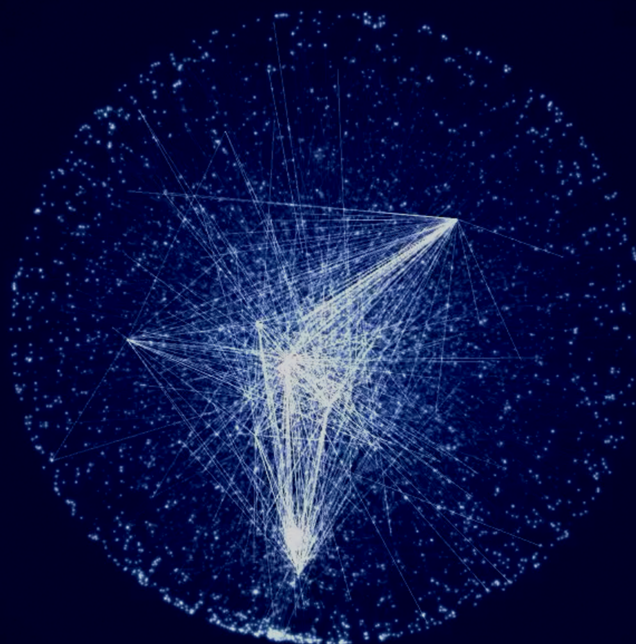
PARTIMOS EN BREVE

MUCHAS GRACIAS

Network Science



Video: Protein-Protein Interactions
Credit: Mauro Martino

Dr. Cristian Candia

Director del Magister en Data Science y del Computational Research in Social Science Laboratory,
Instituto de Data Science, Centro de Inv. en Complejidad Social,
Facultades de Ingeniería y Gobierno
Universidad del Desarrollo, Chile

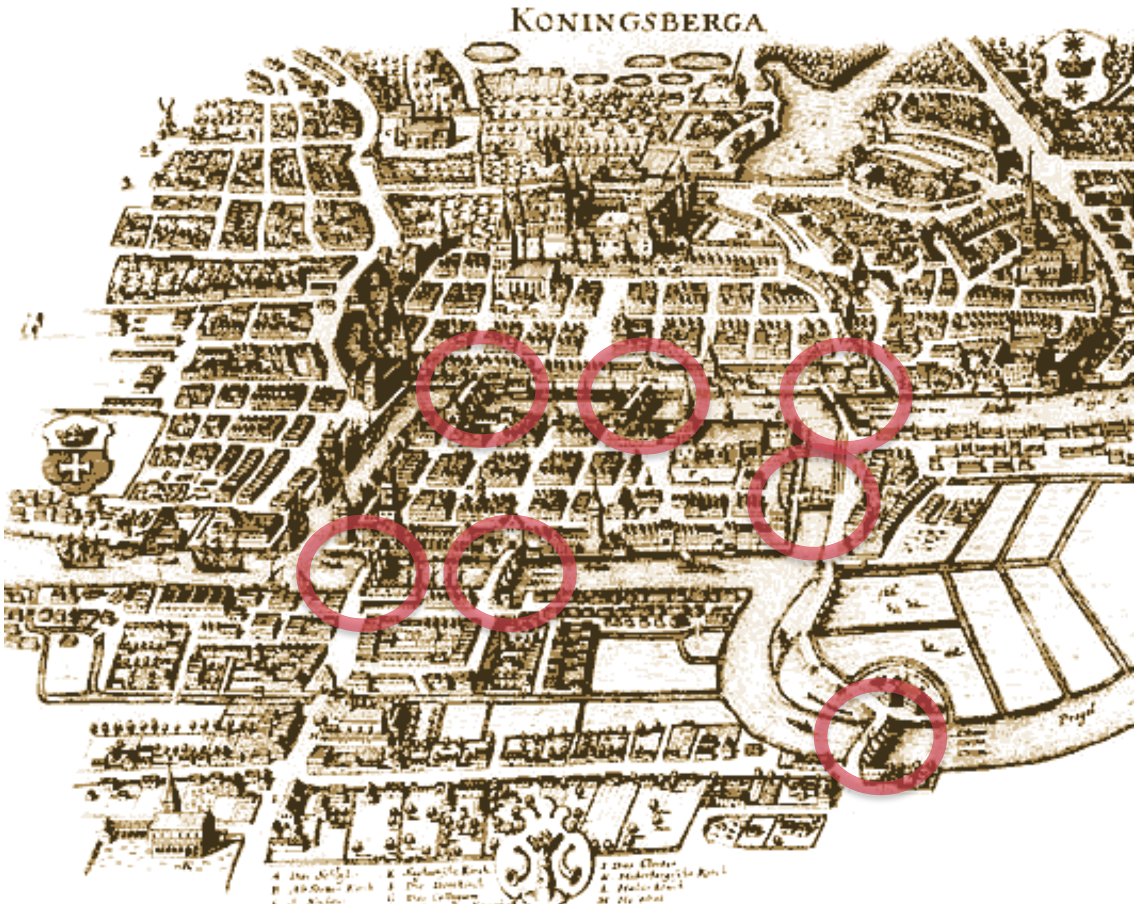
Académico Adjunto,
Northwestern University, United States

Universidad del Desarrollo Facultad de Ingeniería
Universidad del Desarrollo Facultad de Gobierno
dataScience UDD
CICS UDD CENTRO DE INVESTIGACIÓN EN COMPLEJIDAD SOCIAL
CRISSLAB DECODING BEHAVIOR, ADVANCING SOCIETY
NICO NORTHWESTERN INSTITUTE ON COMPLEX SYSTEMS

Teoría de Grafos

Estas diapositivas se basan parcialmente en el curso del
Prof. Albert-László Barabási, de Northeastern University, con autorización.
El contenido ha sido traducido para su uso en este curso.

Los puentes de Konigsberg



1736

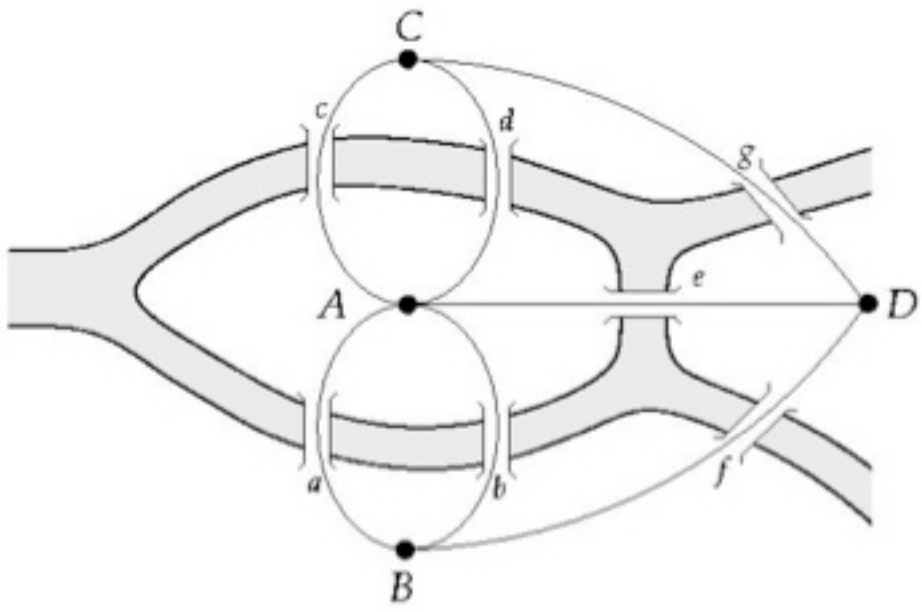
¿Se puede cruzar los siete puentes y nunca cruzar el mismo puente dos veces?

LOS PUENTES DE KONIGSBERG

¿Se puede cruzar los siete puentes y nunca cruzar el mismo puente dos veces?

1736

LOS PUENTES DE KONIGSBERG



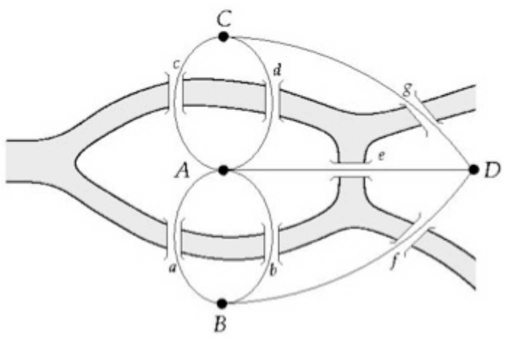
¿Se puede cruzar los siete puentes y nunca cruzar el mismo puente dos veces?

1735: **Teorema de Euler:**

(a) Si un grafo tiene **más de dos nodos de grado impar**, no hay ruta.

(b) Si un grafo está conectado y **no tiene nodos de grados impares**, tiene al menos una ruta.

LOS PUENTES DE KONIGSBERG



¿Se puede cruzar los siete puentes y nunca cruzar el mismo puente dos veces?

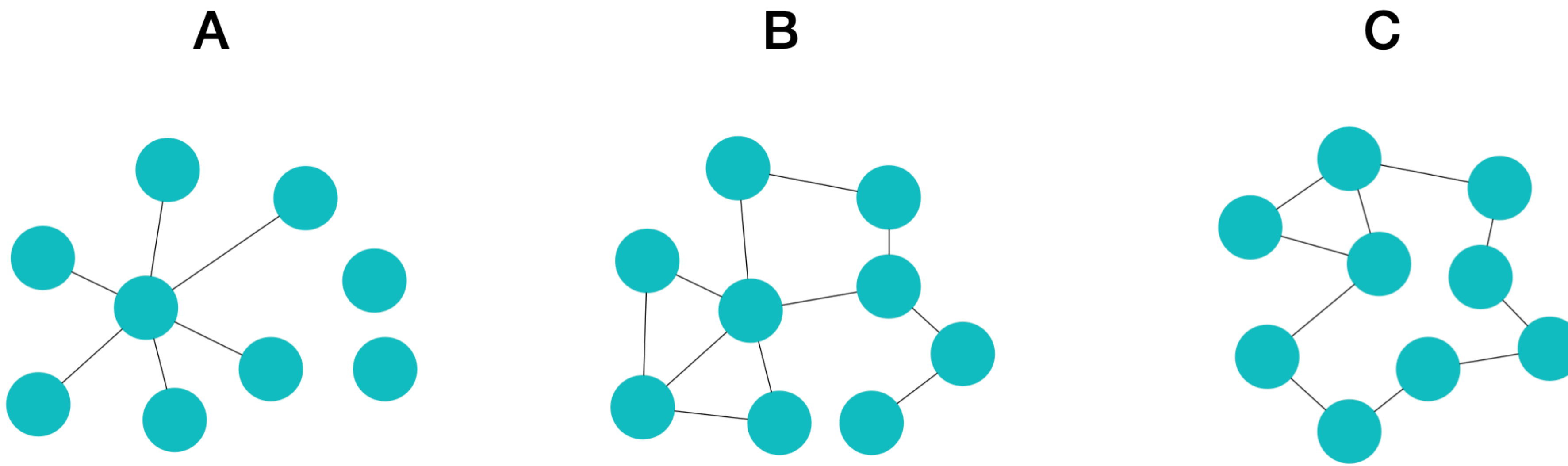
Hoy recordamos la prueba de Euler porque fue la primera vez que alguien resolvió un problema matemático convirtiéndolo en un grafo. En retrospectiva, la prueba tiene dos mensajes importantes:

1. **Algunos problemas se vuelven más simples y tratables si se representan como un grafo.**

2. **La existencia del camino no depende de nuestro ingenio para encontrarlo.** Más bien, es una propiedad del grafo. De hecho, dado el diseño de los puentes Konigsberg, no importa lo inteligentes que seamos, nunca encontraremos el camino deseado.

