Graph Partition Reconstruction

A survey on spectral methods

Dieter Mitsche

JCALM 2011

Overview

Model

- Given some input drawn randomly according to some probability distribution, with some additional **planted** substructure.
- Goal: Without knowledge of this substructure, reconstruct the structure.

- Spectral methods: capture the input in matrix form, apply eigenvalue/eigenvector techniques to find the structure

Substructures considered here:

- Clique
- k-Partition

The hidden clique problem

Definition

- Given a graph $G \in_{u.a.r.} \mathcal{G}(n,p)$ with $0 < p < 1$ together with a randomly chosen set of k vertices forming a clique (k and p might or might not be known as input).
- Goal: Find the clique in polynomial time (whp), without knowing these k vertices.

Fact

- *In $\mathcal{G}(n,p)$, the maximum clique size $\omega(G)$ is $\Theta(\log n)$.*
- *In $\mathcal{G}(n,p)$, all vertex degrees are whp in the interval $(n-1)p \pm C\sqrt{n \log n}$ for some $C > 0$ sufficiently large.*

- If $k < \omega(G)$, one cannot hope to find the planted clique.
- If $k > C'\sqrt{n \log n}$ for some $C' > 0$, then the vertices of the clique are whp the k vertices of highest degree.

The hidden clique problem

Question

Can one find a planted clique of size $k = o(\sqrt{n \log n})$ in polynomial time in $G \in \mathcal{G}(n, \frac{1}{2})$?

Theorem (Alon, Krivelevich, Sudakov 98)

There exists a polynomial time algorithm for finding a clique of size $k = \Theta(\sqrt{n})$.

- Based on spectral methods: capture input graph in matrix form and use eigenvalues, eigenvectors

Background on spectral methods

Definition

Let $M \in \mathbb{R}^{n\times n}$ a real, symmetric matrix. Define the operator norm (2-norm) of M as

$$|M|_2 = \max_{|x|_2=1} |Mx|_2.$$

Letting the eigenvalues be $\lambda_1(M) \geq \ldots \geq \lambda_n(M)$, we have

$$|M|_2 = \max\{|\lambda_1|, |\lambda_n|\}.$$

- How to bound $|M|_2$ for a random matrix?

Model

Let $M \in \mathbb{R}^{n\times n}$ now a real, symmetric matrix, where for $i \leq j$, all M_{ij} are i.i.d. random variables with $\mathbb{E}(M_{ij}) = 0$, $\sigma(M_{ij}) = 1$ and all entries bounded from above by K in absolute value for some constant $K > 0$.

Background on spectral methods

How to bound $|M|_2$ for a random matrix?
Lower bound: find particular vector $x \in \mathbb{R}^n$ and calculate $\frac{x^T M x}{x^T x}$.

Usually more interested in upper bound

- Method 1: maximal $\epsilon-$net argument

Definition

Σ is a **maximal** ϵ**-net** of the sphere $\mathcal{S}$ if for any $x, y \in \Sigma$, $|x - y|_2 \geq \epsilon$, and moreover Σ is maximal with respect to set inclusion.

Background on spectral methods

How to bound $|M|_2$ for a random matrix?
Lower bound: find particular vector $x \in \mathbb{R}^n$ and calculate $\frac{x^T M x}{x^T x}$.

Usually more interested in upper bound

- Method 1: maximal $\epsilon-$net argument

Definition

Σ is a **maximal** ϵ**-net** of the sphere $\mathcal{S}$ if for any $x, y \in \Sigma, |x-y|_2 \geq \epsilon$, and moreover Σ is maximal with respect to set inclusion.

$$\mathbb{P}(|M|_2 > A\sqrt{n}) \leq Ce^{-cAn}.$$

Background on spectral methods

Alternative for an upper bound on $|M|_2$ for a random matrix M

- Method 2: trace method

Idea

Compute trace of a high even power of the adjacency matrix. Use the fact that $tr(M) = \sum_i M_{ii} = \sum_i \lambda_i(M)$. For a high even power, $\lambda_1(M) \sim \sum_i \lambda_i(M)$.

Background on spectral methods

Alternative for an upper bound on $|M|_2$ for a random matrix M

- Method 2: trace method

Idea

Compute trace of a high even power of the adjacency matrix. Use the fact that $tr(M) = \sum_i M_{ii} = \sum_i \lambda_i(M)$. For a high even power, $\lambda_1(M) \sim \sum_i \lambda_i(M)$.

$$\mathbb{E}(|M|^k) \leq C_{k/2} n^{k/2+1},$$

where $C_{k/2}$ is the $k/2$-th Catalan number.

