Role extraction for digraphs via neighbourhood pattern similarity

Giovanni Barbarino [1] Vanni Noferini [1] Paul Van Dooren [2]

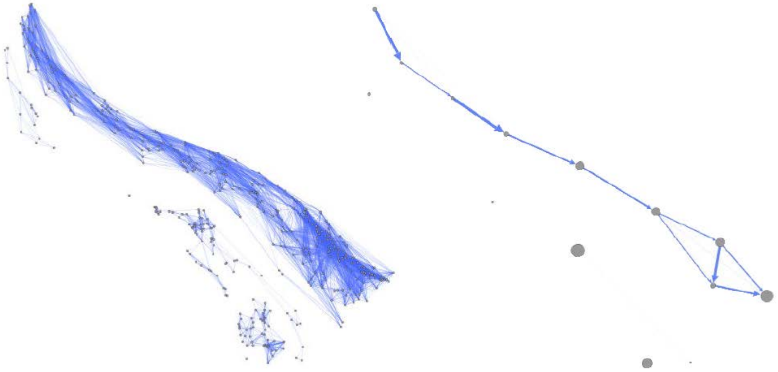
Foundations of Computational Mathematics
Sorbonne University - 21 June 2023

[1]Department of Mathematics and Systems Analysis, Aalto University
[2]Department Mathematical Engineering, Université Catholique de Louvain

A motivating example

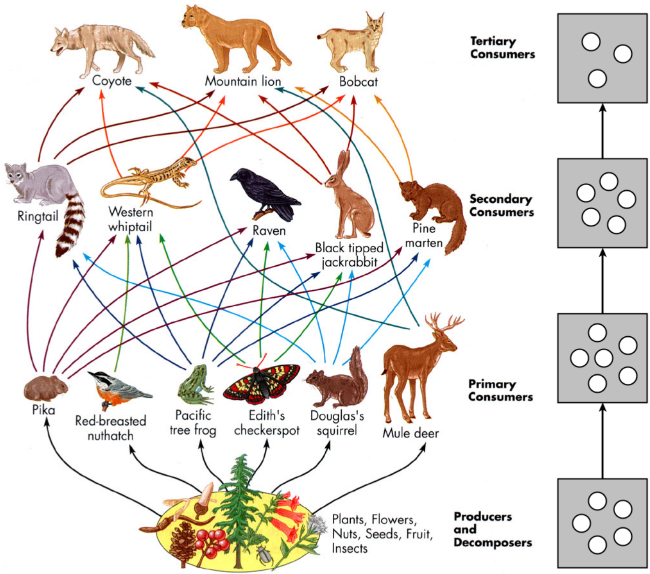
Once a year the corals in the Great Barrier Reef reproduce.
Corals release their eggs and sperm into the water at the same time.
Clouds of coral eggs and sperm float in all directions, carried by the currents, winds, and waves.
When egg and sperm meet, the resulting larvae continue to drift to find the perfect spot to settle.

A motivating example



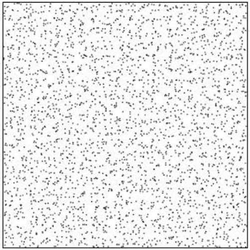
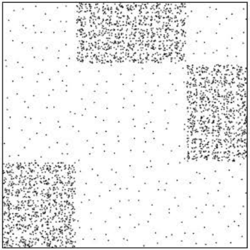
Understanding the underlying directed structure lets us control and predict the growth of the reef

A motivating example



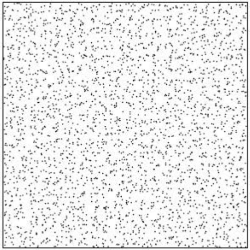
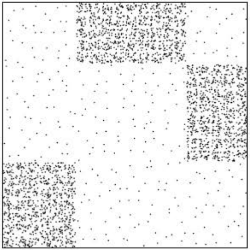
Roles of Directed Graph

Given a graph with adjacency matrix A we want to find an assignment function, that partitions the graph into **Roles**

$A =$  $\xRightarrow{P?} PAP^T =$ 

Roles of Directed Graph

Given a graph with adjacency matrix A we want to find an assignment function, that partitions the graph into **Roles**

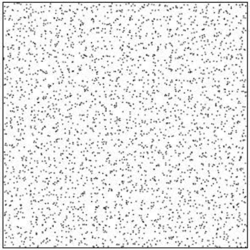
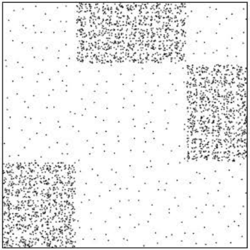
$A =$  $\overset{P?}{\Longrightarrow} PAP^T =$ 

We also want a matrix B telling us how the roles are connected

$$B = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

Roles of Directed Graph

Given a graph with adjacency matrix A we want to find an assignment function, that partitions the graph into **Roles**

$A =$  $\overset{P?}{\Longrightarrow} PAP^T =$ 

We also want a matrix B telling us how the roles are connected

$$B = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

In the ideal case, all nodes in the same role are **structurally equivalent**:

their children are the same

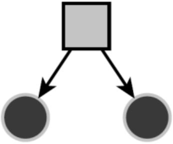
their parents are the same

Neighbourhood Pattern Similarity Measure

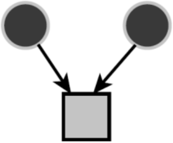
The **Neighbourhood Pattern Similarity Measure** between two nodes (i, j) takes into consideration all their common 'ancestors', 'descendants'... and 'relatives'

Neighbourhood Pattern Similarity Measure

The **Neighbourhood Pattern Similarity Measure** between two nodes (i, j) takes into consideration all their common 'ancestors', 'descendants'... and 'relatives'



$$A^T A$$



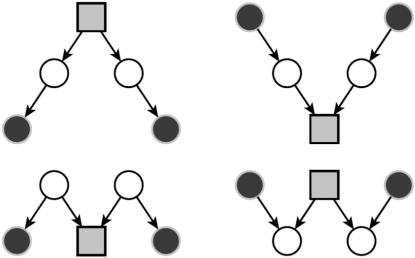
$$A A^T$$

Neighbourhood Pattern Similarity Measure

The **Neighbourhood Pattern Similarity Measure** between two nodes (i, j) takes into consideration all their common 'ancestors', 'descendants'... and 'relatives'

$A^T \ A^T \ A \ A$

$A^T \ A \ A^T \ A$

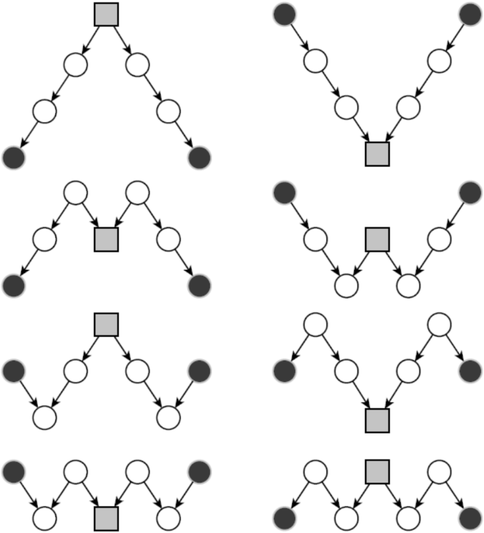


$A \ A \ A^T \ A^T$

$A \ A^T \ A \ A^T$

Neighbourhood Pattern Similarity Measure

The **Neighbourhood Pattern Similarity Measure** between two nodes (i, j) takes into consideration all their common 'ancestors', 'descendants'... and 'relatives'



Neighbourhood Pattern Similarity Measure

The **Neighbourhood Pattern Similarity Measure** between two nodes (i, j) takes into consideration all their common 'ancestors', 'descendants'... and 'relatives'

$$L_1 = AA^T + A^T A = \begin{bmatrix} A & A^T \end{bmatrix} \begin{bmatrix} A^T \\ A \end{bmatrix}$$

$$L_2 = AAA^T A^T + AA^T AA^T + A^T AAA^T + A^T A^T AA = \begin{bmatrix} A & A^T \end{bmatrix} \begin{bmatrix} L_1 & \\ & L_1 \end{bmatrix} \begin{bmatrix} A^T \\ A \end{bmatrix}$$

$$\vdots$$

$$\Gamma_A^{k+1}[I] = \Gamma_A[L_k] := L_{k+1} = \begin{bmatrix} A & A^T \end{bmatrix} \begin{bmatrix} L_k & \\ & L_k \end{bmatrix} \begin{bmatrix} A^T \\ A \end{bmatrix}$$