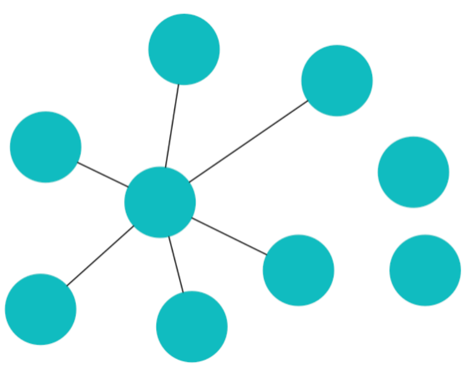
¿En cuál de estas redes es posible idear un paseo que atraviese cada enlace una vez y solo una vez?

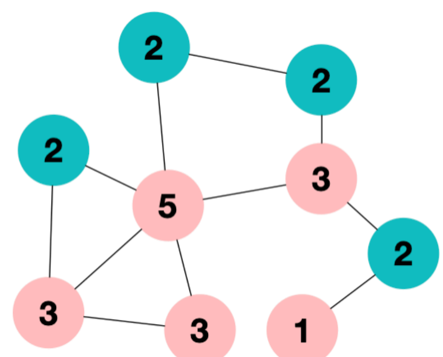
Es decir, ¿en qué red es posible realizar un Camino Euleriano?

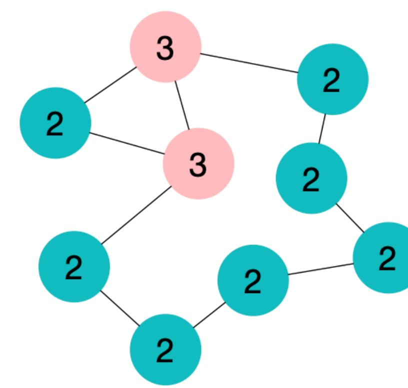


1. ¿Está conectado?

2. Cuente el número de nodos de grados impares, ¿es 0 o 2?



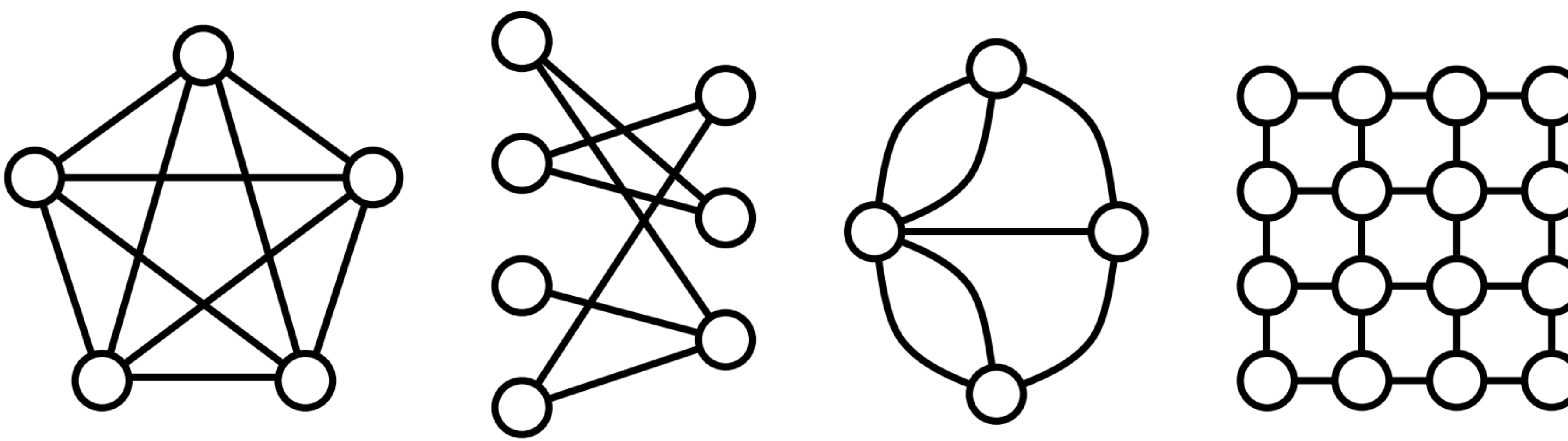




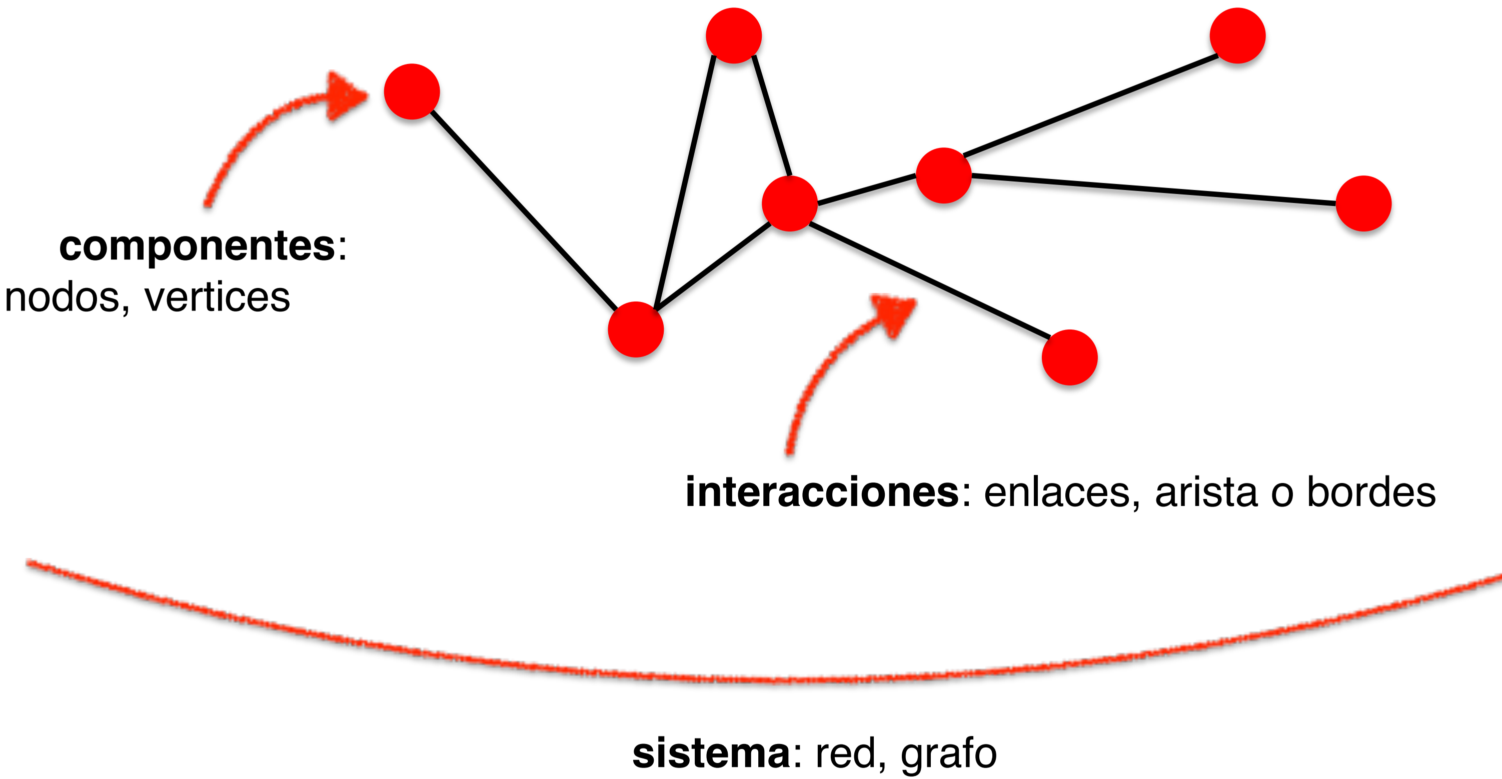
Redes y grafos

¿Qué es un grafo?

Estructura matemática que consta de "nodos" (o vértices) y "enlaces" (aristas o borde) que conectan a los nodos.



COMPONENTES DE UN SISTEMA COMPLEJO



Notación básica de una red

- **Conjunto de nodos:** Representados como V, donde V es un conjunto de nodos, vertices, agentes, actores, etc. en la red.

Por ejemplo: V = {A, B, C, D}

- **Conjunto de enlaces:** Representados como E, donde E es un conjunto ordenado de pares de nodos (u, v), denotando una conexión o relación entre ellos.

Por ejemplo: E = {(A, B), (A, C), (B, D), (C, D)}

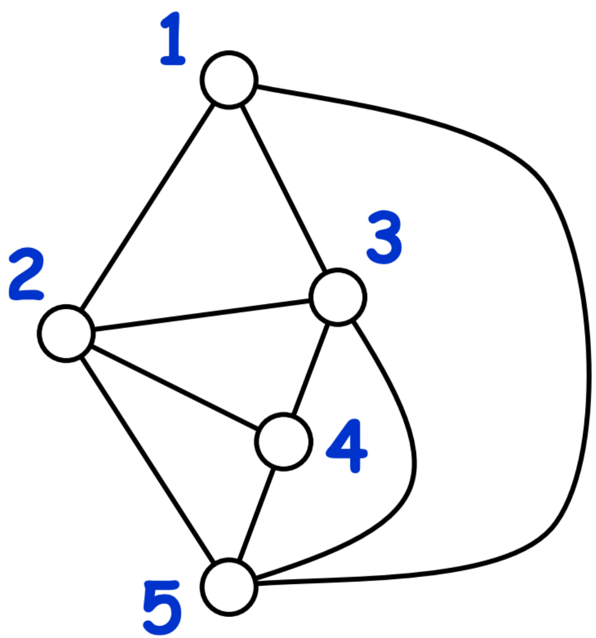
Cada par ordenado (u,v) en E representa un enlace entre nodo u y nodo v.

- **Red o grafo:** La red típicamente se representa como G(V,E), donde G es la red o grado, V es el set de nodos o vertices, y E es el conjunto de enlaces.

