

# Detección de comunidades en muestras de redes

Jesús Arroyo

10 de Septiembre, 2020



# Contenido

- 1 Introducción
- 2 Detección de comunidades no supervisada
- 3 Detección de comunidades supervisada

## 1 Introducción

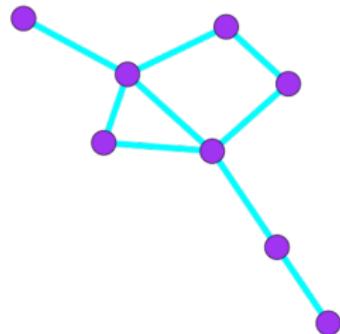
2 Detección de comunidades no supervisada

3 Detección de comunidades supervisada

# Redes

Una **red** o **grafo**  $G = (V, E)$  es un conjunto de

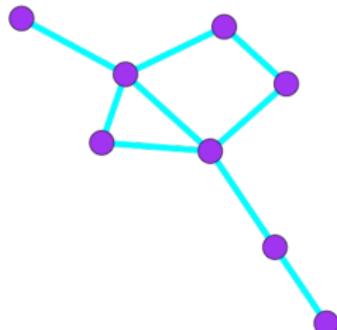
- $V$  = nodos / vertices / actores
- $E$  = aristas / arcos / conexiones



# Redes

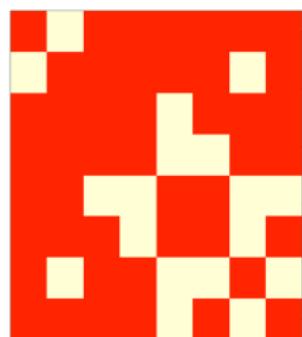
Una **red** o **grafo**  $G = (V, E)$  es un conjunto de

- $V$  = nodos / vertices / actores
- $E$  = aristas / arcos / conexiones

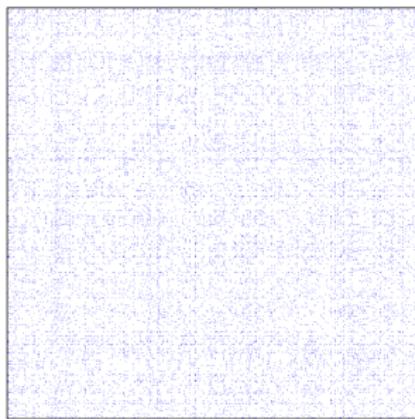
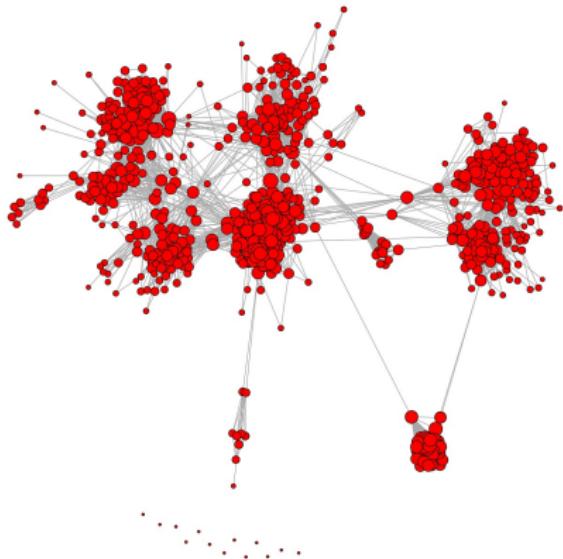


Las redes pueden representarse usando una **matriz de adyacencia**

- $A_{u,v} = 1$  indica que existe una arista entre los nodos  $u$  y  $v$

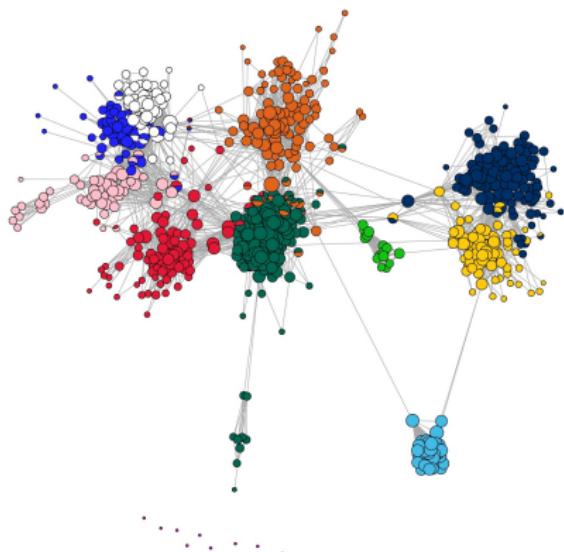


# Grafo de amistad en Facebook

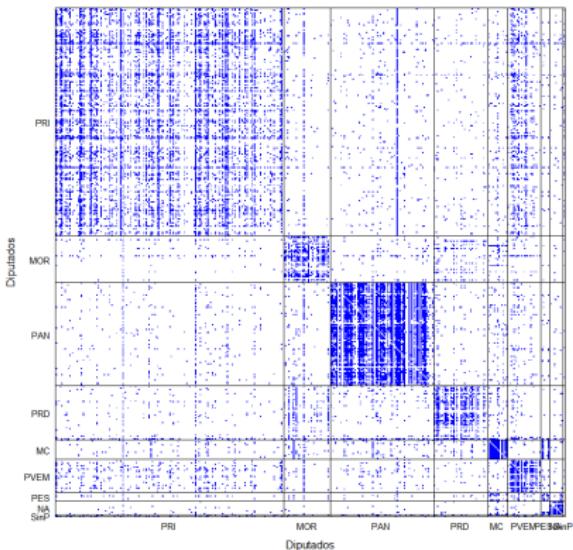
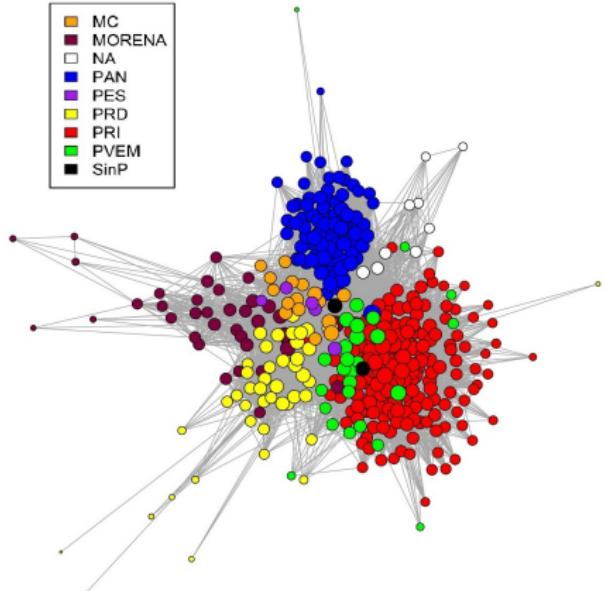


# Comunidades en una red

Las **comunidades** son grupos de nodos con patrones de conectividad similares.

