Chapter 5

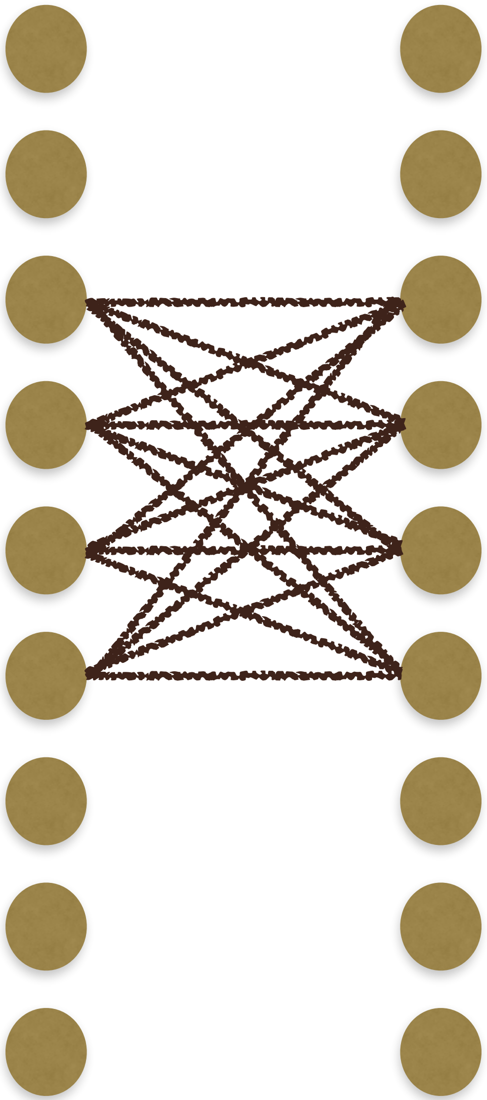
Decompositions for Combinatorial Structures