Por ejemplo: G(V, E) = ({A, B, C, D}, {(A, B), (A, C), (B, D), (C, D)})

Ejemplo

- **Conjunto de nodos:** V = {1, 2, 3, 4, 5}
- **Conjunto de enlaces:** E = {(u,v)}
- **Red o grafo:** G(V, E)



Nodos

1, 2, 3, 4, 5

Enlaces

1<->2, 1<->3, 1<->5,
2<->3, 2<->4, 2<->5,
3<->4, 3<->5, 4<->5

Los nodos pueden tener estados; los enlaces pueden tener direcciones y pesos

REDES O GRAFOS?

red a menudo se refiere a sistemas reales

- www,
- social network (red social)
- metabolic network. (red metabólica)

Lenguaje: (Red (Network), nodo (node), enlace (link))

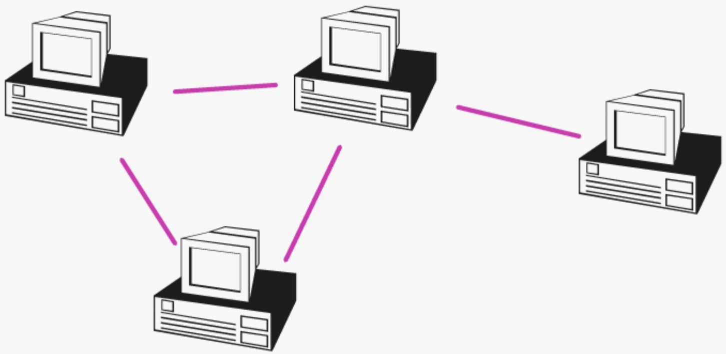
grafo: representación matemática de una red

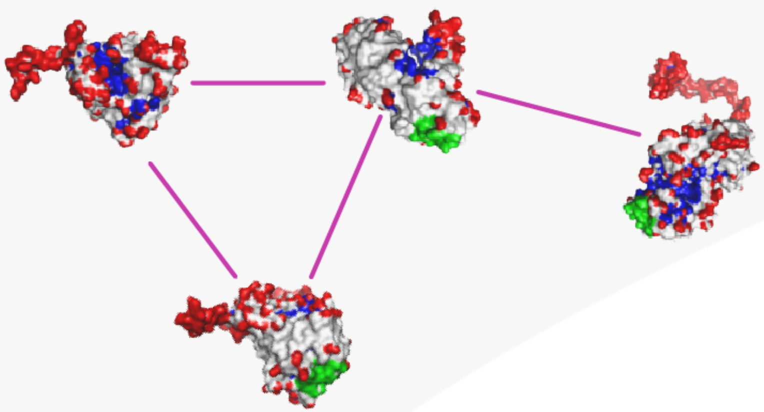
- web graph,
- social graph (a Facebook term)

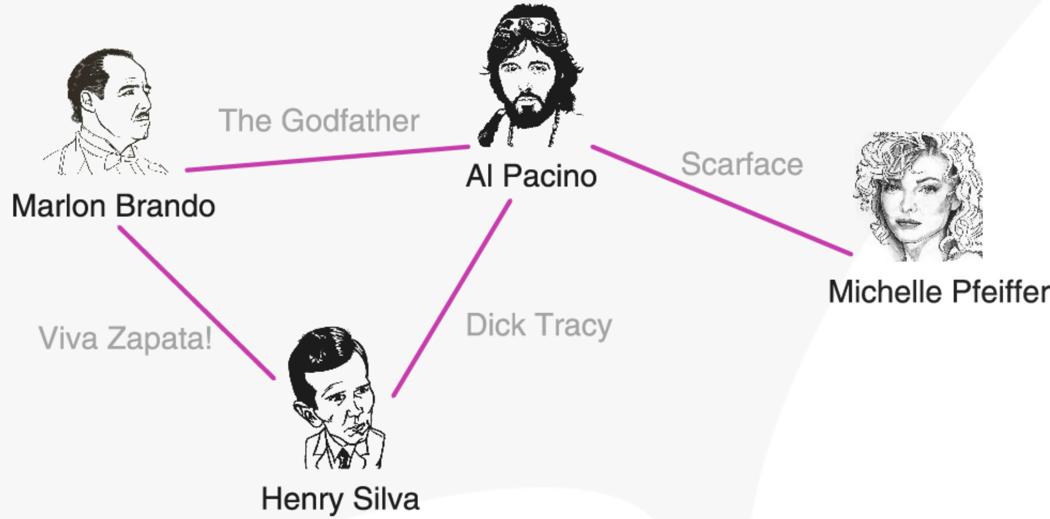
Lenguaje: (Grafo (graph), vertice (vertex), enlace (edge))

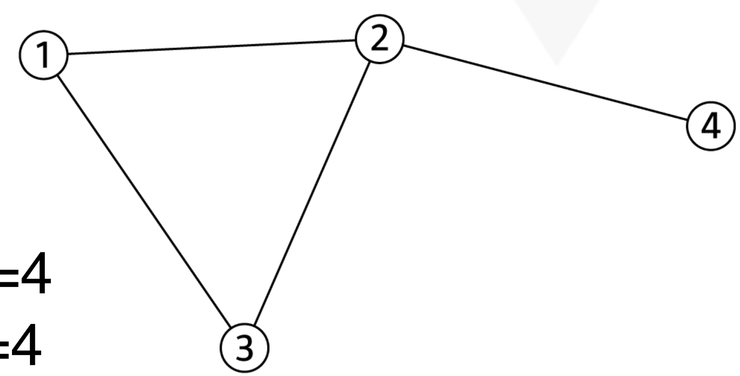
Trataremos de hacer esta distinción siempre que sea apropiado, pero en la mayoría de los casos usaremos los dos términos indistintamente.

UN LENGUAJE COMÚN









N=4

L=4

CHOOSING A PROPER REPRESENTATION

La elección de la representación de red adecuada determina nuestra capacidad para utilizar la teoría de redes con éxito.

En algunos casos hay una representación única e inequívoca.
En otros casos, la representación no es de ninguna manera única.

Por ejemplo, la forma en que asignamos los vínculos entre un grupo de individuos determinará la naturaleza de la pregunta que podemos estudiar.

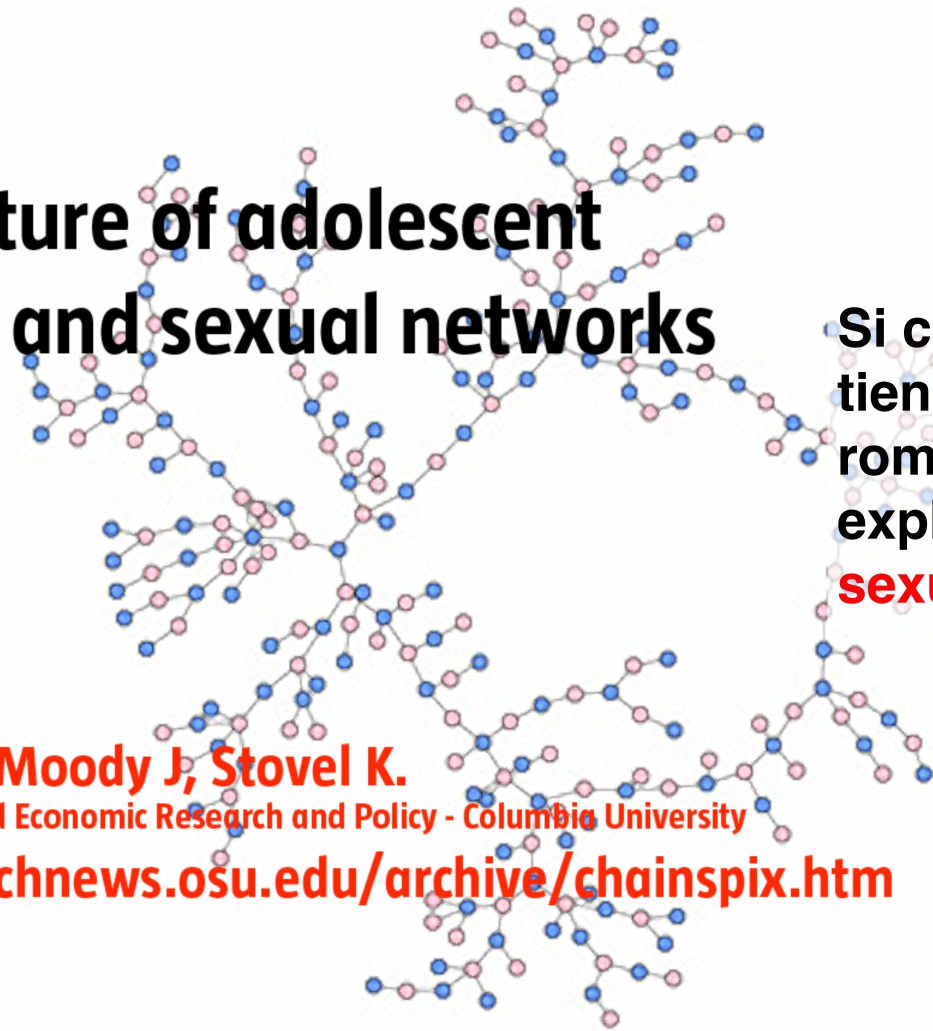
ELIGIENDO UNA REPRESENTACIÓN ADECUADA

They Rule

Si conectas personas que trabajan entre sí, explorarás La red profesional.

Josh On (2004)
http://www.theyrule.net

ELIGIENDO UNA REPRESENTACIÓN ADECUADA



The structure of adolescent romantic and sexual networks

Si conectas a aquellos que tienen una relación romántica y sexual, estarás explorando las **redes sexuales**.

Bearman PS, Moody J, Stovel K.

Institute for Social and Economic Research and Policy - Columbia University

http://researchnews.osu.edu/archive/chainspix.htm

ELIGIENDO UNA REPRESENTACIÓN ADECUADA

Si conectas a personas en función de su primer nombre (todos los Juanes se conectan entre sí), ¿qué explorarás?

Sin embargo, es una red.

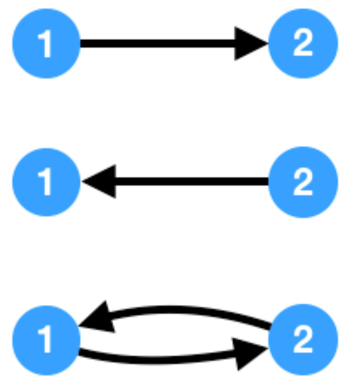
Representación de una red

- **Nodos, vertices, agentes, actores:** unidad de la red.
- **Aristas, enlaces, lazos:** conexiones entre nodos.
  - Los enlaces pueden tener **dirección**: la relación o interés va de un nodo a otro
    - Una persona recibe instrucciones o mentoría de otra persona con más experiencia
    - Seguidores en redes sociales
  - Los enlaces pueden tener **intensidad** (peso):
    - El numero de colaboraciones entre dos autores
    - Cuanto PIB es invertido de un país a otro

