Aprendizaje en grafos, optimización robusta y la métrica de Wasserstein nuclear.

Daniel De Roux (Universidad de los Andes, Colombia)
Mauricio Velasco* (Universidad de los Andes, Colombia)

Escuela de Física matemática 2019.

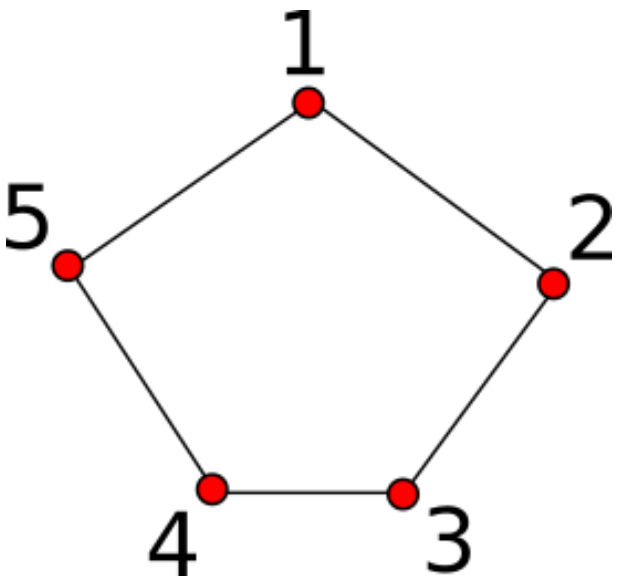
Plan de la charla

1. Grafos y Grafos Aleatorios.
2. **Problema**: Cómo construir (aprender) resúmenes deterministas de grafos aleatorios a partir de muestras.
3. Preliminares:
    1. Optimización robusta
    2. Distancia de Wasserstein.
4. Resultados:
    1. Un nuevo algoritmo usando la norma espectral de Wasserstein.
    2. Certificados probabilísticos de desempeño del algoritmo en el problema de detección de comunidades.
    3. Performance del nuevo algoritmo en simulaciones.

Grafos

Definición

Un grafo G consiste de un conjnto finito $V(G)$ de vértices y un conjunto de aristas $E(G) \subseteq V(G) \times V(G)$.



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

Grafos

Son una estructura matemática fundamental que aparece en todas partes:

1. (Votos de congresistas) Los vértices son los congresistas y conectamos dos de ellos si votan de la misma manera en un proyecto específico (G tiene 268 vértices).
2. (Protein-Protein interaction networks) Los vértices son las proteinas de un organismo y conectamos dos de ellas si interactúan en algun proceso biológico del organismo (para H. Pilori son 1515 proteinas y se conocen (2014) 3004 aristas).
3. (Amigos de Facebook) Los vértices son personas y dos se conectan si son "amigos" en Facebook (tiene 2230 millones de vértices).

Grafos

Muchos de los grafos anteriores cambian en el tiempo y según varios factores a veces de manera muy complicada.

Qué estructura matemática nos sirve para modelar el comportamiento de redes complejas?

Grafos Aleatorios

Definición

Un grafo aleatorio con vértices $1, \ldots, n$ *es una variable aleatoria* $\mathcal{G}$ *que toma valores en el conjunto de los grafos con vértices* $1, 2 \ldots, n$.

Para cualquier propiedad (P) *(por ejemplo ser conexo o contener a la arista* 12*) podemos saber sólamente*

$$\mathbb{P}\{\mathcal{G} \text{ satisface } (P)\} = \alpha$$

Para que nuestro modelo probabilístico sea adecuado basta con que refleje las propiedades estadísticas de nuestras muestras.

Ejemplo: Stochastic block model

Dados:

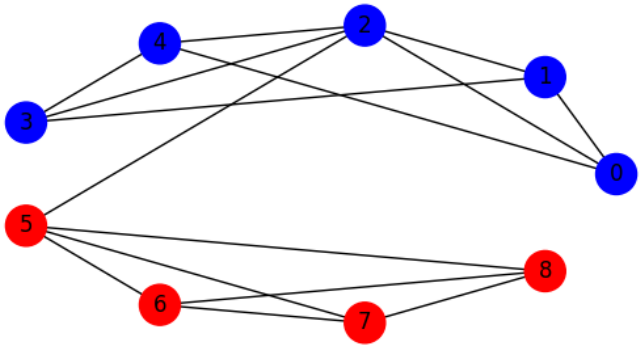
1. Una partición de $[n]$ en dos conjuntos disyuntos (clusters) C_1, C_2
2. Números reales p_1, p_2, q con $0 \leq q < \frac{1}{2} < p_1, p_2 \leq 1$

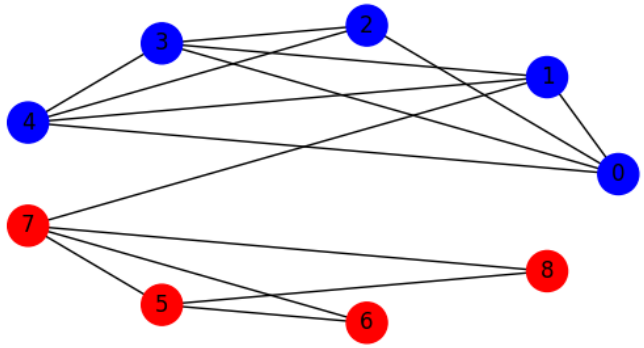
Definición

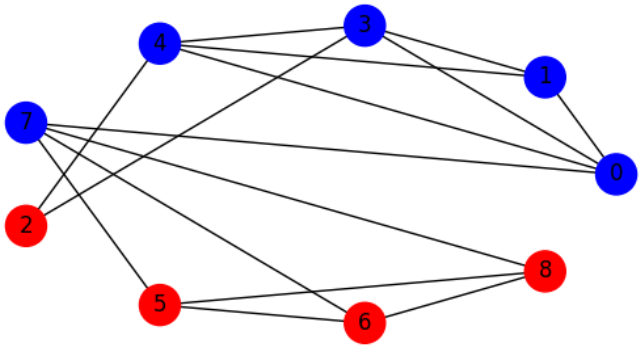
Producimos una instancia de nuestro grafo aleatorio $\mathcal{G}$ poniendo aristas de manera independiente. La probabilidad de poner una arista entre los vértices i, j es

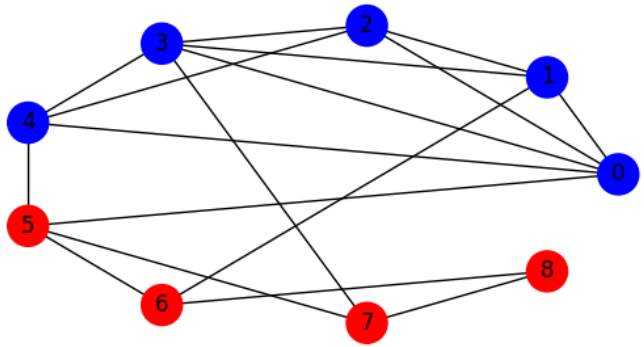
$$p_{ij} := \begin{cases} p_1, \text{ si } i,j \in C_1 \\ p_2, \text{ si } i,j \in C_2 \\ q, \text{ de lo contrario.} \end{cases}$$

