FACULTAD DE INGENIERÍA

UNIVERSIDAD DE LA REPÚBLICA URUGUAY

# Aprendizaje Automático para Datos en Grafos

## Aprendiendo Grafos desde Datos

Marcelo Fiori
Muy basado en transparencias de **Gonzalo Mateos**

mfiori@fing.edu.uy
http://www.fing.edu.uy/~mfiori/

20 de octubre de 2022

# Aprendiendo grafos a partir de señales suaves

1. Métodos estadísticos para inferencia de topología del grafo

2. Aprendiendo grafos a partir de observaciones de señales suaves

# Formulación del problema

## Motivación

- Buscamos grafos sobre el cual las señales admitan ciertas regularidades
  - Predicción por vecino más cercano (a.k.a. graph smoothing)
  - Aprendizaje semi-supervisado
- Muchos datos de redes del mundo real son suaves
  - Grafos basados en similitudes entre atributos de vértices
  - A menudo, formación de redes se basa en homofilia, proximidad en espacio latente

## Planteo de problema

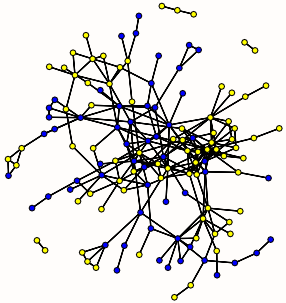
Dadas observaciones $\mathcal{X} := \{\mathbf{x}_p\}_{p=1}^{P}$, identificar un grafo $G$ tal que las señales $\mathcal{X}$ sean suaves en $G$.

- Criterio: Energía de Dirichlet en el grafo $G$ con Laplaciana $\mathbf{L}$

$$\mathrm{TV}(\mathbf{x}) = \mathbf{x}^\top \mathbf{L} \mathbf{x}$$

# Ejemplo: Predecir funciones de proteínas

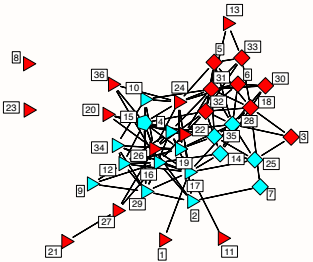
- Datos de levadura, formalmente *Saccharomyces cerevisiae*
  - Grafo: 134 vértices (proteínas) y 241 aristas (interacciones entre proteínas)



- Señal: anotación funcional cascadas de señalización intracelular(ICSC)
  - $x_i = 1$ si proteína $i$ anotada ICSC (amarillo), $x_i = 0$ si no(azul)

# Ejemplo: Prácticas de abogados

- Relaciones laborales entre abotados [Lazega'01]
  - Grafo: 36 abogados, aristas indican colaboraciones




- Señal: varios atributos a nivel de nodo $\mathbf{x} = \{x_i\}_{i \in \mathcal{V}}$ incluyendo

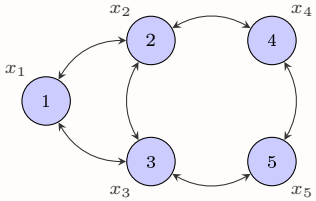
  $\Rightarrow$ Tipo de práctica, i.e., litigante (red) y corporativo (cyan)

- Intuimos que abogados colaboran más con pares de su misma clase de práctica legal

  $\Rightarrow$ El conocimiento de colaboraciones es útil para predecir el tipo de práctica

# Digresión II: Graph signal processing (GSP)

- Grafo $G$ con matriz de adyacencia $\mathbf{A} \in \mathbb{R}^{N \times N}$
  $\Rightarrow A_{ij} =$ proximidad entre $i$ y $j$
- Tenemos una señal $\mathbf{x} \in \mathbb{R}^N$ en el grafo
  $\Rightarrow x_i =$ valor de la señal en el nodo $i$

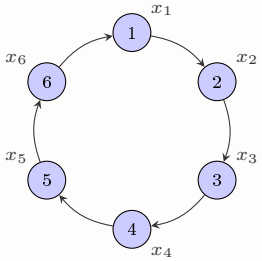



- Graph Signal Processing $\rightarrow$ Explotar la estructura codificada en $\mathbf{A}$ para procesar $\mathbf{x}$

# Digresión II: Importancia de la estructura temporal de una señal

- El procesamento de señales se basa en explotar la estructura de la señal

- Tiempo discreto descripto a través de un grafo cíclico
  - $\Rightarrow$ El instante de tiempo $n$ le sigue al $n-1$
  - $\Rightarrow$ El valor $x_n$ está relacionado a $x_{n-1}$
- Formalizado bajo la noción de frecuencia




- Estructura cíclica $\Rightarrow$ Transformada de Fourier $\Rightarrow \tilde{\mathbf{x}} = \mathbf{F}^H \mathbf{x}$ $\left( F_{kn} = \frac{e^{j2\pi kn/N}}{\sqrt{N}} \right)$
- **Transformada de Fourier** $\Rightarrow$ Proyección en el *eigenvector space* del ciclo