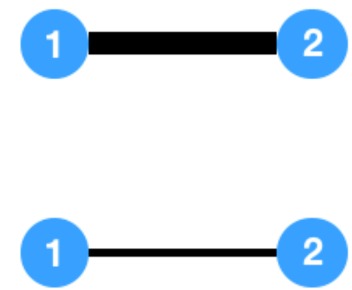
Redes dirigidas vs no dirigidas

Enlaces/Bordes

No-dirigida

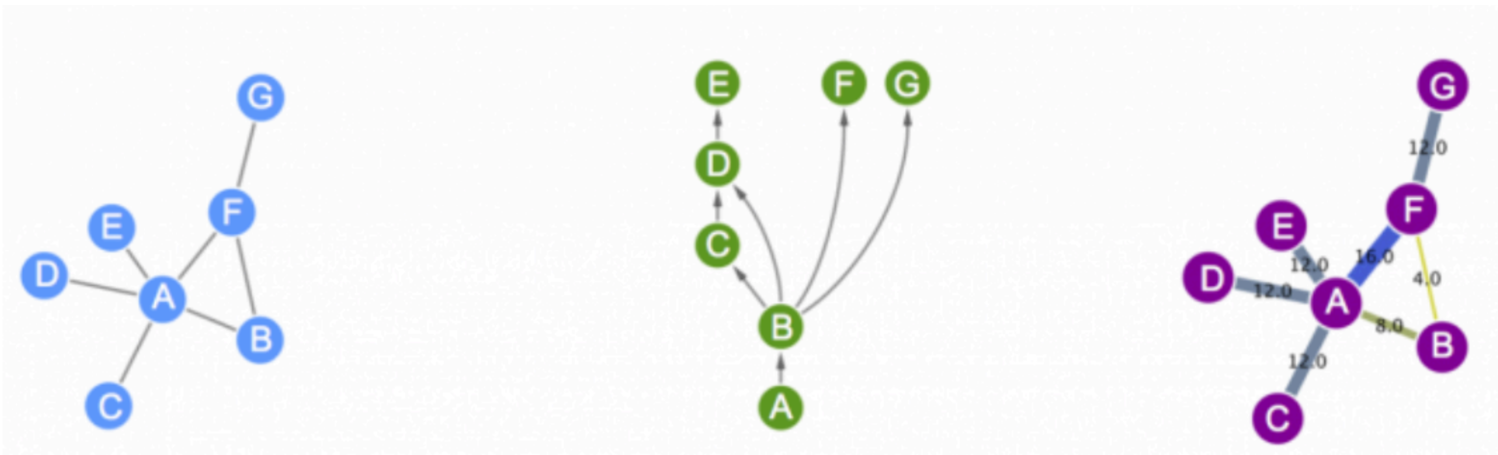


Dirigida



Con pesos

Tipos de Redes



Friendships, Influence

Parenthood, Dependences

Similarity, Financial Ties

No-dirigida

Dirigida

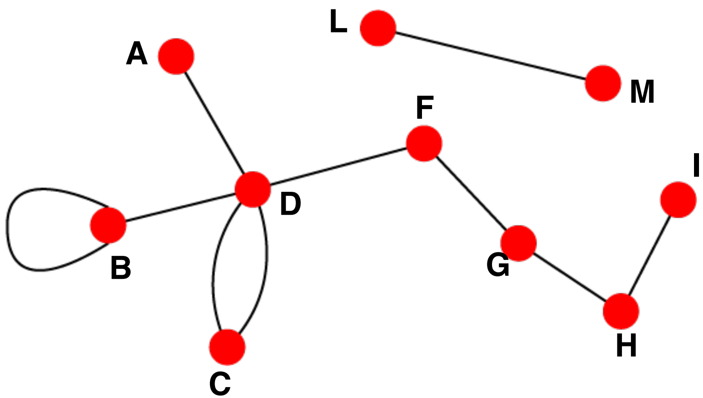
Con pesos

REDES DIRIGIDAS VS REDES NO-DIRIGIDAS

No-dirigidas

Links: no-dirigidos (*simétricos*)

Grafo:

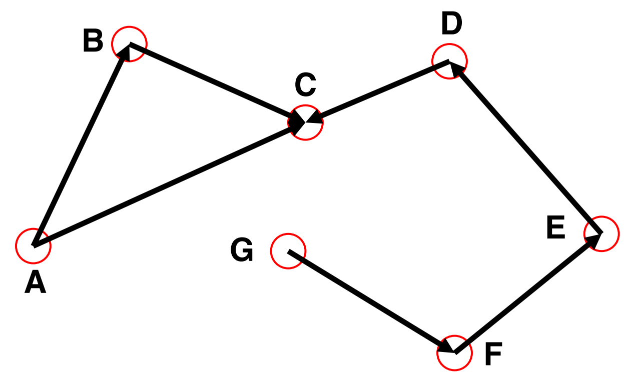


Links no-dirigidos:
Links de co-autores
Red de actores
Interacción de proteínas

Dirigidas

Links: dirigidos (*arcos*).

Digrafo = grafo dirigido:



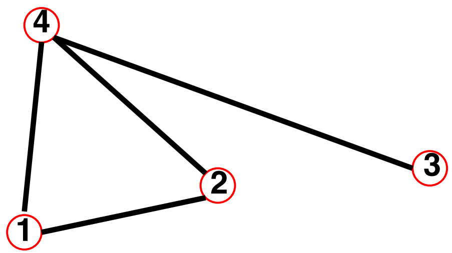
Un enlace no dirigido es la superposición de dos enlaces dirigidos opuestos.

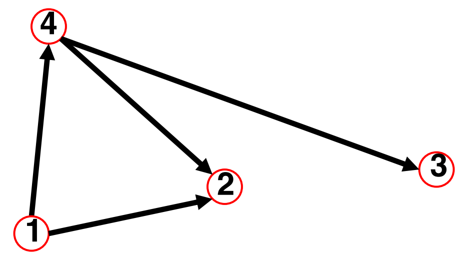
Links dirigidos:
URLs in el WWW
Llamdas de teléfono
Reacciones metabólicas

NETWORK	NODES	LINKS	DIRECTED UNDIRECTED	N	L
Internet	Routers	Internet connections	Undirected	192,244	609,066
WWW	Webpages	Links	Directed	325,729	1,497,134
Power Grid	Power plants, transformers	Cables	Undirected	4,941	6,594
Mobile Phone Calls	Subscribers	Calls	Directed	36,595	91,826
Email	Email addresses	Emails	Directed	57,194	103,731
Science Collaboration	Scientists	Co-authorship	Undirected	23,133	93,439
Actor Network	Actors	Co-acting	Undirected	702,388	29,397,908
Citation Network	Paper	Citations	Directed	449,673	4,689,479
E. Coli Metabolism	Metabolites	Chemical reactions	Directed	1,039	5,802
Protein Interactions	Proteins	Binding interactions	Undirected	2,018	2,930

Matriz de adyacencia

MATRIZ DE ADYACENCIA





$\mathbf{A_{ij}=1}$ si hay un link entre el nodo *i* y *j*

$\mathbf{A_{ij}=0}$ si los nodos *i* y *j* no están conectados entre si

$$A_{ij} = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix} \qquad A_{ij} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

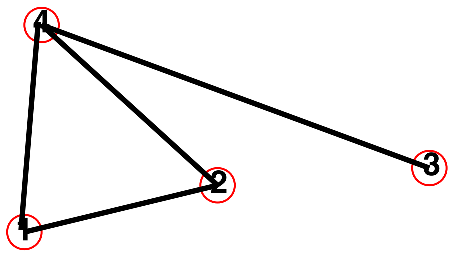
Tenga en cuenta que para un grafo dirigido (derecha) la matriz no es simétrica.

$A_{ij} = 1$ Si hay un link apuntando desde el nodo *j* al *i*

$A_{ij} = 0$ Si hay un link apuntando desde el nodo *i* al *j*

MATRIZ DE ADYACENCIA

No-dirigido



$$A_{ij} = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$

$$A_{ij} = A_{ji}$$

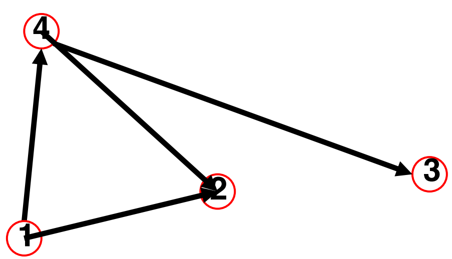
$$A_{ii} = 0$$

$$k_i = \sum_{j=1}^{N} A_{ij}$$

$$k_j = \sum_{i=1}^{N} A_{ij}$$

$$L = \frac{1}{2}\sum_{i=1}^{N} k_i = \frac{1}{2}\sum_{ij}^{N} A_{ij}$$

Dirigido



$$A_{ij} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

$$A_{ij} \neq A_{ji}$$

$$A_{ii} = 0$$

$$k_i^{in} = \sum_{j=1}^{N} A_{ij}$$

$$k_j^{out} = \sum_{i=1}^{N} A_{ij}$$

$$L = \sum_{i=1}^{N} k_i^{in} = \sum_{j=1}^{N} k_j^{out} = \sum_{i,j}^{N} A_{ij}$$

MATRIZ DE ADYACENCIA

N+1 elements

$$r = \begin{pmatrix} 1 \\ k_1 + 1 \\ k_2 + k_1 + 1 \\ \cdots \\ 1 + \sum_{i=1}^{j} k_i \\ \cdots \\ 2L + 1 \end{pmatrix}$$

2L elements

$$j = \begin{pmatrix} \cdots \\ \cdots \\ \cdots \\ \cdots \\ \cdots \\ \cdots \\ \cdots \end{pmatrix} \begin{array}{l} \text{index} \\ r(1) \\ r(2)-1 \\ r(2) \\ r(3)-1 \\ r(j) \\ r(j+1)-1 \end{array}$$

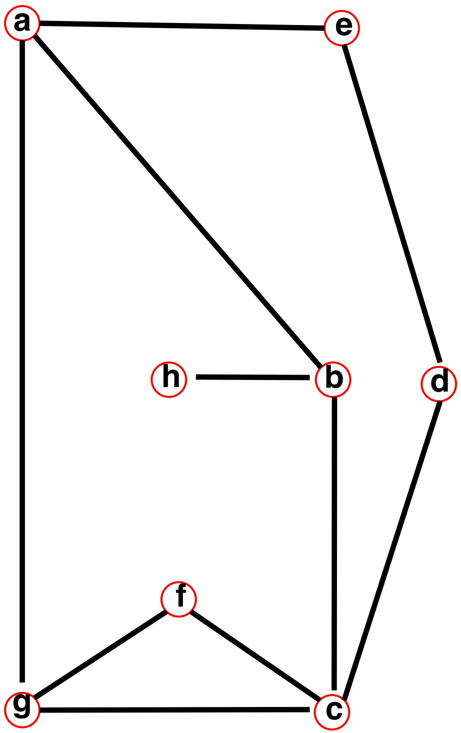
k_1 neighbors of node 1

k_2 neighbors of node 2

k_j neighbors of node j

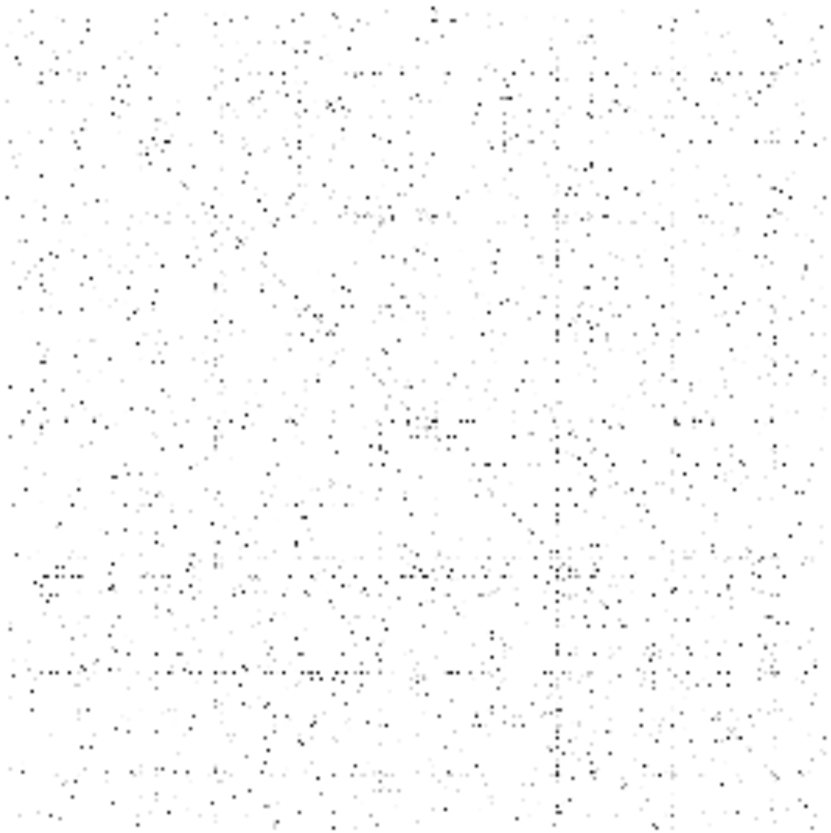
If directed, two representations: one for outgoing neighbors, one for incoming neighbors

