FACULTAD DE INGENIERÍA

UNIVERSIDAD DE LA REPÚBLICA URUGUAY

Aprendizaje Automático para Datos en Grafos

Introducción y Motivación

Paola Bermolen

paola@fing.edu.uy

29 de agosto de 2022

Presentación

- Equipo docente: Paola Bermolen (IMERL), Marcelo Fiori (IMERL), Federico La Rocca (IIE), Bernardo Marenco (IMERL), Gonzalo Mateos (University of Rochester)

- Antecedentes:
  - curso corto dictado por Gonzalo Mateos en 2021 (curso semestral en UR)
  - seminario Redes Complejas organizado por CICADA (M.Arim y P.Bermolen).

Detalles administrativos

- Horarios de clase:
  - Teórico: Martes y Jueves de 9:30 a 11:00 [11:30...]
  - Práctico: horarios a convenir con Bernardo
- Curso de posgrado Facultad de Ingeniería. Reglamento: ver página de Fing –> Enseñanza–> Posgrado y Educación Permanente.
- Por inscripciones extracurriculares o exoneración de arancel completar encuesta en EVA indicando nombre completo y cédula de identidad en el correr de la semana.
- Matricularse al curso en EVA (AAgrafos)

Aprobación del curso

- Entrega de 5 laboratorios en phyton: 40 % (aproximadamente cada 2 semanas)
- Proyecto Final: se entrega propuesta 10 % y se defiende oralmente 50 %

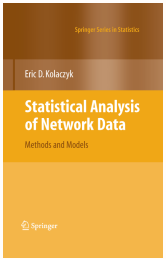
  En el proyecto final, los estudiantes podrán investigar y aplicar algoritmos de aprendizaje automático en grafos del estado del arte, a una aplicación de su interés.
- Las fechas de entregas aproximadas serán comunicadas en el EVA
- Ambas actividades en grupos de aexactamente 2 integrantes

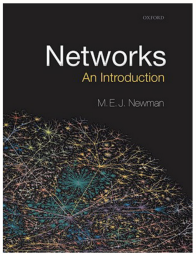
Principales referencias

- Vamos a usar tanto *slides* como pizarrón

  $\Rightarrow$ Otros materiales extras en la página del curso

- Dos libros de cabecera, en especial para la primera parte:
  - M. E. J. Newman *"Networks: An Introduction"* Oxford U. Press
  - Eric D. Kolaczyk *"Statistical Analysis of Network Data: Methods and Models"* Springer





Objetivo del curso

Objetivo

El objetivo general del curso es que los estudiantes puedan afrontar un problema de aprendizaje automático donde los datos se encuentran en forma de grafos.
Se brindarán los conceptos teóricos fundamentales y las herramientas prácticas necesarias para ello.
Al finalizar el curso los estudiantes serán capaces de implementar y entender distintas técnicas del estado del arte en inferencia y predicción en grafos.

→ Foco en: pensar, leer, preguntar, implementar...

Prerequisitos

(I) Teoría de grafos e inferencia estadística

- Los grafos son abstracciones matemáticas de las redes
- La inferencia estadística es útil para “aprender” de los datos de las redes
- Algunos conocimientos básicos esperados. Repaso previstos

(II) Teoría de la probabilidad y álgebra lineal

- Variables aleatorias, distribuciones, valor esperado, varianza
- Notación vectorial/matricial, sistemas de ecuaciones lineales, valores propios

(III) Programación

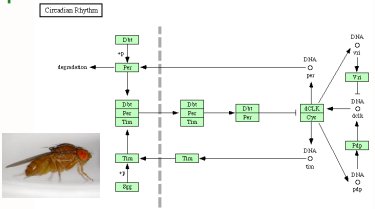
- Phyton! muchas referencias en EVA y Bernardo para ayudar

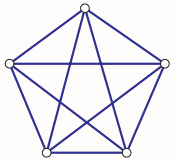
Be nice

- ➤ **We** work hard for this course, expect you to do the same
- ✓ Come to class, be on time, pay attention, ask
- ✓ Check out the additional suggested readings
- ✓ Play with network analysis software
- ✓ Search for datasets
- ✓ Do all of your homework
- × Do not hand in as yours the solution of others
- ➤ Let me know of your interests. I can adjust topics accordingly
- ➤ Come and learn. Useful down the road.

