

PARTIMOS EN BREVE

MUCHAS GRACIAS

# Network Science

dataScience UDD



**Cristian Candia-Castro Vallejos, Ph.D.**

[cristiancandia@udd.cl](mailto:cristiancandia@udd.cl)

Director Magister en Data Science UDD  
Profesor Investigador, Facultad de Ingeniería, UDD  
External Faculty Northwestern Institute on Complex Systems, Kellogg School of  
Management, Northwestern University

# Los puentes de Konigsberg

# LOS PUENTES DE KONIGSBERG



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

1736

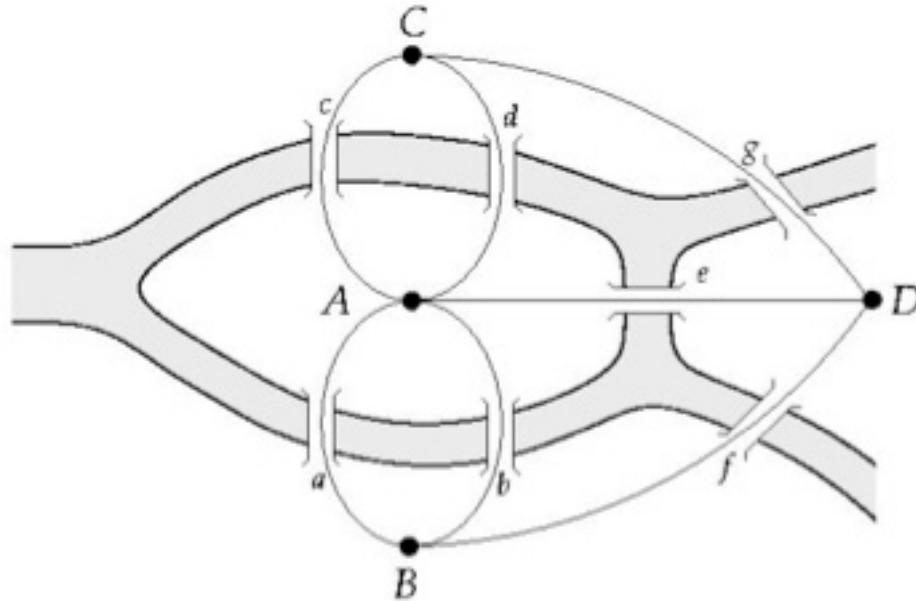
# LOS PUENTES DE KONIGSBERG



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

1736

# LOS PUENTES DE KONIGSBERG

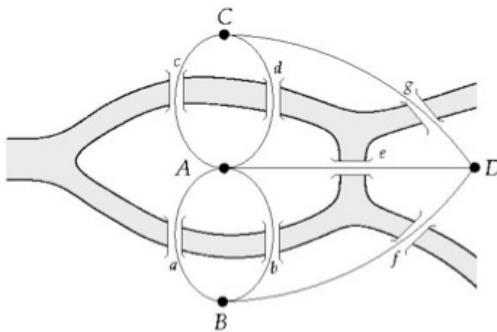


¿Se puede cruzar los siete puentes y nunca cruzar el mismo puente dos veces?

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?**

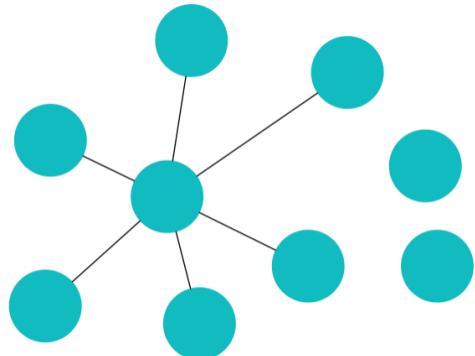
Hoy recordamos la prueba de Euler porque fue la primera vez que alguien resolvió un problema matemático convirtiéndolo en un grafo. En retrospectiva, la prueba tiene dos mensajes importantes:

- 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.