Neighbourhood Pattern Similarity Measure

The **Neighbourhood Pattern Similarity Measure** between two nodes (i, j) takes into consideration all their common 'ancestors', 'descendants'... and 'relatives'

$$L_1 = AA^T + A^T A = \begin{bmatrix} A & A^T \end{bmatrix} \begin{bmatrix} A^T \\ A \end{bmatrix}$$

$$L_2 = AAA^T A^T + AA^T AA^T + A^T AAA^T + A^T A^T AA = \begin{bmatrix} A & A^T \end{bmatrix} \begin{bmatrix} L_1 & \\ & L_1 \end{bmatrix} \begin{bmatrix} A^T \\ A \end{bmatrix}$$

$$\vdots$$

$$\Gamma_A^{k+1}[I] = \Gamma_A[L_k] := L_{k+1} = \begin{bmatrix} A & A^T \end{bmatrix} \begin{bmatrix} L_k & \\ & L_k \end{bmatrix} \begin{bmatrix} A^T \\ A \end{bmatrix}$$

Neighbourhood Pattern Similarity Measure

The **Neighbourhood Pattern Similarity Measure** between two nodes (i, j) takes into consideration all their common 'ancestors', 'descendants'... and 'relatives'

$$L_1 = AA^T + A^T A = \begin{bmatrix} A & A^T \end{bmatrix} \begin{bmatrix} A^T \\ A \end{bmatrix}$$

$$L_2 = AAA^T A^T + AA^T AA^T + A^T AAA^T + A^T A^T AA = \begin{bmatrix} A & A^T \end{bmatrix} \begin{bmatrix} L_1 & \\ & L_1 \end{bmatrix} \begin{bmatrix} A^T \\ A \end{bmatrix}$$

$$\vdots$$

$$\Gamma_A^{k+1}[I] = \Gamma_A[L_k] := L_{k+1} = \begin{bmatrix} A & A^T \end{bmatrix} \begin{bmatrix} L_k & \\ & L_k \end{bmatrix} \begin{bmatrix} A^T \\ A \end{bmatrix}$$

Neighbourhood Pattern Similarity Measure

The **Neighbourhood Pattern Similarity Measure** between two nodes (i, j) takes into consideration all their common 'ancestors', 'descendants'... and 'relatives'

$$L_1 = AA^T + A^T A = \begin{bmatrix} A & A^T \end{bmatrix} \begin{bmatrix} A^T \\ A \end{bmatrix}$$

$$L_2 = AAA^T A^T + AA^T AA^T + A^T AAA^T + A^T A^T AA = \begin{bmatrix} A & A^T \end{bmatrix} \begin{bmatrix} L_1 & \\ & L_1 \end{bmatrix} \begin{bmatrix} A^T \\ A \end{bmatrix}$$

$$\vdots$$

$$\Gamma_A^{k+1}[I] = \Gamma_A[L_k] := L_{k+1} = \begin{bmatrix} A & A^T \end{bmatrix} \begin{bmatrix} L_k & \\ & L_k \end{bmatrix} \begin{bmatrix} A^T \\ A \end{bmatrix}$$

$$S_{k+1} := \Gamma_A[I] + \beta^2 \Gamma_A^2[I] + \cdots + \beta^{2(k+1)} \Gamma_A^{k+1}[I] = \Gamma_A[I + \beta^2 S_k]$$

$$S_{k+1} - S_k = \beta^{2(k+1)} \Gamma_A^{k+1}[I] \succeq 0 \implies S_{k+1} \succeq S_k$$

Neighbourhood Pattern Similarity Measure

$$S_1 := \begin{bmatrix} A & A^T \end{bmatrix} \begin{bmatrix} A^T \\ A \end{bmatrix} \qquad S_{k+1} = \Gamma_A[I + \beta^2 S_k]$$

- S_k are always PSD matrices and $S_{k+1} \succeq S_k$
- If $\beta^2 < \frac{1}{4\|A\|^2}$, then $S_k \to S^*$ with

$$vec(S^*) = \left[I - \beta^2 \left(A \otimes A + (A \otimes A)^T \right) \right]^{-1} vec(S_1)$$

- In the ideal case, if $[B \ B^T]$ has maximum rank, then the rank of each S_k is the number of roles and a spectral method on S_k let us recover the roles

Neighbourhood Pattern Similarity Measure

$$S_1 := \begin{bmatrix} A & A^T \end{bmatrix} \begin{bmatrix} A^T \\ A \end{bmatrix} \qquad S_{k+1} = \Gamma_A[I + \beta^2 S_k]$$

- S_k are always PSD matrices and $S_{k+1} \succeq S_k$
- If $\beta^2 < \frac{1}{4||A||^2}$, then $S_k \to S^*$ with

$$vec(S^*) = \left[I - \beta^2 \left(A \otimes A + (A \otimes A)^T \right) \right]^{-1} vec(S_1)$$

- In the ideal case, if $[B \; B^T]$ has maximum rank, then the rank of each S_k is the number of roles and a spectral method on S_k let us recover the roles

Neighbourhood Pattern Similarity Measure

$$S_1 := \begin{bmatrix} A & A^T \end{bmatrix} \begin{bmatrix} A^T \\ A \end{bmatrix} \qquad S_{k+1} = \Gamma_A[I + \beta^2 S_k]$$

- S_k are always PSD matrices and $S_{k+1} \succeq S_k$
- If $\beta^2 < \frac{1}{4||A||^2}$, then $S_k \to S^*$ with

$$vec(S^*) = \left[I - \beta^2 \left(A \otimes A + (A \otimes A)^T\right)\right]^{-1} vec(S_1)$$

- In the ideal case, if $[B \ B^T]$ has maximum rank, then the rank of each S_k is the number of roles and a spectral method on S_k let us recover the roles

Neighbourhood Pattern Similarity Measure

$$S_1 := \begin{bmatrix} A & A^T \end{bmatrix} \begin{bmatrix} A^T \\ A \end{bmatrix} \qquad S_{k+1} = \Gamma_A[I + \beta^2 S_k]$$

- S_k are always PSD matrices and $S_{k+1} \succeq S_k$
- If $\beta^2 < \frac{1}{4||A||^2}$, then $S_k \to S^*$ with

$$vec(S^*) = \left[I - \beta^2 \left(A \otimes A + (A \otimes A)^T\right)\right]^{-1} vec(S_1)$$

- In the ideal case, if $[B\ B^T]$ has maximum rank, then the rank of each S_k is the number of roles and a spectral method on S_k let us recover the roles

Warning: The number of roles may not be linked to the rank of A or B

$$B = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix} \implies \text{rk}(A) = \text{rk}(B) = 2$$

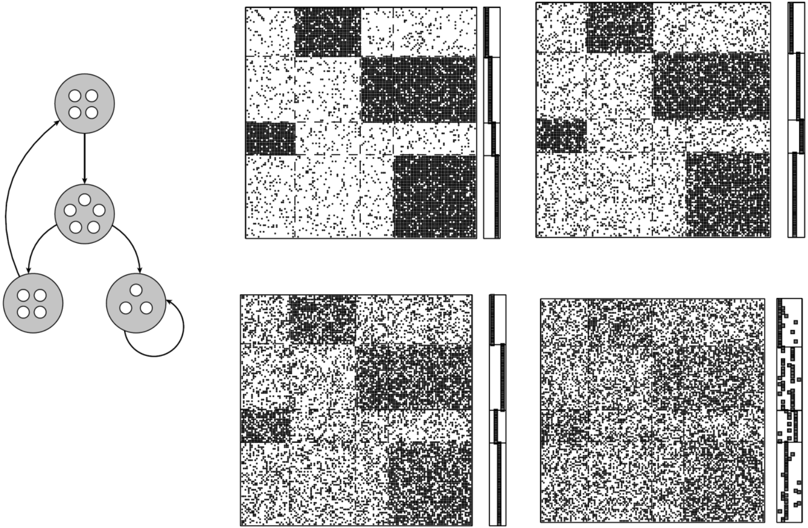
$$[B\ B^T] = \begin{bmatrix} 0 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 1 \end{bmatrix} \implies \text{rk}(S_k) = \text{rk}([B\ B^T]) = 3$$