Algunos ejemplos de Redes

- Diccionario inglés: *A collection of inter-connected things*
- En español: *Conjunto de elementos organizados para determinado fin*
- Ok. Hay múltiples cosas, y están conectadas. Dos extremos





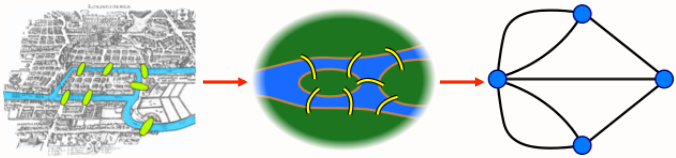
i) Un sistema real (complejo) de componentes interconectadas.

ii) Un grafo represneentando un sistema

- Entender sistemas complejos $\Leftrightarrow$ entender redes subyacentes

Un poco de historia...

- El análisis basado en redes tiene una larga historia en las ciencias
- Fundamentos matemáticos de la teoría de grafos “nacen” con L. Euler en 1735



  - Famoso problema de los siete puentes de Königsberg
- Leyes de circuitos eléctricos (G. Kirchoff, 1845)
- Estructura molecular en química (A. Cayley, 1874)
- Representación en red de interacciones sociales (J. Moreno, 1930)
- Redes de potencia (1910), telecomunicaciones y por supuesto Internet (1960)
- Google (1997), Facebook (2004), Twitter (2006), . . .

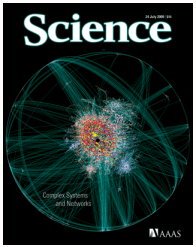
¿Qué pasa ahora?

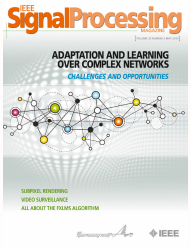
- Entender sistemas complejos ⇔ entender redes subyacentes

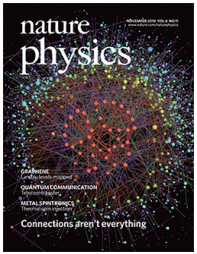
- Existe desde hace mucho pero relativamente reducido hasta ∼ mediados de los 90
- Explosión de interés “reciente". Algunas razones:
  - Perspectiva a nivel de *sistema* en ciencias, alejada del reduccionismo
  - Dispionibilidad de enormes bases de datos y de poder de cómputo
  - Globalización, the Internet, interconexión de las sociedades modernas

Network Science

- Estudio de sistema complejos a través de su representación como redes
  - Complejo: dependencias, estructuras, heterogeneidad, alta dimensión...
  - Ejemplos: economy, metabolism, brain, society, Web, ...
- Lenguaje universal para describir sistemas y datos complejos
  - Sorprendentes similaridades en redes entre ciencia, naturaleza, tecnología
- Desde Biología hasta físcia, economía a estadística, computación a sociología y sus impactos! ...





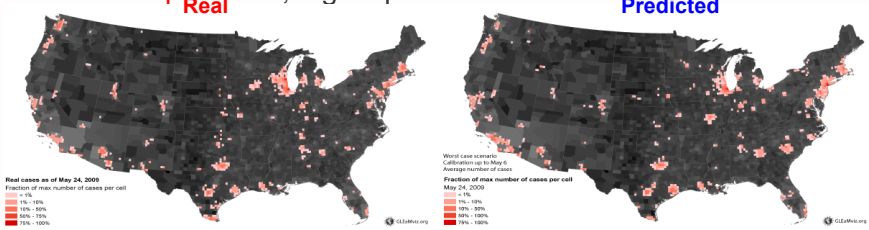


Impacto económico

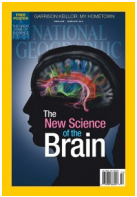
- Google Market cap: $1.24 trillion
- Facebook Market cap: $736 billion
- Cisco Market cap: $188 billion
- Apple Market cap: $2.22 billion