MATRIZ DE ADYACENCIA

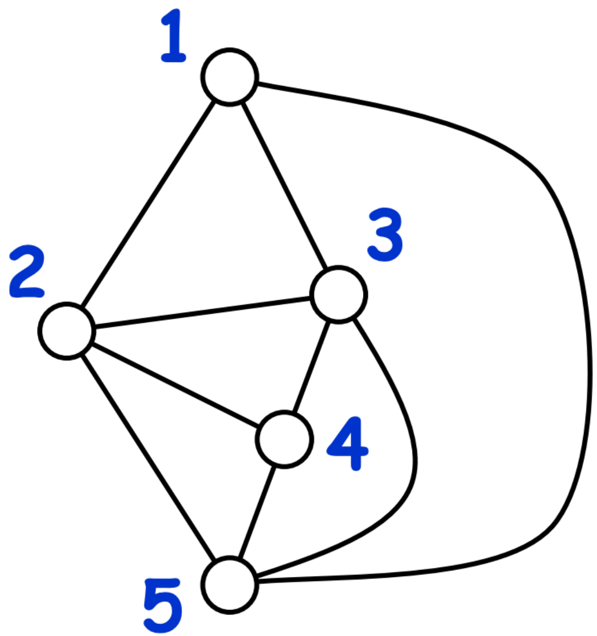


	a	b	c	d	e	f	g	h
a	0	1	0	0	1	0	1	0
b	1	0	1	0	0	0	0	1
c	0	1	0	1	0	1	1	0
d	0	0	1	0	1	0	0	0
e	1	0	0	1	0	0	0	0
f	0	0	1	0	0	0	1	0
g	1	0	1	0	0	0	0	0
h	0	1	0	0	0	0	0	0

LAS MARICES DE ADYACENCIA SON "SPARSE"



Ejercicio



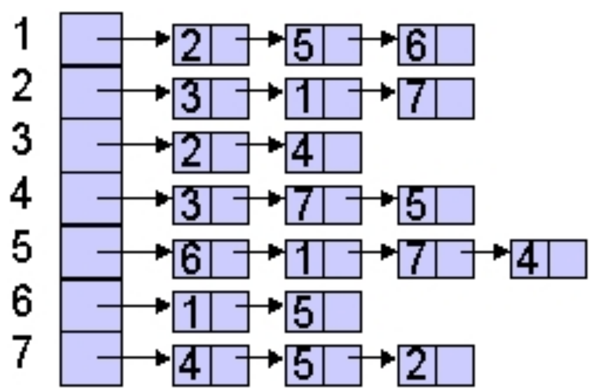
Representa el grafo como una matriz de adyacencia

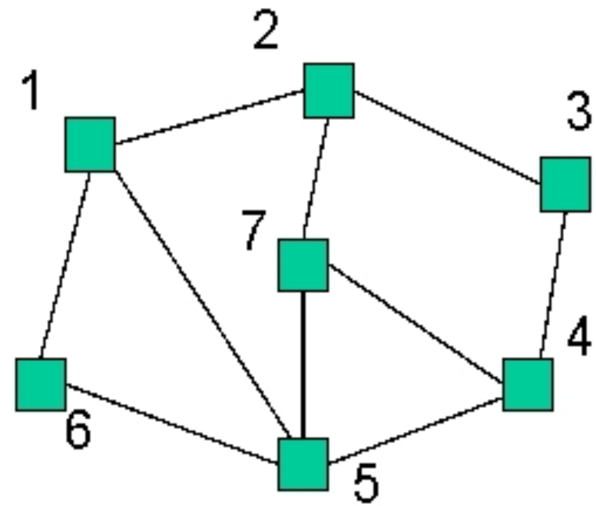
Lista de enlaces y lista de adyacencia

- List of edges

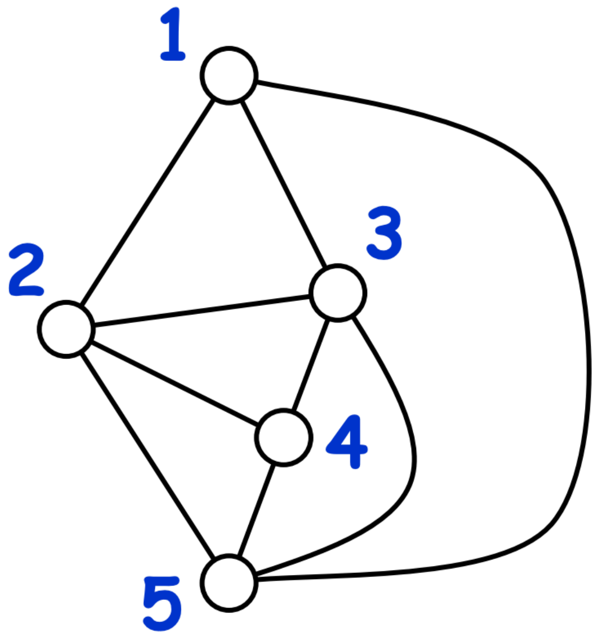
1	5	1	2	2	3	5	7	5	5
2	1	6	7	3	4	6	4	7	4

- Adjacency lists





Ejercicio



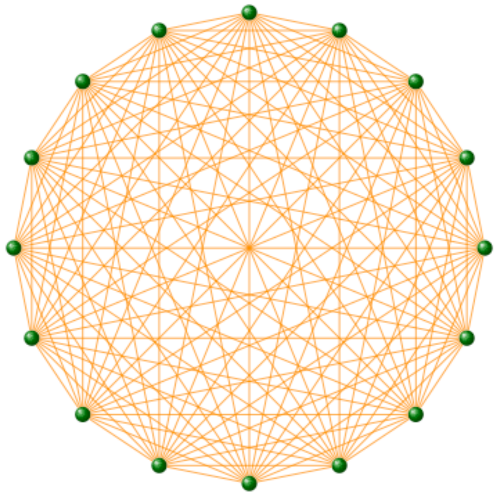
Representa el grafo como una lista de adyacencia
Representa el grafo como una lista de enlaces

Redes reales son poco densas (sparse)

GRAFO COMPLETO

El número máximo de enlaces que puede tener una red de N nodos es :

$$L_{max} = \binom{N}{2} = \frac{N(N-1)}{2}$$



Un grafo con grado $L=L_{max}$ es llamado **grafo completo**, y su grado promedio es **<k>=N-1**

$$C_n^p = \binom{n}{p} = \frac{n!}{p!(n-p)!} .$$

REDES REALES SON ¨SPARSE¨

La mayoría de las redes observadas en sistemas reales son poco densas:

$$L \ll L_{max}$$

o

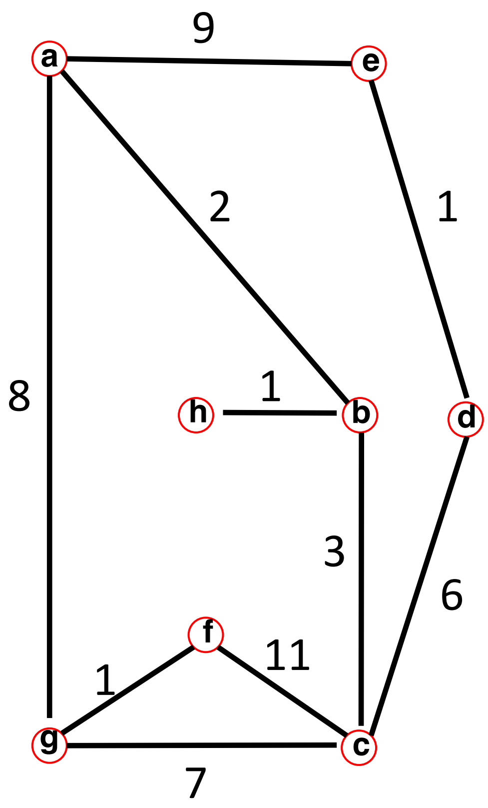
$$\langle k \rangle \ll N-1.$$

WWW (ND Sample):	$N=325{,}729$;	$L=1.4\ 10^{6}$	$L_{max}=10^{12}$	$\langle k \rangle=4.51$
Protein (*S. Cerevisiae*):	$N=\ \ 1{,}870$;	$L=4{,}470$	$L_{max}=10^{7}$	$\langle k \rangle=2.39$
Coauthorship (Math):	$N=\ 70{,}975$;	$L=2\ 10^{5}$	$L_{max}=3\ 10^{10}$	$\langle k \rangle=3.9$
Movie Actors:	$N=212{,}250$;	$L=6\ 10^{6}$	$L_{max}=1.8\ 10^{13}$	$\langle k \rangle=28.78$

(Source: Albert, Barabasi, RMP2002)

REDES CON Y SIN PESOS

MATRIZ DE ADYACENCIA CON PESOS

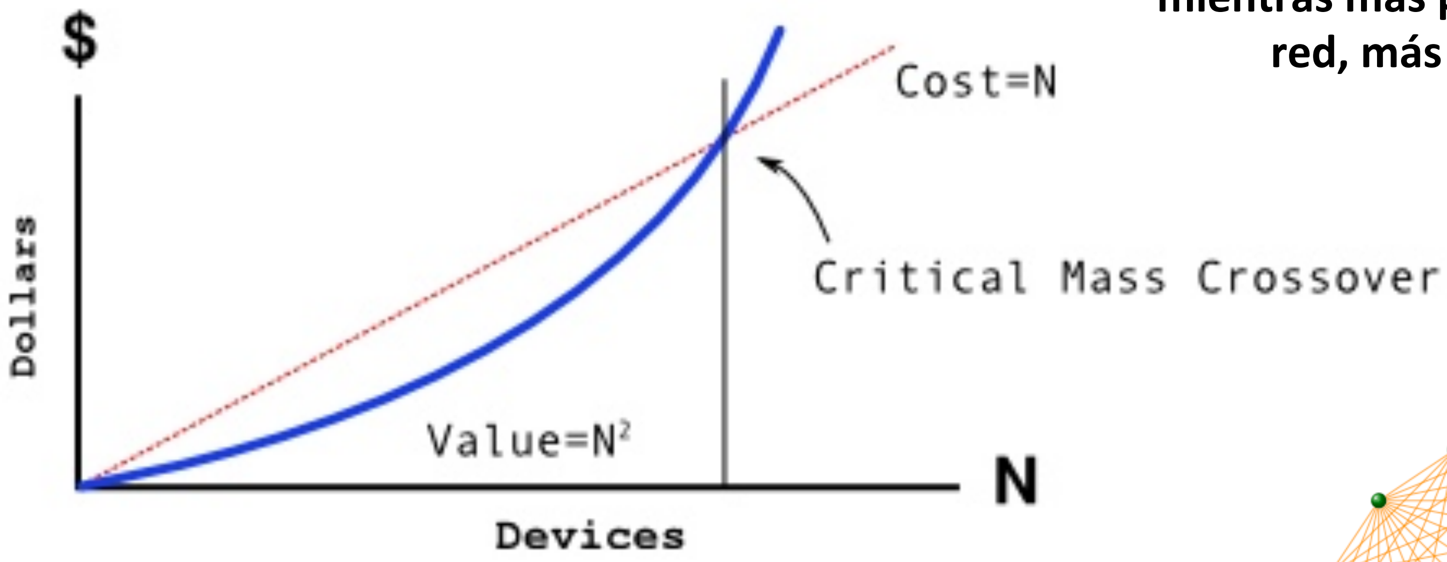


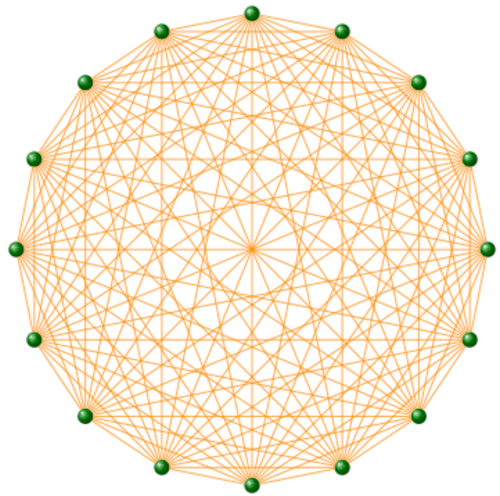
$$A_{ij} = w_{ij}$$

	a	b	c	d	e	f	g	h
a	0	2	0	0	9	0	8	0
b	2	0	3	0	0	0	0	1
c	0	3	0	6	0	11	7	0
d	0	0	6	0	1	0	0	0
e	9	0	0	1	0	0	0	0
f	0	0	11	0	0	0	1	0
g	8	0	7	0	0	1	0	0
h	0	1	0	0	0	0	0	0