Erdös-Renyi Random Graphs

Suppose now that G is a random graph induced by B where the probability of an edge between nodes from role i to role j is p if $B_{i,j} = 1$ and $1 - p$ if $B_{i,j} = 0$

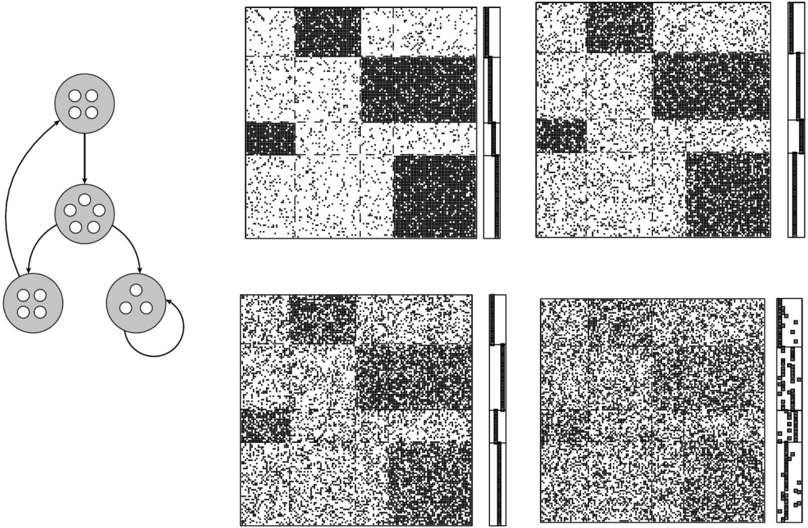
Erdös-Renyi Random Graphs

Suppose now that G is a random graph induced by B where the probability of an edge between nodes from role i to role j is p if $B_{i,j} = 1$ and $1 - p$ if $B_{i,j} = 0$



Erdös-Renyi Random Graphs

Suppose now that G is a random graph induced by B where the probability of an edge between nodes from role i to role j is p if $B_{i,j} = 1$ and $1 - p$ if $B_{i,j} = 0$



NPS Measure can empirically identify the roles

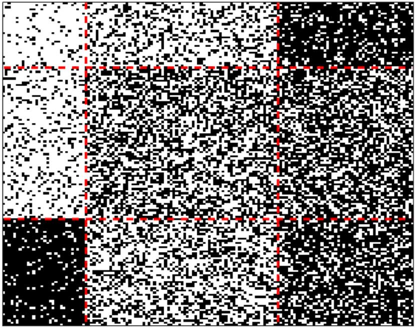
Stochastic Block Model

- The nodes are partitioned in q roles $\mathcal{C}_i$ of size $m_i n$, $m := \sum_i m_i$
- There is an edge between nodes in $\mathcal{C}_i$ and $\mathcal{C}_j$ independently with probability $B_{i,j}$

Stochastic Block Model

- The nodes are partitioned in q roles $\mathcal{C}_i$ of size $m_i n$, $m := \sum_i m_i$
- There is an edge between nodes in $\mathcal{C}_i$ and $\mathcal{C}_j$ independently with probability $B_{i,j}$

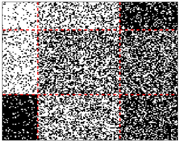
$$B = \begin{bmatrix} .1 & .3 & .8 \\ .2 & .5 & .6 \\ .9 & .4 & .7 \end{bmatrix}$$



Stochastic Block Model

- The nodes are partitioned in q roles $\mathcal{C}_i$ of size $m_i n$, $m := \sum_i m_i$
- There is an edge between nodes in $\mathcal{C}_i$ and $\mathcal{C}_j$ independently with probability $B_{i,j}$

$$B = \begin{bmatrix} .1 & .3 & .8 \\ .2 & .5 & .6 \\ .9 & .4 & .7 \end{bmatrix}$$



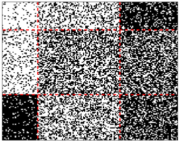
We work asymptotically in n, so we add some additional hypotheses:

- $0 < m_{min} \leq m_i \leq m_{max}$ for any i, and m_{min}, m_{max} are independent from n
- $B_{i,j} = \Psi_{i,j} f(n)$, $0 \leq \Psi_{i,j} \leq 1$ where the matrix Ψ is independent from n
- $[B\ B^T] = [\Psi\ \Psi^T] f(n)$ has full rank, equal to the number of roles q
- $nf(n) \to \infty$ that agrees with standard results for which $f(n) = \Omega(1/n)$ for **exact recovery**

Stochastic Block Model

- The nodes are partitioned in q roles $\mathcal{C}_i$ of size $m_i n$, $m := \sum_i m_i$
- There is an edge between nodes in $\mathcal{C}_i$ and $\mathcal{C}_j$ independently with probability $B_{i,j}$

$$B = \begin{bmatrix} .1 & .3 & .8 \\ .2 & .5 & .6 \\ .9 & .4 & .7 \end{bmatrix}$$



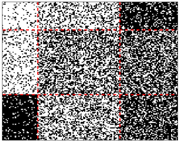
We work asymptotically in n, so we add some additional hypotheses:

- $0 < m_{min} \leq m_i \leq m_{max}$ for any i, and m_{min}, m_{max} are independent from n
- $B_{i,j} = \Psi_{i,j} f(n)$, $0 \leq \Psi_{i,j} \leq 1$ where the matrix Ψ is independent from n
- $[B \ B^T] = [\Psi \ \Psi^T] f(n)$ has full rank, equal to the number of roles q
- $nf(n) \to \infty$ that agrees with standard results for which $f(n) = \Omega(1/n)$ for **exact recovery**

Stochastic Block Model

- The nodes are partitioned in q roles $\mathcal{C}_i$ of size $m_i n$, $m := \sum_i m_i$
- There is an edge between nodes in $\mathcal{C}_i$ and $\mathcal{C}_j$ independently with probability $B_{i,j}$

$$B = \begin{bmatrix} .1 & .3 & .8 \\ .2 & .5 & .6 \\ .9 & .4 & .7 \end{bmatrix}$$



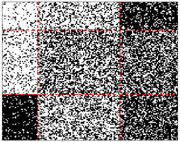
We work asymptotically in n, so we add some additional hypotheses:

- $0 < m_{min} \leq m_i \leq m_{max}$ for any i, and m_{min}, m_{max} are independent from n
- $B_{i,j} = \Psi_{i,j} f(n)$, $0 \leq \Psi_{i,j} \leq 1$ where the matrix Ψ is independent from n
- $[B\ B^T] = [\Psi\ \Psi^T] f(n)$ has full rank, equal to the number of roles q
- $nf(n) \to \infty$ that agrees with standard results for which $f(n) = \Omega(1/n)$ for **exact recovery**

Stochastic Block Model

- The nodes are partitioned in q roles $\mathcal{C}_i$ of size $m_i n$, $m := \sum_i m_i$
- There is an edge between nodes in $\mathcal{C}_i$ and $\mathcal{C}_j$ independently with probability $B_{i,j}$

$$B = \begin{bmatrix} .1 & .3 & .8 \\ .2 & .5 & .6 \\ .9 & .4 & .7 \end{bmatrix}$$



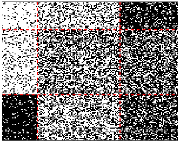
We work asymptotically in n, so we add some additional hypotheses:

- $0 < m_{min} \leq m_i \leq m_{max}$ for any i, and m_{min}, m_{max} are independent from n
- $B_{i,j} = \Psi_{i,j} f(n)$, $0 \leq \Psi_{i,j} \leq 1$ where the matrix Ψ is independent from n
- $[B \ B^T] = [\Psi \ \Psi^T] f(n)$ has full rank, equal to the number of roles q
- $nf(n) \to \infty$ that agrees with standard results for which $f(n) = \Omega(1/n)$ for **exact recovery**

Stochastic Block Model

- The nodes are partitioned in q roles $\mathcal{C}_i$ of size $m_i n$, $m := \sum_i m_i$
- There is an edge between nodes in $\mathcal{C}_i$ and $\mathcal{C}_j$ independently with probability $B_{i,j}$

$$B = \begin{bmatrix} .1 & .3 & .8 \\ .2 & .5 & .6 \\ .9 & .4 & .7 \end{bmatrix}$$



We work asymptotically in n, so we add some additional hypotheses:

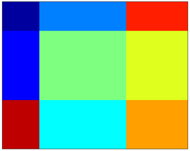
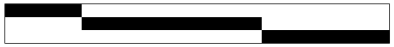
- $0 < m_{min} \leq m_i \leq m_{max}$ for any i, and m_{min}, m_{max} are independent from n
- $B_{i,j} = \Psi_{i,j} f(n)$, $0 \leq \Psi_{i,j} \leq 1$ where the matrix Ψ is independent from n
- $[B \;\; B^T] = [\Psi \;\; \Psi^T] f(n)$ has full rank, equal to the number of roles q
- $nf(n) \to \infty$ that agrees with standard results for which $f(n) = \Omega(1/n)$ for **exact recovery**

Spectral Method on the Average Case

Idea: the NPS of A is similar to the NPS of $M := \mathbb{E}[A]$

Spectral Method on the Average Case

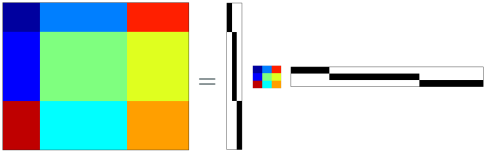
Idea: the NPS of A is similar to the NPS of $M := \mathbb{E}[A]$

$M = \mathbb{E}[A] =$  $=$    $= ZBZ^T$

Spectral Method on the Average Case

Idea: the NPS of A is similar to the NPS of $M := \mathbb{E}[A]$

$$M = \mathbb{E}[A] =$$



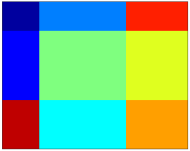
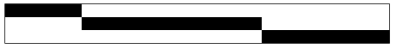
$$= ZBZ^T$$

$$T_1 := \begin{bmatrix} M & M^T \end{bmatrix} \begin{bmatrix} M^T \\ M \end{bmatrix} = Z \begin{bmatrix} B & B^T \end{bmatrix} \begin{bmatrix} Z^T Z & \\ & Z^T Z \end{bmatrix} \begin{bmatrix} B^T \\ B \end{bmatrix} Z^T = Z \hat{T}_1 Z^T$$

where $\hat{T}_1$ is a PD $q \times q$ matrix

Spectral Method on the Average Case

Idea: the NPS of A is similar to the NPS of $M := \mathbb{E}[A]$

$M = \mathbb{E}[A] =$  $=$    $= ZBZ^T$

$$T_1 := \begin{bmatrix} M & M^T \end{bmatrix} \begin{bmatrix} M^T \\ M \end{bmatrix} = Z \begin{bmatrix} B & B^T \end{bmatrix} \begin{bmatrix} Z^T Z & \\ & Z^T Z \end{bmatrix} \begin{bmatrix} B^T \\ B \end{bmatrix} Z^T = Z \hat{T}_1 Z^T$$

where $\hat{T}_1$ is a PD $q \times q$ matrix

$$T_{k+1} := \begin{bmatrix} M & M^T \end{bmatrix} \begin{bmatrix} I + \beta^2 T_k & \\ & I + \beta^2 T_k \end{bmatrix} \begin{bmatrix} M^T \\ M \end{bmatrix} = \cdots = Z \hat{T}_{k+1} Z^T$$

where all $\hat{T}_k$ are PD $q \times q$ matrices

$\Longrightarrow$ **The number of roles can be inferred by the rank of all T_k**

Spectral Method on the Average Case

$$T_{k+1} := \begin{bmatrix} M & M^T \end{bmatrix} \begin{bmatrix} I + \beta^2 T_k & \\ & I + \beta^2 T_k \end{bmatrix} \begin{bmatrix} M^T \\ M \end{bmatrix} = \cdots = Z \hat{T}_{k+1} Z^T$$

where all $\hat{T}_k$ are PD $q \times q$ matrices

$\implies$ **The number of roles can be inferred by the rank of all** T_k

Spectral Method on the Average Case

$$T_{k+1} := \begin{bmatrix} M & M^T \end{bmatrix} \begin{bmatrix} I + \beta^2 T_k & \\ & I + \beta^2 T_k \end{bmatrix} \begin{bmatrix} M^T \\ M \end{bmatrix} = \cdots = Z \hat{T}_{k+1} Z^T$$

where all $\hat{T}_k$ are PD $q \times q$ matrices

$\implies$ **The number of roles can be inferred by the rank of all T_k**

If $D = \operatorname{diag}(\sqrt{nm_i})$ and $U \Sigma U^T$ is the eigendecomposition of $D \hat{T}_k D$, then

$$T_k = Z \hat{T}_k Z^T = \left(Z D^{-1}\right) D \hat{T}_k D \left(Z D^{-1}\right)^T = \left(Z D^{-1} U\right) \Sigma \left(Z D^{-1} U\right)^T$$

is the reduced eigendecomposition of T_k

Spectral Method on the Average Case

$$T_{k+1} := \begin{bmatrix} M & M^T \end{bmatrix} \begin{bmatrix} I + \beta^2 T_k & \\ & I + \beta^2 T_k \end{bmatrix} \begin{bmatrix} M^T \\ M \end{bmatrix} = \cdots = Z \hat{T}_{k+1} Z^T$$

where all $\hat{T}_k$ are PD $q \times q$ matrices

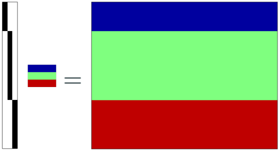
$\implies$ **The number of roles can be inferred by the rank of all T_k**

If $D = \operatorname{diag}(\sqrt{nm_i})$ and $U \Sigma U^T$ is the eigendecomposition of $D \hat{T}_k D$, then

$$T_k = Z \hat{T}_k Z^T = \left(Z D^{-1}\right) D \hat{T}_k D \left(Z D^{-1}\right)^T = \left(Z D^{-1} U\right) \Sigma \left(Z D^{-1} U\right)^T$$

is the reduced eigendecomposition of T_k

The reduced orthogonal matrix $Z D^{-1} U \in \mathbb{R}^{nm \times q}$ has only q distinct rows, since

$Z(D^{-1}U) =$  $\qquad D^{-1}U \in \mathbb{R}^{q \times q}$

$\implies$ **The assignation function can be computed from the EVD of all T_k**

Perturbation Result

Given $M = \mathbb{E}[A]$, $T_1 = \begin{bmatrix} M & M^T \end{bmatrix} \begin{bmatrix} M^T \\ M \end{bmatrix}$, and $T_{k+1} = \Gamma_M[I + \beta^2 T_k]$,

- The number of roles is the rank of any T_k
- The assignment corresponds to the different rows of Q_k in the reduced EVD of $T_k = Q_k \Sigma_k Q_k^T$

Perturbation Result

Given $M = \mathbb{E}[A]$, $T_1 = \begin{bmatrix} M & M^T \end{bmatrix} \begin{bmatrix} M^T \\ M \end{bmatrix}$, and $T_{k+1} = \Gamma_M[I + \beta^2 T_k]$,

- The number of roles is the rank of any T_k
- The assignment corresponds to the different rows of Q_k in the reduced EVD of $T_k = Q_k \Sigma_k Q_k^T$

We need to work on A, $S_1 = \begin{bmatrix} A & A^T \end{bmatrix} \begin{bmatrix} A^T \\ A \end{bmatrix}$, and $S_{k+1} = \Gamma_A[I + \beta^2 S_k]$ by