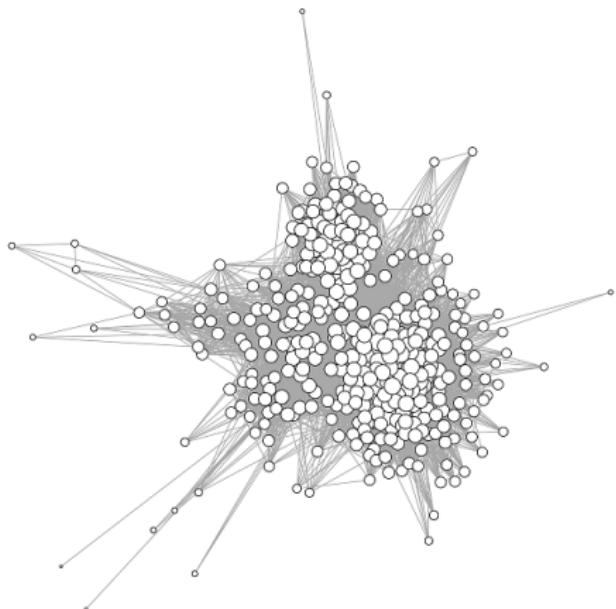


# Red de seguimiento en Twitter de diputados en 2018



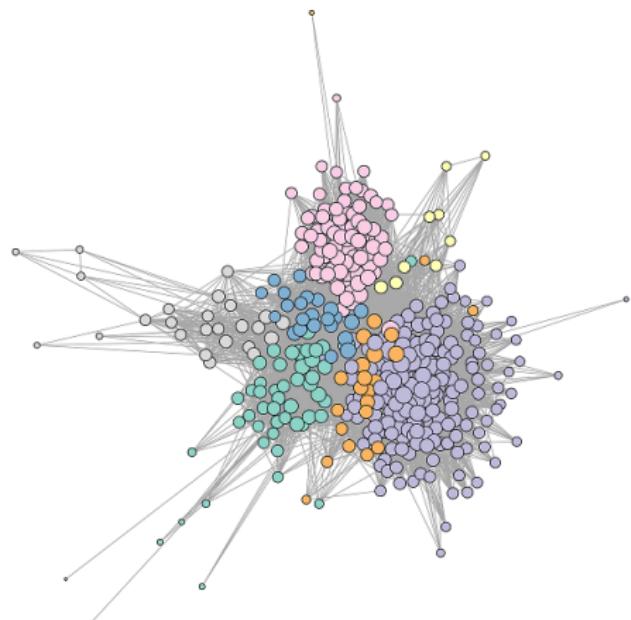
# Detección de comunidades

¿Cómo encontrar grupos significativos en los nodos de una red?



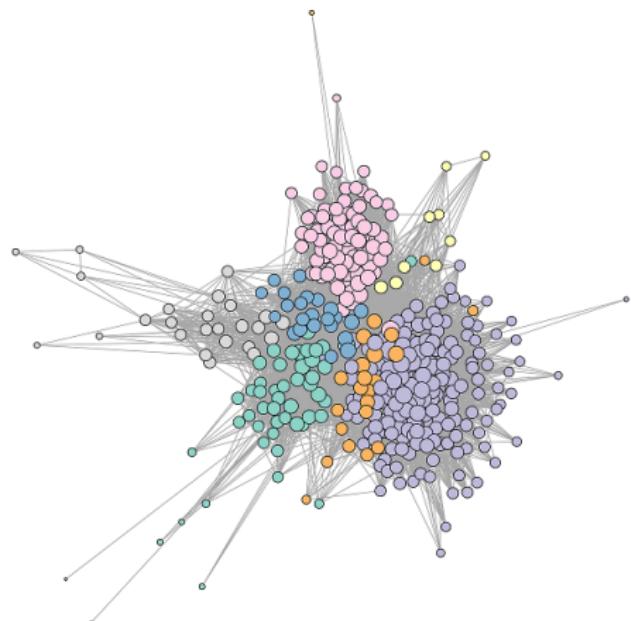
# Detección de comunidades

¿Cómo encontrar grupos significativos en los nodos de una red?

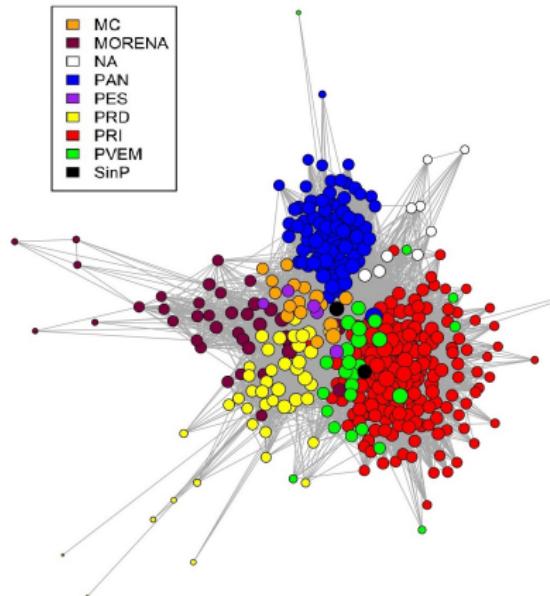


# Detección de comunidades

¿Cómo encontrar grupos significativos en los nodos de una red?

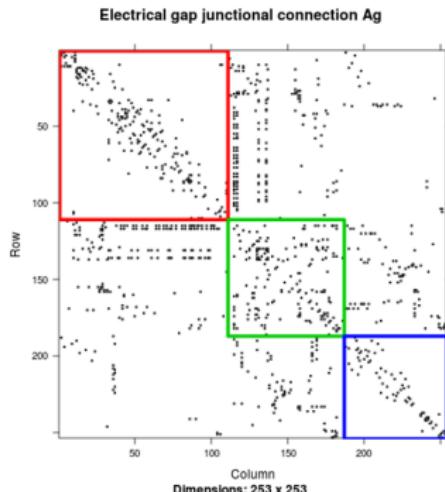
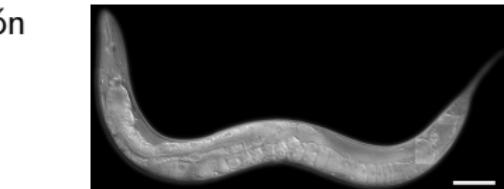
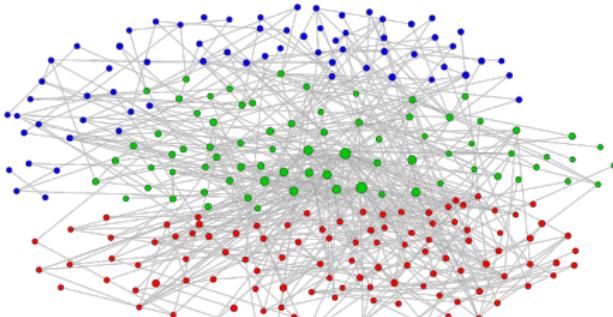