Part III: Finding planted patterns

max planck institut
informatik

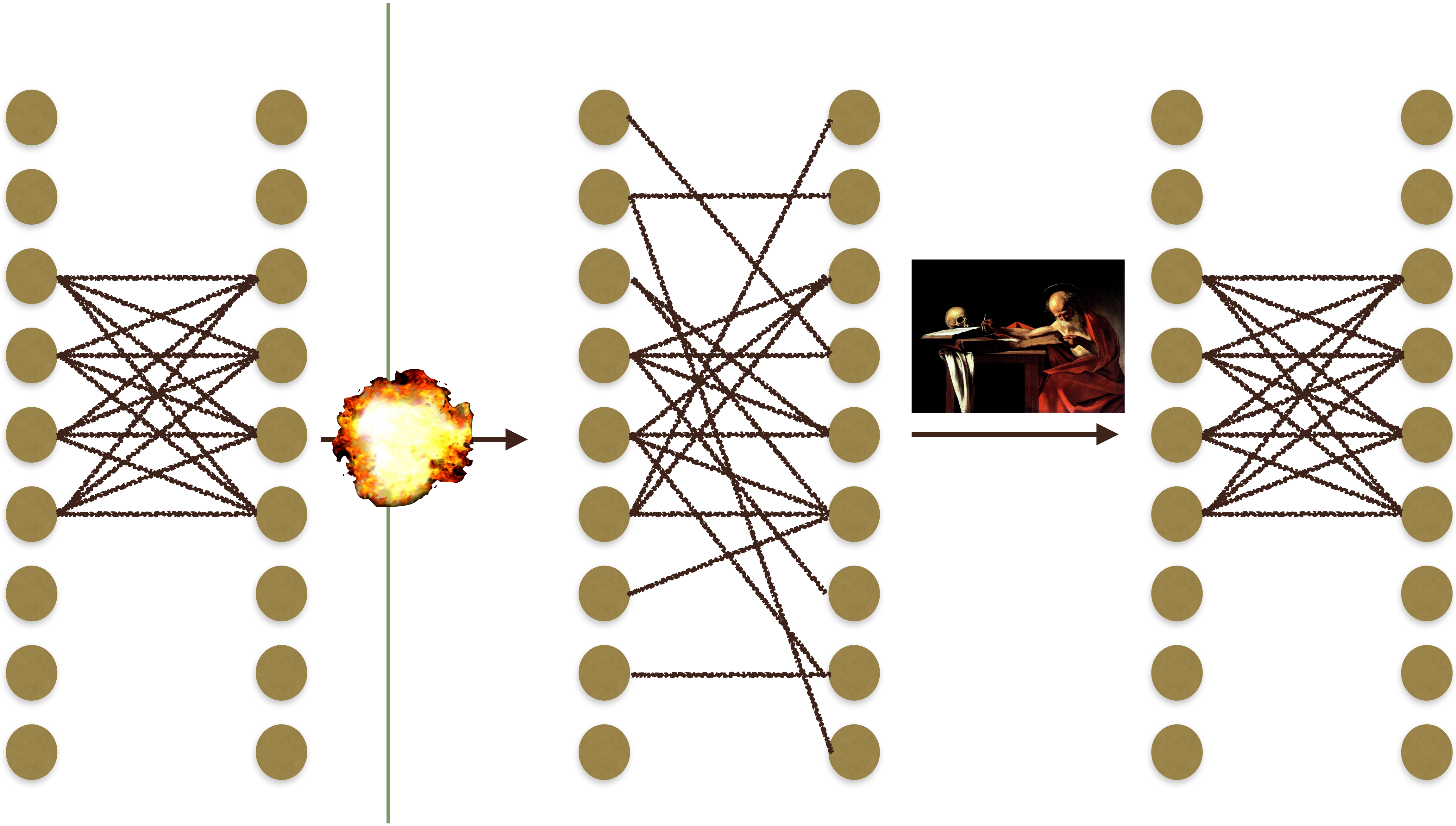
Motivation

Assume a perfect pattern



Detail from
The Creation of Adam
Michelangelo c. 1512

Can we find the original pattern?



To find *a* pattern
or
To find *the* pattern

Public domain / The Guardian

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 Partitions

McSherry 2001

Random graph models with planted partitions

- **Bisection**: Include every intra-part edge with probability q and each inter-part edge with probability $p < q$
- ***k*-Coloring**: Include each edge with probability p and then remove all intra-color edges
- **Clique**: Include each edge with probability p and complete the clique

Planted partition model

- Let $G(\varphi, \boldsymbol{P})$ be a random graph distribution where $\varphi : V \to \{1,\ldots,k\}$ partition the vertices to k classes and $\boldsymbol{P}=(p_{ij})$ is a k-by-k matrix with $p_{ij} \in [0,1]$. Include edge (i, j) with probability $p_{\varphi(i)\varphi(j)}$.
- **Example**: planted clique. Let $\varphi(v) = 1$ iff v is in the clique. Set $p_{11} = 1$ and $p_{ij} = p$ elsewhere
- **Problem**. Given a sample $\boldsymbol{G}'$ from $G(\varphi, \boldsymbol{P})$, find a partition φ' s.t. $\varphi'(\boldsymbol{v}) = \varphi'(\boldsymbol{u})$ iff $\varphi(\boldsymbol{v}) = \varphi(\boldsymbol{u})$

Approach to planted partition

- If we know $\boldsymbol{G} = E[\boldsymbol{G}']$, finding φ is easy
  - Cluster the columns $\boldsymbol{g}_u$ of $\boldsymbol{G}$
- We do not have $\boldsymbol{G}$ but we have the following
  - Matrix $\boldsymbol{G}$ has rank k
  - If $\boldsymbol{P}_{\boldsymbol{G}}$ is the projection on the column space of $\boldsymbol{G}$, then $||\boldsymbol{P}_{\boldsymbol{G}}(\boldsymbol{g}_u) - \boldsymbol{g}_u|| = 0$ and $||\boldsymbol{P}_{\boldsymbol{G}}(\boldsymbol{g}_u - \boldsymbol{g}'_u)||$ is small
    - By the triangle inequality $\boldsymbol{P}_{\boldsymbol{G}}(\boldsymbol{g}'_u)$ is almost $\boldsymbol{g}_u$

