How to find a hidden clique

Brendan Ames

Institute for Mathematics and its Applications
University of Minnesota

MOPTA 2013 Lehigh University
Semidefinite Optimization
Friday August 16, 2013
Agenda

• Present a convex relaxation for the maximum clique problem based on nuclear norm relaxation for rank minimization.

• Extend this relaxation to one for the densest $k$-subgraph problem.

• Phase transitions: tradeoff between on size of cliques and density ensuring recovery from the relaxations.

• Experimental results and open problems.

• Joint work with Stephen Vavasis, University of Waterloo.
Graph clustering

- **Similarity Graph**: represent data set as a graph
  - items = nodes
  - edges indicate similarity

- Cluster the data set by dividing the graph into dense subgraphs.

- Dense = large average degree
Example: Communities in Social Networks

- Nodes = users
- Edges = “friendship”.
- Densely connected groups = communities
Cliques of a graph

• Given graph $G = (V, E)$, a clique of $G$ is a pairwise adjacent subset of $V$.

• $C \subseteq V$ is a clique of $G$ if $uv \in E$ for all $u, v \in C$.

• The subgraph $G(C)$ induced by $C$ is complete.
Cliques of a graph

- Given graph $G = (V, E)$, a clique of $G$ is a pairwise adjacent subset of $V$.

- $C \subseteq V$ is a clique of $G$ if $uv \in E$ for all $u, v \in C$.

- The subgraph $G(C)$ induced by $C$ is complete.
The Maximum Clique problem

- **Optimization version**: Find the size of the largest clique of $G$; size of the largest clique is the **clique number** $\omega(G)$.

- **Decision version**: Given graph $G$, integer $k$: does $G$ contain a clique of cardinality at least $k$?

- **Complexity**: NP-complete, NP-hard to approximate $\omega(G)$ within a ratio of $N^{1-\epsilon}$ for any $\epsilon > 0$, ($N = |V|$)

- **Many applications**: communication, power, and social networks, mathematical biology, cryptography.
Matrix representation of cliques

- **Characteristic vector of** $C$: vector $\mathbf{v} \in \{0, 1\}^V$ with
  \[ v_i = 1 \text{ if } i \in C \quad \text{and} \quad v_i = 0 \text{ otherwise} \]

- If $C$ is a clique with characteristic vector $\mathbf{v}$, let $X = \mathbf{v}\mathbf{v}^T$.

- Nonzero entries of $X$ form a $|C| \times |C|$ all-ones block in $A_G + I$. 
Clique as rank minimization

- **G** has a $k$-clique if and only if there exists rank-one symmetric binary matrix $X$ such that

  $$
  \sum \sum X_{ij} = k^2
  $$

  $$
  X_{ij} = 0 \quad \forall \ ij \notin E, \ i \neq j.
  $$

- **Clique** is equivalent to the rank minimization problem:

  $$
  \min_{\substack{X \in \{0,1\}^{V \times V} \\
  X \in \Sigma^V}} \left\{ \text{rank}(X) : e^T X e = k^2, X_{ij} = 0 \text{ if } (i,j) \in \tilde{E} \right\}
  $$

  where $\tilde{E} = V \times V - \{ E \cup \{(u,u) : u \in V\} \}$. 
Nuclear norm relaxation of rank

- **Affine rank minimization problem**: find matrix with minimum rank satisfying linear constraints:

  \[
  \min \{ \text{rank}(X) : A(X) = b \}.
  \]

  Well-known to be NP-hard.

- Relax \( \text{rank}(X) \) with nuclear norm \( \|X\|_* \):

  \[
  \text{rank}(X) = \|\sigma(X)\|_0, \quad \|X\|_* = \|\sigma(X)\|_1.
  \]

  If \( A \) is “nice” then the minimum nuclear norm solution is the minimum rank solution.