■	MC
■	MORENA
□	NA
■	PAN
■	PES
■	PRD
■	PRI
■	PVEM
■	SinP



# Redes cerebrales

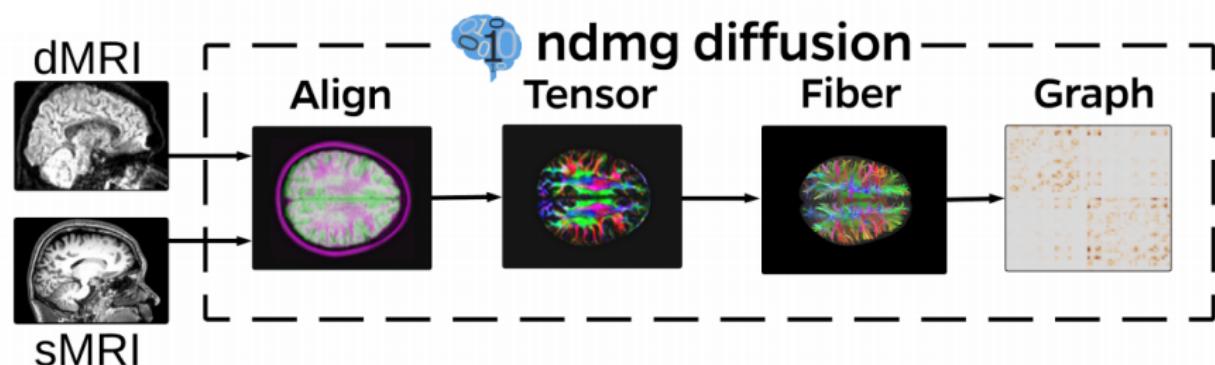
- Las redes cerebrales son una representación de la conectividad en el cerebro
- Ejemplo: red de sinapsis neuronales en el gusano *c. elegans* (aprox. 300 neuronas)



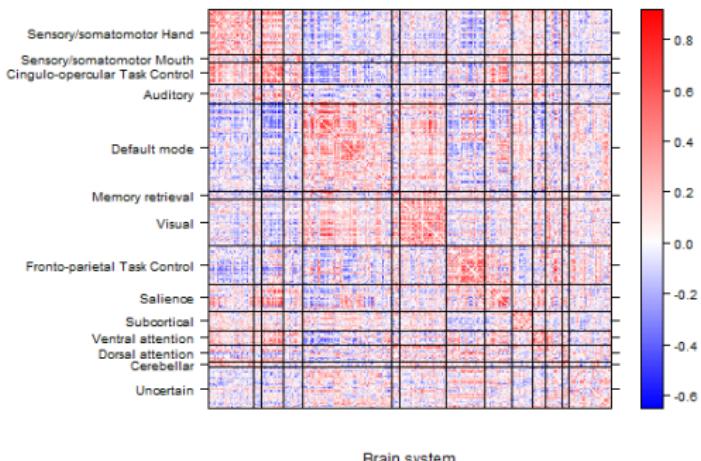
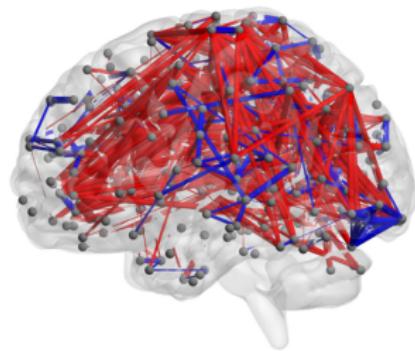
# Redes cerebrales humanas

La conectividad entre distintas regiones del cerebro es estimada a través de *imágenes de resonancia magnética*.

- Nodos: grupos de voxels en la imagen
- Aristas: medida de conectividad entre nodos



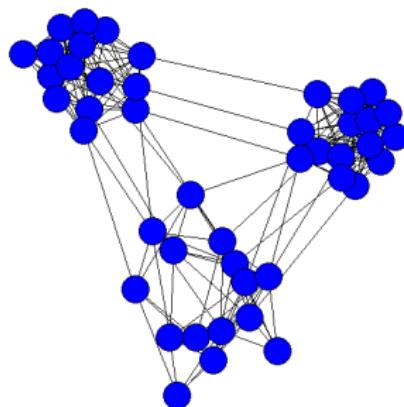
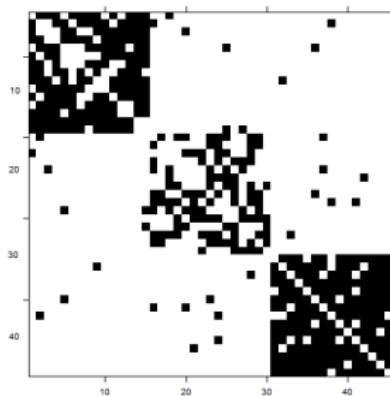
# Red de conectividad funcional de un cerebro humano



Brain system

# Detección de comunidades

- La detección de comunidades es un problema de agrupamiento no supervisado
- La estructura de comunidad es comúnmente estudiada usando el modelo de bloques estocástico (SBM).
- Existen muchos métodos para resolver este problema en una sola red: modularidad, verosimilitud, métodos bayesianos, inferencia espectral.

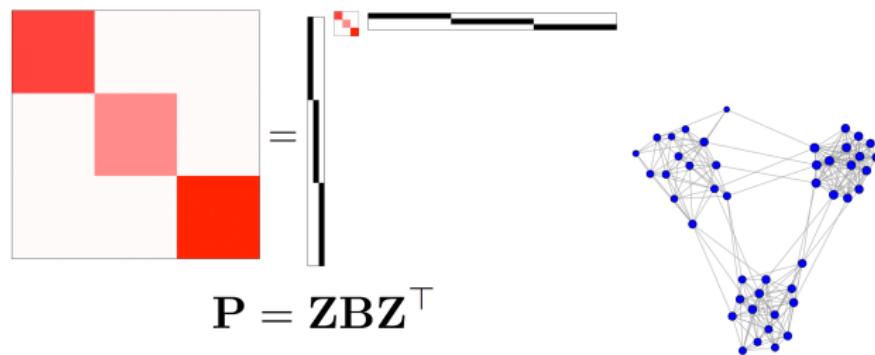


# Modelo de bloques estocástico

- $\mathbf{A}$  es una matriz binaria simétrica de tamaño  $n \times n$
- Cada entrada de  $\mathbf{A}$  tiene distribución Bernoulli independiente, con probabilidades  $\mathbf{P} \in [0, 1]^{n \times n}$

$$\mathbb{P}(\mathbf{A}_{uv} = 1) = \mathbf{P}_{uv}.$$

- La probabilidad de que dos nodos se conecten depende únicamente de sus comunidades



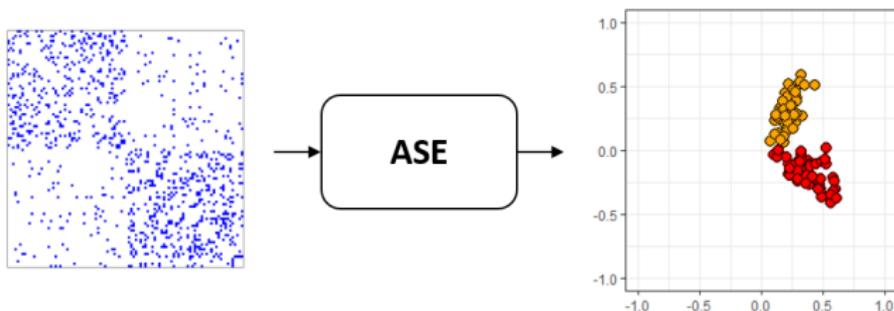
- ▶  $\mathbf{Z} \in \{0, 1\}^{n \times K}$  indica las comunidades de los nodos
- ▶  $\mathbf{B} \in [0, 1]^{K \times K}$ ; indica las probabilidades de las aristas
- ▶  $K$  es el número de comunidades