- Extracting the number of roles as the leading rank of S_k
- Extracting the assignment by K-means on the rows of V_k in the reduced EVD of $S_k = V_k \Sigma_k V_k^T$

Perturbation Result

Given $M = \mathbb{E}[A]$, $T_1 = \begin{bmatrix} M & M^T \end{bmatrix} \begin{bmatrix} M^T \\ M \end{bmatrix}$, and $T_{k+1} = \Gamma_M[I + \beta^2 T_k]$,

- The number of roles is the rank of any T_k
- The assignment corresponds to the different rows of Q_k in the reduced EVD of $T_k = Q_k \Sigma_k Q_k^T$

We need to work on A, $S_1 = \begin{bmatrix} A & A^T \end{bmatrix} \begin{bmatrix} A^T \\ A \end{bmatrix}$, and $S_{k+1} = \Gamma_A[I + \beta^2 S_k]$ by

- Extracting the number of roles as the leading rank of S_k
- Extracting the assignment by K-means on the rows of V_k in the reduced EVD of $S_k = V_k \Sigma_k V_k^T$

Theorem (Bai, Silverstein (2010))

For the matrix $Y := A - M$, almost surely

$$\|Y\|^2 \leq \delta^2 := 4mnf(n) \qquad \|[Y \;\; Y^T]\|^2 \leq 2\delta^2$$

where $B = f(n)\Psi$, Ψ independent on n, and $A \in \{0,1\}^{mn}$

Conjecture: $\|[Y \;\; Y^T]\|^2 \leq \left(\frac{1+\sqrt{2}}{2}\right)^2 \delta^2$ as suggested by MP distribution

Error Propagation

$$Y = A - M \qquad \|[Y \;\; Y^T]\|^2 \leq 2\delta^2 = 8mnf(n) \qquad \|M\| \sim \|B\|\|Z\|^2 = \Theta(\delta^2)$$

Recall that $S_{k+1} = \Gamma_A[I + \beta^2 S_k]$ and $T_{k+1} = \Gamma_M[I + \beta^2 T_k]$ and notice that

- $\Gamma_X[N] = \begin{bmatrix} X & X^T \end{bmatrix} \begin{bmatrix} N & \\ & N \end{bmatrix} \begin{bmatrix} X^T \\ X \end{bmatrix} \implies \|\Gamma_X\| = \left\|[X \; X^T]\right\|^2 \leq 2\|X\|^2$
- $\|\Gamma_A - \Gamma_M\| \sim \left\|[M \;\; M^T]\right\| \left\|[Y \;\; Y^T]\right\| = O(\delta^3)$
- If $\beta^2 < 1/(6\|A\|^2)$, then $\beta^2\|\Gamma_A\| \leq 1/3$ and $\beta^2\|\Gamma_M\| \leq 1/2$ almost surely

More precisely,

Theorem (B., N., V-D. (2022))

- $\|\Gamma_A - \Gamma_M\| \leq \delta^3\|[\Psi \;\; \Psi^T]\|/\sqrt{2} + 2\delta^2$
- $\beta^2 < 1/(6\|A\|^2) \implies \gamma := \beta^2 \max\{\|\Gamma_M\|, \|\Gamma_A\|\} \leq 1/2$ *a.s.*

and they imply that

$$\|S_k - T_k\| \leq \left(\sum_{i=0}^{k-1} \gamma^k\right)^2 \|\Gamma_A - \Gamma_M\| \leq 4\|\Gamma_A - \Gamma_M\| = O(\delta^3)$$

Error Propagation

$$Y = A - M \qquad \|[Y \ Y^T]\|^2 \leq 2\delta^2 = 8mnf(n) \qquad \|M\| \sim \|B\|\|Z\|^2 = \Theta(\delta^2)$$

Recall that $S_{k+1} = \Gamma_A[I + \beta^2 S_k]$ and $T_{k+1} = \Gamma_M[I + \beta^2 T_k]$ and notice that

- $\Gamma_X[N] = \begin{bmatrix} X & X^T \end{bmatrix} \begin{bmatrix} N & \\ & N \end{bmatrix} \begin{bmatrix} X^T \\ X \end{bmatrix} \implies \|\Gamma_X\| = \left\|[X \ X^T]\right\|^2 \leq 2\|X\|^2$
- $\|\Gamma_A - \Gamma_M\| \sim \left\|[M \ M^T]\right\| \left\|[Y \ Y^T]\right\| = O(\delta^3)$
- If $\beta^2 < 1/(6\|A\|^2)$, then $\beta^2\|\Gamma_A\| \leq 1/3$ and $\beta^2\|\Gamma_M\| \leq 1/2$ almost surely

More precisely,

Theorem (B., N., V-D. (2022))

- $\|\Gamma_A - \Gamma_M\| \leq \delta^3\|[\Psi \ \Psi^T]\|/\sqrt{2} + 2\delta^2$
- $\beta^2 < 1/(6\|A\|^2) \implies \gamma := \beta^2 \max\{\|\Gamma_M\|, \|\Gamma_A\|\} \leq 1/2$ *a.s.*

and they imply that

$$\|S_k - T_k\| \leq \left(\sum_{i=0}^{k-1} \gamma^k\right)^2 \|\Gamma_A - \Gamma_M\| \leq 4\|\Gamma_A - \Gamma_M\| = O(\delta^3)$$

Error Propagation

$$Y = A - M \qquad \|[Y\ \ Y^T]\|^2 \leq 2\delta^2 = 8mnf(n) \qquad \|M\| \sim \|B\|\|Z\|^2 = \Theta(\delta^2)$$

Recall that $S_{k+1} = \Gamma_A[I + \beta^2 S_k]$ and $T_{k+1} = \Gamma_M[I + \beta^2 T_k]$ and notice that

- $\Gamma_X[N] = \begin{bmatrix} X & X^T \end{bmatrix} \begin{bmatrix} N & \\ & N \end{bmatrix} \begin{bmatrix} X^T \\ X \end{bmatrix} \implies \|\Gamma_X\| = \left\|[X\ X^T]\right\|^2 \leq 2\|X\|^2$
- $\|\Gamma_A - \Gamma_M\| \sim \|[M\ \ M^T]\|\ \|[Y\ \ Y^T]\| = O(\delta^3)$
- If $\beta^2 < 1/(6\|A\|^2)$, then $\beta^2\|\Gamma_A\| \leq 1/3$ and $\beta^2\|\Gamma_M\| \leq 1/2$ almost surely

More precisely,

Theorem (B., N., V-D. (2022))

- $\|\Gamma_A - \Gamma_M\| \leq \delta^3\|[\Psi\ \ \Psi^T]\|/\sqrt{2} + 2\delta^2$
- $\beta^2 < 1/(6\|A\|^2) \implies \gamma := \beta^2 \max\{\|\Gamma_M\|, \|\Gamma_A\|\} \leq 1/2$ *a.s.*

and they imply that

$$\|S_k - T_k\| \leq \left(\sum_{i=0}^{k-1} \gamma^k\right)^2 \|\Gamma_A - \Gamma_M\| \leq 4\|\Gamma_A - \Gamma_M\| = O(\delta^3)$$

Number of Roles

$$Y = A - M \qquad \|[Y\ Y^T]\|^2 \leq 2\delta^2 = 8mnf(n) \qquad \|S_k - T_k\| = O(\delta^3)$$

The number of roles q is $\text{rk}(T_k)$, so how can we infer q from S_k?

- q is the rank of $[M\ M^T]$ and all its non-zero singular values are $\Theta(\delta^2)$
- The q non-zero eigenvalues of T_k are $\Theta(\delta^4)$
- The biggest q eigenvalues of S_k are $\Theta(\delta^4)$ almost surely
- All other eigenvalues of S_k are $O(\delta^2)$

Number of Roles

$$Y = A - M \qquad \|[Y\ Y^T]\|^2 \leq 2\delta^2 = 8mnf(n) \qquad \|S_k - T_k\| = O(\delta^3)$$

The number of roles q is $\mathrm{rk}(T_k)$, so how can we infer q from S_k?