Ejemplo: $p_1 = 0.9$ $p_2 = 0.9$, $q = 0.04$

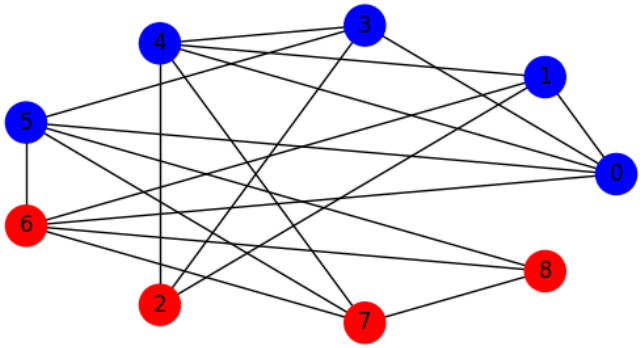


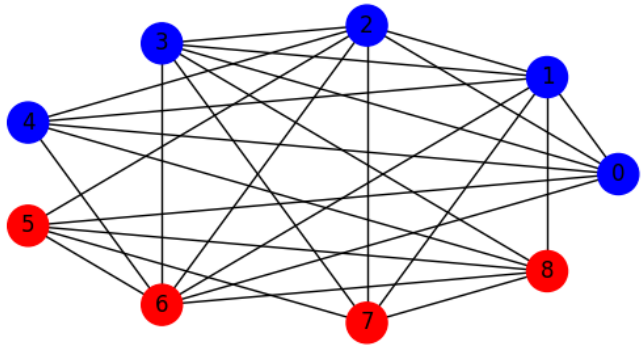


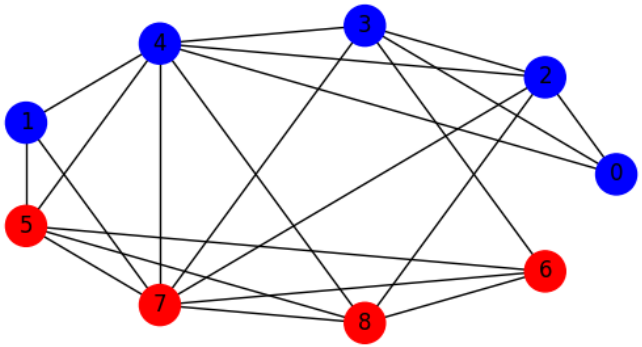


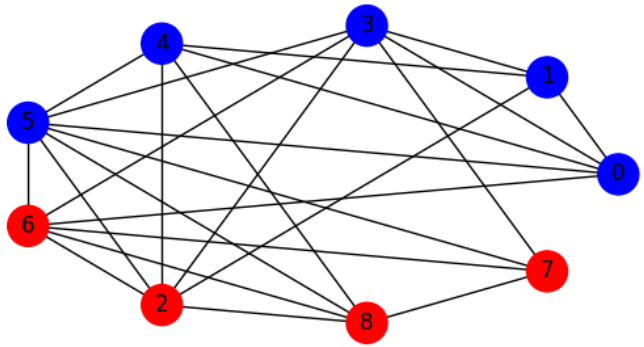


Ejemplo: $p_1 = 0.8$ $p_2 = 0.9$, $q = 0.4$



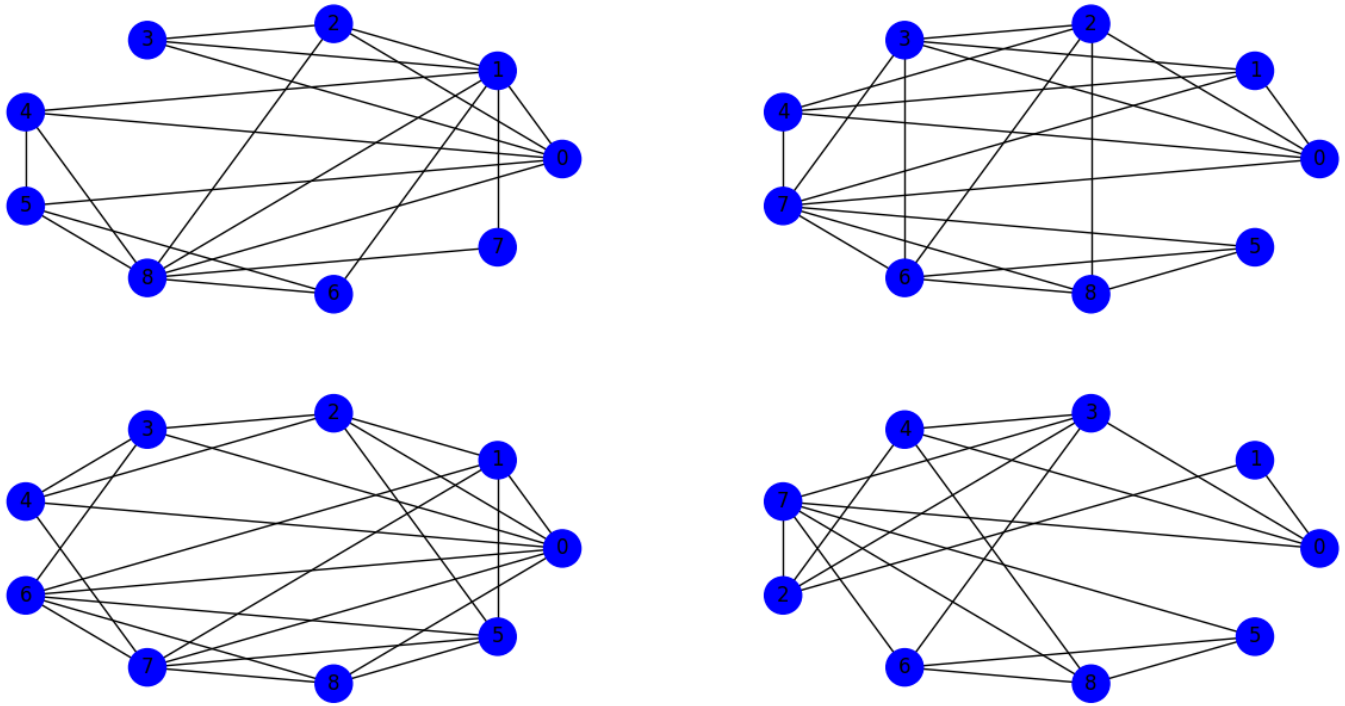






Aprendiendo Grafos Aleatorios

Generalmente no conocemos la distribución y lo que sabemos de un grafo aleatorio es una colección de observaciones del mismo:



Aprendizaje. *Como encontrar (aprender) propiedades estadísticas de un grafo aleatorio a partir de una muestra* $B_1, \ldots, B_N$ *de* $\mathcal{G}$*?*

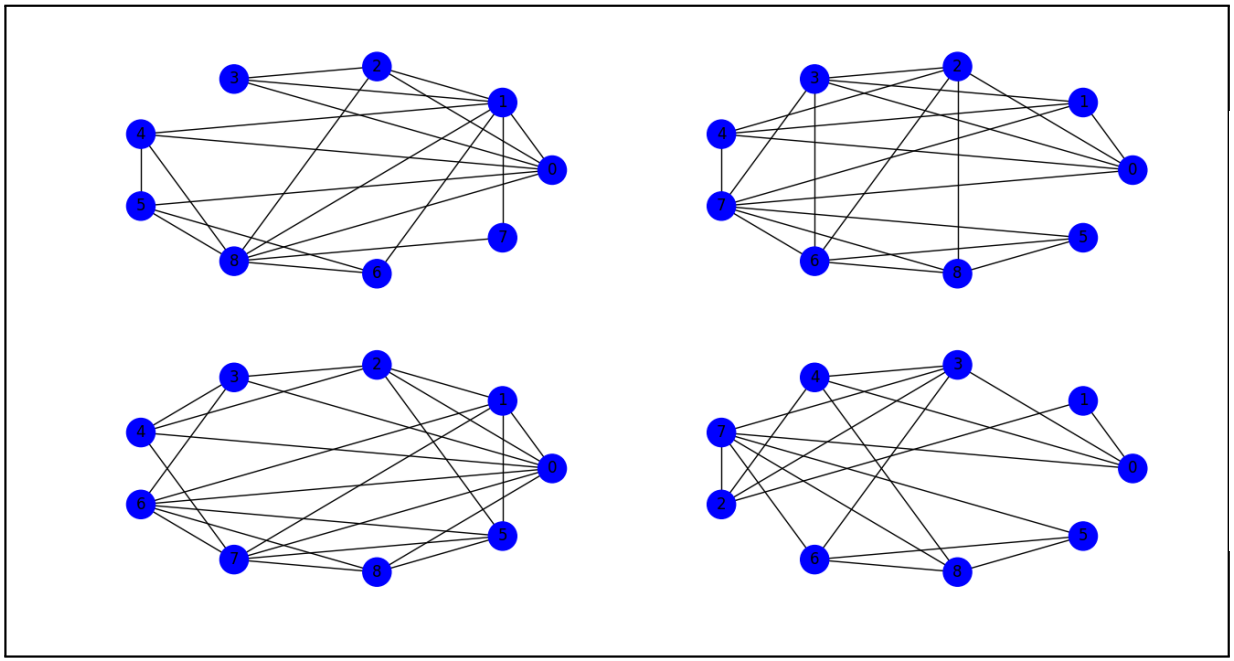
El Problema

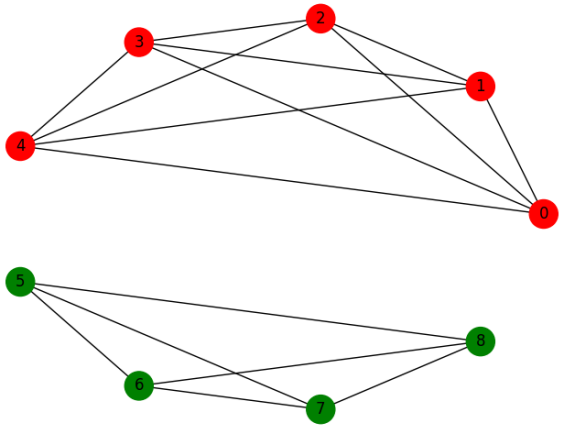
Definición

Un **resumen determinista** *de un grafo aleatorio $\mathcal{G}$ es un grafo A^* que, en promedio, difiere de $\mathcal{G}$ en la menor cantidad posible de aristas.*

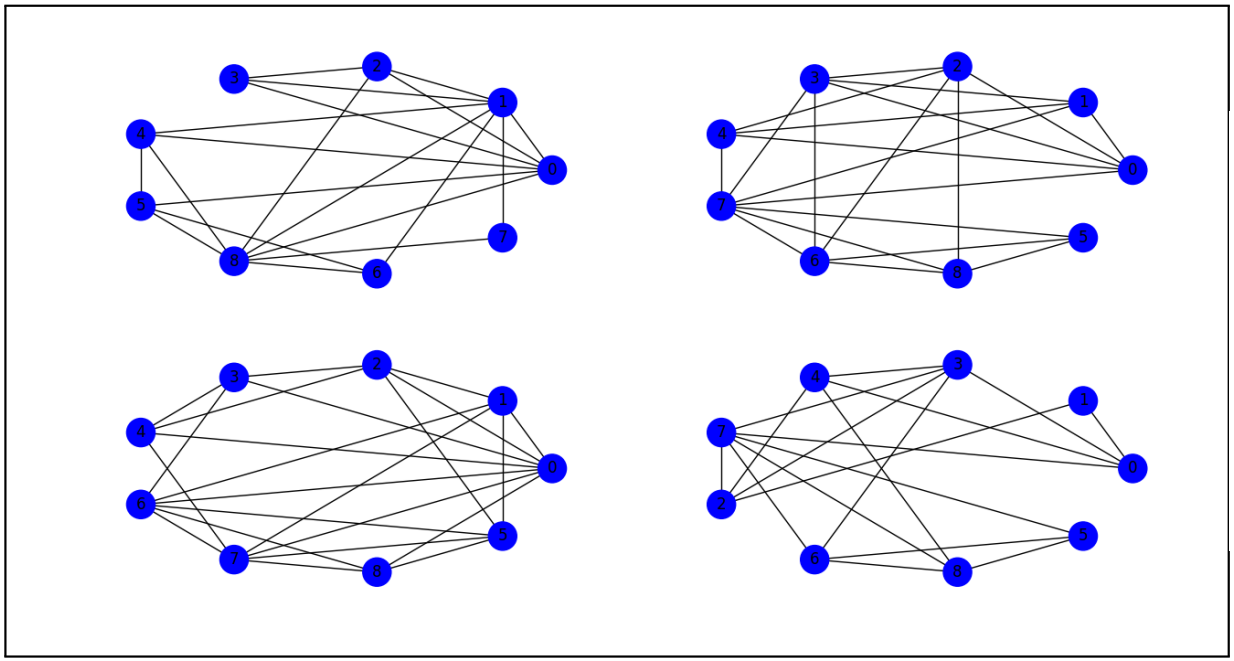
Problema. *Cómo encontrar el resumen determinista A^* de un grafo aleatorio $\mathcal{G}$ a partir de una muestra independiente $B_1, \ldots, B_N$?*

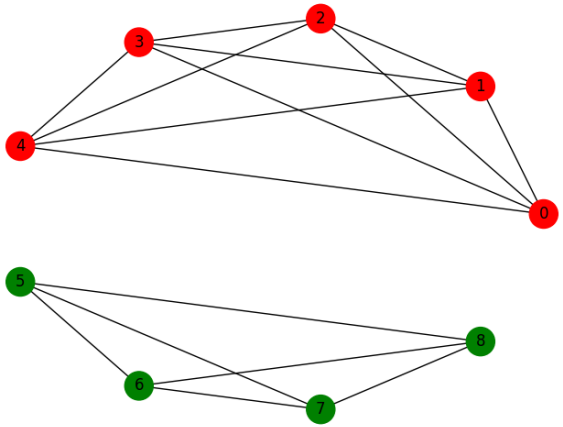
Ejemplo: Qué queremos?