More on projections

- Recall: If the columns of $\boldsymbol{Q}$ are the orthonormal basis of a subspace S, $\boldsymbol{P}_S = \boldsymbol{Q}\boldsymbol{Q}^T$ is the **orthogonal projection onto** S
  - $\boldsymbol{P}_S\boldsymbol{x}$ is the closest vector of $\boldsymbol{x}$ that's in S
- We do not know $\boldsymbol{P}_{\boldsymbol{G}}$, but any projection $\boldsymbol{P}$ suffices if it satisfies
  - $||\boldsymbol{P}(\boldsymbol{g}_u) - \boldsymbol{g}_u||$ is small and $||\boldsymbol{P}(\boldsymbol{g}_u - \boldsymbol{g}'_u)||$ is small
  - Now $\boldsymbol{P}(\boldsymbol{g}'_u) \approx \boldsymbol{g}_u$, and we can find φ

SVD gives projections

- Let $\boldsymbol{P}$ be the projection onto the first k left singular vectors, $\boldsymbol{P} = \boldsymbol{U}_k\boldsymbol{U}k^T$
- With probability at least $1 - \delta$
  - $||\boldsymbol{P}(\boldsymbol{g}_u) - \boldsymbol{g}_u|| \leq 8\sigma(nk/s_u)^{1/2}$
  - $||\boldsymbol{P}(\boldsymbol{g}_u - \boldsymbol{g}'_u)|| \leq (2k \log(n/\delta))^{1/2}$
    - σ is an upper bound on the variance of the entries in $\boldsymbol{G}$, n is the number of vertices, k is the number of classes, and s_u is the size of the class vertex u belongs to

Better projections

- Now if $||\boldsymbol{g}_u - \boldsymbol{g}_v||$ is large enough when $\varphi(v) \neq \varphi(u)$, we can find φ
  - Depends on the above error bounds
- With more complicated error bounds we get:
  - If s is the size of a planted clique, then there is a constant c s.t. for sufficiently large n we can recover φ with probability $1 - \delta$ if

$$\frac{1-p}{p} > c\left(\frac{n}{s^2} + \frac{\log(n/\delta)}{s}\right)$$

Planted Bicliques and Nuclear Norms

Ames & Vavasis 2011

Schatten norms

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

Maximum clique as rank minimization

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

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

A clique is a rank-1 submatrix of size n-by-n

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:

$$\begin{aligned} \min \quad & \|\boldsymbol{X}\|_* \\ \text{s.t.} \quad & \sum_{i \in V} \sum_{j \in V} x_{ij} \geq n^2 \\ & x_{ij} = 0 \quad \text{if } \{i, j\} \notin E \text{ and } i \neq j \end{aligned}$$

← can be replaced with 1

  - 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 ϵn^2 edges, $\epsilon < 1/2$

    o/w there's a larger clique
  - The vertices not in the clique are adjacent to at most δn vertices in the clique for some $0 < \delta < 1$

    o/w the clique is enlarged
- The original clique is still the unique optimizer

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 $N \to \infty$
  - α depends on p, n is the size of the clique, and N is the size of the graph

Bipartite graphs and bicliques

- Recall: A **biclique** is a binary rank-1 submatrix of the binary **bi-adjacency matrix**
  - Biclique of size *n*-by-*m* can be found solving

$$
\begin{aligned}
\min \quad & \operatorname{rank}(\boldsymbol{X}) \\
\text{s.t.} \quad & \sum_{i \in V} \sum_{j \in V} x_{ij} \geq nm \\
& x_{ij} = 0 \quad \text{if } \{i, j\} \in (U \times V) \setminus E \\
& \boldsymbol{X} \in [0, 1]^{V \times V}
\end{aligned}
$$