METCALFE'S LAW

"mientras más personas usan una red, más valiosa se vuelve"

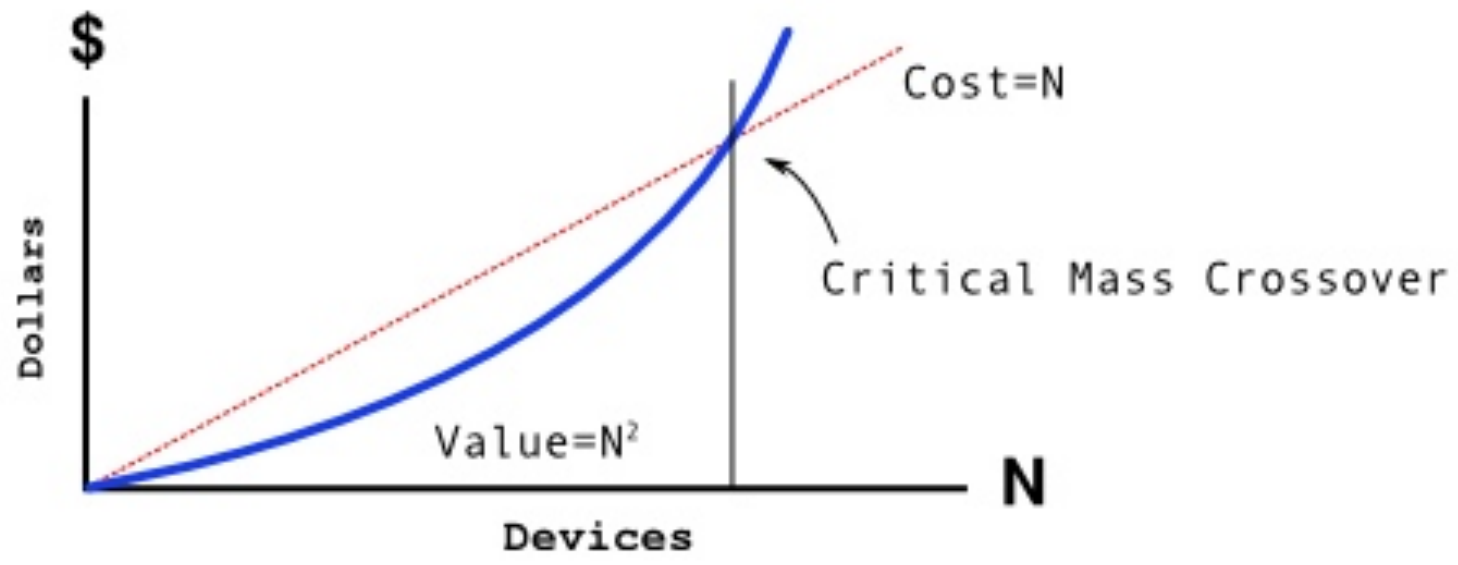




El número de enlaces máximos que una red de N nodos puede tener es:

$$L_{\max} = \binom{N}{2} = \frac{N(N-1)}{2}$$

METCALFE'S LAW

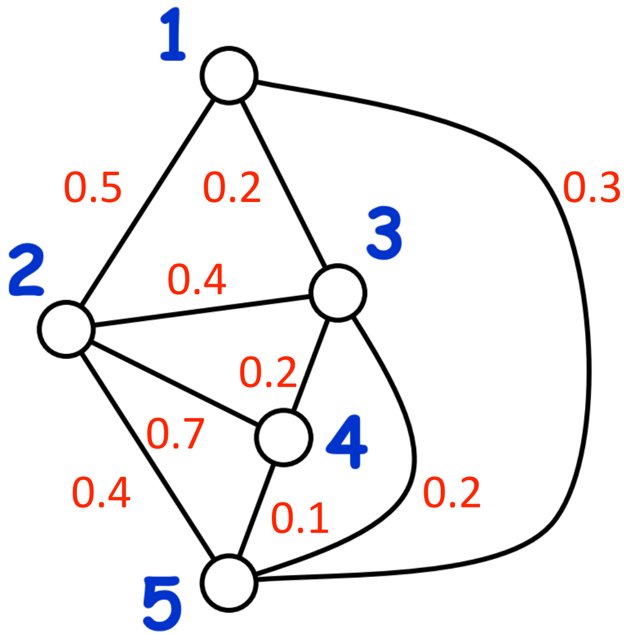


Hay dos problemas fundamentales con la ley de Metcalfe:

Si bien todos los enlaces son posibles, en redes reales no todos los enlaces están presentes. De hecho, la mayoría de las redes reales son dispersas, lo que significa que solo una pequeña fracción de los enlaces están presentes. Si asignamos un valor a cada enlace, entonces el valor total de la red crecerá más lentamente que N ^ 2, como veremos en los próximos capítulos.

No todos los enlaces son de igual valor (pesos). Algunos enlaces se usan mucho mientras que la mayoría de los enlaces son 'débiles', es decir, rara vez se utilizan.

Ejercicio



Representa el grafo como una lista de adyacencia.
El peso del enlace está indicado en rojo

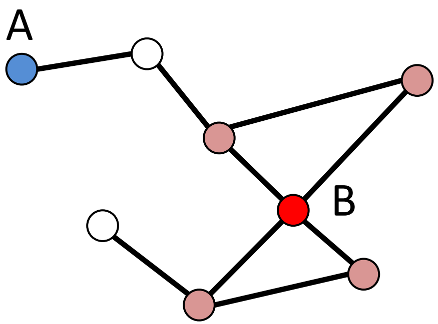
Preguntas básicas

- Qué se puede investigar a partir de la dirección y peso de los enlaces?
- Qué significa que un enlace sea *no* dirigido?
- Qué implica que un enlace tenga peso?
- Cuál es la relevancia de los enlaces propios (self-loops)?
- Como piensas que se debe estructurar la data en la práctica para construir una red?

Grado, Grado promedio y Distribución de grado

GRADOS DE NODOS

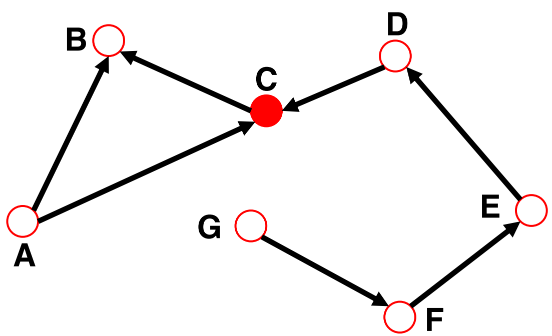
No-dirigido



Grado del nodo: el número de links conectados al nodo.

$$k_A = 1 \qquad k_B = 4$$

Dirigido



En ***redes dirigidas*** Podemos definer un grado de entrada (in-degree) y un grado de salida (out-degree). El grado total es la suma de ambos.

$$k_C^{in} = 2 \qquad k_C^{out} = 1 \qquad k_C = 3$$

Fuente: un nodo con $k^{in} = 0$; Sumidero: un nodo con $k^{out} = 0$.

UN POCO DE ESTADISTICAS

Breve revisión estadística

Four key quantities characterize a sample of N values $x_1, \ldots, x_N$:

Promedio:

$$\langle x \rangle = \frac{x_1 + x_2 + \ldots + x_N}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

El n-esimo momento:

$$\langle x^n \rangle = \frac{x_1^n + x_2^n + \ldots + x_N^n}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i^n$$

Desviación estándar

$$\sigma_x = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \langle x \rangle\right)^2}$$

Distribución de *x*

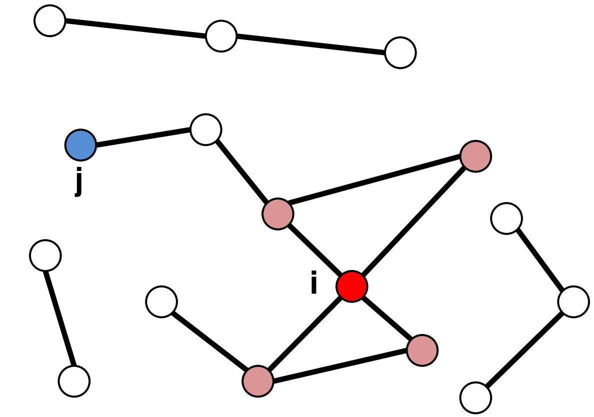
$$p_x = \frac{1}{N}\sum_i \delta_{x,x_i}$$

Donde *x* sigue:

$$\sum_i p_x = 1 \left(\int p_x \, dx = 1\right)$$

GRADO PROMEDIO

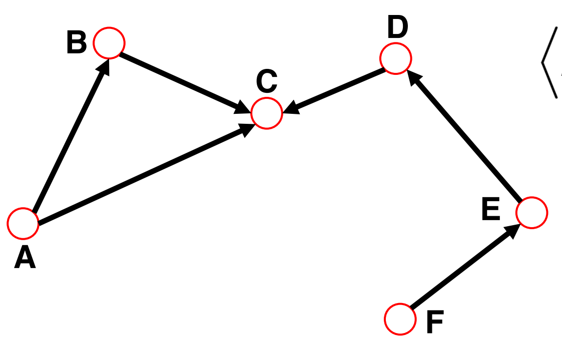
No-dirigido



$$\langle k \rangle \equiv \frac{1}{N}\sum_{i=1}^{N} k_i \qquad \langle k \rangle \equiv \frac{2L}{N}$$

N – el número de nodos en el grafo

Dirigido



$$\langle k^{in} \rangle \equiv \frac{1}{N}\sum_{i=1}^{N} k_i^{in}, \quad \langle k^{out} \rangle \equiv \frac{1}{N}\sum_{i=1}^{N} k_i^{out}, \quad \langle k^{in} \rangle = \langle k^{out} \rangle$$

