FACULTAD DE INGENIERÍA

UNIVERSIDAD DE LA REPÚBLICA URUGUAY

Aprendizaje Automático para Datos en Grafos

Aprendiendo Grafos desde Datos

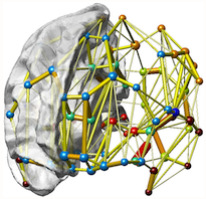
Marcelo Fiori
Muy basado en transparencias de **Gonzalo Mateos**

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

Octubre 2022

¿De qué vamos a hablar?

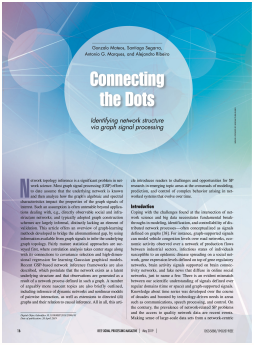
- Aprender grafos desde observaciones en nodos
- Ex: Central en *network neuroscience*
  - $\Rightarrow$ Red funcional a partir de señales de fMRI

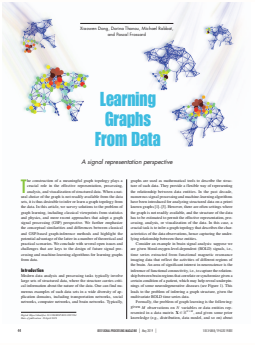


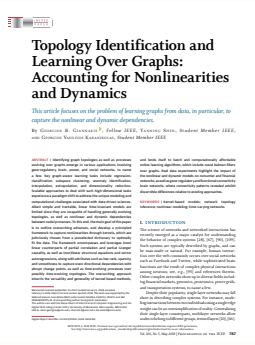
- Mayoría de trabajos GSP: cómo un grafo conocido G afecta señales y filtros
  - Posible para e.g., redes físicas
  - Links son tangibles y directamente observables
- Igual, obtener y actualizar la información de topología es desafiante
  - $\Rightarrow$ por tamaño, reconfiguración, privacidad, seguridad
- Aquí, camino inverso: ¿cómo usar GSP para inferir la topología del grafo?
- **Objetivo:** recuperar una red latente, o, una representación en grafos de datos

Connecting the dots

- Algunos tutoriales recientes en aprendizaje de grafos a partir de datos
  - IEEE Signal Processing Magazine y Proceedings of the IEEE







- IEEE Trans. on Signal and Information Processing over Networks
  - Special issue on Network Topology Inference (Jan. 2020)

Inferencia de topología del grafo

1. Métodos estadísticos para inferencia de topología del grafo
2. Aprendiendo grafos a partir de observaciones de señales suaves
3. Identificando la estructura de procesos de difusión en redes

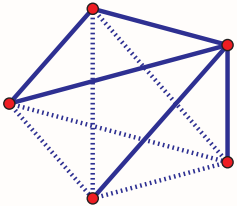
Problemas de inferencia de topología del grafo

- **Q:** Si G (o una porción de él) no es observada, ¿podemos inferirlo a partir de datos?
- Formular como inferencia estadística, i.e. dados
  - Medidas de señales x_i en todos/algunos vértices $i \in \mathcal{V}$
  - Indicadores A_{ij} de estado de aristas para algunos pares de vértices $\{i, j\} \in \mathcal{V}_{obs}^{(2)}$
  - Una colección $\mathcal{G}$ de grafos candidatos G

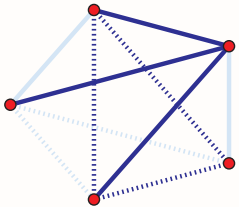
  Objetivo: inferir la topología del grafo $G(\mathcal{V}, \mathcal{E})$
- Poder aprovechar conceptos estadísticos existentes y herramientas

  $\Rightarrow$ Estudiar identificabilidad, consistencia, robustez, complejidad
- Tres problemas canónicos de inferencia de topología del grafo [Kolaczyk'09]
  - (i) Predicción de enlaces
  - (ii) Association network inference $\leftarrow$ Énfasis de estas clases
  - (iii) Tomographic network topology inference