## Incrustación espectral

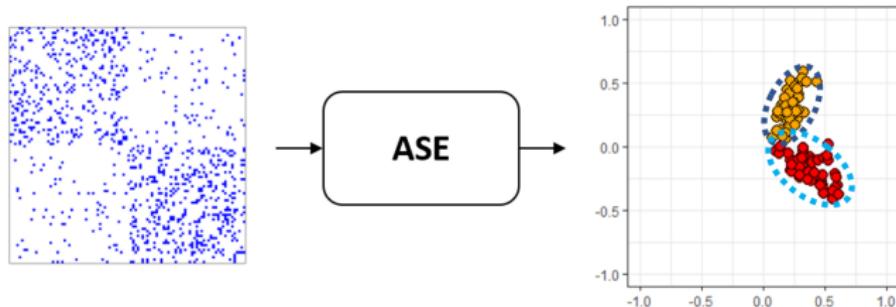
- La **incrustación espectral de una matriz de adyacencia** (ASE) consiste en representar los nodos como vectores en  $\mathbb{R}^d$
- Descomposición de  $\mathbf{A}$  en eigenvalores y eigenvectores

$$\mathbf{A} = \hat{\mathbf{V}}\hat{\mathbf{D}}\hat{\mathbf{V}}^T + \hat{\mathbf{V}}_{\perp}\hat{\mathbf{D}}_{\perp}\hat{\mathbf{V}}_{\perp}^T,$$

- ▶  $\hat{\mathbf{V}} \in \mathbb{R}^{n \times d}$  es la matrix de eigenvectores principales
- ▶  $d$  es la dimensión de la incrustación



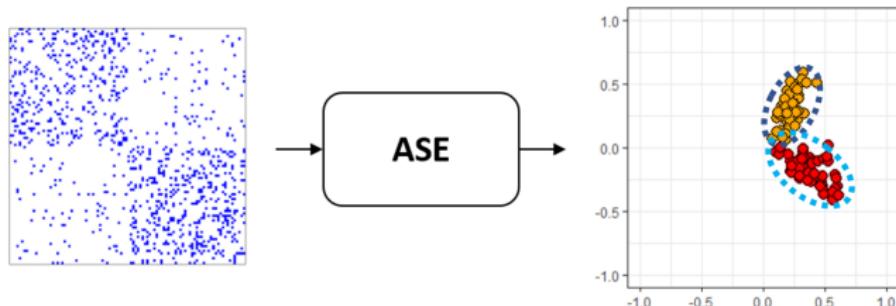
# Agrupamiento espectral



## Algoritmo de agrupamiento espectral

- ① Calcular la matrix  $\hat{V}$  de eigenvectores principales de  $\mathbf{A}$
- ② Agrupar los vectores renglón de  $\hat{V}$  usando  $k$ -medias

# Agrupamiento espectral

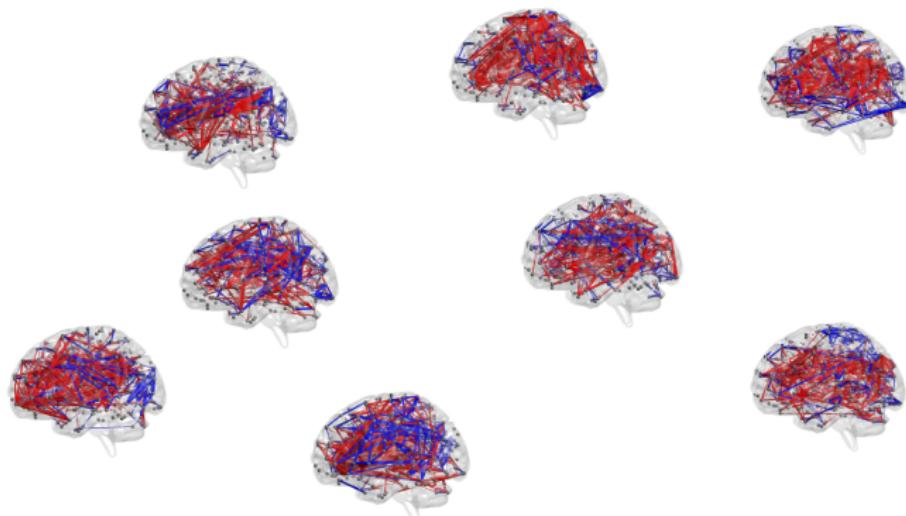


## Algoritmo de agrupamiento espectral

- ① Calcular la matrix  $\hat{V}$  de eigenvectores principales de  $A$
  - ② Agrupar los vectores renglón de  $\hat{V}$  usando  $k$ -medias
- 
- Los métodos de inferencia espectral presentan algunas ventajas
    - ▶ Fundamentos teóricos
    - ▶ Escalabilidad computacional
    - ▶ Robustez

## Muestras de redes

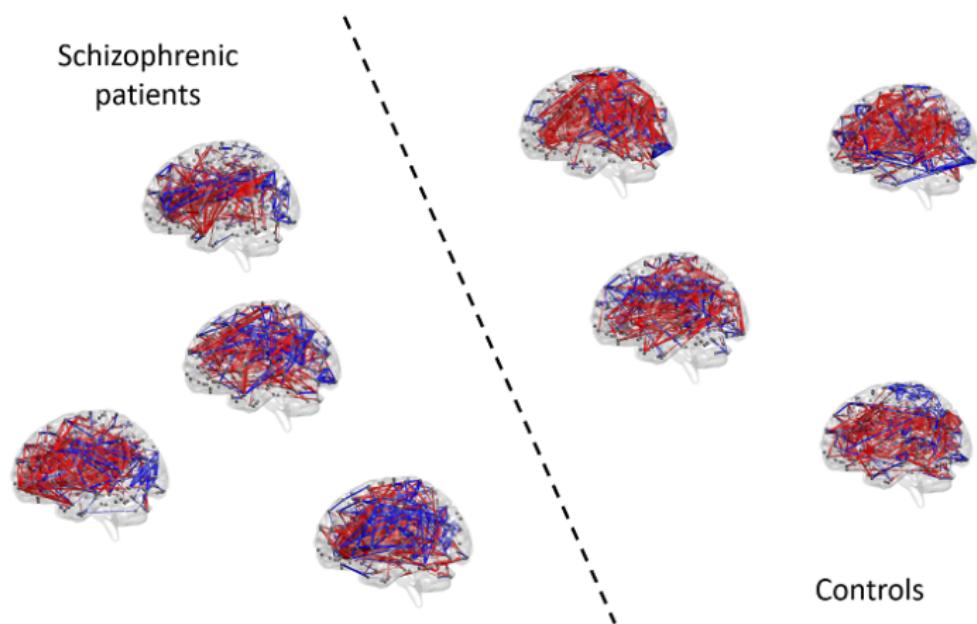
- En muchas aplicaciones, los datos contienen observaciones de más de una red
  - ▶ Redes multicapa
  - ▶ Series de tiempo
  - ▶ Muestras de redes