- q is the rank of $[M\ M^T]$ and all its non-zero singular values are $\Theta(\delta^2)$
- The q non-zero eigenvalues of T_k are $\Theta(\delta^4)$
- The biggest q eigenvalues of S_k are $\Theta(\delta^4)$ almost surely
- All other eigenvalues of S_k are $O(\delta^2)$

$$D = \mathrm{diag}(\sqrt{m_i n})$$

$$[M\ M^T] = Z[B\ B^T]\begin{bmatrix} Z^T & \\ & Z^T \end{bmatrix}$$

$$= UDf(n)[\Psi\ \Psi^T]\begin{bmatrix} D & \\ & D \end{bmatrix} V^T$$

where U, V have orthonormal columns

$$U = ZD^{-1} \qquad V = \begin{bmatrix} ZD^{-1} & \\ & ZD^{-1} \end{bmatrix}$$

and thus

$$m_{min} \leq \frac{\sigma_i([M\ M^T])}{nf(n)\sigma_i([\Psi\ \Psi^T])} \leq m_{max} \quad 1 \leq i \leq q$$

Recall that Ψ, m_{max} and $m_{\min}$ do not depend on n

Number of Roles

$$Y = A - M \qquad \|[Y\ Y^T]\|^2 \leq 2\delta^2 = 8mnf(n) \qquad \|S_k - T_k\| = O(\delta^3)$$

The number of roles q is $\mathrm{rk}(T_k)$, so how can we infer q from S_k?

- q is the rank of $[M\ M^T]$ and all its non-zero singular values are $\Theta(\delta^2)$
- The q non-zero eigenvalues of T_k are $\Theta(\delta^4)$
- The biggest q eigenvalues of S_k are $\Theta(\delta^4)$ almost surely
- All other eigenvalues of S_k are $O(\delta^2)$

$$T_1 = [M\ M^T] \begin{bmatrix} M^T \\ M \end{bmatrix} \implies \|T_1\| \leq 2\|M\|^2 = O(\delta^4)$$

$$\begin{aligned} \|T_{k+1}\| &= \|\Gamma_M[I + \beta^2 T_k]\| \\ &\leq \|\Gamma_M\| + \|T_k\|/2 \\ &\leq \|\Gamma_M\| \left(\sum_{i=0}^{k} \frac{1}{2^i} \right) \leq 4\|M\|^2 = O(\delta^4) \end{aligned}$$

$$T_{k+1} \succeq T_k \succeq \cdots \succeq T_1 \implies \lambda_q(T_k) \geq \lambda_q(T_1)$$

$$\lambda_q(T_1) = \sigma_q([M\ M^T])^2 = \Theta(\delta^4)$$

Number of Roles

$$Y = A - M \qquad \|[Y \; Y^T]\|^2 \leq 2\delta^2 = 8mnf(n) \qquad \|S_k - T_k\| = O(\delta^3)$$

The number of roles q is $\mathrm{rk}(T_k)$, so how can we infer q from S_k?

- q is the rank of $[M \; M^T]$ and all its non-zero singular values are $\Theta(\delta^2)$
- The q non-zero eigenvalues of T_k are $\Theta(\delta^4)$
- The biggest q eigenvalues of S_k are $\Theta(\delta^4)$ almost surely
- All other eigenvalues of S_k are $O(\delta^2)$

By Weyl's perturbation Theorem

$$|\sigma_i(S_k) - \sigma_i(T_k)| \leq \|S_k - T_k\|$$

but

$$\sigma_i(T_k) = \Theta(\delta^4) \gg O(\delta^3) = \|S_k - T_k\|$$

Number of Roles

$$Y = A - M \qquad \|[Y\ Y^T]\|^2 \le 2\delta^2 = 8mnf(n) \qquad \|S_k - T_k\| = O(\delta^3)$$

The number of roles q is $\mathrm{rk}(T_k)$, so how can we infer q from S_k?

- q is the rank of $[M\ M^T]$ and all its non-zero singular values are $\Theta(\delta^2)$
- The q non-zero eigenvalues of T_k are $\Theta(\delta^4)$
- The biggest q eigenvalues of S_k are $\Theta(\delta^4)$ almost surely
- All other eigenvalues of S_k are $O(\delta^2)$

$$\|S_{k+1}\| = \|\Gamma_A[I + \beta^2 S_k]\| \le \|\Gamma_A\| + \|S_k\|/2$$

$$\le \|\Gamma_A\| \left(\sum_{i=0}^{k} \frac{1}{2^i} \right) \le 2\|\Gamma_A\|$$

$$S_{k+1} = \Gamma_A[I + \beta^2 S_k] \preceq (1 + \beta^2 \|S_k\|)\Gamma_A[I] \preceq 2S_1$$

By Weyl on $Y = A - M$ and

$$S_1 = [A\ A^T] \begin{bmatrix} A^T \\ A \end{bmatrix} \implies \lambda_i(S_1) = \sigma_i([A\ A^T])^2$$

we conclude that for $i > q$

$$\lambda_i(S_k) \le 2\lambda_i(S_1) = 2\sigma_i([A\ A^T])^2$$

$$\sigma_i([A\ A^T]) \le \sigma_i([M\ M^T]) + \|[Y\ Y^T]\| = O(\delta)$$

Number of Roles

$$Y = A - M \qquad \|[Y\ Y^T]\|^2 \leq 2\delta^2 = 8mnf(n) \qquad \|S_k - T_k\| = O(\delta^3)$$

The number of roles q is $\mathrm{rk}(T_k)$, so how can we infer q from S_k?

- q is the rank of $[M\ M^T]$ and all its non-zero singular values are $\Theta(\delta^2)$
- The q non-zero eigenvalues of T_k are $\Theta(\delta^4)$
- The biggest q eigenvalues of S_k are $\Theta(\delta^4)$ almost surely
- All other eigenvalues of S_k are $O(\delta^2)$

The dominant q eigenvalues of S_k are well separated from the noise

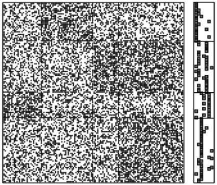
Theorem (B., N., V-D. (2022))

$$\frac{1}{2}\|[\Psi\ \Psi^T]\|^2\delta^4 \geq \|S_k\| = \lambda_1(S_k)$$

$$\lambda_q(S_k) \geq \frac{1}{2}\left[\frac{\sigma_q([\Psi\ \Psi^T])}{4q}\,\frac{m_{min}}{m_{max}}\right]^2\delta^4$$

$$4\delta^2 \geq \lambda_{q+1}(S_k)$$

Notice: if $[\Psi\ \Psi^T]$ is almost singular, it is hard to infer the number of roles



Dominant Subspaces

Recall that rk $T_k = q$ and the clustering assignment corresponds to the different rows of Q_k in the reduced EVD of $T_k = Q_k \Sigma_k Q_k^T$

As a consequence, we work on the q-dominant subspace given by the reduced EVD of $S_k = V_k \tilde{\Sigma}_k V_k^T$, but how are they related?

Theorem (Davis, Kahan (1970))

Given E, F, the q-dominant subspaces of S_k, T_k

$$\|\Pi_E - \Pi_F\| \leq 2 \frac{\|S_k - T_k\|}{\lambda_q(T_k)}$$

If we now plug in our previous estimations

$$\|\Pi_E - \Pi_F\| \leq \frac{4\sqrt{2}\delta^3 \|[\Psi \ \Psi^T]\| + 16\delta^2}{\left[\frac{\sigma_q([\Psi \ \Psi^T])}{4q} \frac{m_{min}}{m_{max}}\right]^2 \delta^4} = O(\delta^{-1}).$$

Notice that an almost singular $[\Psi \ \Psi^T]$ produces a small $\sigma_q([\Psi \ \Psi^T])$ and thus a bigger distance between the subspaces

Dominant Subspaces

Recall that rk $T_k = q$ and the clustering assignment corresponds to the different rows of Q_k in the reduced EVD of $T_k = Q_k \Sigma_k Q_k^T$

As a consequence, we work on the q-dominant subspace given by the reduced EVD of $S_k = V_k \tilde{\Sigma}_k V_k^T$, but how are they related?