Predicción de enlaces



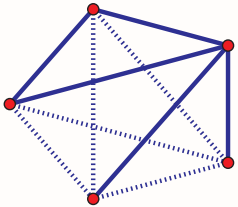
Original graph



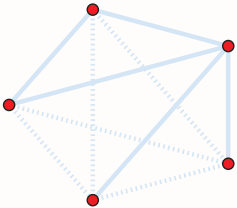
Link prediction

- Supongamos que observamos la señal en el grafo $\mathbf{x} = [x_1, \ldots, x_N]^\top$; y
- Estado de aristas observado solo en algun subconjunto de pares $\mathcal{V}^{(2)}_{obs} \subset \mathcal{V}^{(2)}$
- **Objetivo:** predecir estado de aristas para el resto de los pares, i.e., $\mathcal{V}^{(2)}_{miss} = \mathcal{V}^{(2)} \setminus \mathcal{V}^{(2)}_{obs}$

Association network inference



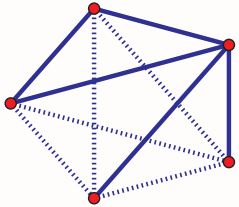
Original graph



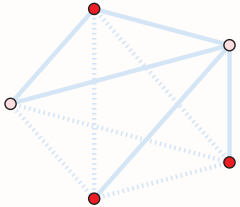
Association network inference

- Supongamos que solo observamos la señal en el grafo $\mathbf{x} = [x_1, \ldots, x_N]^\top$; y
- Asumimos (i, j) definido por un 'nivel de asociación' no trivial entre x_i, x_j
- **Objetivo:** predecir estado de aristas para todos los pares de vértices $\mathcal{V}^{(2)}$

Tomographic network topology inference



Original graph

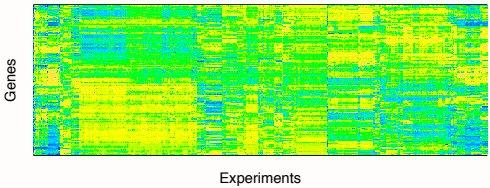


Tomographic inference

- Supongamos que solo observamos x_i para algunos vértices $i \subset \mathcal{V}$ en el 'perímetro' de G
- **Objetivo:** predecir estado de vértices y aristas en el 'interior' de G

Association networks

- **Def**: en association networks los vértices están unidos por aristas si hay un nivel suficiente de 'asociación' entre atributos de los pares de vértices



Ejemplo

- Gene-regulatory networks
- Neuro-functional connectivity networks

Association network inference

- Dada una colección de N elementos representados como vértices $v \in \mathcal{V}$
  - Señal en el grafo $\mathbf{x} = [x_1, \ldots, x_N]^\top \in \mathbb{R}^N$ de atributos observados en los vértices
- Similaridad definida por usuario $\texttt{sim}(i,j) = f(x_i, x_j)$ especifica aristas $(i,j) \in \mathcal{E}$
  - **Q:** ¿Si los valores mismos de `sim` (i.e., estado de aristas) no son observables?

Association network inference

Inferir valores no triviales de `sim` a partir de observaciones i.i.d. $\mathcal{X} := \{\mathbf{x}_p\}_{p=1}^P$

- Hay muchas elecciones a tomar, y entonces muchos acercamientos posibles
  - Elección de `sim`: correlación, correlación parcial, información mutua
  - Elección técnica de inferencia: test de hipótesis, regresión, ad hoc
  - Elección de parámetros: umbrales de test, nivel de significancia, regularización

Redes de correlación

- Coeficiente de correlación de Pearson como `sim` entre pares de vértices

$$\texttt{sim}(i,j) := \rho_{ij} = \frac{\operatorname{cov}[x_i, x_j]}{\sqrt{\operatorname{var}[x_i]\operatorname{var}[x_j]}}, \quad i,j \in \mathcal{V}$$

- **Def:** el grafo de correlación $G(\mathcal{V}, \mathcal{E})$ tiene como aristas

$$\mathcal{E} = \left\{(i,j) \in \mathcal{V}^{(2)} : \rho_{ij} \neq 0\right\}$$

  - Association network inference $\Leftrightarrow$ Inferencia de correlaciones no nulas