Ejemplo: redes cerebrales de un grupo de pacientes

# Redes con atributos

- Problemas supervisados: clasificación o predicción de atributos



1 Introducción

2 Detección de comunidades no supervisada

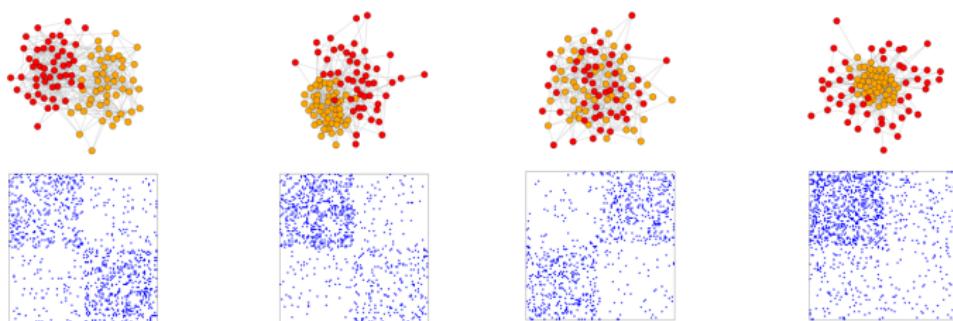
3 Detección de comunidades supervisada

# Modelo de bloques multicapa

El modelo estocástico de bloques se puede generalizar a un **modelo multicapa**

- Muestra de  $m$  redes  $\mathbf{A}^{(1)}, \dots, \mathbf{A}^{(m)}$
- Cada red es modelada con una probabilidad distinta
- Todas las redes comparten la misma división en comunidades

$$\mathbf{P}^{(i)} = \mathbf{Z}\mathbf{B}^{(i)}\mathbf{Z}^T$$

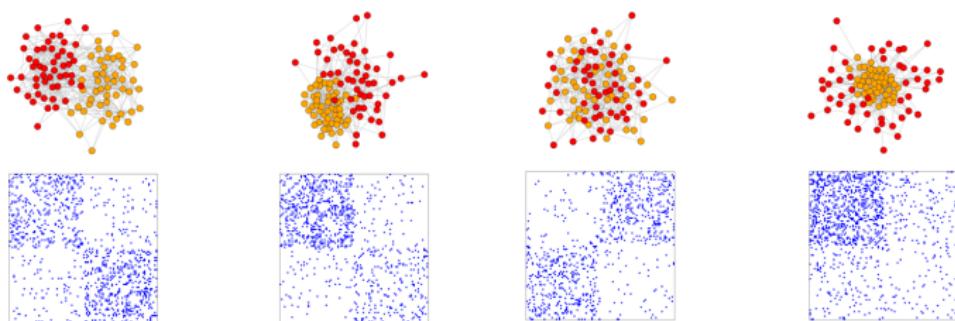


# Modelo de bloques multicapa

El modelo estocástico de bloques se puede generalizar a un **modelo multicapa**

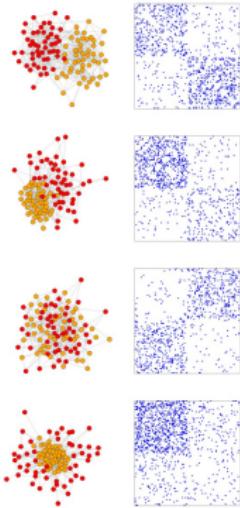
- Muestra de  $m$  redes  $\mathbf{A}^{(1)}, \dots, \mathbf{A}^{(m)}$
- Cada red es modelada con una probabilidad distinta
- Todas las redes comparten la misma división en comunidades

$$\mathbf{P}^{(i)} = \mathbf{Z}\mathbf{B}^{(i)}\mathbf{Z}^T$$

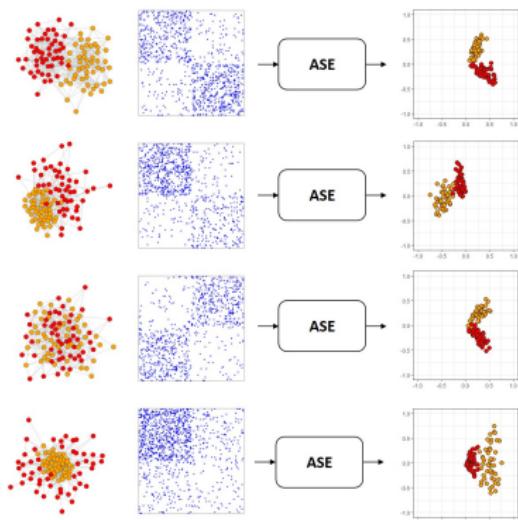


¿Cómo estimar las comunidades usando todas las redes simultáneamente?

## Incrustamiento espectral múltiple (MASE)

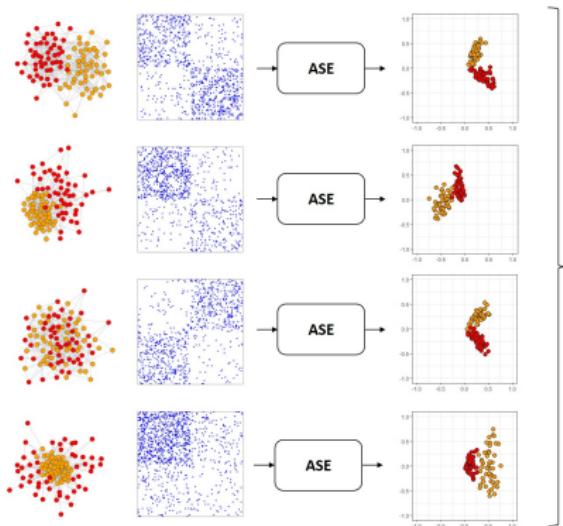


# Incrustamiento espectral múltiple (MASE)



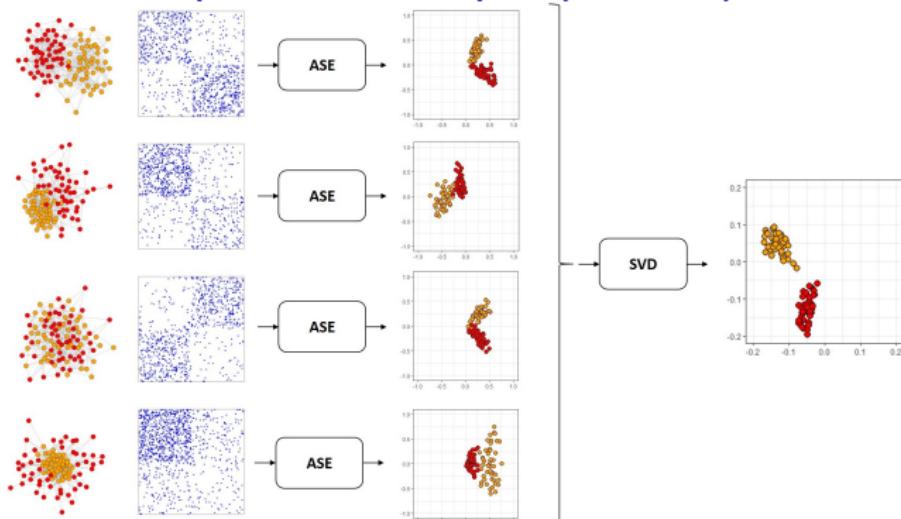
- Para cada red  $i$ , calcular la matriz de eigenvectores principales  $\hat{\mathbf{V}}^{(i)} \in \mathbb{R}^d$ .