Nuclear norm relaxation

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

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

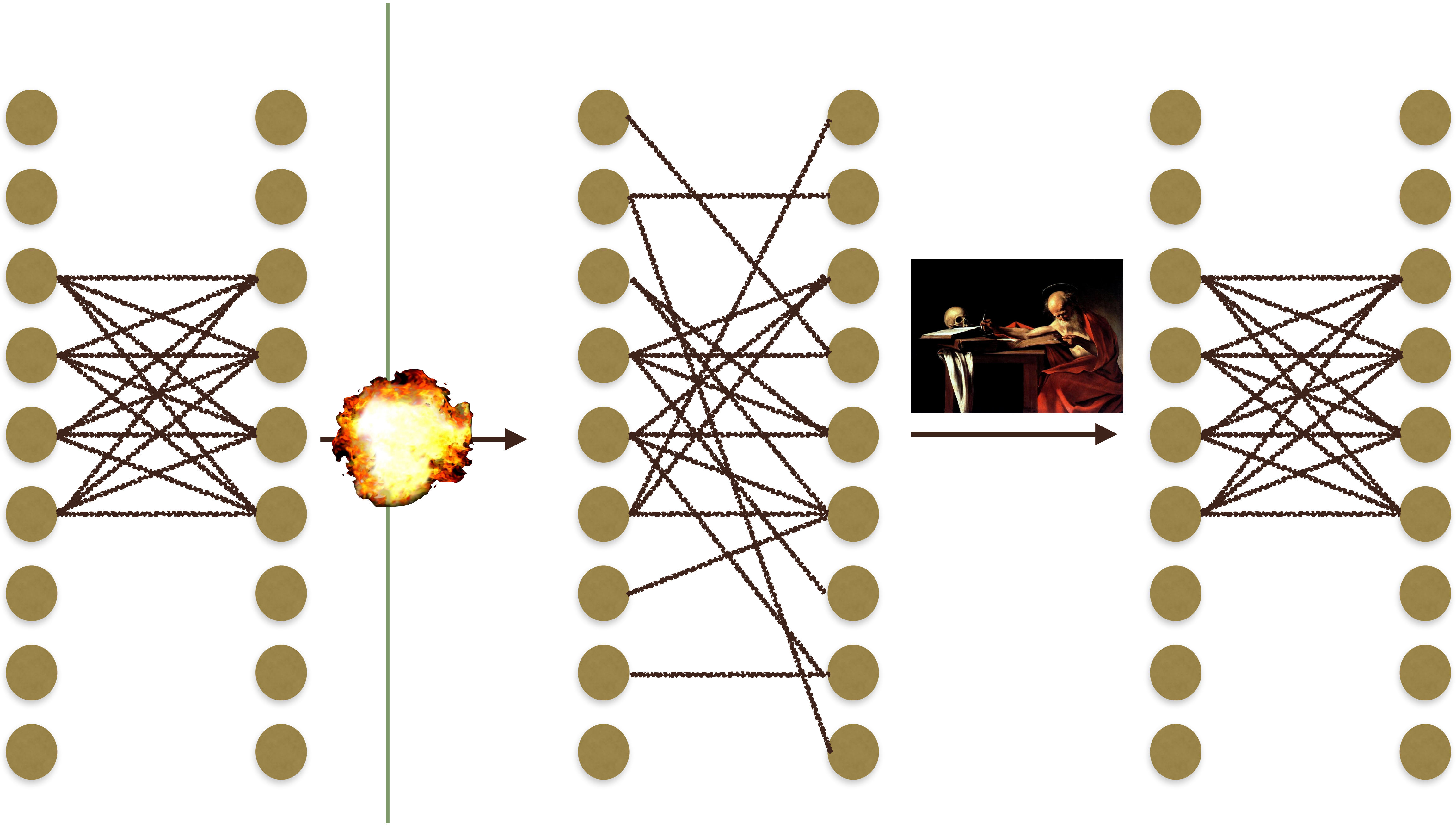
Results

- 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 α depending on p and the relation of n and m and $|V|$ and $|U|$

Bicliques with Destructive Noise

Ramon, Miettinen & Vreeken 2013

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 $\boldsymbol{A}' = \boldsymbol{A} \cup \boldsymbol{N}$
- But in reality the noise can also destroy existing edges
  - Now we have the original biclique matrix $\boldsymbol{A}$, noise matrix $\boldsymbol{N}$, and observed matrix $\boldsymbol{A}' = \boldsymbol{A} \oplus \boldsymbol{N}$

Rebuilding the biclique

- We consider the **maximum-similarity/ minimum-dissimilarity quasi-biclique**
  - I.e. rank-1 binary $\boldsymbol{B}$ minimizing $||\boldsymbol{A}' - \boldsymbol{B}||_F$
- Finding such $\boldsymbol{B}$ 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 $$\mathrm{dist}(G, \widetilde{G}) = \max\{|U \oplus \widetilde{U}|, |V \oplus \widetilde{V}|\}$$