- Inferencia de $\mathcal{E}$ típicamente atacado como problema de test de hipótesis

$$H_0 : \rho_{ij} = 0 \quad \text{versus} \quad H_1 : \rho_{ij} \neq 0$$

Estadísticos para el test

- Una elección usual de estadístico son las correlaciones empíricas

$$\hat{\rho}_{ij} = \frac{\hat{\sigma}_{ij}}{\sqrt{\hat{\sigma}_{ii}\hat{\sigma}_{jj}}}, \quad \text{where } \hat{\mathbf{\Sigma}} = [\hat{\sigma}_{ij}] = \frac{1}{P-1}\sum_{p=1}^{P} \mathbf{x}_p \mathbf{x}_p^\top$$

- Un estadístico alternativo conveniente es el transformado de Fisher

$$\hat{z}_{ij} = \frac{1}{2}\log\left(\frac{1+\hat{\rho}_{ij}}{1-\hat{\rho}_{ij}}\right), \quad i,j \in \mathcal{V}$$

$\Rightarrow$ Bajo H_0, $\hat{z}_{ij} \sim \mathcal{N}(0, \frac{1}{P-3}) \Rightarrow$ Simple de controlar significancia

- Rechazamos H_0 bajo nivel de significancia α, i.e., asignamos arista (i,j) si $|\hat{z}_{ij}| > \frac{z_{\alpha/2}}{\sqrt{P-3}}$

Tasa de control de error: $\mathrm{P}_{H_0}(\text{falsa arista}) = \mathrm{P}_{H_0}\left(|\hat{z}_{ij}| > \frac{z_{\alpha/2}}{\sqrt{P-3}}\right) = \alpha$

Grafos y testeo múltiple

- Surgen desafíos interesantes con grafos de gran escala
  - $\Rightarrow$ Supongamos que testeamos los $\binom{N}{2}$ pares de vértices, cada uno a nivel α
- Incluso si el grafo verdadero G es el grafo vacío, i.e., $\mathcal{E} = \emptyset$
  - $\Rightarrow$ Esperamos encontrar $\binom{N}{2}\alpha$ aristas espúreas solo por azar!
  - $\Rightarrow$ Para un grafo grande, este número puede ser considerable
- Ex: Si G tiene $N = 100$ nodos y testeamos aristas individualmente a nivel $\alpha = 0{,}05$
  - $\Rightarrow$ El número esperado de aristas espúreas es $4950 \times 0{,}05 \approx 250$
- En estadística, este dilema es conocido como el problema de testeo múltiple

Corrección para testeo múltiple

- **Idea:** Controlar los errores a nivel de la colección de test, no de forma individual
- False discovery rate (FDR) controla, i.e., para un nivel dado γ aseguramos

$$\text{FDR} = \mathbb{E}\left[\frac{R_{false}}{R} \,\middle|\, R > 0\right] \text{P}\left(R > 0\right) \leq \gamma$$

  - R es el número total de aristas detectadas; y
  - R_{false} es el número de falsas aristas detectadas

- Método para controlar FDR a nivel γ [Benjamini-Hochberg'94]

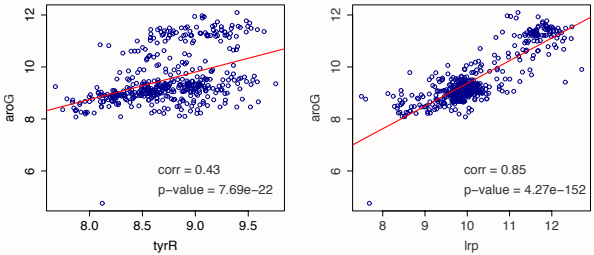
  Paso 1: Ordenar p-valores para los $\bar{N} := \binom{N}{2}$ tests, obtenemos $p_{(1)} \leq \ldots \leq p_{(\bar{N})}$

  Paso 2: Rechazar H_0, i.e., declarar todas las aristas para las cuales

$$p_{(k)} \leq \left(\frac{k}{\bar{N}}\right)\gamma$$

Ejemplo: correlaciones en nivel de expresión de genes

- Datos de microarray para bacteria Escherichia coli (E. coli)
  - Dos TFs $tyrR$ y lrp, potencial target $aroG$ sobre $n = 445$ experimentos
  - Ground truth: $aroG$ es regulado por $tyrR$ pero no por lrp



- Fisher scores: $z_{tyrR}^{aroG} = 0{,}4599$ y $z_{lrp}^{aroG} = 1{,}2562$. Ambos p-valores son chicos
- En base a correlaciones, $aroG$ está fuertemente asociado con ambos TFs $tyrR$ y lrp