Ejemplo: Qué queremos?





Resumenes deterministas como optimización.

Definición

Un **resumen determinista** *de un grafo aleatorio $\mathcal{G}$ es un grafo determinista A^* que, en promedio, difiere de $\mathcal{G}$ en la menor cantidad posible de aristas.*

Problema. *(Resumen Determinista) Encuentre la matriz A^* con entradas en $\{0,1\}$ que minimiza la función*

$$R(A) := \mathbb{E}_{B\sim\mathcal{G}}[\|A - B\|_1]$$

a partir de una muestra independiente $B_1, \ldots, B_N$ de $\mathcal{G}$.

Más precisamente intentamos encontrar un estimado $\overline{A} = \overline{A}(B_1, \ldots, B_N)$ de A^*.

Minimización de Riesgo empírico

Idea 1. *Construir la medida empírica* $\hat{\mu} := \frac{1}{N}\sum_{i=1}^{N} \delta_{B_i}$ *y usarla como aproximación de la distribución de* $\mathcal{G}$*.*

Es decir encontramos $\overline{A}$ minimizando el *riesgo empírico*

$$\hat{R}(A) := \mathbb{E}_{B\sim\hat{\mu}}[\|A - B\|_1] = \frac{1}{N}\sum_{i=1}^{N} \|A - B_i\|_1$$

lo que lleva a un problema de optimización lineal.

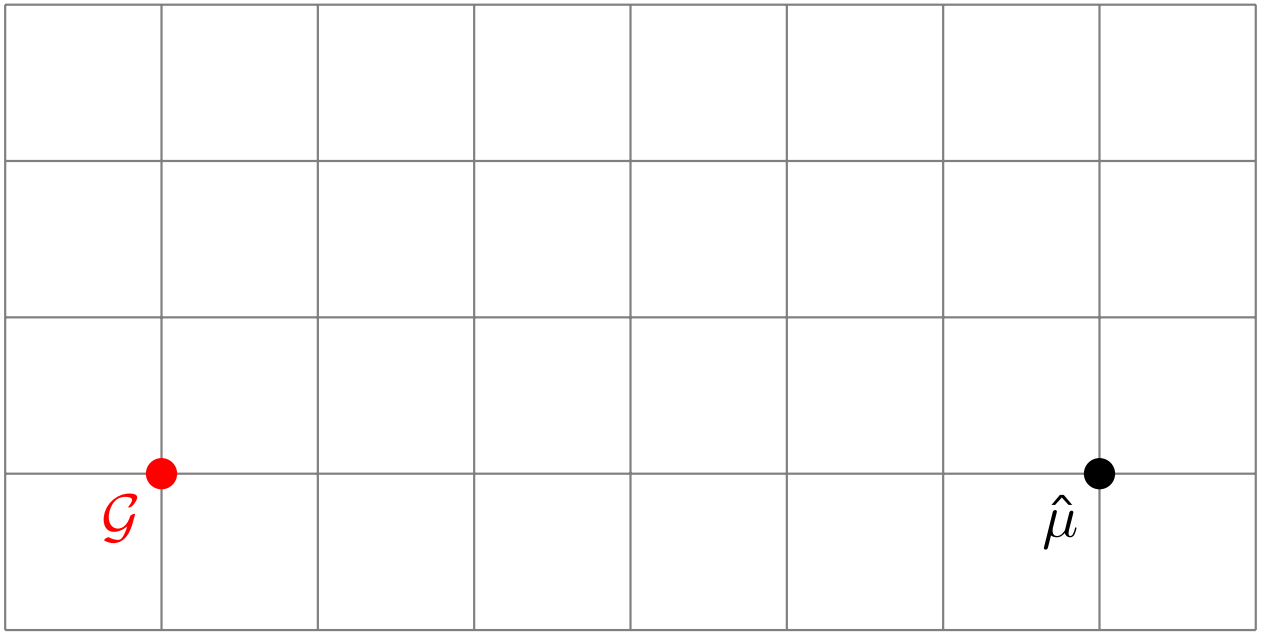
Minimización de riesgo empírico.

Idea 1. *Construir la medida empírica* $\hat{\mu} := \frac{1}{N}\sum_{i=1}^{N} \delta_{B_i}$ *y usarla como aproximación de la distribución de* $\mathcal{G}$*.*

La ventaja de la Idea 1 es que lleva a un algoritmo consistente. Nuestros estimadores $\overline{A}$ convergerán al resumen A^* cuando $N \to \infty$.

El problema con la Idea 1. es que las medidas $\hat{\mu}$ *y* $\mathcal{G}$ *pueden ser muy distintas entre sí cuando el tamaño de la muestra* N *es pequeño. Esto lleva a que nuestro minimizador* $\overline{A}$ *esté lejos de* A^* *para* N *pequeño (over-fitting).*

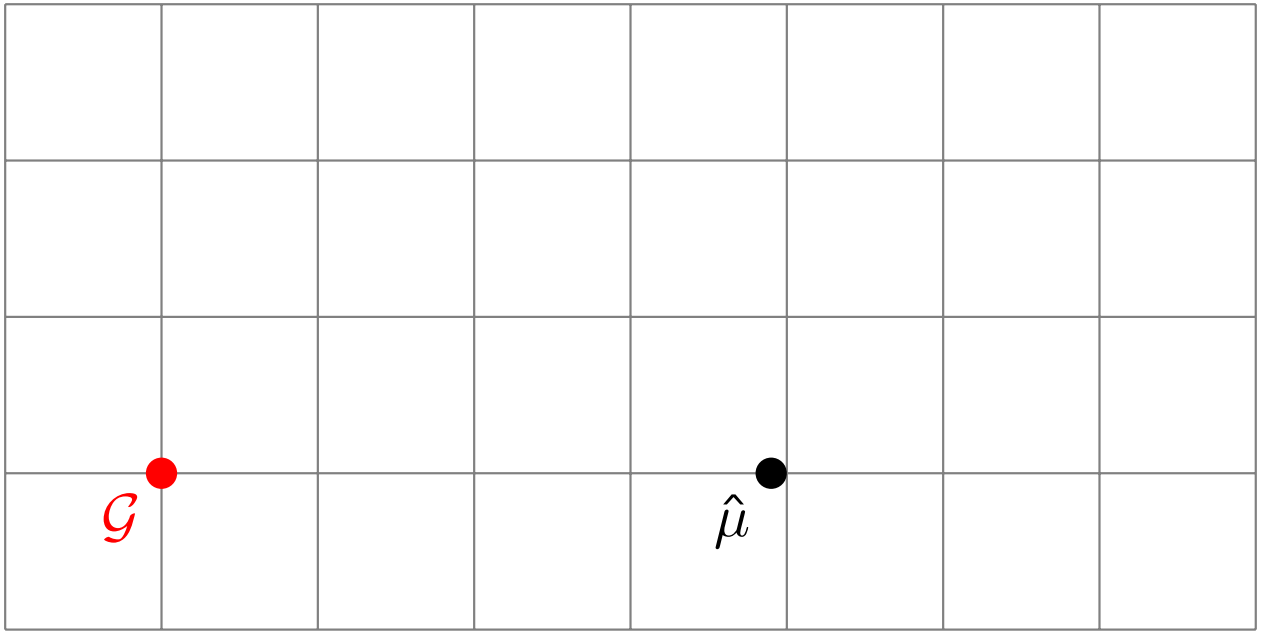
Gráficamente: Riesgo empírico (Idea 1)



$$\hat{R}(A) := \mathbb{E}_{B\sim\hat{\mu}}[\|A - B\|_1]$$

Gráficamente: Riesgo empírico (Idea 1)