Theorem (Davis, Kahan (1970))

Given E, F, the q-dominant subspaces of S_k, T_k

$$\|\Pi_E - \Pi_F\| \leq 2 \frac{\|S_k - T_k\|}{\lambda_q(T_k)}$$

If we now plug in our previous estimations

$$\|\Pi_E - \Pi_F\| \leq \frac{4\sqrt{2}\delta^3 \|[\Psi \ \ \Psi^T]\| + 16\delta^2}{\left[\frac{\sigma_q([\Psi \ \ \Psi^T])}{4q} \frac{m_{min}}{m_{max}}\right]^2 \delta^4} = O(\delta^{-1}).$$

Notice that an almost singular $[\Psi \ \ \Psi^T]$ produces a small $\sigma_q([\Psi \ \ \Psi^T])$ and thus a bigger distance between the subspaces

Dominant Subspaces

Recall that rk $T_k = q$ and the clustering assignment corresponds to the different rows of Q_k in the reduced EVD of $T_k = Q_k \Sigma_k Q_k^T$

As a consequence, we work on the q-dominant subspace given by the reduced EVD of $S_k = V_k \tilde{\Sigma}_k V_k^T$, but how are they related?

Theorem (Davis, Kahan (1970))

Given E, F, the q-dominant subspaces of S_k, T_k

$$\|\Pi_E - \Pi_F\| \leq 2 \frac{\|S_k - T_k\|}{\lambda_q(T_k)}$$

If we now plug in our previous estimations

$$\|\Pi_E - \Pi_F\| \leq \frac{4\sqrt{2}\delta^3 \|[\Psi \;\; \Psi^T]\| + 16\delta^2}{\left[\frac{\sigma_q([\Psi \;\; \Psi^T])}{4q} \frac{m_{min}}{m_{max}}\right]^2 \delta^4} = O(\delta^{-1}).$$

Notice that an almost singular $[\Psi \;\; \Psi^T]$ produces a small $\sigma_q([\Psi \;\; \Psi^T])$ and thus a bigger distance between the subspaces

Misclassification Error

$$\|\Pi_E - \Pi_F\| = O(\delta^{-1}) \qquad \|[Y\ \ Y^T]\|^2 \leq 2\delta^2 = 8mnf(n)$$

Given the q-reduced EVD $T_k = U_k \Sigma_k U_k^T$, K-means on the rows of U_k (whose columns are a basis of F) returns the exact role partition

Given the q-reduced EVD $S_k = V_k \tilde{\Sigma}_k V_k^T$, we apply K-means on the rows of V_k (whose columns are a basis of E) to obtain the roles

Given the exact roles $\mathcal{C}_1, \ldots, \mathcal{C}_q$ and $\mathcal{T}_1, \ldots, \mathcal{T}_q$ the resulting roles from the algorithm on S_k, let the **misclassification error** $\widehat{f}$ be

$$\widehat{f} := \min_{\pi \in \mathcal{S}_q} \max_{i=1,\ldots,q} \frac{|\mathcal{T}_{\pi(i)} \triangle \mathcal{C}_i|}{|\mathcal{C}_i|}$$

where $\triangle$ is the symmetric difference of sets

Theorem

There exists an absolute constant C such that

$$\widehat{f} \leq Cq \frac{m_{max}}{m_{min}} \|\Pi_E - \Pi_F\|^2 \leq C \frac{q^5}{\delta^2} \frac{m_{max}^5}{m_{min}^5} \frac{\|[\Upsilon\ \ \Upsilon^T]\|^2}{\sigma_q([\Upsilon\ \ \Upsilon^T])^4} = O\left(\frac{1}{nf(n)}\right)$$

Misclassification Error

$$\|\Pi_E - \Pi_F\| = O(\delta^{-1}) \qquad \|[Y\ \ Y^T]\|^2 \le 2\delta^2 = 8mnf(n)$$

Given the q-reduced EVD $T_k = U_k \Sigma_k U_k^T$, K-means on the rows of U_k (whose columns are a basis of F) returns the exact role partition

Given the q-reduced EVD $S_k = V_k \tilde{\Sigma}_k V_k^T$, we apply K-means on the rows of V_k (whose columns are a basis of E) to obtain the roles

Given the exact roles $\mathcal{C}_1, \ldots, \mathcal{C}_q$ and $\mathcal{T}_1, \ldots, \mathcal{T}_q$ the resulting roles from the algorithm on S_k, let the **misclassification error** $\widehat{f}$ be

$$\widehat{f} := \min_{\pi \in \mathcal{S}_q} \max_{i=1,\ldots,q} \frac{|\mathcal{T}_{\pi(i)} \triangle \mathcal{C}_i|}{|\mathcal{C}_i|}$$

where $\triangle$ is the symmetric difference of sets

Theorem

There exists an absolute constant C such that

$$\widehat{f} \le Cq \frac{m_{max}}{m_{min}} \|\Pi_E - \Pi_F\|^2 \le C \frac{q^5}{\delta^2} \frac{m_{max}^5}{m_{min}^5} \frac{\|[\Upsilon\ \ \Upsilon^T]\|^2}{\sigma_q([\Upsilon\ \ \Upsilon^T])^4} = O\left(\frac{1}{nf(n)}\right)$$

Misclassification Error

$$\|\Pi_E - \Pi_F\| = O(\delta^{-1}) \qquad \|[Y \;\; Y^T]\|^2 \leq 2\delta^2 = 8mnf(n)$$

Given the q-reduced EVD $T_k = U_k \Sigma_k U_k^T$, K-means on the rows of U_k (whose columns are a basis of F) returns the exact role partition

Given the q-reduced EVD $S_k = V_k \tilde{\Sigma}_k V_k^T$, we apply K-means on the rows of V_k (whose columns are a basis of E) to obtain the roles

Given the exact roles $\mathcal{C}_1, \ldots, \mathcal{C}_q$ and $\mathcal{T}_1, \ldots, \mathcal{T}_q$ the resulting roles from the algorithm on S_k, let the **misclassification error** $\widehat{f}$ be

$$\widehat{f} := \min_{\pi \in \mathcal{S}_q} \max_{i=1,\ldots,q} \frac{|\mathcal{T}_{\pi(i)} \triangle \mathcal{C}_i|}{|\mathcal{C}_i|}$$

where $\triangle$ is the symmetric difference of sets

Theorem

There exists an absolute constant C such that

$$\widehat{f} \leq Cq \frac{m_{max}}{m_{min}} \|\Pi_E - \Pi_F\|^2 \leq C \frac{q^5}{\delta^2} \frac{m_{max}^5}{m_{min}^5} \frac{\|[\Upsilon \;\; \Upsilon^T]\|^2}{\sigma_q([\Upsilon \;\; \Upsilon^T])^4} = O\left(\frac{1}{nf(n)}\right)$$

The Idea Behind

Theorem (Sheffet, Awasthi (2012))

Let U, V be $mn \times q$ matrices, where U has only q distinct rows $\mu_1, \dots, \mu_q$ that identify the roles $\mathcal{C}_i$ and call

$$\Delta_i := \frac{1}{\sqrt{|\mathcal{C}_i|}} \min\{\sqrt{q}\|U - V\|, \|U - V\|_F\}$$

Suppose there exists $\rho \geq 100$ such that $\|\mu_i - \mu_j\| \geq \rho(\Delta_i + \Delta_j)$ for any $i \neq j$. If $\mathcal{T}_i$ are the roles determined by the K-means algorithm on V, then there exists a permutation π and an absolute constant C such that

$$\hat{f} \leq \max_r \frac{|\mathcal{C}_r \triangle \mathcal{T}_{\pi(r)}|}{|\mathcal{C}_r|} \leq \frac{C}{\rho^2}$$

In our case, let $T_k = U_k \Sigma_k U_k^T$ and $S_k = V_k \tilde{\Sigma}_k V_k^T$ be q-reduced EVDs