Impacto sanitario

- Predicción de epidemias, e.g. la pandemia de H1N1 en 2009

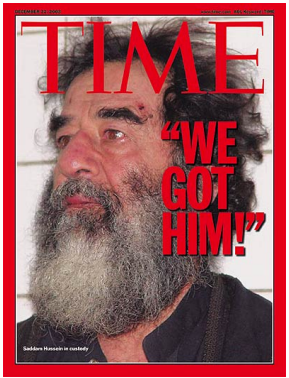


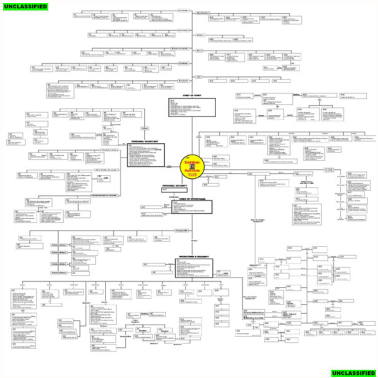
- Human Connectome Project par mapear las conexiones del cerebro



Impactos en seguridad

- Análisis de la red social de allegador para capturar a S. Hussein





Características de Network Science

- ¿Cuáles son los objetivos de NS?
  - Descubir patrones y propiedades estadísticas de los datos en red
  - Entender las fundamentes del comportamiento y la estructura de las redes
  - Diseñar redes más eficientes, robustas y socialmente inteligentes
- Características: interdisciplinaria, empírica, cuantitativa, computacional
- Empírica estudios de datos en grafo para descubrir patrones y principios
  - recolección, toma de medidaas, resumen de características, visualización?
- Modelos matemáticos. Encuentro de la teoría de grafos con la inferencia estadística
  - entender, predecir, clasificar, detectar anomalías?
- Algoritmos para analítica de grafos
  - desafíos computacionales, escalabilidad, tratabilidad vs optimalidad?

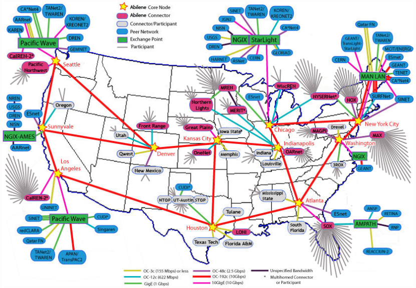
Ejemplos de redes

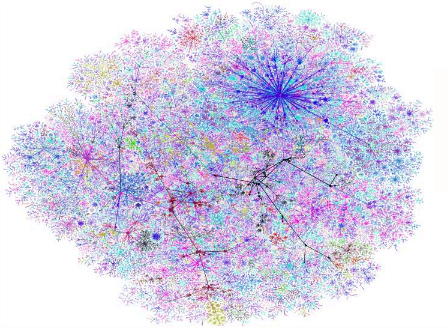
- El análisis de redes atraviesa abarca ciencia, humanidades y artes
- Veremos algunos (pocos) ejemplos de cuatro grandes áreas (ver referencias):
  - Tecnológica
  - Biológica
  - Social
  - Información
- Taxonomía éstandar (no la única!)

  $\Rightarrow$ clasificación “blanda”: una red puede pertenecer a múltiples categorías

Redes tecnológicas

- Ejemplos: comunicación, transporte, energía, redes de sensores

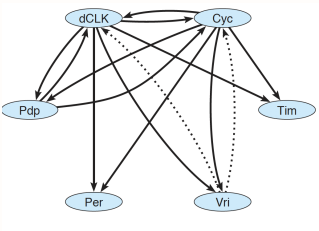


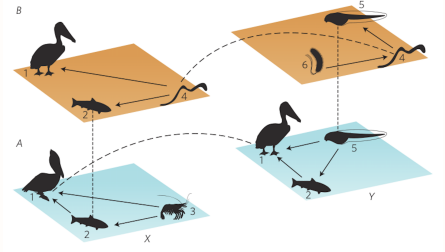


- P1: ¿Cómo luce la Internet hoy? ¿Qué tan grande es?
- P2: ¿Cómo será el tráfico mañana entre Shangrilay Punta Carretas?
- P3: ¿Cómo detectar comportamientos anómalos?