Correlaciones parciales

- Hay que usar la correlación con cuidado: 'correlación no implica causalidad'
  - Nodos $i, j \in \mathcal{V}$ pueden tener alto ρ_{ij} porque se influencian entre sí
- Pero ρ_{ij} podría ser alto si ambos i, j son influenciados por un tercer nodo $k \in \mathcal{V}$

  $\Rightarrow$ Redes de correlación pueden declarar aristas debidas a factores de confusión
- Las correlaciones parciales capturan mejor la influencia directa entre nodos
  - Para $i, j \in \mathcal{V}$ consideremos los nodos latentes $S_m = \{k_1, \ldots, k_m\} \subset \mathcal{V} \setminus \{i, j\}$
- Correlación parcial entre x_i y x_j, ajustada por (o condicionada a) $\mathbf{x}_{S_m} = [x_{k_1}, \ldots, x_{k_m}]^\top$ es

$$\rho_{ij|S_m} = \frac{\operatorname{cov}[x_i, x_j \,|\, \mathbf{x}_{S_m}]}{\sqrt{\operatorname{var}\left[x_i \,|\, \mathbf{x}_{S_m}\right] \operatorname{var}\left[x_j \,|\, \mathbf{x}_{S_m}\right]}}, \quad i, j \in \mathcal{V}$$

- **Q:** ¿Cómo calcular estas correlaciones parciales?

Cálculo de correlaciones parciales

- Dados $\mathbf{x}_{S_m} = [x_{k_1}, \ldots, x_{k_m}]^\top$, la correlación parcial entre x_i y x_j es

$$\rho_{ij|S_m} = \frac{\operatorname{cov}[x_i, x_j \,|\, \mathbf{x}_{S_m}]}{\sqrt{\operatorname{var}\left[x_i \,|\, \mathbf{x}_{S_m}\right] \operatorname{var}\left[x_j \,|\, \mathbf{x}_{S_m}\right]}} = \frac{\sigma_{ij|S_m}}{\sqrt{\sigma_{ii|S_m}\sigma_{jj|S_m}}}$$

- Aquí $\sigma_{ii|S_m}, \sigma_{jj|S_m}$ y $\sigma_{ij|S_m}$ son los elemntos en la diagonal y fuera de la diagonal de

$$\boldsymbol{\Sigma}_{11|2} := \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} \in \mathbb{R}^{2\times 2}$$

- Las matrices $\boldsymbol{\Sigma}_{11}$, $\boldsymbol{\Sigma}_{22}$ y $\boldsymbol{\Sigma}_{21} = \boldsymbol{\Sigma}_{12}^\top$ son los bloques de la matriz de covarianza:

$$\operatorname{cov}\begin{bmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \end{bmatrix} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}, \quad \text{donde } \mathbf{w}_1 := [x_i, x_j]^\top \text{ y } \mathbf{w}_2 := \mathbf{x}_{S_m}$$

Redes de correlaciones parciales

- Hay varias formas de usar correlaciones parciales para definir aristas en G

  Ex: x_i, x_j correlacionados sin importar sobre qué m vértices condicionamos

$$\mathcal{E} = \left\{ (i,j) \in \mathcal{V}^{(2)} : \rho_{ij|S_m} \neq 0, \text{ para todo } S_m \in \mathcal{V}^{(m)}_{\setminus\{i,j\}} \right\}$$

- Inferencia de potencial arista (i,j) como problema de test de hipótesis

$$H_0 : \rho_{ij|S_m} = 0 \text{ para algún } S_m \in \mathcal{V}^{(m)}_{\setminus\{i,j\}}$$
$$H_1 : \rho_{ij|S_m} \neq 0 \text{ para todo } S_m \in \mathcal{V}^{(m)}_{\setminus\{i,j\}}$$

- De nuevo, dadas medidas $\mathcal{X} := \{\mathbf{x}_p\}_{p=1}^{P}$ necesitamos:
  - Seleccionar un estadístico para el test
  - Construir una distribución de referencia
  - Ajustes de test múltiple