Background on spectral methods

Theorem

$$|M|_2 \leq (2 + o(1))\sqrt{n}.$$

Application for random graphs: given $G \in \mathcal{G}(n, \frac{1}{2})$, write $A(G) = P + M$, where $P_{ij} = \frac{1}{2}$ for $i \neq j$, and 0 otherwise. M is a random matrix satisfying the above properties.

Fact

- $\lambda_1(P) = (n-1)\frac{1}{2}, |\lambda_2(P)| = \ldots = |\lambda_n(P)| = \frac{1}{2}$.
- *Using* $\lambda_1(P) + \lambda_n(M) \leq \lambda_1(G) \leq \lambda_1(P) + \lambda_n(M)$, $\lambda_1(G) = (n-1)\frac{1}{2}(1 + o(1))$.
- $\lambda_2(G) \leq \lambda_2(P) + \lambda_1(M) = (2 + o(1))\sqrt{n}$.

The hidden clique problem (cont'd)

Adding a hidden clique of size $k = \Omega(\sqrt{n})$ makes $\lambda_2(G)$ bigger:

Proposition (Alon,Krivelevich,Sudakov 98)

Assume w.l.o.g. that vertices $1, \ldots, k$ *are forced to be a planted clique. Define* $z \in \mathbb{R}^n$ *as* $z_i = n - k$ *for* $1 \leq i \leq k$ *and* $z_i = -k$ *for* $i > k$. *There exists a vector* δ *with* $|\delta|_2^2 \leq \frac{1}{60}|z|_2^2$ *so that* $z - \delta$ *is collinear with the second eigenvector* v_2 *of* $A(G)$, *with corresponding eigenvalue*

$$\frac{k}{2} - \sqrt{\frac{n}{2}} \leq \lambda_2(G) \leq \frac{k}{2} - \sqrt{\frac{n}{2}}.$$

In particular, when $k \geq 10\sqrt{n}$, λ_2 is much bigger than λ_i for $i \geq 3$.

The hidden clique problem (cont'd)

Algorithm (Alon,Krivelevich,Sudakov 98)

Input: A graph $G \in G(n, \frac{1}{2})$ with a planted clique of size $k \geq 10\sqrt{n}$

- Find the second eigenvector v_2 of the adjacency matrix of G
- Sort the vertices of V by decreasing order of the absolute values of their coordinates in v_2. Let W be the first k vertices.
- Postprocess: Let $Q \subseteq V$ be the set of all vertices with at least $\frac{3k}{4}$ vertices in W

Output: The subset $Q \subseteq V$.

Between hidden clique and planted partitions

- Easy extension: Plant a denser subgraph instead of a clique
- If more planted (big) cliques of different sizes are added, look at more eigenvectors and apply the previous algorithm iteratively

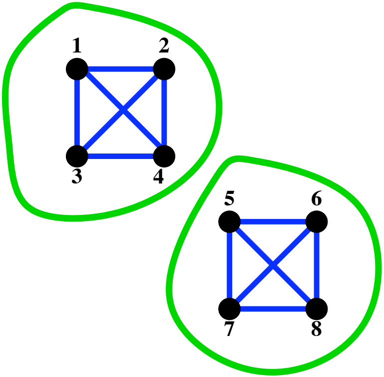
- What if all (some) cliques have the same size? A partition into cliques?
- Eigenvectors corresponding to cliques of different cliques are not "robust" to perturbations anymore

Idea

Although eigenvectors are not robust, eigenspaces are robust. Assuming a partition of G into s planted cliques, define the **projector** $P = \sum_{i=1}^{s} v_i v_i^T$, where v_i is the i-th eigenvector of G. P is stable.

Planted partitions - intuition

- Graph with $n = 8$, $s = 2$



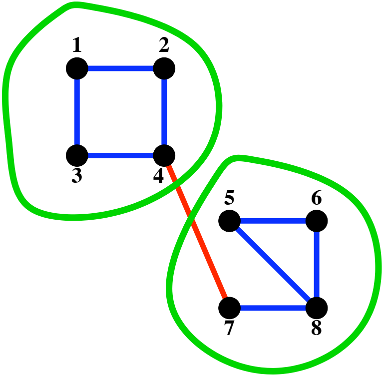
Planted partitions - intuition

- Projector P of 'perfect' graph with $s = 2$ (Values scaled by factor 10):

$$P = \begin{bmatrix} 2.5 & 2.5 & 2.5 & 2.5 & 0 & 0 & 0 & 0 \\ 2.5 & 2.5 & 2.5 & 2.5 & 0 & 0 & 0 & 0 \\ 2.5 & 2.5 & 2.5 & 2.5 & 0 & 0 & 0 & 0 \\ 2.5 & 2.5 & 2.5 & 2.5 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2.5 & 2.5 & 2.5 & 2.5 \\ 0 & 0 & 0 & 0 & 2.5 & 2.5 & 2.5 & 2.5 \\ 0 & 0 & 0 & 0 & 2.5 & 2.5 & 2.5 & 2.5 \\ 0 & 0 & 0 & 0 & 2.5 & 2.5 & 2.5 & 2.5 \end{bmatrix}$$