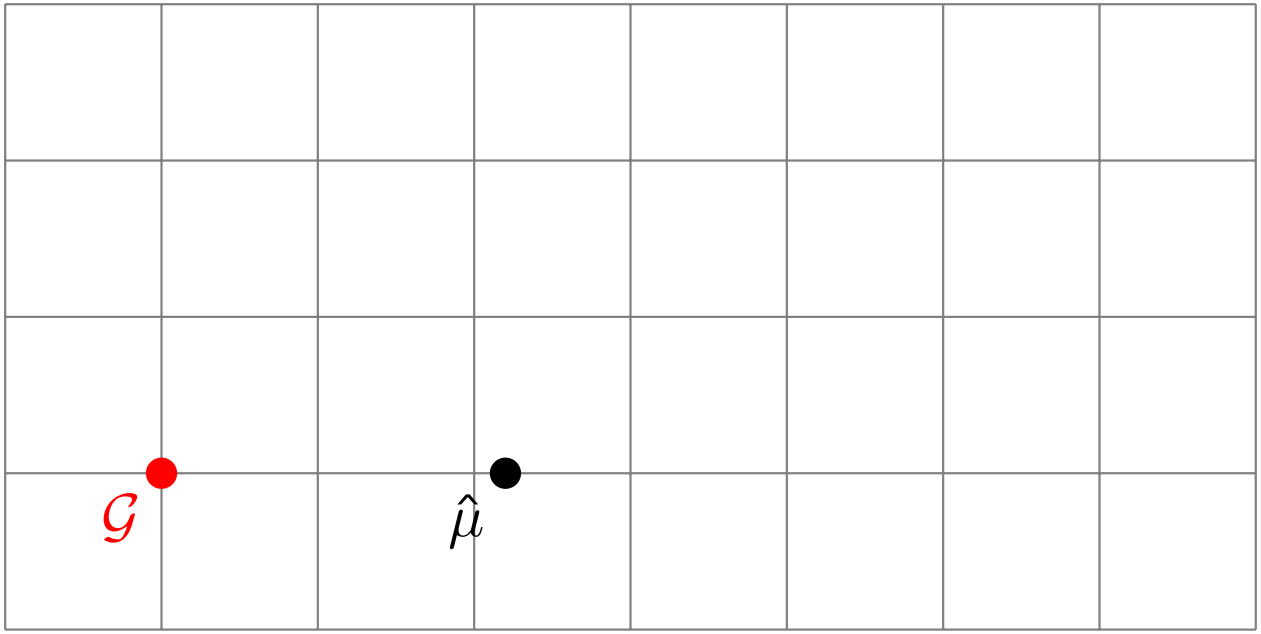
Mientras N crece...



$$\hat{R}(A) := \mathbb{E}_{B\sim\hat{\mu}}[\|A - B\|_1]$$

Gráficamente: Riesgo empírico (Idea 1)

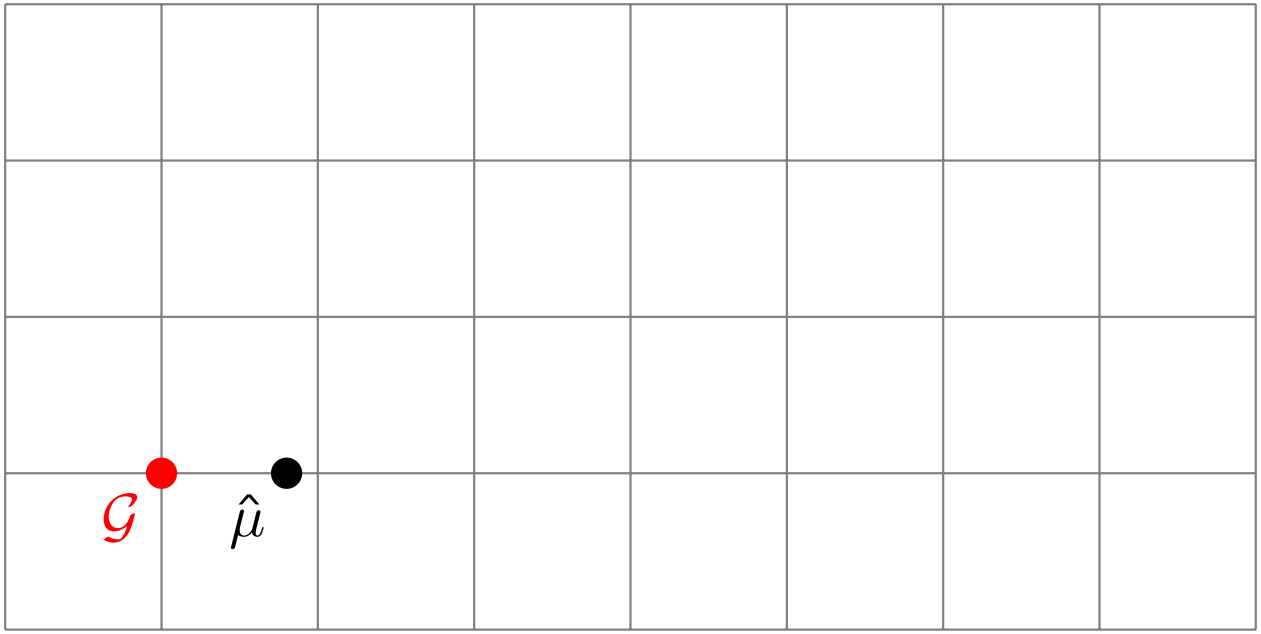
Mientras N crece...



$$\hat{R}(A) := \mathbb{E}_{B\sim\hat{\mu}}[\|A - B\|_1]$$

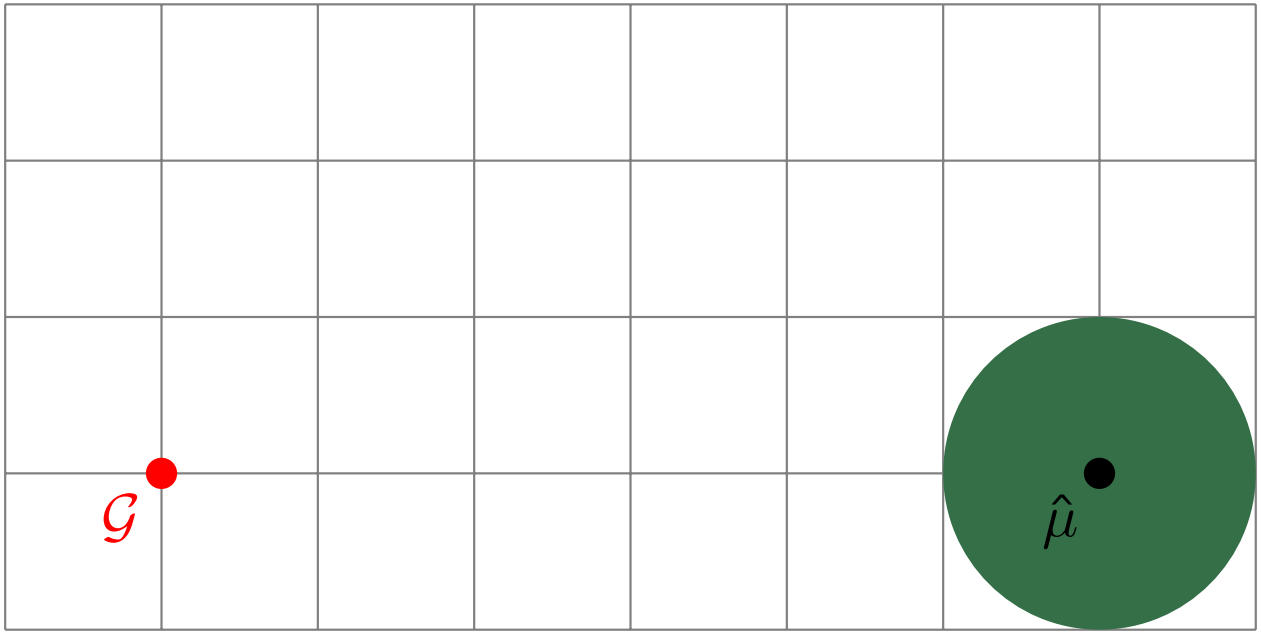
Gráficamente: Riesgo empírico (Idea 1)

Mientras N crece...



$$\hat{R}(A) := \mathbb{E}_{B\sim\hat{\mu}}[\|A - B\|_1]$$

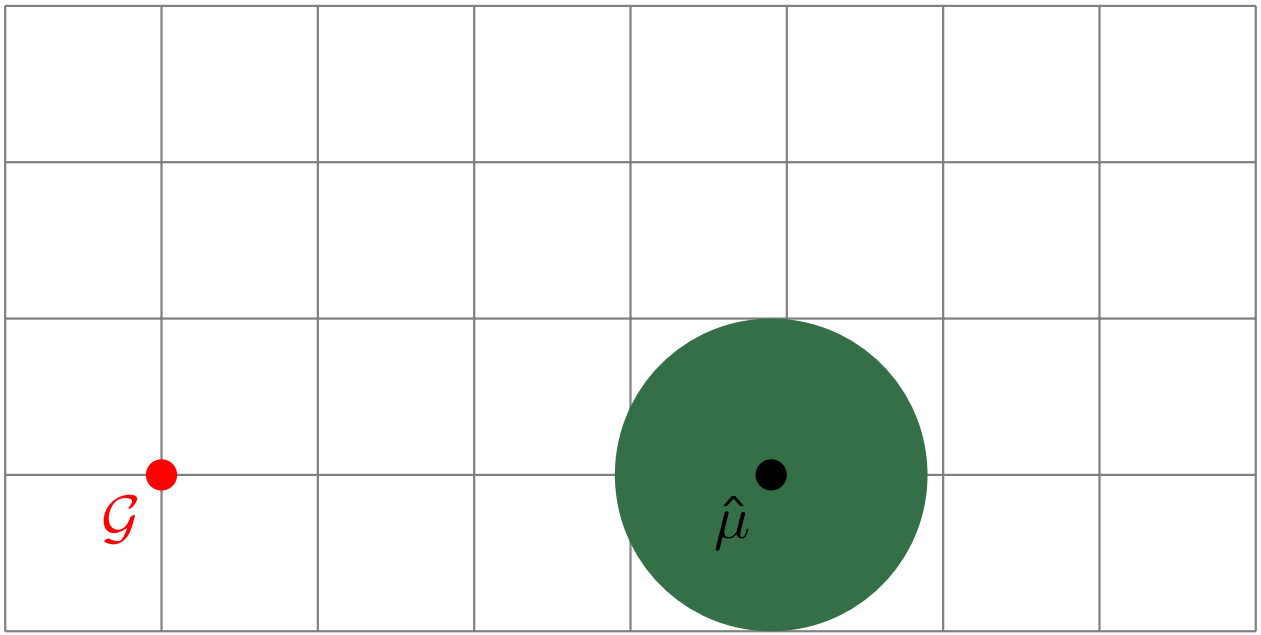
Gráficamente: Riesgo robusto (Idea 2)



$$R(A) := \sup_{\nu \in B_\delta(\hat{\mu})} (\mathbb{E}_{B\sim\nu}[\|A - B\|_1])$$

Gráficamente: Riesgo robusto (Idea 2)