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

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

then

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

with $$\delta_2 = T(\epsilon, \left|U'\right|, \left|V'\right|, \left|U'\right|, \left|V'\right|) T(\epsilon, N, M, \left|U'\right|, \left|V'\right|)$$

where

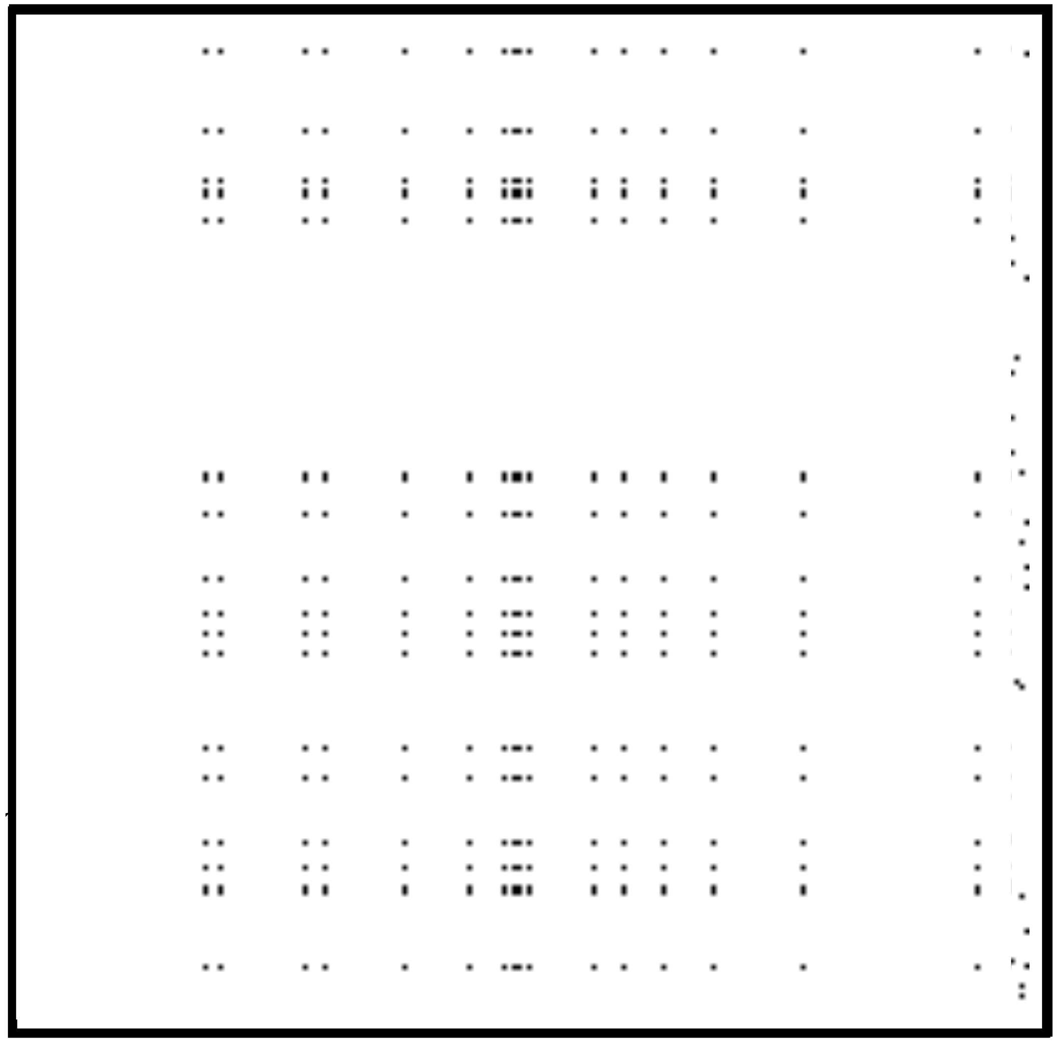
$$T(\epsilon, a, b, c, d) = \frac{\exp\left(\epsilon\left(\log\left(a+1\right) + \log\left(b+1\right) - \min\left(c, d\right)\right) c_{p,q}\right)}{1 - \exp\left(\left(\log(a+1) + \log(b+1) - \min(c,d)\right) c_{p,q}\right)}$$

