's check how does it work...



$n = 2000, \ k = 100$

Spectral algorithm: First attempt

NAIVE SPECTRAL ALGORITHM

Input : Graph $G = (V, E)$, Clique size k
Output : Clique of size k

1: Compute first eigenvector v of matrix $W = W(G)$;
2: Sort vertices by value of $|v_i|$;
3: Check if the k vertices with largest value form a clique;
4: If yes, output them;

Where is the problem?

Spectral algorithm: First attempt

NAIVE SPECTRAL ALGORITHM

Input : Graph $G = (V, E)$, Clique size k
Output : Clique of size k

1: Compute first eigenvector v of matrix $W = W(G)$;
2: Sort vertices by value of $|v_i|$;
3: Check if the k vertices with largest value form a clique;
4: If yes, output them;

Where is the problem?

Spectral algorithm

SPECTRAL ALGORITHM

Input : Graph $G = (V, E)$, Clique size k

Output : Clique of size k

1: Compute first eigenvector v of matrix $W = W(G)$;
2: Sort vertices by value of $|v_i|$;
3: Let $R \subseteq V$ be the set of k vertices with largest value;
4: For $i \in V$
5: If $\deg_R(i) > 3k/4$, let $S \leftarrow S \cup \{i\}$;
6: Output S;

Why is this a good trick?

- By the perturbation bound R is roughly good: $R \cap S > 0.9 \cdot k$.
- All the vertices in S pass the test.
- For $i \notin S$, $\mathbb{E}\mathrm{deg}_R(i) = k/2$ and $\mathrm{deg}_R(i) < 3k/4$ whp.

Improving over the spectral algorithm

We proved this

Theorem

If $k \geq 100\sqrt{n}$ then spectral algorithm finds the clique.

Can we make 100 as small as we want?

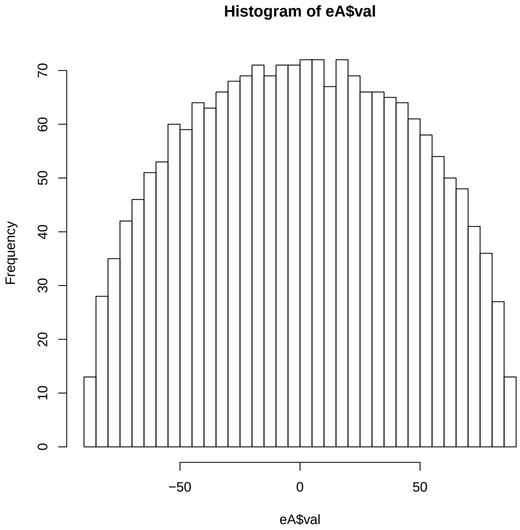
We proved this

Theorem

If $k \geq 100\sqrt{n}$ then spectral algorithm finds the clique.

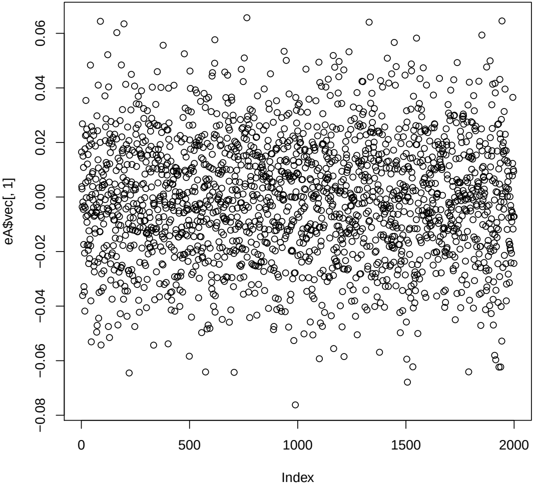
Can we make 100 as small as we want?

Not without a new idea...



$n = 2000,\ k = 30$

Not without a new idea...



