Chapter 6
Spectral Methods
Part II: Finding planted patterns
Assume a perfect pattern
Whoops!

Detail from The Creation of Adam by Michelangelo c. 1512.
Can we find the original pattern?
To find a pattern  
or  
To find the pattern

That is the question
Planted patterns

• Most data mining algorithms promise to find some pattern(s)
  • Or exhaustively list all of them
• Few can promise to find the pattern, even if we’re promised there’s one
  • Data mining concentrates on discovery, not recovery
Planted Bicliques and Nuclear Norms
Schatten norms

- The **Schatten matrix norms** for $p \geq 1$ are defined as $\left( \sum_{i=1}^{\min\{n,m\}} \sigma_i^p \right)^{1/p}$
  - $\sigma_i$ are the singular values of $A = U\Sigma V^T$
- $p = 2 \Rightarrow$ Frobenius norm
- $p = \infty \Rightarrow$ operator norm
- $p = 1 \Rightarrow$ nuclear norm $||A||^*$
  - Also $||A||^* = \text{tr}(\Sigma) = \text{tr}(\sqrt{A^T A})$
Maximum clique as rank minimization

- Maximum $n$-vertex clique in graph $G = (V, E)$ can be found with the following program

\[
\begin{align*}
\text{min} & \quad \text{rank}(X) \\
\text{s.t.} & \quad \sum_{i \in V} \sum_{j \in V} x_{ij} \geq n^2 \\
& \quad x_{ij} = 0 \quad \text{if } \{i, j\} \notin E \text{ and } i \neq j \\
& \quad X = X^T \\
& \quad X \in [0, 1]^{V \times V}
\end{align*}
\]

A clique is a rank-1 submatrix

- Proper submatrix
- Symmetric
- No entry larger than 1
Nuclear norm relaxation

- The rank minimization problem is NP-hard
- We can relax it to nuclear norm minimization:
  \[
  \min_{X} \|X\|_* \quad \text{s.t.} \quad \sum_{i \in V} \sum_{j \in V} x_{ij} \geq n^2 \]  
  \( x_{ij} = 0 \) if \( \{i, j\} \notin E \) and \( i \neq j \)
- The maximum clique is a valid solution and the unique optimizer under certain conditions
- When this is the case, we can find the clique
Adversarial case

• Assume we have a graph that contains only a clique of $n$ nodes

• Adversary adds up to $\epsilon n^2$ edges, $\epsilon < 1/2$

• The vertices not in the clique are adjacent to at most $\delta n$ vertices in the clique for some $0 < \delta < 1$

• The original clique is still the unique optimizer

O/w there's a larger clique

O/w the clique is enlarged
Randomized case

• Assume the extra edges are added i.i.d. with probability $p \in [0, 1)$

• **Thm.** There exists an $\alpha > 0$ s.t. with $n \geq \alpha\sqrt{N}$, the planted clique is the unique optimizer with probability tending exponentially to 1 as $N \to \infty$

• $\alpha$ depends on $p$, $n$ is the size of the clique, and $N$ is the size of the graph
Bipartite graphs and bicliques

• A **biclique** is a binary rank-1 submatrix of the binary **bi-adjacency matrix**

• Biclique of size $n$-by-$m$ can be found solving

$$\min \quad \text{rank}(X)$$

s.t.
$$\sum_{i \in V} \sum_{j \in V} x_{ij} \geq nm$$

$$x_{ij} = 0 \quad \text{if} \quad \{i, j\} \in (U \times V) \setminus E$$

$$X \in [0, 1]^{V \times V}$$
Nuclear norm relaxation

\[
\begin{align*}
\min & \quad \|X\|_* \\
\text{s.t.} & \quad \sum_{i \in V} \sum_{j \in V} x_{ij} \geq nm \\
& \quad x_{ij} = 0 \quad \text{if } \{i, j\} \in (U \times V) \setminus E
\end{align*}
\]

- The maximum biclique is again the (unique) minimizer under certain conditions
- Problem is, when can we show the conditions hold
Results

• Adversary can add at most $O(nm)$ edges

• No new vertex can touch too many vertices in the biclique

• We can add edges i.i.d. as long as the biclique is $\alpha\sqrt{N}$ for some $\alpha$ depending on $p$ and the relation of $n$ and $m$ and $|V|$ and $|U|$
Bicliques with Destructive Noise
Can we find the original pattern?
Destructive noise

• So far we’ve only considered the case where new edges are added
  • New 1s in to the (bi-)adjacency matrix
  • We observe \( A' = A \cup N \)

• But in reality the noise can also destroy existing edges
  • Now we have the original biclique matrix \( A \), noise matrix \( N \), and observed matrix \( A' = A \oplus N \)
Rebuilding the biclique

• We consider the maximum-similarity/minimum-dissimilarity quasi-biclique

  • I.e. rank-1 binary $B$ minimizing $||A' - B||_F$

• Finding such $B$ is NP-hard

  • 2-approximation algorithms for minimum dissimilarity

  • PTAS for maximum similarity
Noise models

• So far we’ve added each edge independently with probability $p$
  • Erdős–Rényi random graph model
• We can also follow the preferential attachment model
  • Barabási–Albert random graph model
  • Some vertices have big changes on neighbors, others less
    • If the noise follows the B–A model, it can’t have large bicliques ⇒ easy
Intimidating Math

Let $\text{dist}(G, \tilde{G}) = \max\{|U \oplus \tilde{U}|, |V \oplus \tilde{V}|\}$

where

$$A \oplus B = (A \setminus B) \cup (B \setminus A)$$

If $\forall X, Y : \Pr[q(X, Y) < q(U', V')] \leq \exp\{|(X, Y) - (U', V')| c\}$

then

$$\forall \varepsilon > 0 \forall U', V' (\min\{|U'|, |V'|\} \geq \zeta) : \Pr(\text{dist}(G, G^*) \leq \varepsilon) \geq 1 - \delta_1 - \delta_2$$

with $\delta_2 = T(\varepsilon, |U'|, |V'|, |U'|, |V'|)T(\varepsilon, N, M, |U'|, |V'|)$

where

$$T(\varepsilon, a, b, c, d) = \frac{\exp (\varepsilon (\log (a + 1) + \log (b + 1) - \min (c, d))c_{p,q})}{1 - \exp ((\log(a + 1) + \log(b + 1) - \min(c, d))c_{p,q})}$$
Results

• Erdős–Rényi: The minimum size of the original biclique $\zeta = \log(NM)$

• Barabási–Albert: $\log N \ll \zeta \ll \sqrt{N}$
Example results

What the algorithm finds

What is the underlying structure
Summary

• We can find planted cliques and bicliques (and other patterns)
  • Under certain conditions
• Spectral methods can be proven to work
• Nuclear norm relaxes rank
• Sometimes we might have to solve NP-hard problems
Literature

