

# Trabajo Práctico 3

Métodos Numéricos

Segundo cuatrimestre - 2013

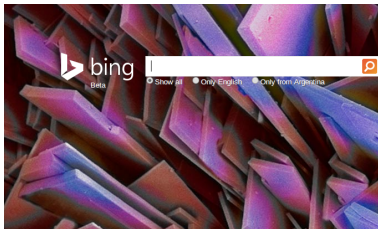
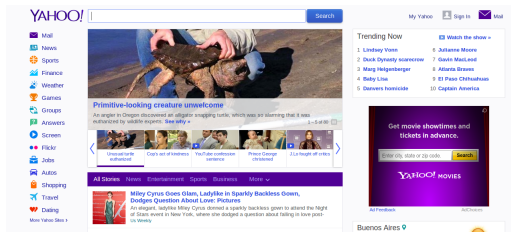
# Hasta ahora

- ▶ TP 1: Fast inverse square root (Quake III).
- ▶ Taller 1: Calcular isoterma alto horno.
- ▶ TP 2: Análisis de estructuras Pratt Truss.
- ▶ Taller 2: Eliminación de ruido en imágenes.

## Objetivo

Seguir viendo aplicaciones reales de MN.

# Motores de búsqueda



# Motores de búsqueda

- ▶ Explorar la red e identificar todas las páginas con acceso público.
- ▶ Almacenar la información obtenida, para realizar búsquedas eficientemente.
- ▶ Determinar un orden de las páginas según su importancia, para presentar la información con un orden de relevancia.


Google

Web Imágenes Videos Noticias Más ▾ Herramientas de búsqueda

---

Cerca de 221.000.000 resultados (0,16 segundos)

[Noticias de messi](#)



[La admiración por Lionel Messi también llegó a Estados Unidos](#)  
Clarín.com - hace 37 minutos  
Quedó séptimo en una encuesta que reunió a los atletas profesionales favoritos de los norteamericanos. Es la primera vez en la historia que ...

[Las bromas antes del clásico español pasan por cómo frenar a Messi](#)  
Clarín.com - hace 37 minutos

[Las lecciones de Messi a Neymar para evitar que lo cosan a patadas cada pa...](#)  
ecodiario - De Javier Martín - hace 3 horas

[Lionel Messi - Wikipedia, la enciclopedia libre](#)  
es.wikipedia.org/wiki/Lionel\_Messi ▾  
Lionel Andrés **Messi** Cuccitini (Rosario, Argentina, 24 de junio de 1987), conocido también como Leo **Messi**, es un futbolista argentino que también posee la ...  
Biografía - Clubes - Selección de fútbol de ... - Estadísticas

[Punto positivo: Lionel Messi volvió con un gol en el empate de ...](#)  
canchallena.lanacion.com.ar > Fútbol > Champions League  
hace 10 horas - Punto positivo: Lionel **Messi** volvió con un gol en el empate de Barcelona ante Milan, Por la Champions, el equipo de Martino igualó 1 a 1 en ...

[Los goles de Messi, decisivos en Champions | Barca | Sport.es](#)  
www.sport.es/es/noticias/.../los-goles-messi-decisivos-champions-2774050

# Outline

Contexto TP3

Cadenas de Markov

Algoritmo PageRank

Enunciado

# Cadenas de Markov

## Definición

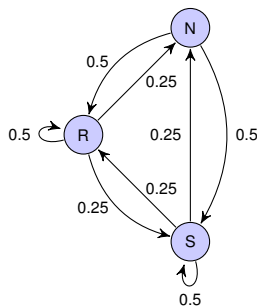
Consideramos un conjunto de estados  $S = \{s_1, s_2, \dots, s_r\}$ . El proceso empieza en alguno de estos estados y se mueve de un estado a otro. A cada movimiento se lo denomina *paso*. Si la cadena se encuentra actualmente en el estado  $s_i$ , en el siguiente paso se mueve al estado  $s_j$  con probabilidad  $p_{ij}$ . Esta probabilidad no depende de los estados anteriores a  $s_i$  en los que se haya encontrado el proceso.