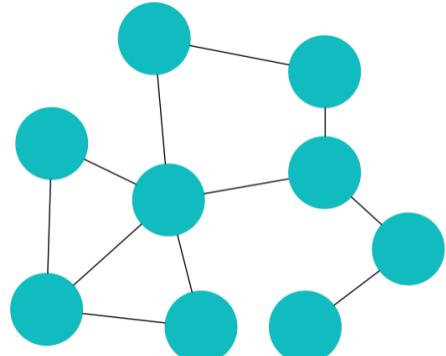
¿En cuál de estas redes es posible idear un paseo que atraviese cada enlace una vez y solo una vez?

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

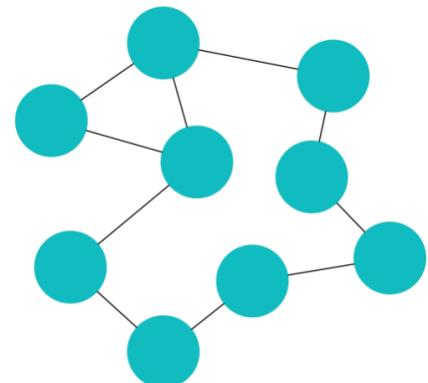
A



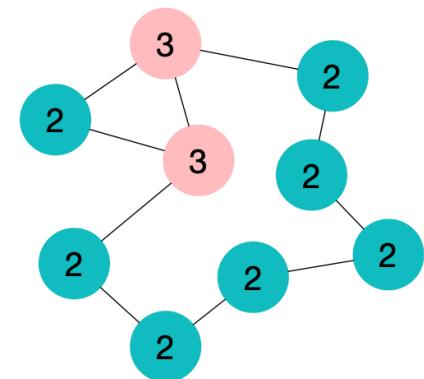
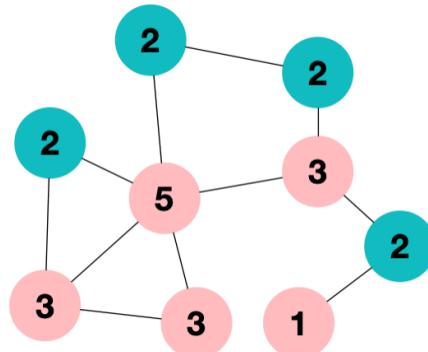
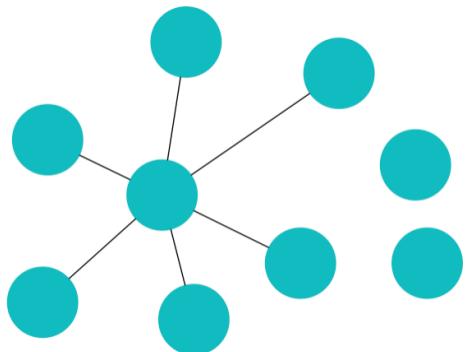
B



C



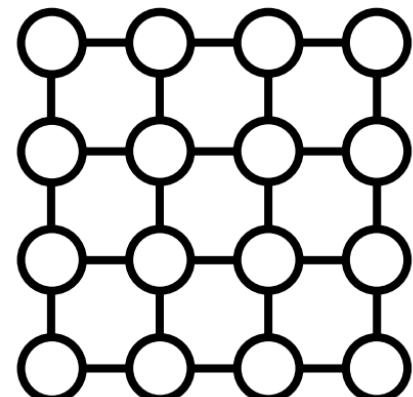
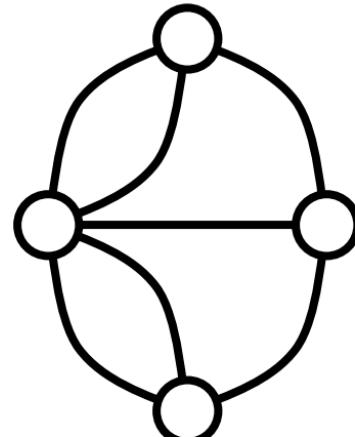
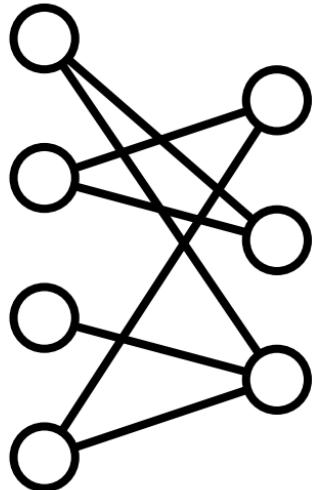
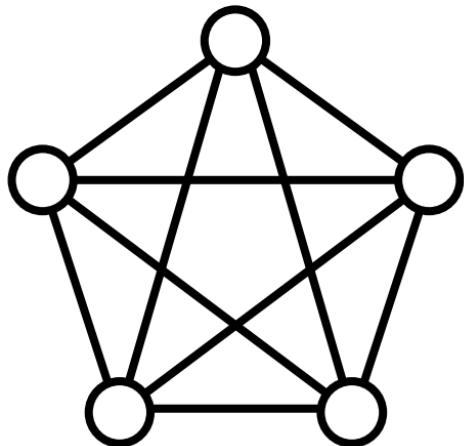
1. ¿Está conectado?
2. Cuente el número de nodos de grados impares, ¿es 0 o 2?