# Incrustamiento espectral múltiple (MASE)



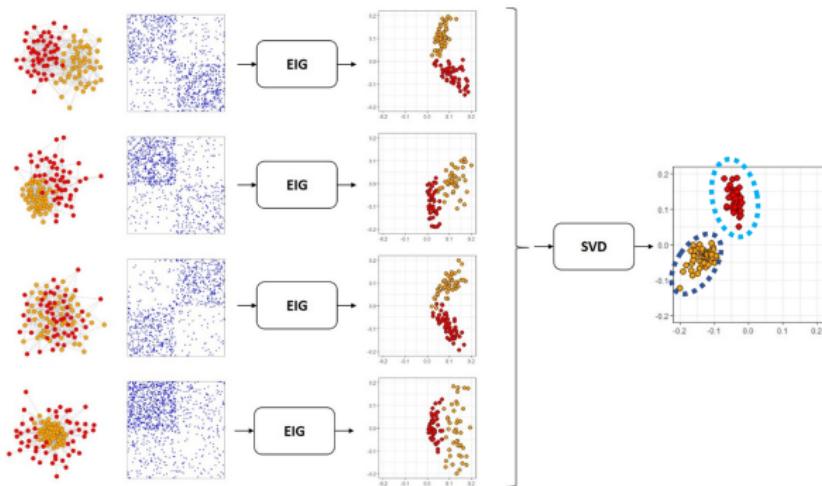
- 1 Para cada red  $i$ , calcular la matriz de eigenvectores principales  $\hat{\mathbf{V}}^{(i)} \in \mathbb{R}^d$ .
- 2 Concatenar todos los vectores en una matriz  $\hat{\mathbf{U}} = [\hat{\mathbf{V}}^{(1)} \dots \hat{\mathbf{V}}^{(m)}] \in \mathbb{R}^{n \times md}$ .

# Incrustamiento espectral múltiple (MASE)



- ① Para cada red  $i$ , calcular la matriz de vectores principales  $\hat{\mathbf{V}}^{(i)} \in \mathbb{R}^d$ .
- ② Concatenar todos los vectores en una matriz  $\hat{\mathbf{U}} = [\hat{\mathbf{V}}^{(1)} \dots \hat{\mathbf{V}}^{(m)}] \in \mathbb{R}^{n \times md}$ .
- ③ Usando la descomposición en valores singulares de  $\hat{\mathbf{U}}$ , definir  $\hat{\mathbf{V}} \in \mathbb{R}^{n \times d}$  como la matriz de vectores principales izquierdos de  $\hat{\mathbf{U}}$ .

# Incrustamiento espectral múltiple (MASE)



- ① Para cada red  $i$ , calcular la matriz de eigenvectores principales  $\hat{\mathbf{V}}^{(i)} \in \mathbb{R}^d$ .
- ② Concatenar todos los vectores en una matriz  $\hat{\mathbf{U}} = [\hat{\mathbf{V}}^{(1)} \dots \hat{\mathbf{V}}^{(m)}] \in \mathbb{R}^{n \times md}$ .
- ③ Usando la descomposición en valores singulares de  $\hat{\mathbf{U}}$ , definir  $\hat{\mathbf{V}} \in \mathbb{R}^{n \times d}$  como la matriz de vectores principales izquierdos de  $\hat{\mathbf{U}}$ .
- ④ Agrupar los vectores renglón de  $\hat{\mathbf{V}}$  usando  $k$ -medias.

# Consistencia del estimador de comunidades

- El error promedio del algoritmo de agrupamiento se define como

$$\frac{\# \text{ nodos en una comunidad incorrecta}}{n}$$

## Teorema

Bajo condiciones de regularidad, el error esperado promedio del algoritmo de detección de comunidades es

$$\mathbb{E}[\text{error promedio}] = O\left(K^2 \left(\frac{1}{nm} + \frac{1}{n^2}\right)\right).$$

# Consistencia del estimador de comunidades

- El error promedio del algoritmo de agrupamiento se define como

$$\frac{\# \text{ nodos en una comunidad incorrecta}}{n}$$

## Teorema

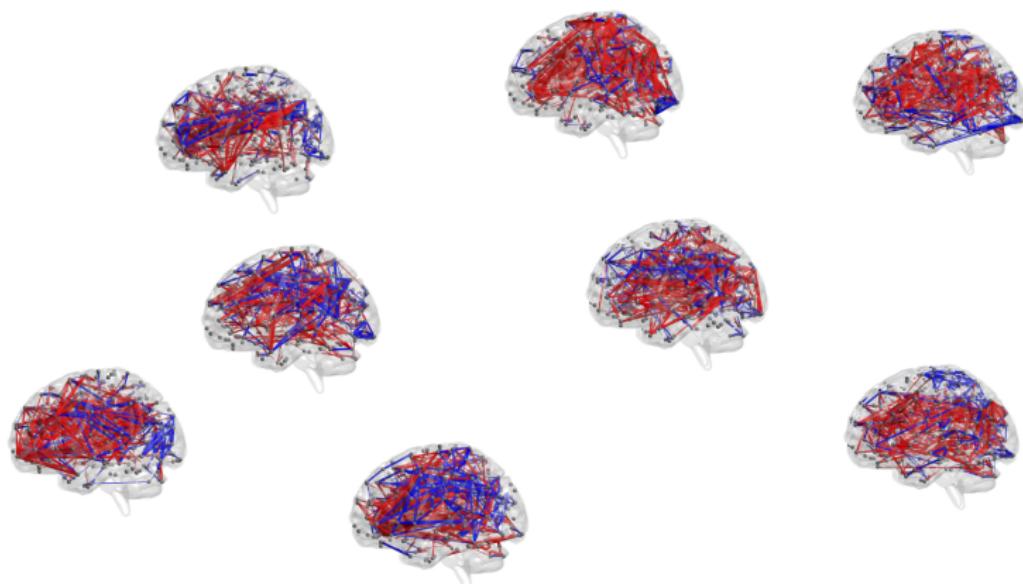
Bajo condiciones de regularidad, el error esperado promedio del algoritmo de detección de comunidades es

$$\mathbb{E}[\text{error promedio}] = O\left(K^2 \left(\frac{1}{nm} + \frac{1}{n^2}\right)\right).$$

- La detección de comunidades espectral en una sola red tiene un error promedio de  $K^2/n$  (Lei & Rinaldo, 2015)

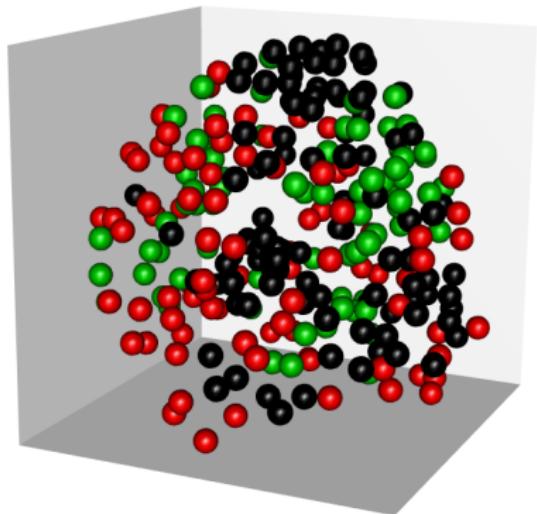
# Redes cerebrales

- La base de datos COBRE contiene una muestra de imágenes cerebrales
  - ▶ 54 pacientes esquizofrénicos, 70 personas sanas ( $n = 124$ )
  - ▶  $n = 263$  regiones de interés (nodos)



# Detección de comunidades cerebrales

$m = 1$



$m = 124$

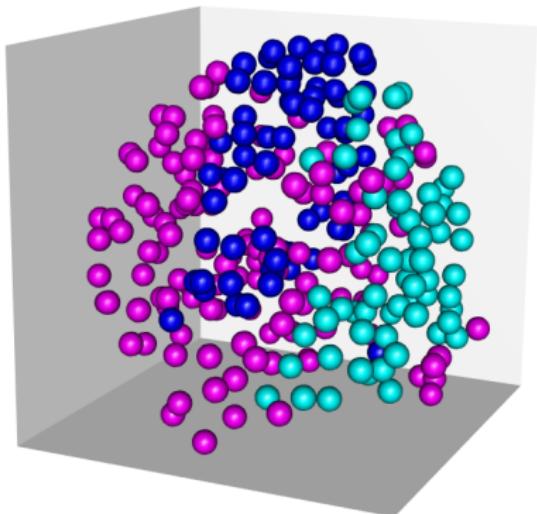


Ilustración de agrupamiento espectral con 3 comunidades.