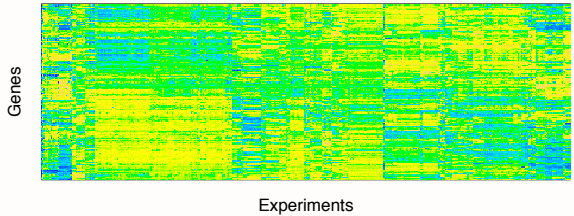
Case study: Inferencia de interacciones de regulación génica

- Genes son segmentos del ADN que codifican información sobre funcionamiento celular
- Esta información se usa en el proceso de expresión de genes
  - ⇒ Creación de productos bioquímicos, i.e., ARN o proteínas
- Regulación de un gen refiere al control de esta expresión

  Ex: regulación durante la transcripción, copia del ADN a ARN
  - ⇒ Los genes que controlan son transcription factors (TFs)
  - ⇒ Los genes controlados se denominan targets
  - ⇒ Tipo de regulación: activación o represión
- Ineracciones de regulación entre genes es fundamental para entender el funcionamiento de oganismos
  - ⇒ Inferencia de interacciones → Encontrar pares de genes TF/target
- Esta información relacional se resume en una red de regulación génica

Interacciones de regulación entre genes en E. coli

- Uso de datos de microarray y métodos de correlación para inferir pares TF/target



- Dataset: nivel de expresión relativa logartítmica de expresión ARN, para genes en E. coli
  - 4,345 genes medidos bajo 445 condiciones experimentales diferentes
- Ground truth: 153 TFs, y pares de TF/target de la base de datos RegulonDB

Métodos para inferir pares de genes TF/target

- Tres métodos basados en correlación para inferir pares de genes TF/target

  $\Rightarrow$ Declaramos interacciones si los p-valores caen debajo de cierto umbral

  Método 1: Correlación de Pearson entre TF y potencial gen destino (target)

  Método 2: Correlaciones parciales, condicionado individualmente a un ($m = 1$) TF, sobre todos los 152 TFs

  Método 3: Correlación parcial completa, condicionando simultáneamente a todos los otros TFs ($m = 152$)

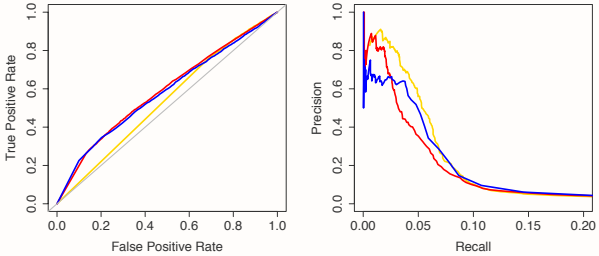
- En todos los casos se aplica la transformación de Fisher para obtener z-scores

  $\Rightarrow$ Distribuciones asintóticas gaussianas para p-values, con $P = 445$

- Comparamos los grafos inferidos con el ground-truth network de RegulonDB

Comparación de performance

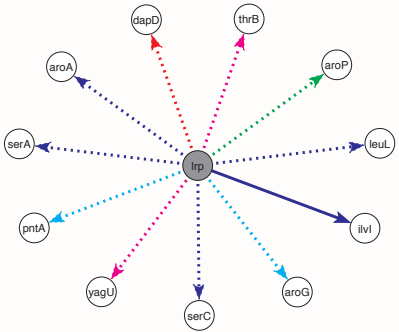
- Curvas ROC y Precision/Recall para los métodos 1, 2, y 3
  - ⇒ Precision: fracción de aristas que se predicen que son efectivamente ciertas
  - ⇒ Recall: fracción de aristas verdaderas que se predicen correctamente



- Método 1 es el peor, pero ninguno es la gran cosa
  - ⇒ Correlación no es un indicador fuerte de regulación en estos datos
- Todos los métodos comparten una región de alta precisión, pero con muy bjo recall
  - ⇒ Limitantes en número y diversidad de perfiles [Faith et al'07]

Predecir nuevos pares de genes TF/target

- En biología, suele haber interés en predecir nuevas interacciones



- 11 interacciones encontradas para TF *lrp*, 10 confirmadas experimentalmente (punteado)
  - $\Rightarrow$ 5 interacciones con genes target son nuevas (magenta, red, cyan)
  - $\Rightarrow$ 4 presentes en RegulonDB (magenta, cyan), pero no como *lrp* targets