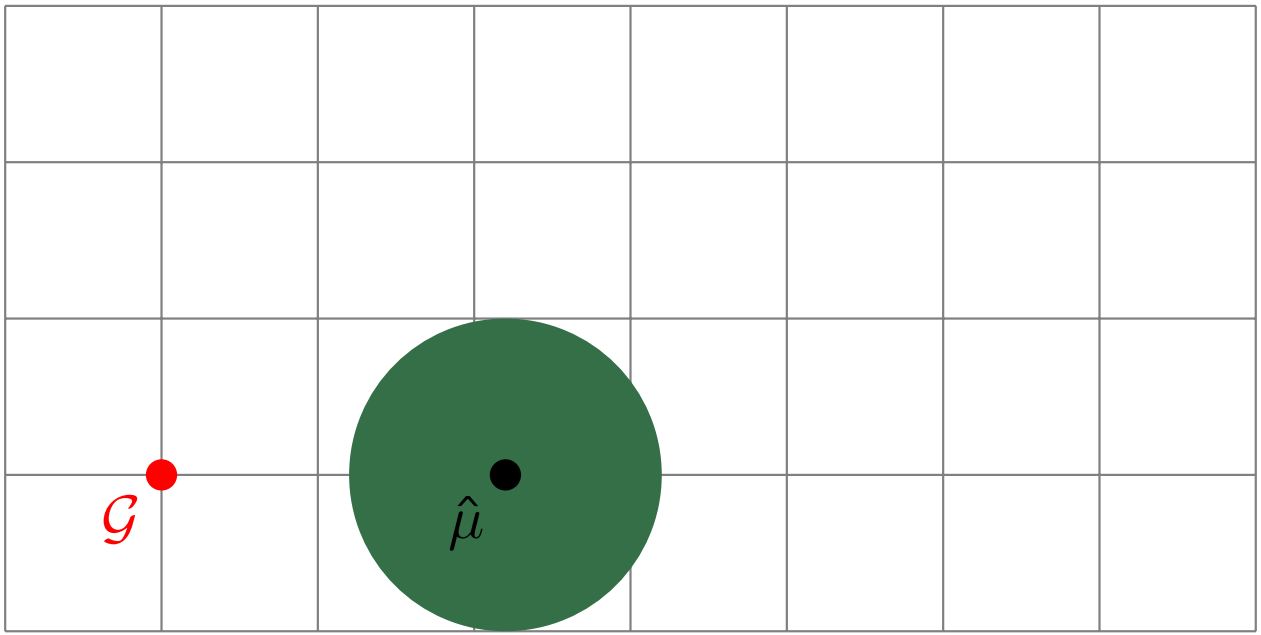
Mientras N crece...



$$R(A) := \sup_{\nu \in B_\delta(\hat{\mu})} (\mathbb{E}_{B\sim\nu}[\|A - B\|_1])$$

Gráficamente: Riesgo robusto (Idea 2)

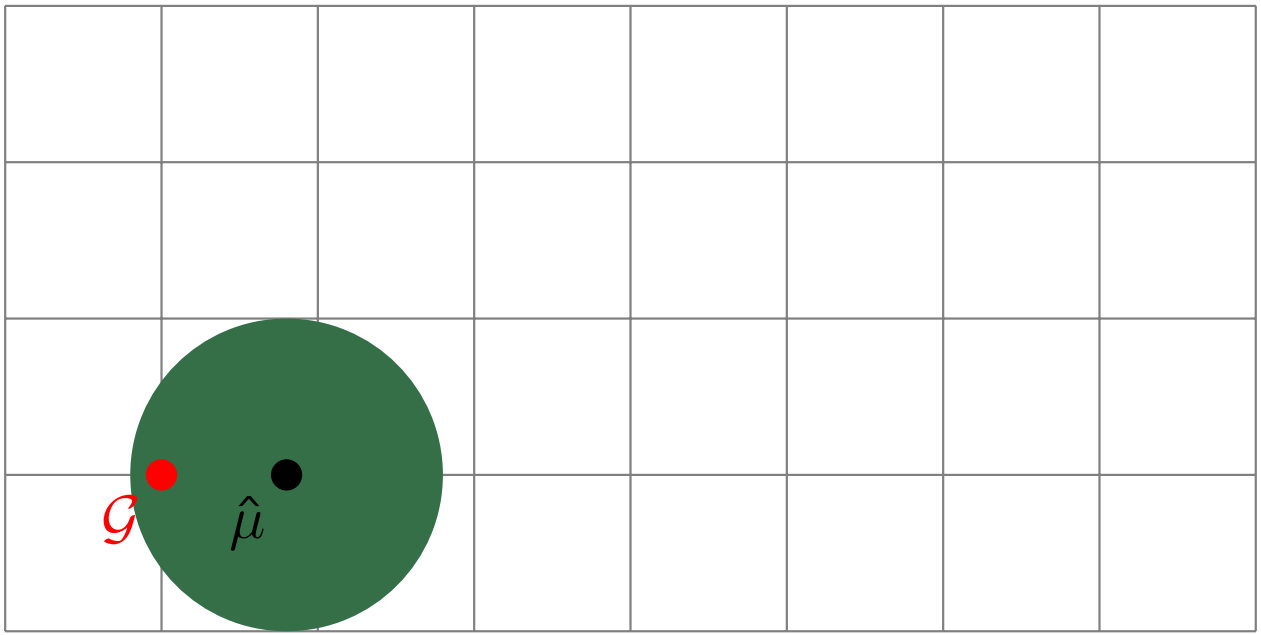
Mientras N crece...



$$R(A) := \sup_{\nu \in B_\delta(\hat{\mu})} (\mathbb{E}_{B\sim\nu}[\|A - B\|_1])$$

Gráficamente: Riesgo robusto (Idea 2)

Mientras N crece...



$$R(A) := \sup_{\nu \in B_\delta(\hat{\mu})} (\mathbb{E}_{B\sim\nu}[\|A - B\|_1])$$

Minimización del riesgo robusto.

Encontramos $\overline{A}$ minimizando el riesgo empírico robusto

$$R(A) := \sup_{\nu \in B_\delta(\hat{\mu})} (\mathbb{E}_{B\sim\nu}[\|A - B\|_1])$$

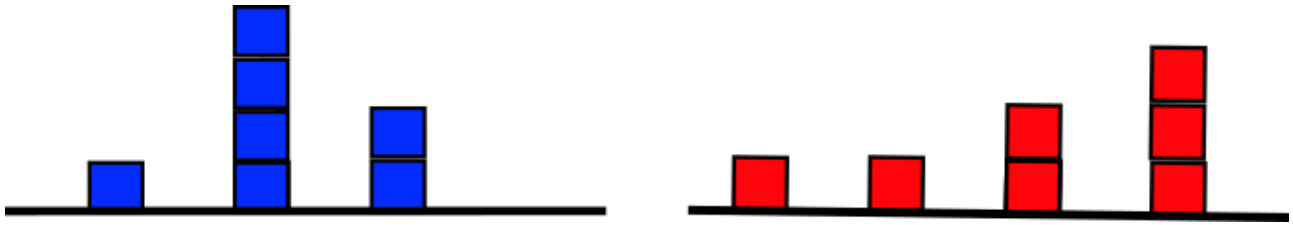
Donde $B_\delta(\hat{\mu})$ es la bola de radio δ centrada en $\hat{\mu}$ en alguna métrica entre distribuciones de probabilidad.

Note que este es un problema de optimización en dimensión infinita y es típicamente muy difícil.

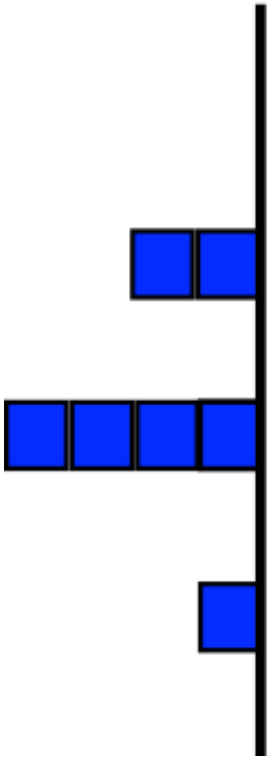
Distancia entre distribuciones de probabilidad

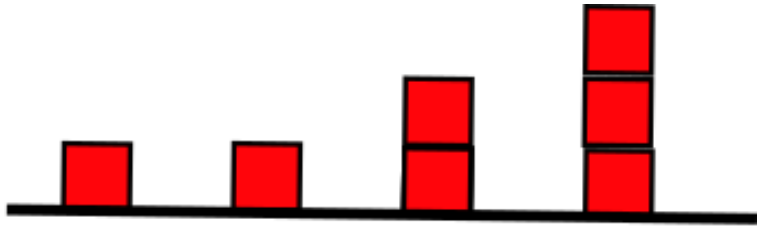
Definición

La métrica de Wasserstein (earth mover's distance) entre dos distribuciones de probabilidad ν_1, ν_2 es la mínima cantidad de trabajo necesario para transformar una en la otra.