- Can be written as an SDP.
We have the rank minimization problem:

$$\min_{X \in \{0,1\}^V \times V, X \in \Sigma^V} \ \left\{ \text{rank}(X) : e^T X e = k^2, \ X_{ij} = 0 \text{ if } (i,j) \in \tilde{E} \right\}$$

Relax rank with the nuclear norm and ignore the binary and symmetry constraints:

$$\min \left\{ \|X\|_* : e^T X e = k^2, \ X_{ij} = 0 \text{ if } (i,j) \in \tilde{E} \right\} \quad (\text{NNR})$$

When does the solution of the relaxation coincide with that of the rank minimization problem?
The planted case

Construction:

- Add edges between nodes in vertex set \( V^* \) of size \( k \).
The planted case

Construction:

- Add edges between nodes in vertex set $V^*$ of size $k$.
- Each remaining edge is added to $E$ independently with fixed probability $p$. 
The planted case

Construction:

• Add edges between nodes in vertex set $V^*$ of size $k$.

• Each remaining edge is added to $E$ independently with fixed probability $p$.

• $V^*$ is a clique of $G$ (called a planted or hidden clique).
The planted case

Construction:

- Add edges between nodes in vertex set $V^*$ of size $k$.
- Each remaining edge is added to $E$ independently with fixed probability $p$.
- $V^*$ is a clique of $G$ (called a planted or hidden clique).
The planted case

Construction:

- Add edges between nodes in vertex set $V^*$ of size $k$.

- Each remaining edge is added to $E$ independently with fixed probability $p$.

- $V^*$ is a clique of $G$ (called a planted or hidden clique).
The planted case

Construction:

• Add edges between nodes in vertex set $V^*$ of size $k$.

• Each remaining edge is added to $E$ independently with fixed probability $p$.

• $V^*$ is a clique of $G$ (called a planted or hidden clique).
Recovery guarantee in the planted case

Theorem (Ames-Vavasis 2009)

There exists scalar $c > 0$ (depending only on $p$) such that if

$$k \geq c\sqrt{N}$$

then

- $V^*$ is the unique maximum clique of $G$, and
- $X^* = vv^T$ is the unique optimal solution of $(NNR)$

with high probability.
What happens if edges are deleted?

- Guarantee does not tolerate edge deletion noise.
- Suppose edge \( uv \) is deleted for some \( u, v \in V^* \).
- Then \( V^* \) is not a clique and \( X^* \) is not feasible for (NRR).
What happens if edges are deleted?

- Guarantee does not tolerate edge deletion noise.
- Suppose edge $uv$ is deleted for some $u, v \in V^*$. 
- Then $V^*$ is not a clique and $X^*$ is not feasible for (NNR).
What if $p$ varies with $N$?

- Guarantee assumes noise probability $p$ is fixed.
- Not a realistic model of many real-world networks.
- Want a bound that allows $p$ to grow/shrink with $N$. 
The densest k-subgraph problem

- Want a **dense** subgraph of size $k$, not necessarily a clique.

- **Densest $k$-subgraph problem (DKS):** find subgraph $H \subseteq G$ on $k$ nodes with maximum density:

  $$d(H) = \frac{|E(H)|}{|V(H)|} = \frac{|E(H)|}{k}.$$  

- **NP-hard:** proof is by reduction to Clique; hard to approximate.

- Maximizing $d(H) = \text{maximizing } |E(H)|$ over all $k$-node subgraphs.
Duality of density and number of missing edges

- Let \( V^* \subseteq V \) be a \( k \)-subset with characteristic vector \( v \).
- Introduce a correction \( Y \) for entries of \( X = vv^T \) that should be 0:
  \[
  Y_{ij} = \begin{cases} 
    -X_{ij}, & \text{if } ij \in \tilde{E} \\
    0, & \text{otherwise}
  \end{cases}
  \]
- If \( V^* \) is almost a clique then \( G(V^*) \) should be very dense and \( Y \) should be very sparse.
- Cardinality of \( Y \) acts as a dual of density of \( G(V^*) \):
  \[
  |E(G(V^*))| = \binom{k}{2} - \frac{\|Y^*\|_0}{2}
  \]
Relaxation as Principal Component Pursuit

- Can relax \((\text{DKS})\) as

\[
\min_{X} \|X\|_* + \gamma \|Y\|_1 \\
\text{st} \quad e^T X e = k^2 \\
X_{ij} + Y_{ij} = 0 \text{ if } ij \in \tilde{E} \\
X \in [0, 1]^{V \times V}
\]

\((\text{DKSR})\)

where \(\gamma\) is a regularization parameter.