# Cadenas de Markov

## Ejemplo: Cambio de clima

- ▶ Tres posibilidades: Bueno (N), Lluvioso (R), Nieve (S).
- ▶  $p_{ij}$  es la probabilidad de que si en un determinado día estamos en un estado  $i$  (i.e., N, R ó S) al día siguiente estemos en el estado  $j$ .
- ▶ Particularidad: no pueden haber dos días buenos (N) seguidos.

Grafo de transiciones:



Matriz de transiciones:

$$P = \begin{matrix} & \begin{matrix} R & N & S \end{matrix} \\ \begin{matrix} R \\ N \\ S \end{matrix} & \begin{pmatrix} 0.5 & 0.25 & 0.25 \\ 0.5 & 0 & 0.5 \\ 0.25 & 0.25 & 0.5 \end{pmatrix} \end{matrix}$$

- ▶ Filas: Estado actual.
- ▶ Columnas: Estado al que podemos movernos.
- ▶ Matriz estocástica por filas.

# Cadenas de Markov

Mirando más allá de un día

## Nuevo problema

Queremos saber cuál es la probabilidad que, si hoy está lluvioso, nieve dentro de dos días. Llamamos a esta probabilidad  $p_{RS}^{(2)}$ .

Esto es la unión disjunta de los siguientes eventos:

1. Lluvioso (R) mañana y nieve (S) pasado.
2. Bueno (N) mañana y nieve (S) pasado.
3. Nieve (S) mañana y nieve (S) pasado.

$$P = \begin{matrix} & \begin{matrix} R & N & S \end{matrix} \\ \begin{matrix} R \\ N \\ S \end{matrix} & \begin{pmatrix} 0.5 & 0.25 & 0.25 \\ 0.5 & 0 & 0.5 \\ 0.25 & 0.25 & 0.5 \end{pmatrix} \end{matrix} \quad p_{RS}^{(2)} = \underbrace{p_{11}p_{13}}_1 + \underbrace{p_{12}p_{23}}_2 + \underbrace{p_{13}p_{33}}_3$$



# Cadenas de Markov

## En general

En el caso anterior,

$$p_{ij}^{(2)} = \sum_{k=1}^r p_{ik} p_{kj} = (P^2)_{ij}.$$

## Propiedad

El resultado de multiplicar dos matrices estocásticas por filas es una matriz estocástica por filas.

## Teorema

Sea  $P$  la matriz de transición de una cadena de Markov. El elemento  $p_{ij}^{(k)}$  de la matriz  $P^k$  es la probabilidad de que la cadena de Markov, empezando en el estado  $i$ , se encuentre en el estado  $j$  después de  $k$  pasos.

# Cadenas de Markov

Y si no conocemos el estado actual?

Hasta ahora, supusimos que conocemos el estado actual. Qué pasa si la cadena se encuentra en algún estado con una probabilidad?

**Definición: vector de probabilidades**

$x \in \mathbb{R}^k$  es un vector (fila) de probabilidades si  $x_i \geq 0$  y  $\sum_{i=1}^k x_i = 1$ .

**Teorema**

Sea  $P$  la matriz de transición de una cadena de Markov, y sea  $u$  el vector que representa la distribución inicial. Entonces, la probabilidad de que la cadena se encuentre en el estado  $s_i$  luego de  $k$  pasos es la componente  $i$ -ésima del vector

$$u^{(k)} = uP^k$$

# Cadenas de Markov

Estado estacionario: qué pasa en el largo plazo

Qué sucede con el sistema si consideramos

$$\lim_{n \rightarrow \infty} P^n?$$

## Definición: Matriz Regular

Una matriz de transiciones  $P$  se dice regular si  $P^k$  tiene solamente entradas positivas para algún entero  $k$ .

## Teorema

Sea  $P$  una matriz de transiciones regular. Entonces:

- ▶  $\lim_{n \rightarrow \infty} P^n = W$ , donde todas las filas de  $W$  son un mismo vector  $w$ .
- ▶  $wP = w$ , y todos los vectores que cumplan  $vP = v$  son un múltiplo de  $w$ .
- ▶  $xP^n \rightarrow w$  con  $n \rightarrow \infty$ .