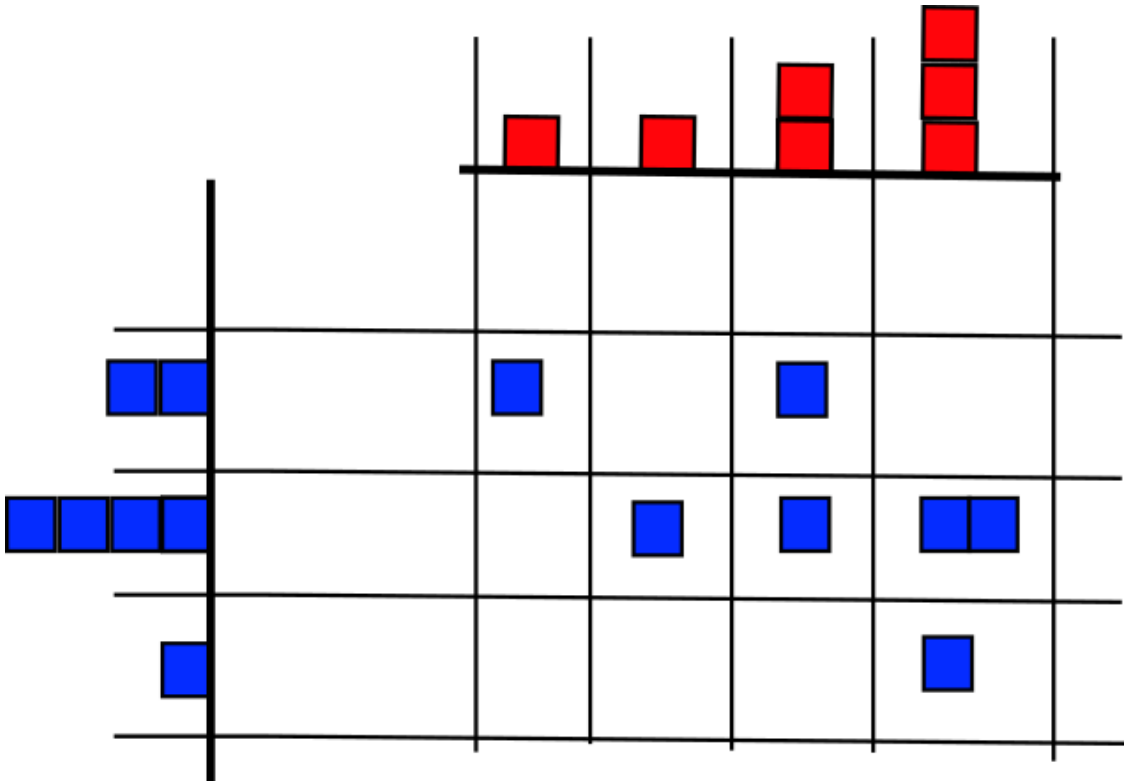


Planes de transporte





Planes de transporte



$$\begin{pmatrix} \frac{1}{7} & 0 & \frac{1}{7} & 0 \\ 0 & \frac{1}{7} & \frac{1}{7} & \frac{2}{7} \\ 0 & 0 & 0 & \frac{1}{7} \end{pmatrix}$$

La métrica de Wasserstein

Sean ν_1, ν_2 distribuciones de probabilidad en K.

Definición

Sea $\Pi(\nu_1, \nu_2)$ el conjunto de planes de transporte, es decir, distribuciones de probabilidad en $K \times K$ cuyas distribuciones marginales coinciden con ν_1 y ν_2 resp.

Definición

la métrica de Wasserstein asociada a una norma $\| \bullet \|$ en K esta dada por:

$$\mathcal{W}(\nu_1, \nu_2) := \inf_{\pi \in \Pi(\nu_1, \nu_2)} \mathbb{E}\|X - Y\|$$

donde $(X, Y) \sim \pi$.

Una formulación precisa.

Dada una muestra $B_1, \ldots, B_N$, definimos $\hat{\mu}$ y definimos $\overline{A}$ como un minimizador de

$$R(A) := \sup_{\nu \in B_\delta(\hat{\mu})} (\mathbb{E}_{B \sim \nu}[\|A - B\|_1])$$

donde $B_\delta(\hat{(}\mu))$ es la bola de radio δ en la métrica de Wasserstein asociada a $\| \bullet \|$.

1. *Cómo encontrar $\overline{A}$?*
2. *Qué propiedades tiene el minimizador $\overline{A}$?:*
   1. *Es consistente (i.e. $\overline{A} \to A^*$ cuando $N \to \infty$)?*
   2. *Para N fijo qué tan lejos esta $\overline{A}$ de A^*?*

Cómo encontrar $\overline{A}$?

Si medimos la distancia entre probabilidades usando la Métrica de Wasserstein asociada a la norma espectral entre matrices entonces:

Teorema. (De Roux, -)

Las siguientes afirmaciones son ciertas:

1. $\overline{A}$ *es el óptimo de un problema de optimización semidefinida.*
2. $\overline{A}$ *nos da un estimados consistente de* A^* *en el sentido en que existe un radio* $\delta(N)$ *tal que* $\overline{A} \to A^*$ *cuando* $N \to \infty$.

Este Teorema se sigue de resultados recientes de [Esfahani-Kuhn] sobre "tratabilidad" de optimización robusta en la métrica de Wasserstein y resultados de [Boyd, Fazel y Parrilo] sobre representabilidad de normas mediante espectraedros.

Teorema. (De Roux, -)

La siguiente desigualdad es cierta para cualquier norma $\|\bullet\|$ y A en K.

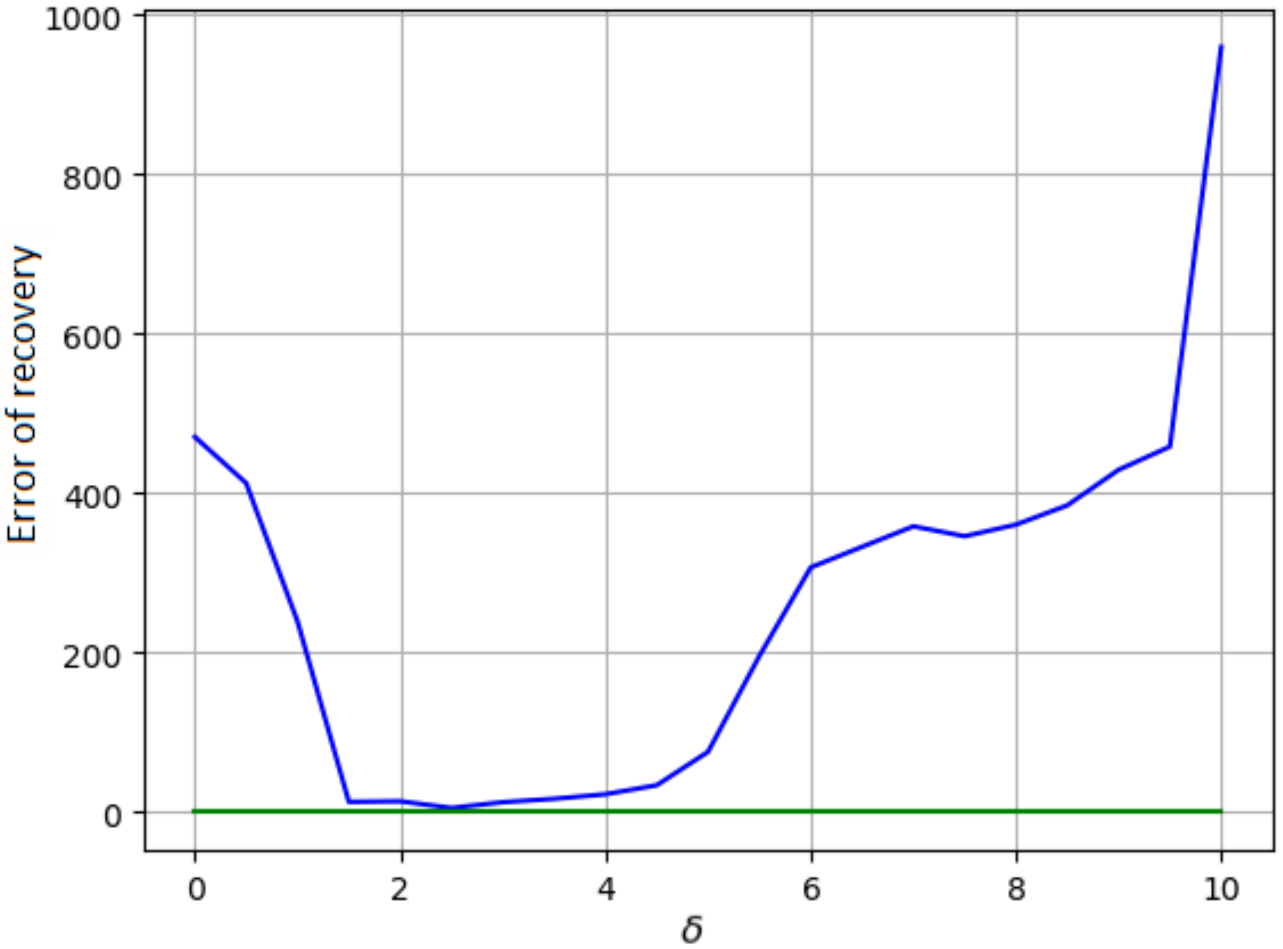
$$\sup_{\nu \in B_\delta(\hat{\mu})} \left(\mathbb{E}_{B\sim\nu}[\|A-B\|_1]\right) \leq \min_{A\in K} \Delta(A)$$

con

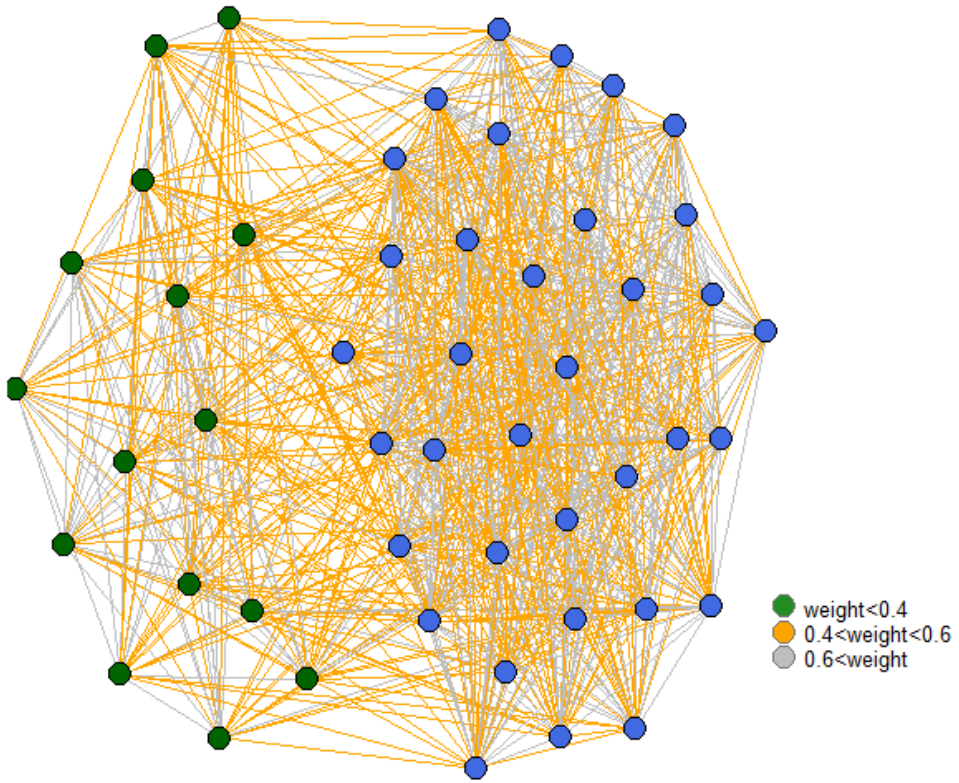
$$\Delta(A) := \frac{1}{N}\sum_{i=1}^{N} \|A-B_i\|_1 + \delta\|2A-11^t\|_*.$$

El problema de la derecha es una forma de riesgo empírico regularizado (y ambos lados coinciden cuando el óptimo tiene entradas en $\{0,1\}$).