Redes biológicas

- Ejemplos: neuronas, regulatorias de genes, interacción entre proteínas, metabólicas, ecológicas, predadores





- P1: ¿Ciertas interacciones entre genes son mayores a las esperadas?
- P2: ¿Qué partes del cerebro se “comunican” durante una tarea dada?
- P3: ¿ Podemos predecir funciones biológicas de proteínas a partir de sus interacciones?

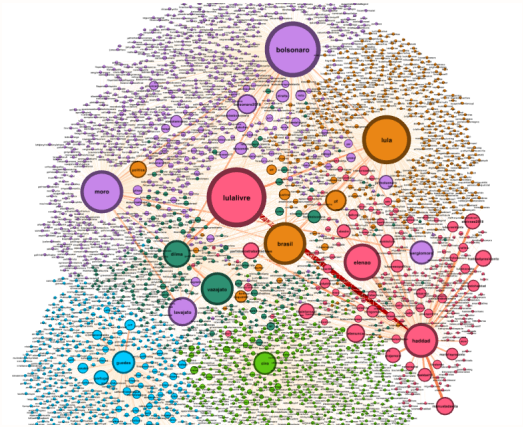
Redes sociales

- Ejemplos: amistad, corporativas, intercambio de mails, relaciones internacionales, financieras

- P1: What are the mechanisms underpinning friendship formation?
- P3: Can we identify overlapping communities?

Redes sociales

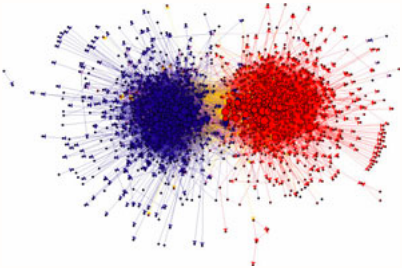
- Ejemplos: amistad, corporativas, intercambio de mails, relaciones internacionales, financieras



- P: ¿Qué elementos son centrales y cuáles son periféricos?

Rede de Información

- Ex: WWW, Twitter, citas comunes entre revistas académicas, blogosphere, co-autoría de papers, peer-to-peer networks



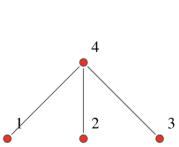
- P1: ¿Cómo cambia el tamaño y la estructura de la WWW en el tiempo?
- P2: ¿Podemos reconocer estilos de autores?
- P3: ¿Cómo se puede rastrear información en redes sociales?

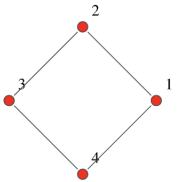
Esquema del curso

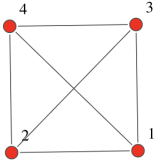
1. Introducción, motivación y desafíos del área.
2. **Teoría básica de grafos**. (Paola - 3 clases clases). Laboratorio 1.
3. **La estructura de las grandes redes de datos**.(Paola - 3 clases). Laboratorio 2.
4. **Modelos de grafos**. (Federico - 4 clases).
5. **Inferencia en redes**. (Marcelo - 4/5 clases). Laboratorio 3.
6. **Graph Neuronal Networks (GNN)**. (Federico - 4/5 semanas). Laboratorio 4.
7. **Estimación y Clasificación para datos en grafos**. (Marcelo - 4/5 clases). Laboratorio 5.

Teoría básica de grafos

- Vértices, aristas, grados, subgrafos, familias de grafos . . .
- Conectividad, paseos y caminos en grafos, componentes gigantes. . .
- Teoría algebraica de grafos: matriz de adyacencia y laplaciano del grafo, espectro . .