Planted partitions - intuition

- Graph with $n = 8$, $s = 2$



Planted partitions - intuition

- Adjacency matrix A of previous graph:

$$A = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \end{bmatrix}$$

Planted partitions - intuition

- Projector $P(A) =: P$ of previous graph with $s = 2$ (Values scaled by factor 10):

$$P = \begin{bmatrix} 2.0 & 2.1 & 2.1 & 2.4 & -0.6 & -0.6 & 0.9 & -0.3 \\ 2.1 & 2.2 & 2.2 & 2.5 & -0.4 & -0.4 & 1.0 & -0.1 \\ 2.1 & 2.2 & 2.2 & 2.5 & -0.4 & -0.4 & 1.0 & -0.1 \\ 2.4 & 2.5 & 2.5 & 3.0 & 0.0 & 0.0 & 1.5 & 0.5 \\ -0.6 & -0.4 & -0.4 & 0.0 & 2.7 & 2.7 & 1.4 & 3.1 \\ -0.6 & -0.4 & -0.4 & 0.0 & 2.7 & 2.7 & 1.4 & 3.1 \\ 0.9 & 1.0 & 1.0 & 1.5 & 1.4 & 1.4 & 1.4 & 1.8 \\ -0.3 & -0.1 & -0.1 & 0.5 & 3.1 & 3.1 & 1.8 & 3.7 \end{bmatrix}$$

Planted partitions - the related classical s-partition problem

- The classical s-Partition-Problem: Given some graph G, partition $V(G)$ into s sets of equal size, s.t. number of edges between sets is minimized
- Applications: Parallel scheduling, mesh partitioning, clustering
- s-Partition-Problem is NP-hard (even for $s = 2$)
- Here: Planted partition problem

Planted partitions - the model (McSherry '01)

- A distribution $\mathcal{G}(\phi, p, q)$
  - ϕ: $V \rightarrow \{1, \ldots, s\}$; include edge $\{v, w\} \in E, v \neq w$, independently from other edges with probability:

$$\mathbb{P}(\{v, w\} \in E) = \begin{cases} p, & \text{if } \phi(v) = \phi(w), \\ q, & \text{if } \phi(v) \neq \phi(w), \end{cases}$$

    where $p > q$
- For fixed p, q, $p > q$, the problem is more difficult if s is bigger

- Problem: Given G drawn from $\mathcal{G}(\phi, p, q)$ distribution, find a partition $\bar{\phi}$ such that $\bar{\phi}(v)=\bar{\phi}(w)$ iff $\phi(v)=\phi(w)$, $\forall v, w \in V$.

Planted partitions - the model

- Planted Partition Model: see e.g. Frieze, McDiarmid (1996)
- Algorithms for Planted Partition Model
  - Spectral approach: McSherry (2001)
  - Nonspectral approach: Shamir, Tsur (2002)
- Both reconstruct up to $O(\sqrt{n/\log n})$ partition classes

Question

Can we reconstruct $\Theta(\sqrt{n})$ partition classes, matching the $\Theta(\sqrt{n})$ bound for one clique?

partial solution to this ...

Planted partitions - the model

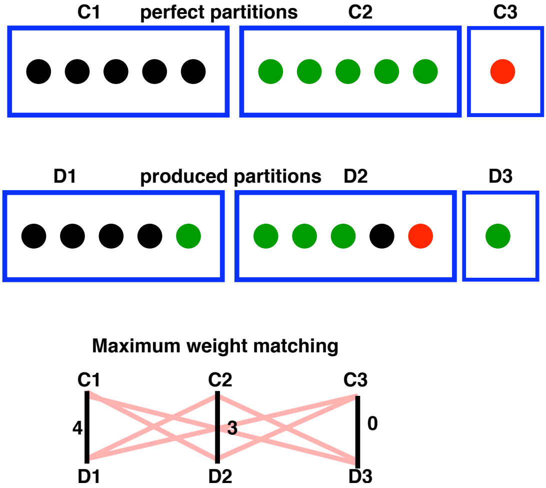
- Problem: Given G drawn from $\mathcal{G}(\phi, p, q)$ distribution, find a partition $\bar{\phi}$ such that $\bar{\phi}(v)=\bar{\phi}(w)$ iff $\phi(v)=\phi(w)$, $\forall v, w \in V$.
- Measures of success
  - Goal 1: Perfect reconstruction
    - Every vertex is correctly classified
  - Goal 2: Good reconstruction
    - Allows some misclassified vertices