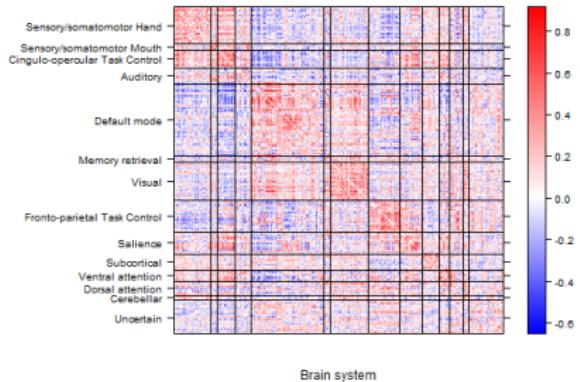
1 Introducción

2 Detección de comunidades no supervisada

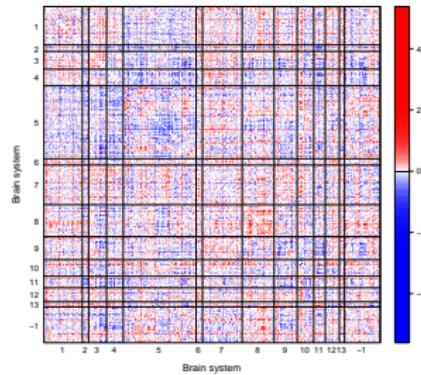
3 Detección de comunidades supervisada

# Comunidades no supervisadas

- Las comunidades no supervisadas son útiles para modelar cada red particular
- Sin embargo, no capturan las relaciones entre comunidades y atributos de la red



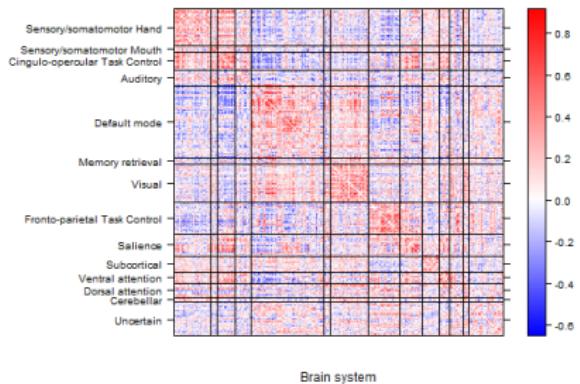
Red cerebral individual  
de un paciente



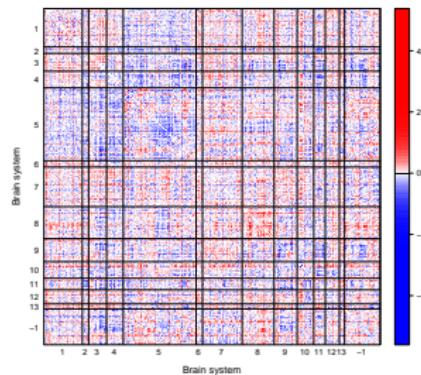
Diferencia promedio entre pacientes esquizofrénicos y sanos

# Comunidades no supervisadas

- Las comunidades no supervisadas son útiles para modelar cada red particular
- Sin embargo, no capturan las relaciones entre comunidades y atributos de la red



Red cerebral individual  
de un paciente



Diferencia promedio entre pacientes  
esquizofrénicos y sanos

¿Cómo encontrar comunidades que expliquen los atributos de una red?

## Modelo lineal para redes

- $(\mathbf{A}^{(1)}, y_1), \dots, (\mathbf{A}^{(m)}, y_m)$  es una muestra de redes con atributos
- Modelo lineal:

$$\begin{aligned}y_i &= \left\langle \mathbf{A}^{(i)}, \mathbf{B} \right\rangle + \epsilon_i \\&= \sum_{u=1}^n \sum_{v=1}^n \mathbf{A}_{uv}^{(i)} \mathbf{B}_{uv} + \epsilon_i\end{aligned}$$

- ▶  $\mathbf{B} \in \mathbb{R}^{n \times n}$  es la matriz de coeficientes
- ▶  $\epsilon_i$  es el error de regresión con varianza  $\sigma$ .

## Modelo lineal para redes

- $(\mathbf{A}^{(1)}, y_1), \dots, (\mathbf{A}^{(m)}, y_m)$  es una muestra de redes con atributos
- Modelo lineal:

$$\begin{aligned}y_i &= \langle \mathbf{A}^{(i)}, \mathbf{B} \rangle + \epsilon_i \\&= \sum_{u=1}^n \sum_{v=1}^n \mathbf{A}_{uv}^{(i)} \mathbf{B}_{uv} + \epsilon_i\end{aligned}$$

- ▶  $\mathbf{B} \in \mathbb{R}^{n \times n}$  es la matriz de coeficientes
- ▶  $\epsilon_i$  es el error de regresión con varianza  $\sigma$ .
- **Problema de alta dimensionalidad:** el número de parámetros ( $n^2$ ) usualmente es mayor al número de observaciones ( $m$ )
- Regularizaciones típicas: penalización de norma  $\ell_1$  o  $\ell_2$ .

# Comunidades supervisadas

- La matriz de coeficientes  $\mathbf{B}$  es dividida en  $K$  comunidades

$$\mathbf{B} = \mathbf{Z}\mathbf{C}\mathbf{Z}^T$$

- ▶ Comunidades:  $\mathbf{Z} \in \{0, 1\}^{n \times K}$ ,  
 $\sum_{k=1}^K \mathbf{Z}_{jk} = 1$
- ▶ Coeficientes:  $\mathbf{C} \in \mathbb{R}^{K \times K}$

# Comunidades supervisadas

- La matriz de coeficientes  $\mathbf{B}$  es dividida en  $K$  comunidades

$$\mathbf{B} = \mathbf{ZCZ}^T$$

- ▶ Comunidades:  $\mathbf{Z} \in \{0, 1\}^{n \times K}$ ,  
 $\sum_{k=1}^K \mathbf{Z}_{jk} = 1$
- ▶ Coeficientes:  $\mathbf{C} \in \mathbb{R}^{K \times K}$

- Los coeficientes (aristas) son agrupados en *celdas* con valores similares
- Las comunidades disminuyen la dimensión del problema (de  $n^2$  a  $K^2$  coeficientes) y simplifican la interpretación

## Agrupamiento espectral supervisado

- Calcular el estimador inicial  $\hat{\mathbf{B}}^{(0)}$ , definido como

$$\hat{\mathbf{B}}_{uv}^{(0)} = \text{Cov}(Y, \mathbf{A}_{uv}) = \frac{1}{m} \sum_{i=1}^m y_i \mathbf{A}_{uv}^{(i)}$$

- Aplicar agrupamiento espectral a  $\hat{\mathbf{B}}^{(0)}$
- El error promedio esperado bajo condiciones de regularidad es

$$\mathbb{E}[\text{error promedio}] = O\left(K^2 \left(\frac{1 + \sigma^2/\|\mathbf{B}\|_F^2}{mn}\right)\right).$$