Results

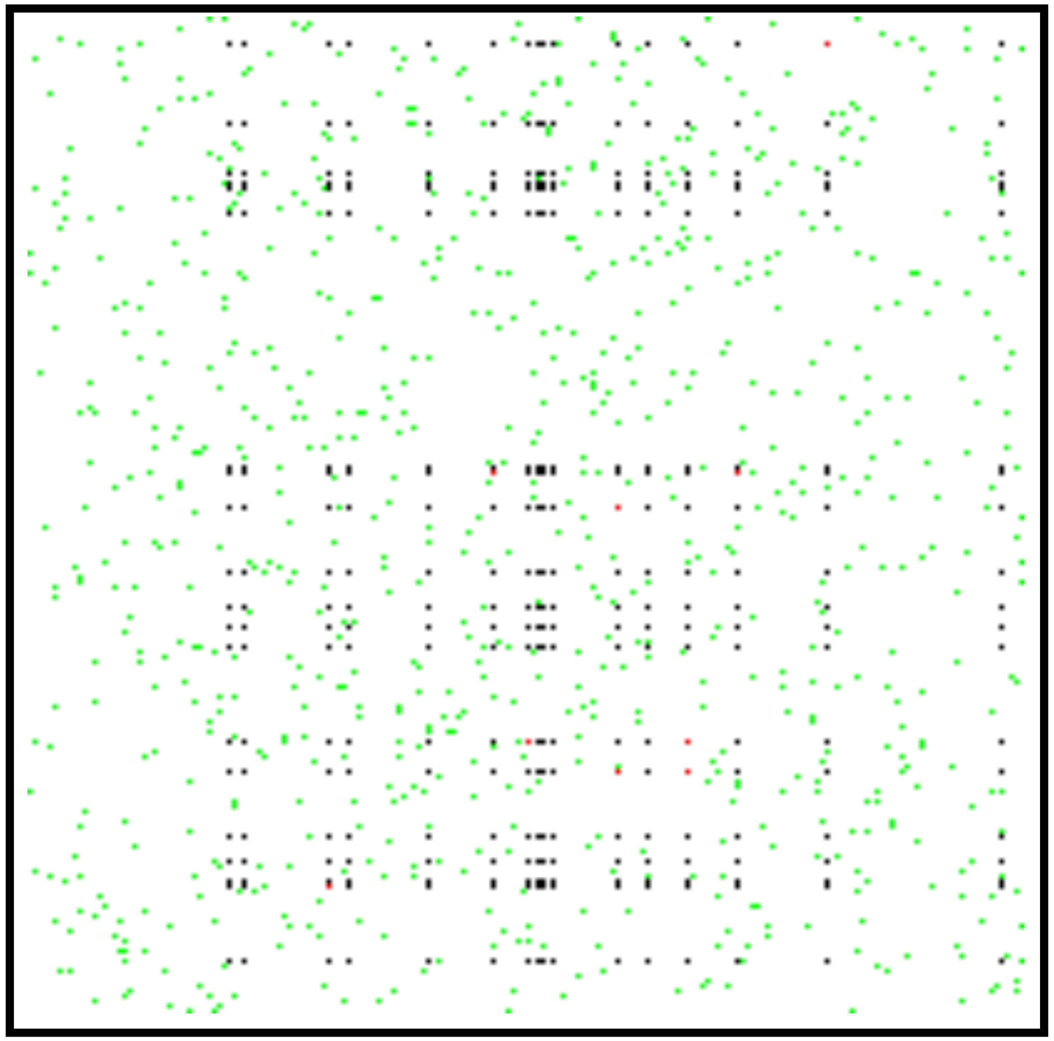
- Erdős–Rényi: The minimum size of the original biclique $\zeta = \log(NM)$
- Barabási–Albert: $\log N \ll \zeta \ll \sqrt{N}$

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

- McSherry, F., 2001. *Spectral partitioning of random graphs*. In 24th IEEE Symposium on Foundations of Computer Science, pp. 529–537.
- Ames, B.P.W. & Vavasis, S.A., 2011. *Nuclear norm minimization for the planted clique and biclique problems*. Mathematical Programming, Series B, 129(1), pp.69–89.
- Ramon, J., Miettinen, P. & Vreeken, J., 2013. *Detecting Bicliques in GF[q]*. In 2013 European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, pp. 509–524.

Next week: Ask Me Anything

- Next week's lecture: wrap up, random thoughts, and Ask Me Anything
  - Related to the course or not
    - I don't promise to answer to everything
  - Questions sent beforehand by email have higher changes of getting (sensible) answers