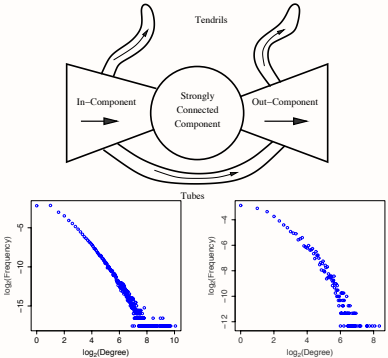




$$\begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix} \quad \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{pmatrix} \quad \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$

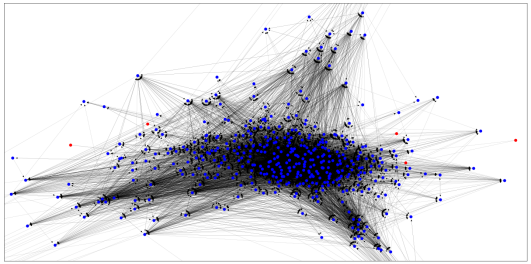
La estructura de las grandes bases de datos

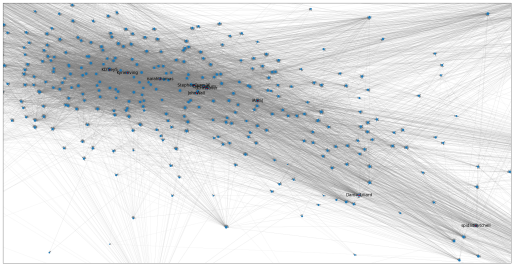
- Análisis descriptivo y propiedades de las redes
- La WWW y otras grandes redes reales muestran una estructura “bowtie”
- Distribuciones de grado “power-law” y small-world presentes también en redes reales



- De interés: construcción de grafos asociados a redes, visualización, medidas de centralidad (Google’s PageRank) , detección de comunidades, muestreo, etc.

Análisis descriptivo

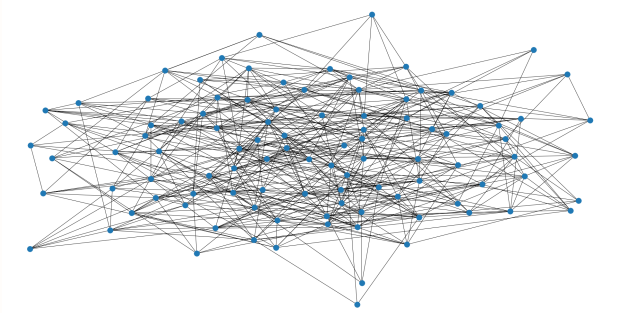


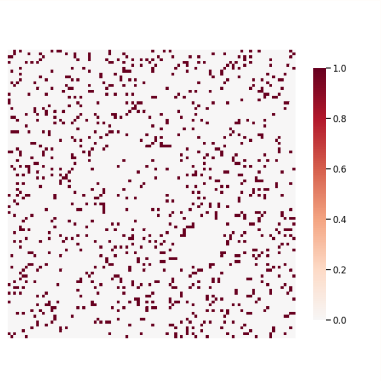


- Seguidores Twitter jugadores NBA
- Medidas de importancia de los nodos (King James, Sephen Curry...)

Modelos de Grafos

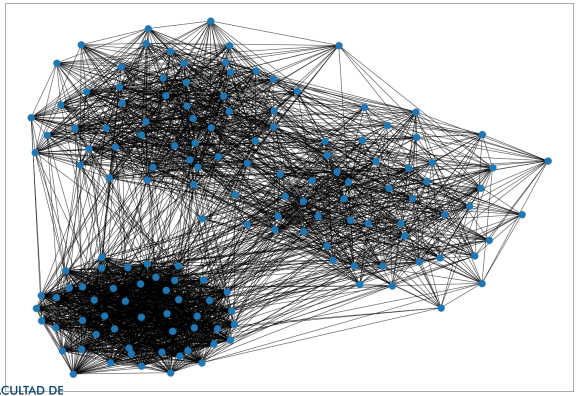
- Grafos aleatorios (ER), Stochastic Block Models (SBM), Preferential Attachment, Random Dot Product Graphs (RDPG)