- Relax \(\|Y\|_0\) using the \(\ell_1\)-norm \(\|Y\|_1\), \(\text{rank}(X)\) with the nuclear norm \(\|X\|_*\).
Robust PCA

- **Robust PCA**: decompose matrix $M$ as

  $$M = L + S$$

  where $L$ has low-rank, $S$ is sparse.


  $$\arg\min \{ \text{rank}(L) + \gamma \|S\|_0 : P_\Omega(L + S) = P_\Omega(M) \}$$

  $$= \arg\min \{ \|L\|_* + \gamma \|S\|_1 : P_\Omega(L + S) = P_\Omega(M) \}.$$

  under certain assumptions on the support of $S$, column/row space of $L$, choice of $\gamma$, and the set of observed entries $\Omega$. 
**Planted case**

- Start with set of $N$ nodes $V$.
- Add all edges between nodes in $V^* \subseteq V$.
- Add noise:
  - Add potential edges with probability $p(N)$
  - Delete some edges in $V^* \times V^*$ with probability $q(N)$
Planted case

• Start with set of $N$ nodes $V$.
• Add all edges between nodes in $V^* \subseteq V$.
• Add noise:
  • Add potential edges with probability $p(N)$
  • Delete some edges in $V^* \times V^*$ with probability $q(N)$
Planted case

- Start with set of $N$ nodes $\mathcal{V}$.
- Add all edges between nodes in $\mathcal{V}^* \subseteq \mathcal{V}$.
- Add noise:
  - Add potential edges with probability $p(N)$
  - Delete some edges in $\mathcal{V}^* \times \mathcal{V}^*$ with probability $q(N)$
Recovery guarantee

Theorem (Ames 2013)

Under certain technical assumptions on \( p, q, k, N \), there exist absolute constants \( c_1, c_2, c_3 > 0 \) such that if

\[
(1 - p - q)(1 - p)k \geq c_1 \max \left\{ \sqrt{p}, 1 / \sqrt{(1 - p)k} \right\} \cdot \sqrt{N \log N}
\]

and

\[
\gamma \in \left( \frac{c_2}{(1 - p - q)k}, \frac{c_3}{(1 - p - q)k} \right)
\]

then

- \( G(V^*) \) is the unique densest \( k \)-subgraph of \( G \), and
- \( X^* = vv^T \) is the unique optimal solution of \((\text{DKSR})\) with high probability.
Translating the Recovery guarantee

• If $p, q$ are fixed:
  • Get the same bound as before $k = \Omega(\sqrt{N \log N})$ (with added log term).

• If $p, q \to 0$ as $N \to \infty$ we can find smaller planted cliques.
  • e.g., if $p, q = \Theta(\log k/k)$ can find the planted clique if
    $$k = \Omega(N^{1/3} \log N)$$
Proof Idea

• Apply KKT conditions and SDP duality to derive conditions ensuring optimality and uniqueness of $X^*$. 

• Solve for a choice of multipliers corresponding to $X^*$. 

• Use random matrix theory to show optimality and uniqueness conditions are satisfied (w.h.p.).
Numerical results

- Randomly generated $N$-node graphs containing planted dense $k$-subgraphs.
- Probability of deletion $q = 0.25$ used in each trial, $p$ varied.
- 10 trials for each $(p, k)$ pair.
- Solved (DKSR) using ADMM:
  - Move constraints to objective and split the variables so the new objective is separable.
  - Alternately minimize the augmented Lagrangian with respect to each decision variable.
- Compared obtained solution with the planted solution.
Numerical results

$N = 250$

$N = 500$
Conclusions

- New convex relaxations for the Clique and Densest $k$-subgraph problems.
- Theoretical guarantees for exact recovery in random case.
- Analogous recovery guarantees for deterministic construction.
- Identical results for bipartite graphs.
Conclusions

• Open problems:
  • How to efficiently solve the relaxations?
  • Different sparsity inducing penalties?
  • Many hidden cliques?
  • Are the random bounds tight? Can we relax $\Omega(N^{1/2})$ to $\Omega(N^{1/2-\epsilon})$ for fixed $p$?

• References:
  • B. Ames. Robust convex relaxation for the planted clique and densest k-subgraph problem. 2013. arxiv.org/abs/1305.4891