- $T_k = Z \hat{T}_k Z^T = \cdots = \left(ZD^{-1}W_k\right) \Sigma_k \left(ZD^{-1}W_k\right)^T \implies U_k = ZD^{-1}W_k$ so U_k has q distinct rows of the form $\nu_i / \sqrt{m_i n}$ where ν_i are orthonormal
- There exists a $q \times q$ orthogonal Q s.t. $\|U_k Q - V_k\|_F \leq \sqrt{2q}\|\Pi_E - \Pi_F\|$

$$\rho = \min_{i \neq j} \frac{\left\| \frac{\mu_i}{nm_i} - \frac{\mu_j}{nm_j} \right\|}{\Delta_i + \Delta_j} \geq \frac{\sqrt{\frac{1}{m_i} + \frac{1}{m_j}}}{\frac{1}{\sqrt{m_i}} + \frac{1}{\sqrt{m_j}}} \frac{1}{\sqrt{2q}\|\Pi_E - \Pi_F\|} = \Omega(\delta) \to \infty$$

$$\implies \hat{f} = O(\rho^{-2}) = O(\delta^{-2}) = O\left(1/(nf(n))\right)$$

The Idea Behind

Theorem (Sheffet, Awasthi (2012))

Let U, V be $mn \times q$ matrices, where U has only q distinct rows $\mu_1, \dots, \mu_q$ that identify the roles $\mathcal{C}_i$ and call

$$\Delta_i := \frac{1}{\sqrt{|\mathcal{C}_i|}} \min\{\sqrt{q}\|U - V\|, \|U - V\|_F\}$$

Suppose there exists $\rho \geq 100$ such that $\|\mu_i - \mu_j\| \geq \rho(\Delta_i + \Delta_j)$ for any $i \neq j$. If $\mathcal{T}_i$ are the roles determined by the K-means algorithm on V, then there exists a permutation π and an absolute constant C such that

$$\hat{f} \leq \max_r \frac{|\mathcal{C}_r \triangle \mathcal{T}_{\pi(r)}|}{|\mathcal{C}_r|} \leq \frac{C}{\rho^2}$$

In our case, let $T_k = U_k \Sigma_k U_k^T$ and $S_k = V_k \tilde{\Sigma}_k V_k^T$ be q-reduced EVDs

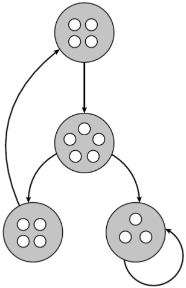
- $T_k = Z\hat{T}_k Z^T = \cdots = \left(ZD^{-1}W_k\right)\Sigma_k\left(ZD^{-1}W_k\right)^T \implies U_k = ZD^{-1}W_k$ so U_k has q distinct rows of the form $\nu_i/\sqrt{m_i n}$ where ν_i are orthonormal
- There exists a $q \times q$ orthogonal Q s.t. $\|U_k Q - V_k\|_F \leq \sqrt{2q}\|\Pi_E - \Pi_F\|$

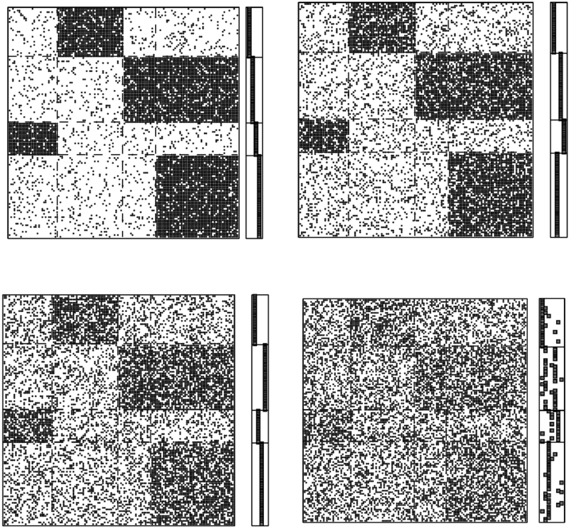
$$\rho = \min_{i \neq j} \frac{\left\|\frac{\mu_i}{nm_i} - \frac{\mu_j}{nm_j}\right\|}{\Delta_i + \Delta_j} \geq \frac{\sqrt{\frac{1}{m_i} + \frac{1}{m_j}}}{\frac{1}{\sqrt{m_i}} + \frac{1}{\sqrt{m_j}}} \frac{1}{\sqrt{2q}\|\Pi_E - \Pi_F\|} = \Omega(\delta) \to \infty$$

$$\implies \hat{f} = O(\rho^{-2}) = O(\delta^{-2}) = O\left(1/(nf(n))\right)$$

Numerical Example

We have already seen some: for $p = .9, .8, .7, .6$ and S_{10} we have

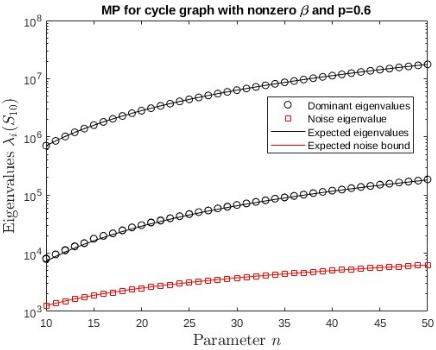


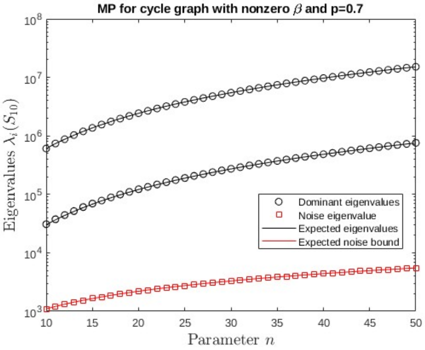


Numerical Example

Here instead we compare the eigenvalues of T_{10} with those of S_{10} where the matrix dimension is $30n$ and

$$B = p \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} + (1-p) \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}$$



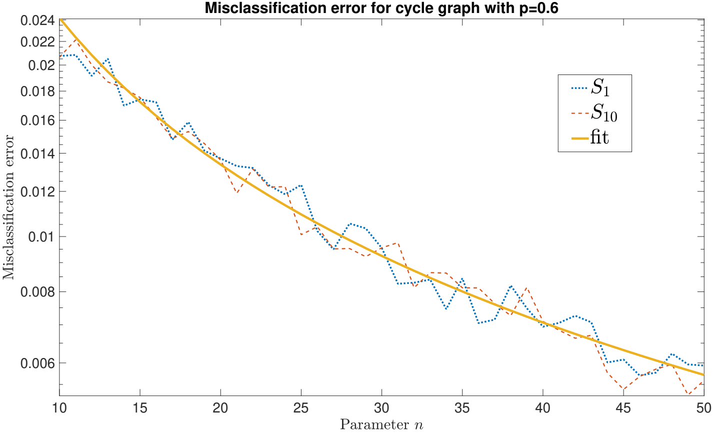


The eigenvalues estimations are more accurate when taking into consideration the **conjecture:** $\|[Y \;\; Y^T]\|^2 \leq \left(\frac{1+\sqrt{2}}{2}\right)^2 \delta^2$

Numerical Example

Here is the misclassification error for S_1 and S_{10} where the matrix dimension is $30n$, the yellow line is a fit for the estimated bound $\hat{f} \leq C/n$ and

$$B = p \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} + (1-p) \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}$$



Thank You!

- Abbe E., Fan J., Wang K., and Zhong Y. **Entrywise eigenvector analysis of random matrices with low expected rank.** *The Annals of Statistics*, 48(3):1452 – 1474, 2020.
- Barbarino G., Noferini V., and Van Dooren P. **Role extraction for digraphs via neighborhood pattern similarity.** *Physical Review E*, 106:054301, 2022.
- Qing H. and Wang J. **Community detection for weighted bipartite networks.** *Knowledge-Based Systems*, 274:110643, 2023.
- Ilse C.F. Ipsen. **Absolute and relative perturbation bounds for invariant subspaces of matrices.** *Linear Algebra and its Applications*, 309(1):45–56, 2000.
- Marchand M., Gallivan K., Huang W., and Van Dooren P. **Analysis of the neighborhood pattern similarity measure for the role extraction problem.** *SIAM Journal on Mathematics of Data Science*, 3(2):736–757, 2021.