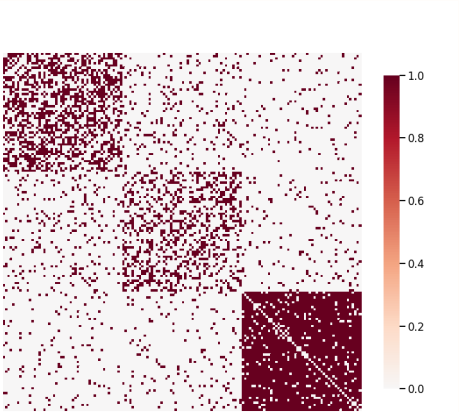




Modelos de Grafos

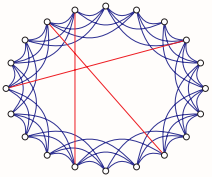
- Grafos aleatorios (ER), Stochastic Block Models (SBM), Preferential Attachment, Random Dot Product Graphs (RDPG)

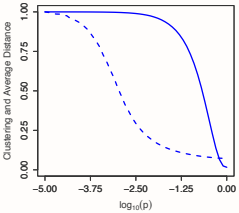




Inferencia en redes

- Modelo de Watts-Strogatz captura el efecto de estructura small-world
  - muy estructurado localmente y
  - “poco” estructurado globalmente (similar a grafos aleatorios puros)

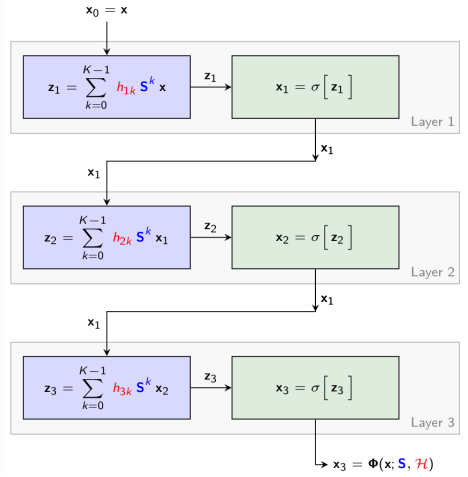




- De interés: modelos aleatorios de grados, inferencia de la topología de la rede, modelos de crecimiento para redes dinámicas, prefential attachment.

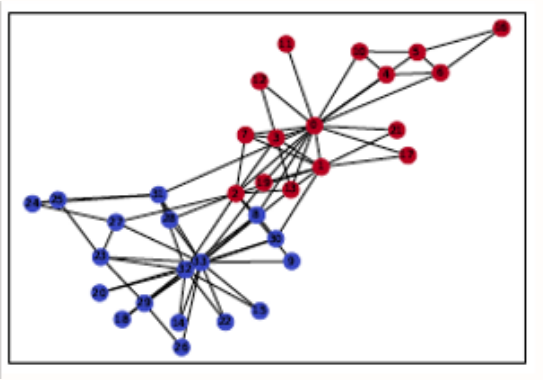
Graph Neural Networks

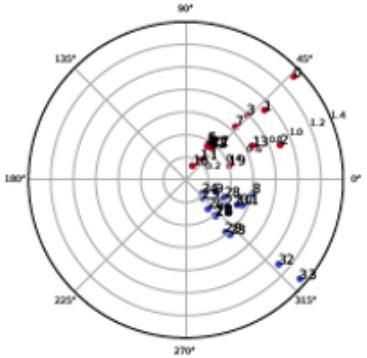
- Agregamos datos en los vértices



Estimación y Clasificación para Datos en Grafos

- Detección de comunidades: Zachary Karate Club





Estimación y Clasificación para Datos en Grafos

- Detección de comunidades: Blog políticos en Francia