Planted partitions - the model

- How do we count the number of misclassifications?
  - Given planted partitions $V_1,\ldots,V_s$, 'produced' partitions $\bar{V}_1, \ldots, \bar{V}_s$, define a bipartite graph B with vertex set $\{V_1, \ldots, V_s, \bar{V}_1, \ldots, \bar{V}_s\}$, and weighted edge w_{ij} between V_i and $\bar{V}_j$ with weight $|V_i \cap \bar{V}_j|$
  - Let M be a maximum weight matching on B
- Number of misclassifications: $\sum_{i=1}^{s} |V_i|$ - $\sum_{(i,j)\in M} w_{ij}$

Planted partitions - the model

Example: misclassifications on graph with $s = 3$



Planted partitions - results

Theorem (Giesen, M.)

- *For fixed p and q, with $p > q$, and large n, we can whp*
  - (i) *reconstruct in polynomial time up to s partitions correctly, for $s \leq c\frac{\sqrt{n}}{\log\log n}$.*
  - (ii) *in polynomial time bound the number of misclassifications by $cs\sqrt{n}$, for some $c > 0$, for $s \leq c'\sqrt{n}$.*
  - (iii) *if $s \geq \frac{cn}{\log n}$, then the planted s-partition is not a minimum partition anymore.*

Planted partitions - algorithm and analysis

Outline of the simplest algorithm - for item (ii)

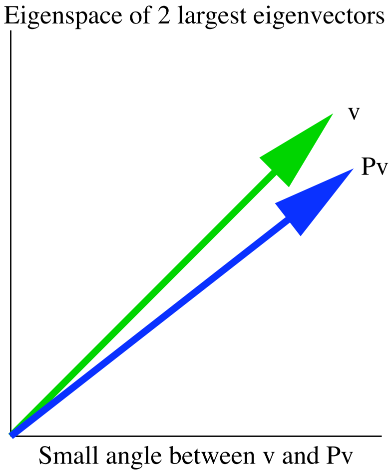
1. INPUT: Adjacency matrix A of a graph from $\mathcal{G}(\phi, p, q)$
2. Estimate $s :=$ number of partitions
3. Compute projector P onto s largest eigenvectors of A.
4. Compute vectors $c_i \in \{0, 1\}^n$ of each column of P:
   $c_i(j) = 1$ iff j-th entry of P_i among n/s largest entries
5. (Basically) put i and j in the same partition C_ℓ if the Hamming Distance between c_i and c_j is small

Planted partitions - algorithm and analysis

- Compare projector with 'perfect' projector
  - ‘Perfect’ projector has (after permutation) blockdiagonal structure
  - Subspace of 'perfect' projector = space of all piecewise constant vectors (vectors constant on each partition class)
- Want indicator vectors for reconstruction
- Indicator vectors = special type of piecewise constant vectors v

Planted partitions - algorithm and analysis

- Let v be an indicator vector found. Matrix perturbation theory guarantees:



Planted partitions - algorithm and analysis

Theorem (Stewart)

Let M and $\hat{M}$ two symmetric matrices $\in \mathbb{R}^{n \times n}$ and let P (and $\hat{P}$, respectively) be the projection matrices onto the space anned by the eigenvectors corresponding to the k largest eigenvectors of M ($\hat{M}$, respectively). Then

$$|P - \hat{P}|_2 \leq \frac{2|M - \hat{M}|_2}{|\lambda_k(M) - \lambda_{k+1}(M)| - 2|M - \hat{M}|_2},$$

if $|\lambda_k(M) - \lambda_{k+1}(M)| > 4|M - \hat{M}|_2$.

- Projectors are close in norm if corresponding matrices have small norm difference and spectral gap is high

Planted partitions - algorithm and analysis

Remarks about improved algorithm for correct reconstruction (item (i))

- partition input matrix into several parts
- boost algorithm by pruning already reconstructed parts

Remarks about lower bound on non-reconstructability (item (iii))

- Non-spectral. Compute probability of existence of smaller s-partition than the planted one (second moment method)

Remarks

- Spectral methods work well, give best known bounds for (planted) partition problems
- Does an angle-based approach of the s largest eigenvectors give the same bounds? Seems to perform as well as projector ...

$$A_1 = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix}$$

$$A_2 = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \end{bmatrix}$$

Open questions

- In reconstruction problems less theoretical work about different matrices (Laplacian, normalized Laplacian). Do similar bounds hold?
- Can one find $\Theta(\sqrt{n})$ planted partitions?
- Can one find cliques of size $k = o(\sqrt{n})$? Major drawback: The 2-norm is a very global measure, and hard to use algorithmically (in the analysis of an algorithm)