$n = 2000, \ k = 30$

Tight analysis

$$W = u_S u_S^{\mathsf{T}} + Z - Z_{S,S} \approx u_S u_S^{\mathsf{T}} + Z$$

Low-rank deformation of a random matrix (e.g. Knowles, Yin 2011)

Proposition

If $k > (1+\epsilon)\sqrt{n}$, then $\langle u_S, v\rangle \geq \min(\epsilon, \sqrt{\epsilon})/2$.
Viceversa, if $k < (1-\epsilon)\sqrt{n}$, then $|\langle u_S, v\rangle| \leq n^{-1/2+\delta}$.

The test set idea

ENHANCED SPECTRAL ALGORITHM

Input : Graph $G = (V, E)$, Clique size k
Output : Clique of size k
1: For $T \subseteq V$, $|T| = 2$, T is a clique;
2: Let $V_T = \{i \in V \setminus T : \deg_T(i) = 2\}$
and let G_T be the induced graph;
3: Run SPECTRAL ALGORITHM($G_T, k-2$);
4: If a clique S of size $k-2$ is found, return $S \cup T$;

Why is this a good trick?

$$V'_T \approx \frac{n}{4}$$

Looking for clique of size $k-2$.

If SPECTRAL succeeds for $k \geq c\sqrt{n}$, then ENHANCED SPECTRAL works for $k - 2 \geq c\sqrt{n/4}$.

Equivalently for $k \geq c'\sqrt{n}$ with

$$c' = c/2 + 0.000001$$

Why is this a good trick?

$$V'_T \approx \frac{n}{4}$$

Looking for clique of size $k-2$.

If SPECTRAL succeeds for $k \geq c\sqrt{n}$, then ENHANCED SPECTRAL works for $k-2 \geq c\sqrt{n/4}$.

Equivalently for $k \geq c'\sqrt{n}$ with

$$c' = c/2 + 0.000001$$

Why is this a good trick?

$$V'_T \approx \frac{n}{4}$$

Looking for clique of size $k - 2$.

If SPECTRAL succeeds for $k \geq c\sqrt{n}$, then ENHANCED SPECTRAL works for $k - 2 \geq c\sqrt{n/4}$.

Equivalently for $k \geq c'\sqrt{n}$ with

$$c' = c/2 + 0.000001$$

Iterating the same idea

ENHANCED SPECTRAL ALGORITHM

Input : Graph $G = (V, E)$, Clique size k
Output : Clique of size k
1: For $T \subseteq V$, $|T| = s$, T is a clique;
2: Let $V_T = \{i \in V \setminus T : \deg_T(i) = 2\}$
and let G_T be the induced graph;
3: Run SPECTRAL ALGORITHM$(G_T, k - s)$;
4: If a clique S of size $k - s$ is found, return $S \cup T$;

The state of the art for polynomial algorithms

Theorem

ENHANCED SPECTRAL*(s) finds cliques of size* $c2^{-s/2}\sqrt{n}$ *in time* $O(n^{s+c_0})$.

What about really efficient algorithms?

Theorem (Dekel, Gurel-Gurevitch, Peres, 2011)

If $k \geq 1.261\sqrt{n}$, then there exists an algorithm with complexity $O(n^2)$ that finds the clique with high probability.

Theorem (Deshpande, Montanari, 2013)

If $k \geq (1+\epsilon)\sqrt{n/e}$, then there exists an algorithm with complexity $O(n^2 \log n)$ that finds the clique with high probability.

What about really efficient algorithms?

Theorem (Dekel, Gurel-Gurevitch, Peres, 2011)

If $k \geq 1.261\sqrt{n}$, then there exists an algorithm with complexity $O(n^2)$ that finds the clique with high probability.

Theorem (Deshpande, Montanari, 2013)

If $k \geq (1+\epsilon)\sqrt{n/e}$, then there exists an algorithm with complexity $O(n^2 \log n)$ that finds the clique with high probability.

What about really efficient algorithms?

Theorem (Dekel, Gurel-Gurevitch, Peres, 2011)

If $k \geq 1.261\sqrt{n}$, then there exists an algorithm with complexity $O(n^2)$ that finds the clique with high probability.

Theorem (Deshpande, Montanari, 2013)

If $k \geq (1+\epsilon)\sqrt{n/e}$, then there exists an algorithm with complexity $O(n^2 \log n)$ that finds the clique with high probability.