# Cadenas de Markov

Estado estacionario: qué pasa en el largo plazo

Si  $P$  es una matriz de transiciones regular, entonces:

- ▶ 1 es un autovalor de  $P$ .
- ▶ Hay un único vector de probabilidades que es el autovector asociado al autovalor 1, y es  $w$ .
- ▶ Se demuestra que los demás autovalores cumplen  $1 = \lambda_1 > |\lambda_2| \geq \dots \geq |\lambda_r|$ .

## Interpretación

Al vector de probabilidades  $w$  se lo denomina *estado estacionario*. La componente  $w_i$  representa la proporción de tiempo que la cadena se encuentra, en el largo plazo, en el estado  $s_i$ .

## En la práctica

Como  $wP = w$ , entonces  $P^t w^t = w^t$ . Podemos intentar usar el método de la potencia para calcular  $w^t$ .

# PageRank

## Problema

- ▶ Tenemos un conjunto de páginas  $Web = \{1, \dots, n\}$ .
- ▶ El objetivo es asignar a cada una de ellas un puntaje que determine la importancia relativa de la página respecto de las demás.
- ▶ Vamos a trabajar directamente sobre la matriz traspuesta.
- ▶ Si definimos una cadena de Markov regular, entonces el estado estacionario nos dará la proporción de tiempo que el navegante aleatorio pasará en cada página.

# PageRank

## Modelo inicial

### Modelo mediante cadenas de Markov

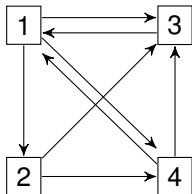
Consideramos el modelo del *navegante aleatorio*, que comienza en una página cualquiera del conjunto y va navegando a través de sus links.

- ▶ Cada página representa un estado de la cadena.
- ▶ Podemos pasar de una página  $j$  a otra  $i$  si hay un link de  $j$  a  $i$ . Definimos  $W \in \{0, 1\}^{n \times n}$  como  $w_{ij} = 1$  si hay un link de  $j$  a  $i$ , y  $w_{ij} = 0$  en caso contrario.
- ▶  $n_j = \sum_{i=1}^n w_{ij}$  es el grado de la página  $j$  (cantidad de links salientes).
- ▶ Definimos  $P \in \mathbb{R}^{n \times n}$  como  $P_{ij} = 1/n_j$  como la probabilidad de ir de la página  $j$  a la  $i$ , dado que existe un link de  $j$  a  $i$ .

# PageRank

Ejemplo (Bryan y Leise)

$$n_1 = 3, n_2 = 2, n_3 = 1, n_4 = 2$$



$$P = \begin{bmatrix} 0 & 0 & 1 & \frac{1}{2} \\ \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{2} & 0 & 0 \end{bmatrix}$$

$P$  es estocástica por columnas.

Pregunta:

Qué pasa si una página  $i$  no tiene links salientes (i.e.,  $n_i = 0$ , denominado *dangling node*)?

# PageRank

## Solución a *dangling nodes*

Definimos:

- ▶  $v \in \mathbb{R}^n$ ,  $v_i = 1/n$ .
- ▶  $d \in \{0, 1\}^n$ ,  $d_i = 1$  si  $n_i = 0$ ,  $d_i = 0$  en caso contrario.
- ▶  $D = vd^t$
- ▶  $P_1 = P + D$

## Idea

Si estamos en una página sin links salientes, entonces con probabilidad uniforme  $1/n$  el navegante pasa a cualquiera de las páginas en *Web*.

## Pregunta:

Ahora la matriz es estocástica por columnas. Es regular?



# PageRank

## Asegurando regularidad

Depende del grafo de conectividad. Sin embargo, podemos extender la idea anterior en general a todas las páginas. A este fenómeno se lo denomina *teletransportación*.

- ▶  $\vec{1} = (1, \dots, 1) \in \mathbb{R}^n$ .
- ▶  $E = v\vec{1}^t$ .
- ▶  $P_2 = cP_1 + (1 - c)E$ ,  $c \in (0, 1)$ .
- ▶  $P_2$  es estocástica por columnas y  $(P_2)_{ij} > 0$ ,  $1 \leq i, j \leq n$ .