# Digresión II: Graph Fourier Transform

- Adyacencia $\mathbf{A}$, Laplaciana $\mathbf{L}$, o genéricamente graph shift $\mathbf{S} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1}$
  - $\Rightarrow S_{ij} = 0$ para $i \neq j$ y $(i,j) \notin \mathcal{E}$ (captura la estructura local en $G$)
- La Graph Fourier Transform (GFT) de $\mathbf{x}$ se define como

$$\tilde{\mathbf{x}} = \mathbf{V}^{-1}\mathbf{x}$$

- Mientras que la GFT inversa (iGFT) de $\tilde{\mathbf{x}}$ se define como

$$\mathbf{x} = \mathbf{V}\tilde{\mathbf{x}}$$

  - $\Rightarrow$ Los vectores propios $\mathbf{V} = [\mathbf{v}_1, ..., \mathbf{v}_N]$ son la base frecuencial (átomos)
- Estructura adicional
  - $\Rightarrow$ Si $\mathbf{S}$ es normal, entonces $\mathbf{V}^{-1} = \mathbf{V}^H$ y $\tilde{x}_k = \mathbf{v}_k^H \mathbf{x} = \langle \mathbf{v}_k, \mathbf{x} \rangle$
  - $\Rightarrow$ Vale Parseval: $\|\mathbf{x}\|^2 = \|\tilde{\mathbf{x}}\|^2$
- **GFT** $\Rightarrow$ Proyección sobre el espacio de vectores propios de la matriz de shift $\mathbf{S}$

# Digresión II: Modos frecuenciales de la Laplaciana

- Total variation de la señal $\mathbf{x}$ con respecto a $\mathbf{L}$

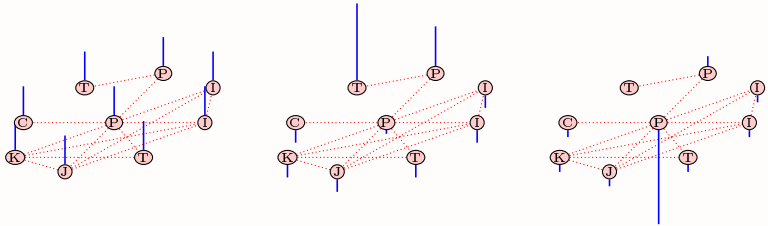
$$\mathrm{TV}(\mathbf{x}) = \mathbf{x}^T \mathbf{L} \mathbf{x} = \sum_{i,j=1, j>i}^{N} A_{ij}(x_i - x_j)^2$$

  $\Rightarrow$ Medida de suavidad de la señal sobre el grafo $G$ (energía de Dirichlet)

- Para los valores propios de la Laplaciana $\mathbf{V} = [\mathbf{v}_1, \cdots, \mathbf{v}_N]$ $\Rightarrow$ $\mathrm{TV}(\mathbf{v}_k) = \lambda_k$

  $\Rightarrow$ Podemos ver a $0 = \lambda_1 < \cdots \leq \lambda_N$ como frecuencias

- Ejemplo: grafo con $N = 10$, $k = 1$, $k = 2$, $k = 9$



# Modelo de análisis factorial basado en la Laplaciana

- Consideremos un grafo desconocido $G$ con Laplaciana $\mathbf{L} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top$
  - $\Rightarrow$ Adoptamos la base $\mathbf{V}$ GFT como as signal representation matrix
- Factor-analysis model para la señal observada en el grafo [Dong et al'16]

$$\mathbf{x} = \mathbf{V}\boldsymbol{\chi} + \boldsymbol{\epsilon}$$

  - $\Rightarrow$ Variables latentes $\boldsymbol{\chi} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Lambda}^\dagger)$ ($\approx$ coeficientes GFT)
  - $\Rightarrow$ Término de error isotrópico $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$
  - $\Rightarrow$ ¿Cuáles serán los coeficientes más grandes en $\boldsymbol{\chi}$?
  - $\Rightarrow$ ¿asociados a qué valores propios? ¿cómo repercute eso en $\mathbf{x}$?
- Suavidad: esto favorece señales "pasa bajo" $\mathbf{x}$
  - $\Rightarrow$ Valores propios chicos de $\mathbf{L}$ (baja frecuencia) $\rightarrow$ Peso alto en los factores

# Inferencia como denoising via graph kernel regression

- Estimador *Maximum a posteriori* (MAP) de las variables $\boldsymbol{\chi}$

$$\hat{\boldsymbol{\chi}}_{\text{MAP}} = \arg\min_{\boldsymbol{\chi}} \left\{ \|\mathbf{x} - \mathbf{V}\boldsymbol{\chi}\|^2 + \alpha \boldsymbol{\chi}^\top \boldsymbol{\Lambda} \boldsymbol{\chi} \right\}$$