## Problema de optimización

- Para estimar la matriz de coeficientes, resolvemos el siguiente problema de optimización:

$$\min_{\mathbf{B}, \mathbf{Z}, \mathbf{C}} \ell(\mathbf{B}) + \Omega(\mathbf{B})$$

$$\begin{aligned} \text{sujeto a } & \mathbf{Z} \in \{0, 1\}^{n \times K}, \quad \mathbf{Z}\mathbf{1}_K = \mathbf{1}_n \\ & \mathbf{C} \in \mathbb{R}^{K \times K} \\ & \mathbf{B} = \mathbf{Z}\mathbf{C}\mathbf{Z}^T. \end{aligned}$$

- $\ell$  es la función de mínimos cuadrados (o alguna otra función de pérdida)
- $\Omega$  es una penalización opcional (como norma  $\ell_1$ )

## Problema de optimización

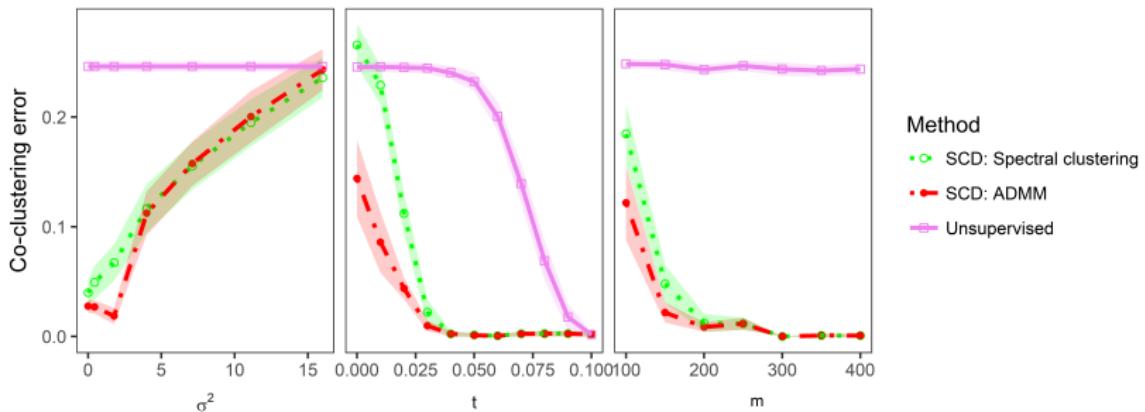
- Para estimar la matriz de coeficientes, resolvemos el siguiente problema de optimización:

$$\min_{\mathbf{B}, \mathbf{Z}, \mathbf{C}} \ell(\mathbf{B}) + \Omega(\mathbf{B})$$

$$\begin{aligned} \text{sujeto a } & \mathbf{Z} \in \{0, 1\}^{n \times K}, \quad \mathbf{Z}\mathbf{1}_K = \mathbf{1}_n \\ & \mathbf{C} \in \mathbb{R}^{K \times K} \\ & \mathbf{B} = \mathbf{Z}\mathbf{C}\mathbf{Z}^T. \end{aligned}$$

- $\ell$  es la función de mínimos cuadrados (o alguna otra función de pérdida)
- $\Omega$  es una penalización opcional (como norma  $\ell_1$ )
- El problema no es convexo, pero el algoritmo de agrupamiento espectral es una buena inicialización.
- La solución del problema es aproximada usando un algoritmo ADMM.

# Resultados en redes sintéticas



# Clasificación de redes cerebrales

- Regresión logística para encontrar comunidades supervisadas y distinguir pacientes esquizofrénicos y sanos



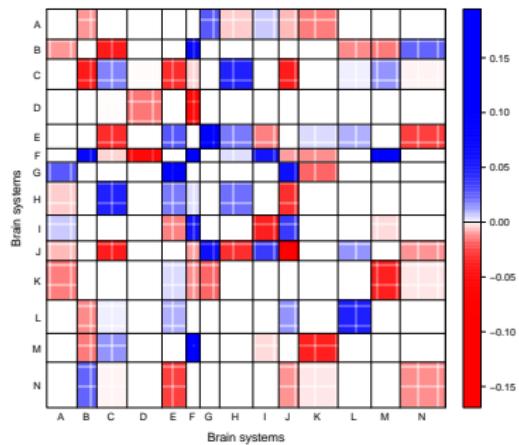
Comunidades de Power et al. (2011)  
CV accuracy: 62 %



Comunidades supervisadas  
CV accuracy: 79 %

# Comunidades supervisadas

- La *red neuronal por defecto* (comunidad 5 de Power et al. (2011)) es comúnmente asociada a esquizofrenia.
- El algoritmo supervisado proporciona una mejor partición de esta región



Coeficientes estimados

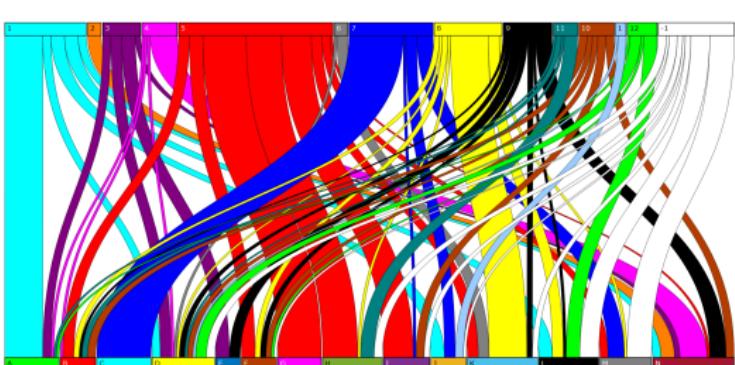
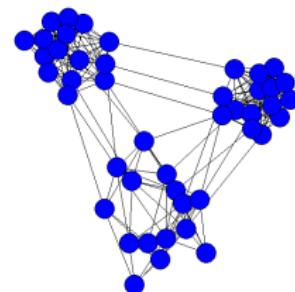
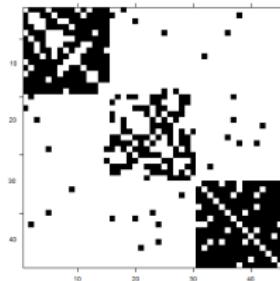


Diagrama de Sankey diagram:  
Power (top) vs DCS (bottom)

# Conclusiones

- El aprendizaje estadístico en muestras de redes necesita el desarrollo de nuevos métodos
- Las comunidades son una estructura importante para entender e interpretar datos de redes.
- La inferencia espectral es una herramienta útil en el análisis de redes.



# Referencias

- **Inferencia no supervisada en muestras de redes:**

Arroyo, J., Athreya, A., Cape, J., Chen, G., Priebe, C. E., Vogelstein, J. T. (2019). *Inference for multiple heterogeneous networks with a common invariant subspace*. arXiv preprint arXiv:1906.10026, aceptado en Journal of Machine Learning Research.

- **Detección de comunidades supervisada:**

Arroyo, J., Levina, E. (2020). *Simultaneous prediction and community detection for networks with application to neuroimaging*. arXiv preprint arXiv:2002.01645.

# Referencias

- **Inferencia no supervisada en muestras de redes:**  
Arroyo, J., Athreya, A., Cape, J., Chen, G., Priebe, C. E., Vogelstein, J. T. (2019). *Inference for multiple heterogeneous networks with a common invariant subspace*. arXiv preprint arXiv:1906.10026, aceptado en Journal of Machine Learning Research.
- **Detección de comunidades supervisada:**  
Arroyo, J., Levina, E. (2020). *Simultaneous prediction and community detection for networks with application to neuroimaging*. arXiv preprint arXiv:2002.01645.

¡Gracias!