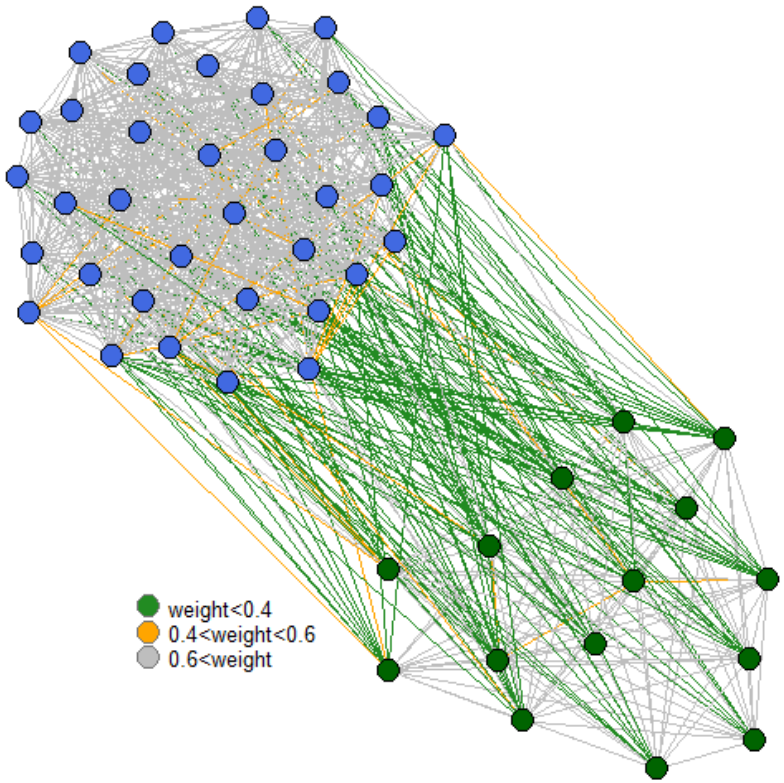
Ejemplo: Error de recuperación para diferentes valores de δ



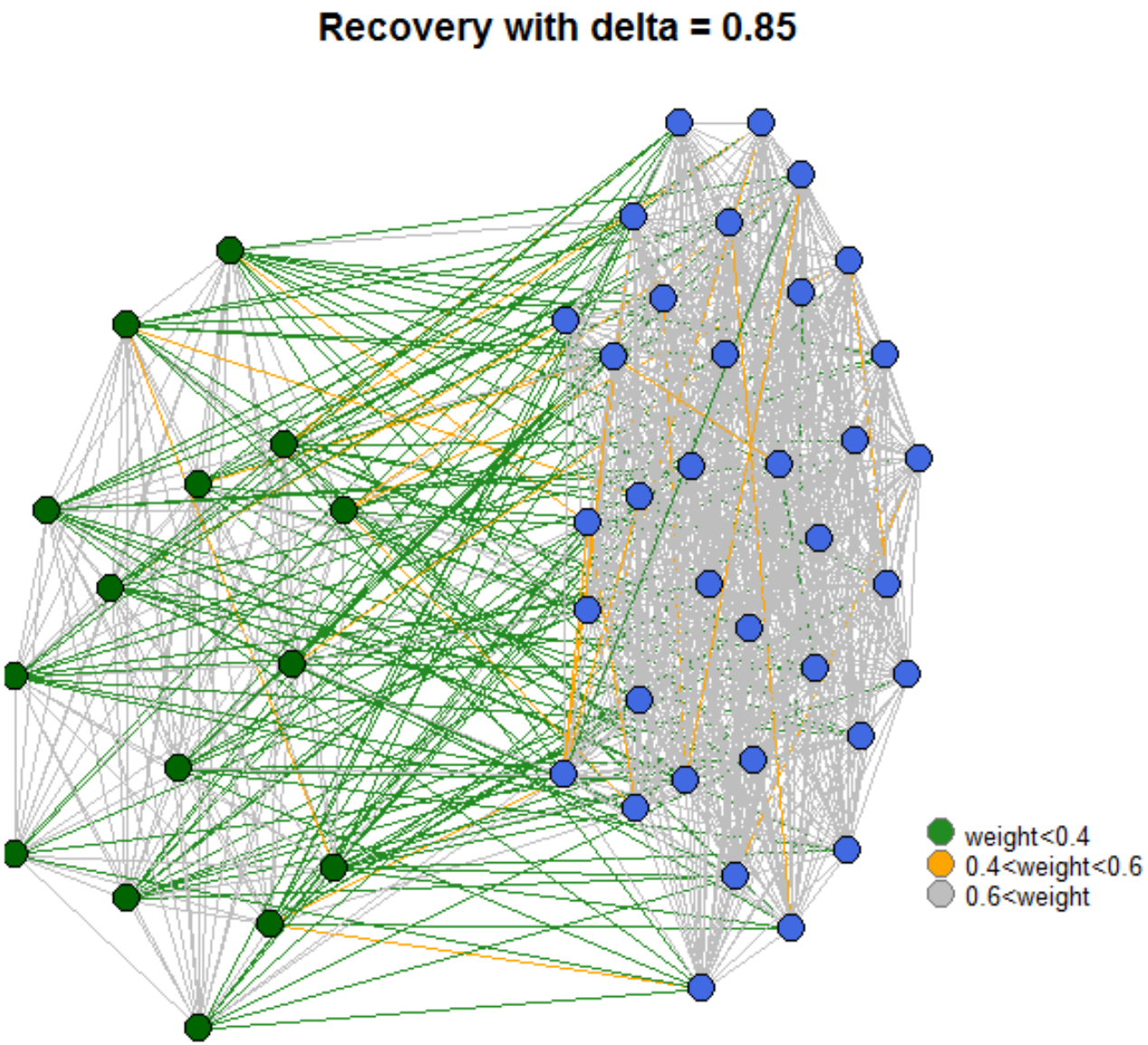
Grafos recuperados para diferentes valores de δ



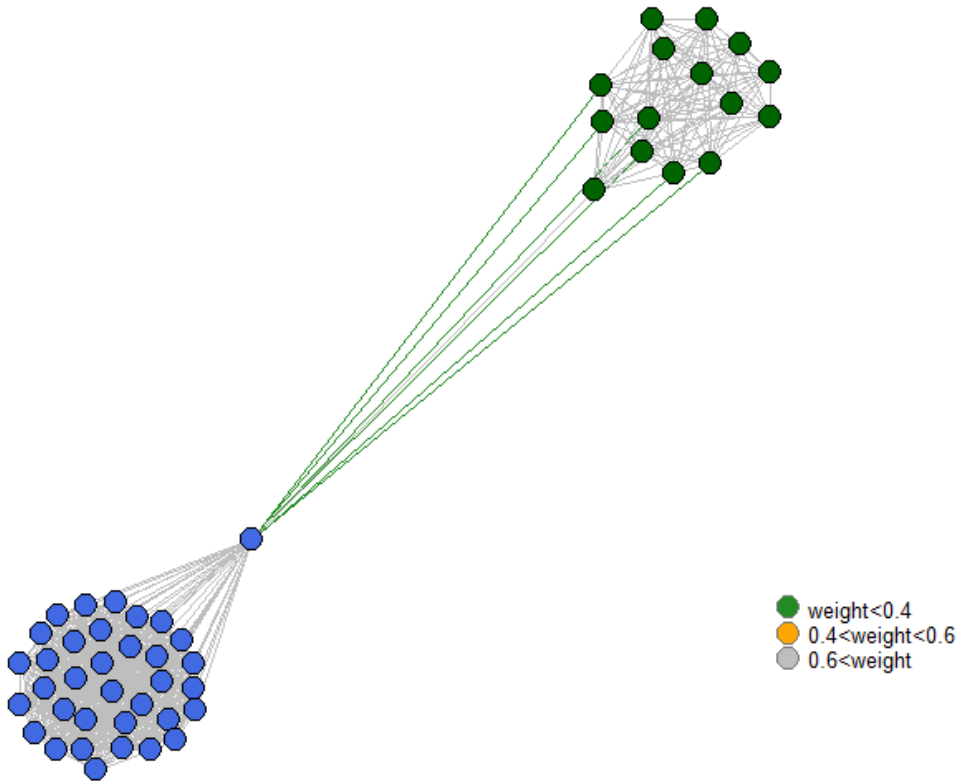
Grafos recuperados para diferentes valores de δ



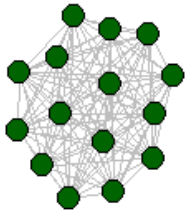
Grafos recuperados para diferentes valores de δ