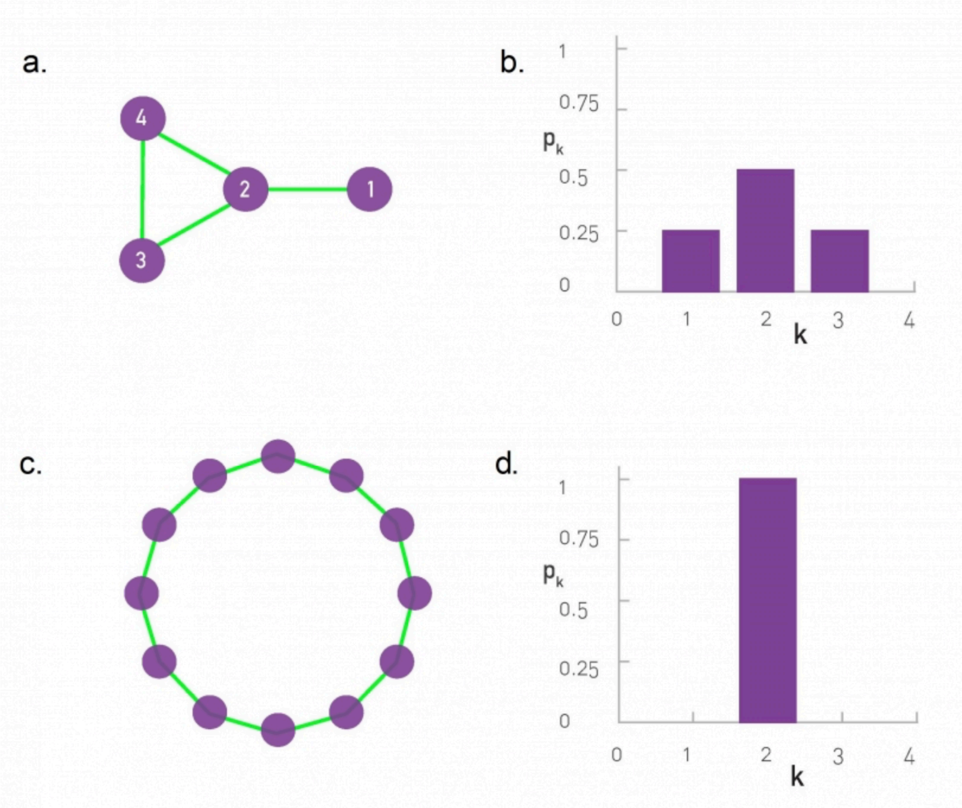
$$\langle k \rangle \equiv \frac{L}{N}$$

GRADO PROMEDIO

NETWORK	NODES	LINKS	DIRECTED UNDIRECTED	N	L	⟨k⟩
Internet	Routers	Internet connections	Undirected	192,244	609,066	6.33
WWW	Webpages	Links	Directed	325,729	1,497,134	4.60
Power Grid	Power plants, transformers	Cables	Undirected	4,941	6,594	2.67
Mobile Phone Calls	Subscribers	Calls	Directed	36,595	91,826	2.51
Email	Email addresses	Emails	Directed	57,194	103,731	1.81
Science Collaboration	Scientists	Co-authorship	Undirected	23,133	93,439	8.08
Actor Network	Actors	Co-acting	Undirected	702,388	29,397,908	83.71
Citation Network	Paper	Citations	Directed	449,673	4,689,479	10.43
E. Coli Metabolism	Metabolites	Chemical reactions	Directed	1,039	5,802	5.58
Protein Interactions	Proteins	Binding interactions	Undirected	2,018	2,930	2.90

Distribución de grado

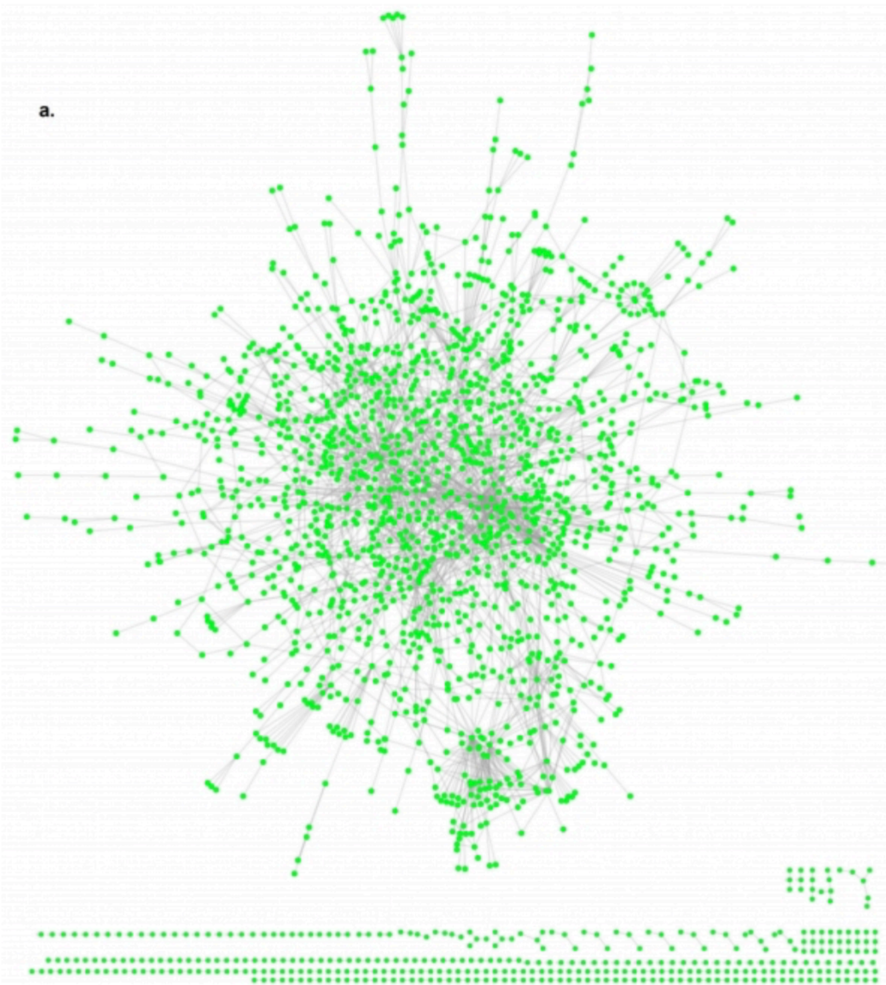
P(k): probabilidad de que un nodo elegído aleatoriamente tenga grado *k*



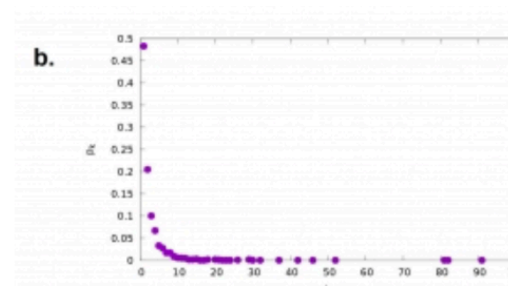
N_k = # nodos con grado k

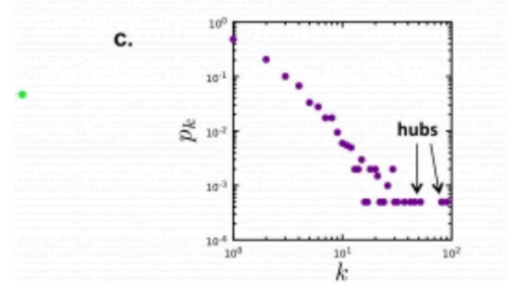
$P(k) = N_k / N$ ➔ plot

DISTRIBUCION DE GRADO



En muchas redes reales, el grado de nodo puede variar considerablemente.
Por ejemplo, como indica la distribución de grados (a), los grados de las proteínas en la red de interacción de proteínas que se muestran en (b) varían entre k = 0 (nodos aislados) y k = 92, que es el grado del nodo más grande, llamado un centro. También hay grandes diferencias en el número de nodos con diferentes grados: como muestra (a), casi la mitad de los nodos tienen grado uno (es decir, p1 = 0,48), mientras que solo hay una copia del nodo más grande, por lo tanto, p_92 = 1 / N = 0.0005. (c) La distribución de grados a menudo se muestra en el llamado gráfico log-log, en el que trazamos log p_k en función de log k, o, como hicimos en (c), usamos ejes logarítmicos.





DISTRIBUCION DE GRADO

Representación discreta: $\mathbf{p_k}$ Es la probabilidad de que un nodo tenga grado **k**.

Descripción continua: **p(k)** es la pdf de los grados, donde

$$\int_{k_1}^{k_2} p(k)\,dk$$

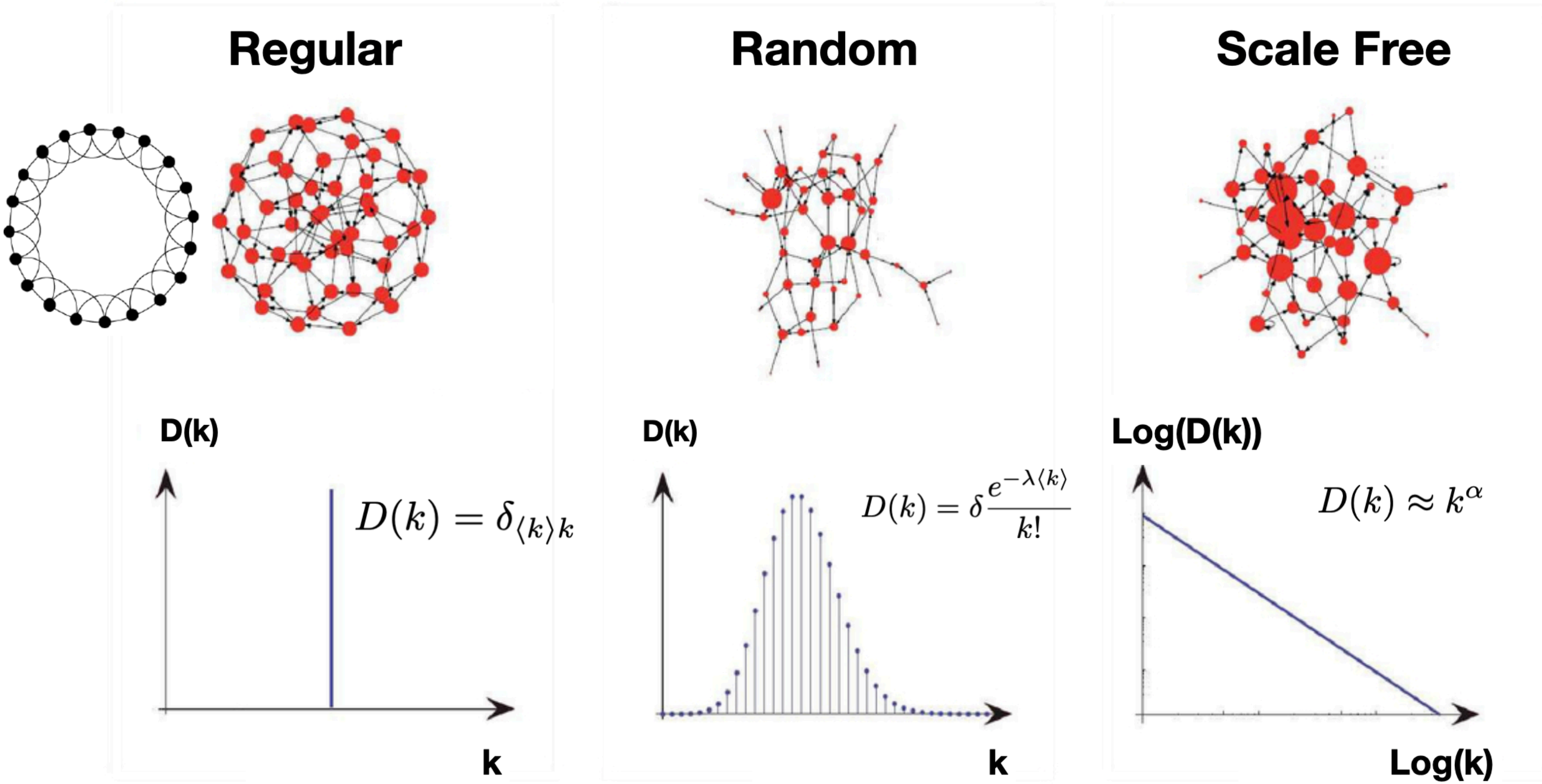
representa la probabilidad de que el grado de un nodo esté entre $\mathbf{k_1}$ y $\mathbf{k_2}$.