  $\Rightarrow$ Parameterizado por $\mathbf{V}$ y $\boldsymbol{\Lambda}$, que no conocemos

- Definamos el predictor $\mathbf{y} := \mathbf{V}\boldsymbol{\chi}$, entonces el regularizador queda:

$$\boldsymbol{\chi}^\top \boldsymbol{\Lambda} \boldsymbol{\chi} = \mathbf{y}^\top \mathbf{V} \boldsymbol{\Lambda} \mathbf{V}^\top \mathbf{y} = \mathbf{y}^\top \mathbf{L} \mathbf{y} = \text{TV}(\mathbf{y})$$

  $\Rightarrow$ Denoiser (basado en TV Laplaciano) de $\mathbf{x}$, prior de suavidad en $\mathbf{y}$

- **Idea:** buscar simultáneamente $\mathbf{L}$ y la representación sin ruido (denoised) $\mathbf{y} = \mathbf{V}\boldsymbol{\chi}$

$$\min_{\mathbf{L}, \mathbf{y}} \left\{ \|\mathbf{x} - \mathbf{y}\|^2 + \alpha \mathbf{y}^\top \mathbf{L} \mathbf{y} \right\}$$

# Formulación y algoritmo

- Dadas señales $\mathcal{X} := \{\mathbf{x}_p\}_{p=1}^P$ en $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_P] \in \mathbb{R}^{N \times P}$, resolver

$$\min_{\mathbf{L},\mathbf{Y}} \left\{ \|\mathbf{X} - \mathbf{Y}\|_F^2 + \alpha \text{trace}\left(\mathbf{Y}^\top \mathbf{L} \mathbf{Y}\right) + \frac{\beta}{2} \|\mathbf{L}\|_F^2 \right\}$$
$$\text{s. to} \quad \text{trace}(\mathbf{L}) = N,\ \mathbf{L}\mathbf{1} = \mathbf{0},\ L_{ij} = L_{ji} \leq 0,\ i \neq j$$

  - $\Rightarrow$ Hay que fijar "la escala" ($trace(\mathbf{L}) = N$). Si no $\mathbf{L} = 0$ y $\mathbf{X} = \mathbf{Y}$ es solución
  - $\Rightarrow$ Es fácil ver que $trace(L) = N$ fija la norma $\ell_1$ de $\mathbf{L}$
  - $\Rightarrow$ Función objetivo: Ajuste a datos + suavidad + sparsity en aristas
  - $\Rightarrow$ No conjuntamente convexa en $\mathbf{L}$ y $\mathbf{Y}$, pero bi-convexa

- Enfoque algorítmico: minimización alternada (AM)

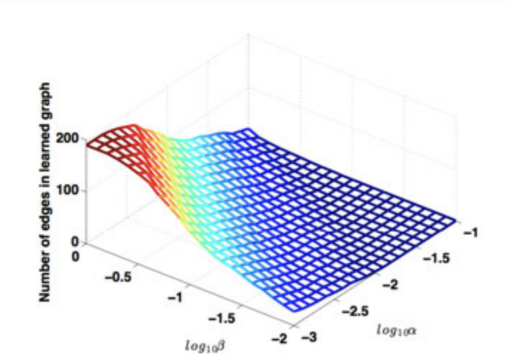
  (S1) Fijado $\mathbf{Y}$: resolver para $\mathbf{L}$ via interior-point methods, ADMM

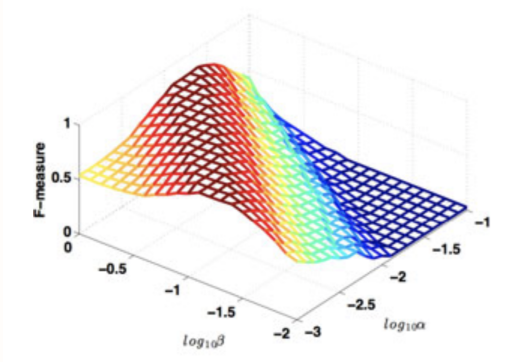
  (S2) Fijado $\mathbf{L}$, solución cerrada: filtro pasa-bajos, versión suavizada de $\mathbf{X}$

$$\mathbf{Y} = (\mathbf{I} + \alpha \mathbf{L})^{-1} \mathbf{X}$$

# Impacto de los parámetros en esparsidad y accuracy

- Se generan muchas señales en un grafo sintético Erdős-Rényi
  $\Rightarrow$ Se infiere el grafo para distintos valores de $\alpha$ and $\beta$




- Hay más aristas si crece $\beta$ y si decrece $\alpha$
- Si hay poco ruido, el cociente $\beta/\alpha$ determina la performance