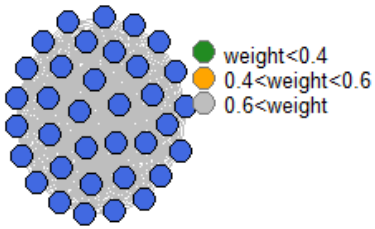


Grafos recuperados para diferentes valores de δ

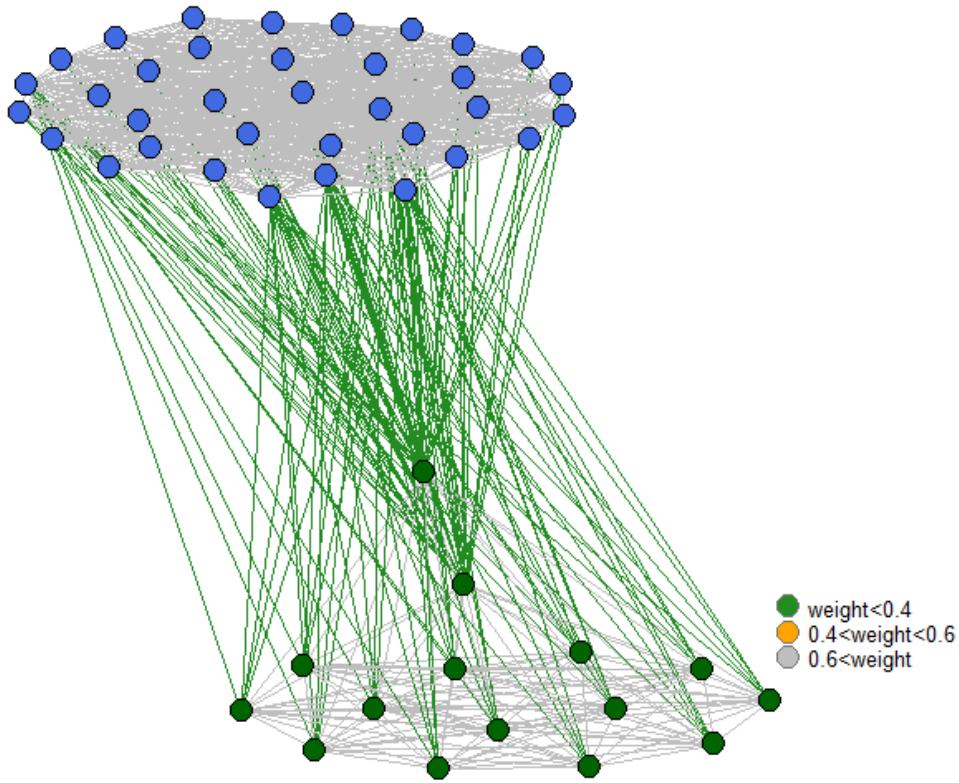


Grafos recuperados para diferentes valores de δ

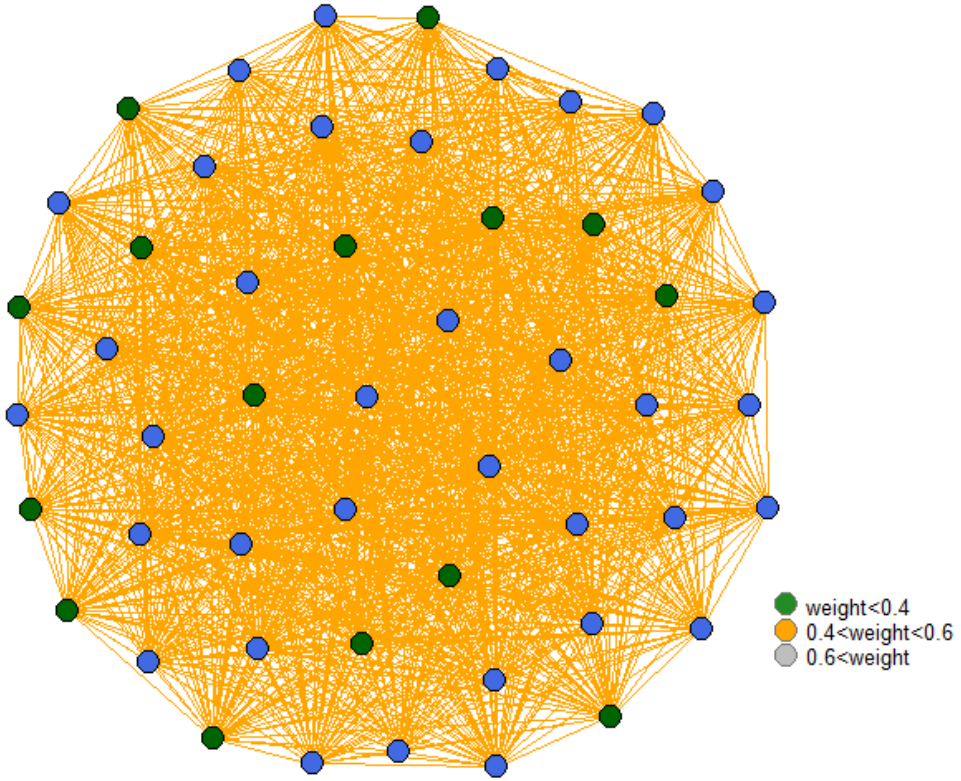




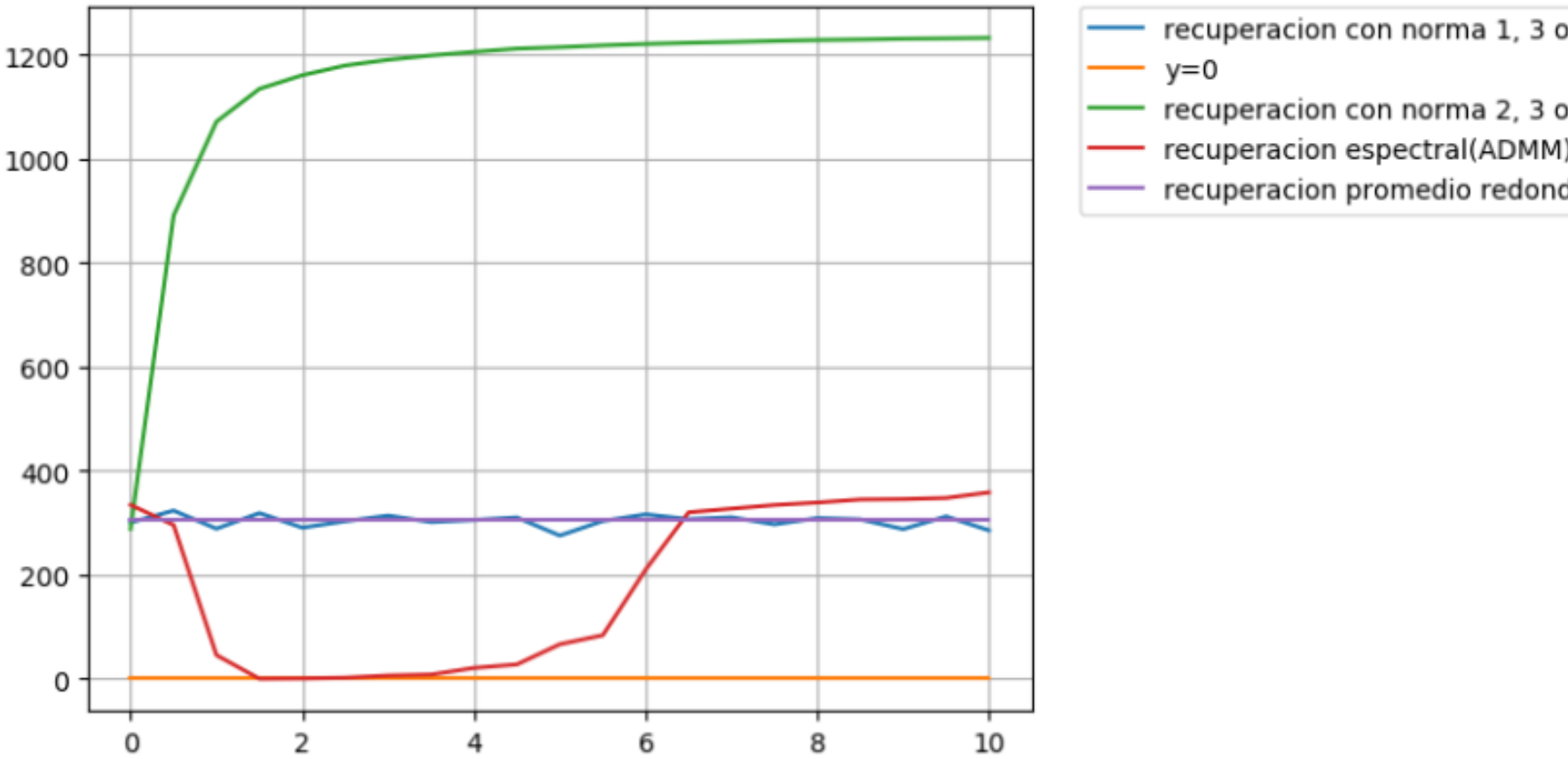
Grafos recuperados para diferentes valores de δ



Grafos recuperados para diferentes valores de δ



Error de recuperación para distintas métricas de Wasserstein con $N = 3$.



Qué propiedades tiene $\overline{A}$?

Suponga que $\overline{A}$ es un óptimo del problema de optimización producido por nuestro algoritmo.

Teorema. (De Roux, -)

Si $\mathcal{G}$ tiene distribución dada por el stochastic block model entonces existe un radio $\delta(N)$ para el cual

$$\mathbb{P}\{\overline{A} \neq A^*\} \leq c \exp(-Nd)$$

En palabras, el algoritmo aprende *de manera exacta* la estructura de clusters en pocas muestras (usa como información adicional que tal estructura existe y eso le permite encontrarla con pocos samples).

Algoritmos

Un solver industrial (MOSEK) no puede resolver nuestro problema de optimización semidefinida para grafos con $n = 40$, $N = 4$.

Teorema. (De Roux, -)

Propusimos e implementamos una versión de un algoritmo ADMM (Alternating Directions Method of Multipliers) que hace que nuestro método pueda aplicarse a grafos de hasta 10000 vértices.