Condición de normalización:

$$\sum_{0}^{\infty} p_k = 1 \qquad\qquad \int_{K_{\min}}^{\infty} p(k)\,dk = 1$$

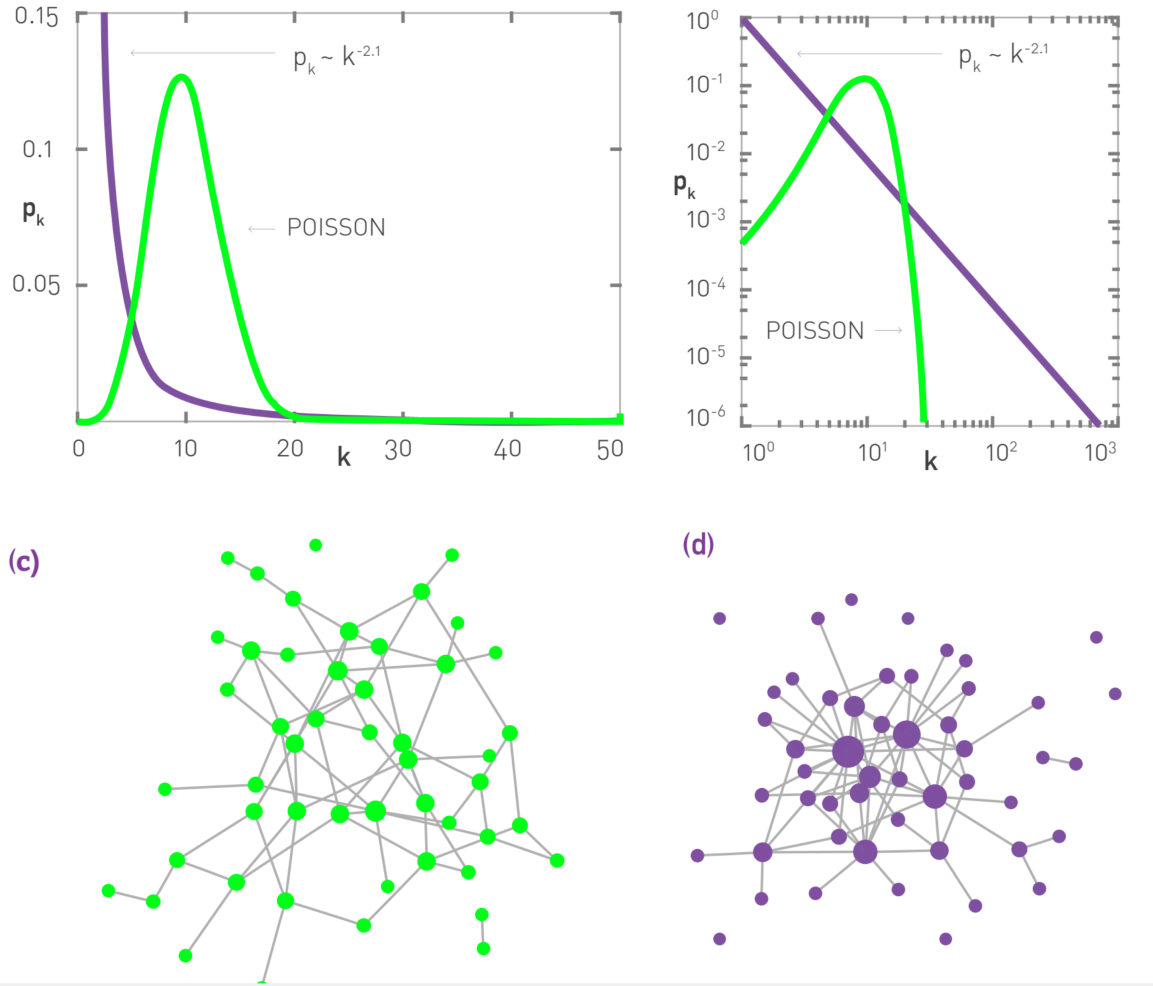
donde K_{min} es el mínimo grado de la red.

DISTRIBUCION DE GRADO



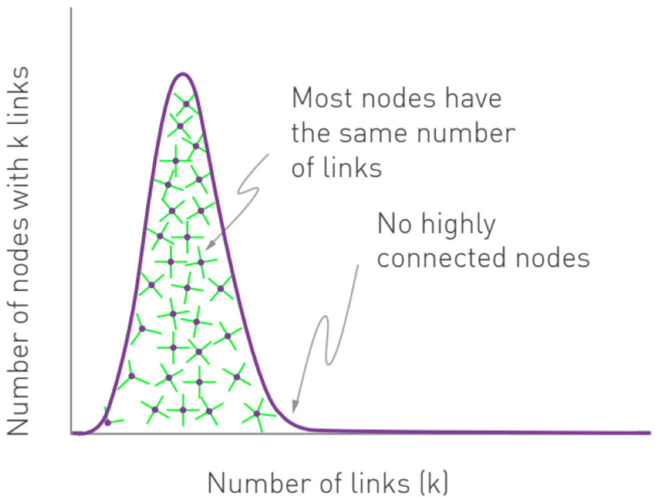
Piensen en los histogramas de frecuencias de grado
¿Cómo se comparan los promedios con los valores máximos?

Diferencia entre red aleatoria y libre de escala

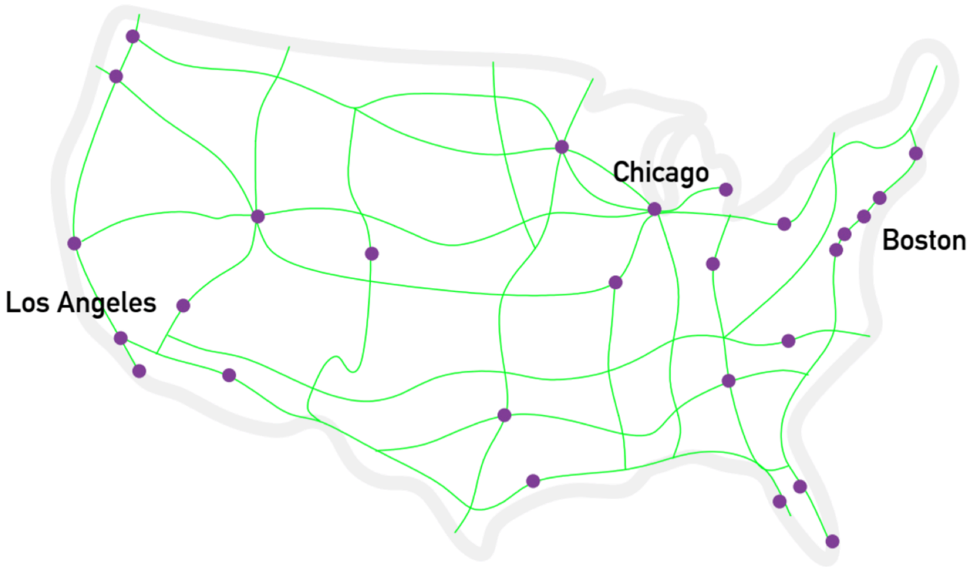


Diferencia entre red aleatoria y libre de escala

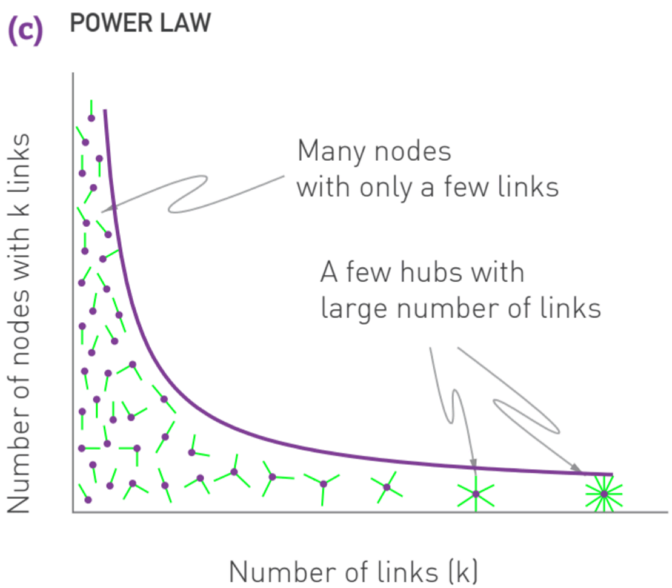
(a) POISSON

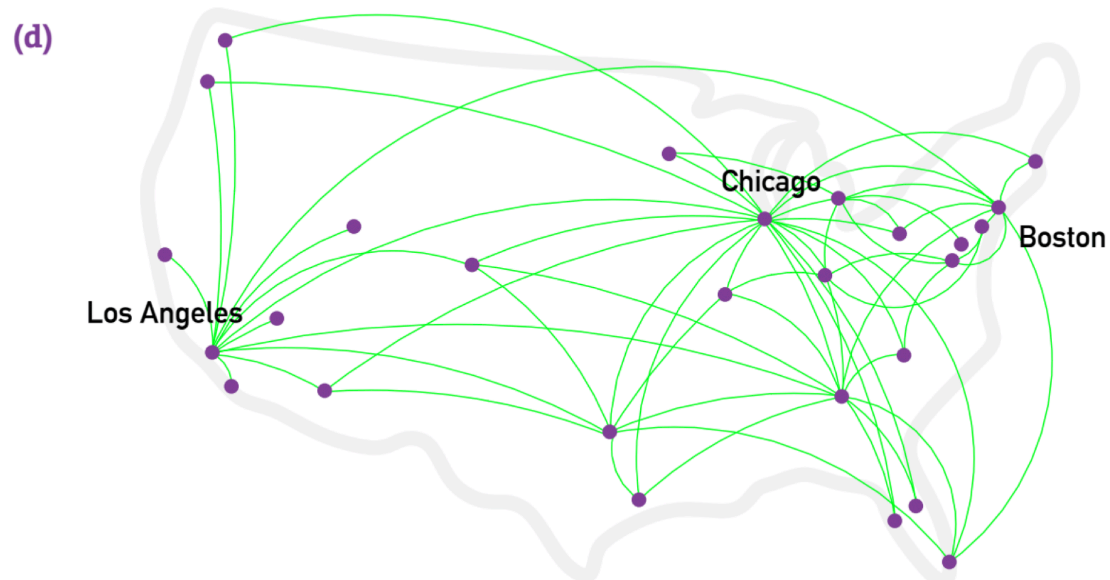


(b)



Diferencia entre red aleatoria y libre de escala





Estas estructuras tienen un impacto significativo en cómo se propaga la información en los sistemas físicos, biológicos, sociales, etc.

Preguntas básicas

- Qué indica el grado promedio de una red?
- Piensa en un ejemplo de la vida real donde se aplique Metcalfe's law
- Qué tipo de información *no* se puede inferir de la representación de la lista de enlaces de la red que se puede inferir de la matriz de adyacencia?
- Qué diferencias estructurales puedes observar de la visualización de una red aleatoria respecto de una de libre escala?