## Finalmente

Tenemos una cadena de Markov que modela el problema y cumple todas las condiciones. Para generar el ranking de las páginas, buscamos un autovector  $w$  asociado al autovalor 1 de  $P_2$ , tal que  $P_2 w = w$ , y  $w$  sea un vector de probabilidades.

# TP3

## Objetivos generales

- ▶ Trabajar sobre una aplicación real, implementando el algoritmo PageRank.
- ▶ Considerar dos métodos distintos para su resolución. El método de la potencia y un variante particular propuesta en Kamvar et al.
- ▶ Simular un trabajo de investigación mediante la implementación de uno de los métodos propuestos en Kamvar et al., estudiando el desarrollo del mismo y comparando los resultados con el método tradicional.

# TP3

## Enunciado: Punto 1.

1. En base a su definición,  $P_2$  no es una matriz esparsa. Sin embargo, en Kamvar et al. (Algoritmo 1) se propone una forma alternativa para computar  $x^{(k+1)} = P_2 x^{(k)}$ . Mostrar que el cómputo propuesto es equivalente. Utilizarlo para mejorar el espacio requerido en memoria para el almacenamiento de la matriz  $P_2$  y el tiempo de ejecución requerido para hacer la multiplicación entre matrices y vectores.

# TP3

## Enunciado: Punto 2.

2. Basándose en el análisis del punto anterior, implementar el método de la potencia para calcular el autovector principal de la matriz  $P_2$ .

# TP3

## Enunciado: Punto 3.

3. Implementar la variante del Método de la Potencia propuesta en Kamvar et al. (Sección 5), denominada Extrapolación Cuadrática. El método de Cuadrados Mínimos involucrado debe ser resuelto utilizando la Factorización QR de la matriz mediante alguno de los métodos vistos en la materia.

# TP3

## Enunciado: Punto 4.

4. Realizar experimentos considerando distintas instancias de prueba. Para ello, se podrá utilizar el código adjuntada para la generación de instancias en base a datos reales, o cualquier otra herramienta que el grupo considere necesaria. Evaluar también los algoritmos alguno de los conjuntos de instancias provistos en SNAP. Para cada algoritmo, analizar el tiempo de ejecución, la evolución del error entre iteraciones consecutivas y considerar distintos criterios de parada. Además, analizar la calidad del ordenamiento obtenido en términos de la relevancia de las páginas.

# TP3

## Material extra (optativo)

Para generar las instancias, se adjunta un código Python que, dada una lista de direcciones de páginas web, parsea el código html de cada una de ellas y genera el grafo de conectividad.

## Algunas aclaraciones

- ▶ Se restringe a links entre las páginas de la lista. El resto de los links son descartados.
- ▶ El chequeo para decidir si un link es o no a una página de la lista es básico (ejemplo: `www.example.com`, ó `example.com`, ó `example.com.ar` son considerados links distintos)
- ▶ Links que aparezcan dos o más veces son contados una única vez.
- ▶ Pueden tomar este código y modificarlo según sus necesidades.
- ▶ Si encuentran algún error en el código, por favor contacten a los docentes.

# TP3

## Material extra (optativo)

### Utilización

#### El comando

```
python webparser.py weblist.in graph.out
```

toma como entrada la lista de páginas y genera el grafo, con el formato indicado en el enunciado del trabajo, en el archivo graph.out.



# TP3

## Recomendaciones

- ▶ Lunes 28/10: Lectura artículos, implementación matriz rala, generación matriz  $P$ .
- ▶ Lunes 04/11: Método de la potencia, demostración, implementación CM.
- ▶ Lunes 11/11: Metodo de Karmar et al., experimentos, informe.

# Trabajo Práctico

Fecha de entrega

- ▶ **Formato Electrónico:** Martes 12 de Noviembre de 2013, hasta las 23:59 hs, enviando el trabajo (informe + código) a la dirección **metnum.lab@gmail.com**. El subject del email debe comenzar con el texto **[TP3]** seguido de la lista de apellidos de los integrantes del grupo.
- ▶ **Formato físico:** Miercoles 13 de Noviembre de 2013, de 17 a 18 hs. en la clase teórica.

## Importante

El horario es estricto. Los correos recibidos después de la hora indicada serán considerados re-entrega.