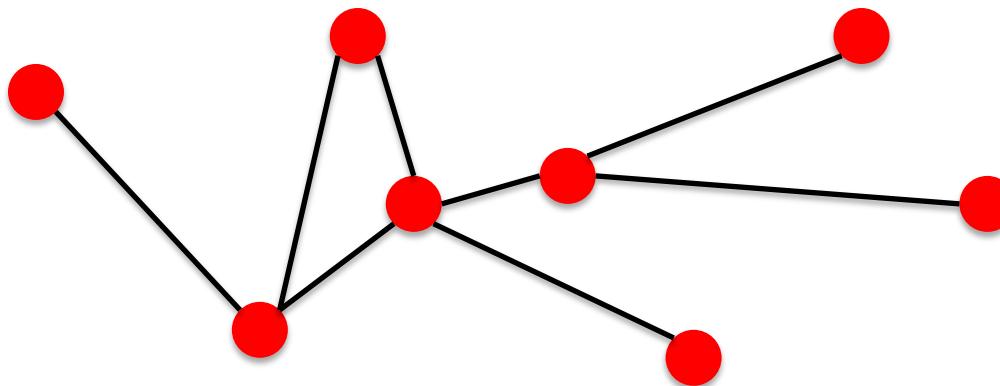
# Redes y grafos

# ¿Qué es un grafo?

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



# COMPONENTES DE UN SISTEMA COMPLEJO



■ **componentes:** nodos, vertices

N

■ **interacciones:** enlaces, bordes

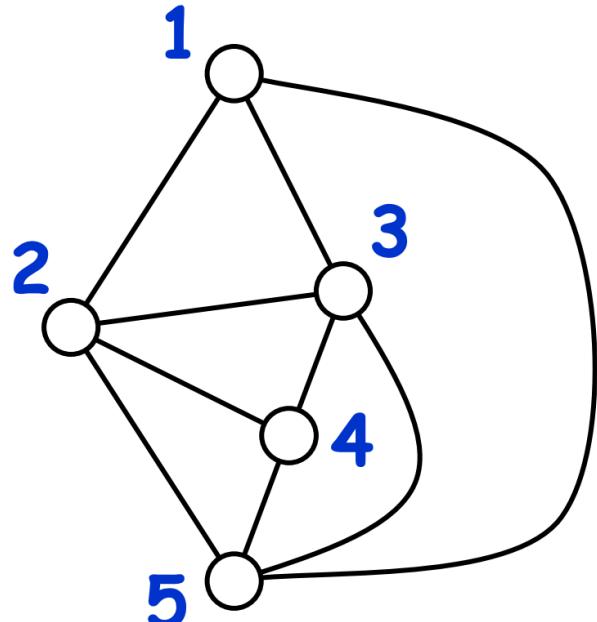
L

■ **sistema:** red, grafo

(N,L)

# Ejemplo

- $G(V, E)$ : graph (network)  
V: vertices (nodes), E: edges (links)



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))

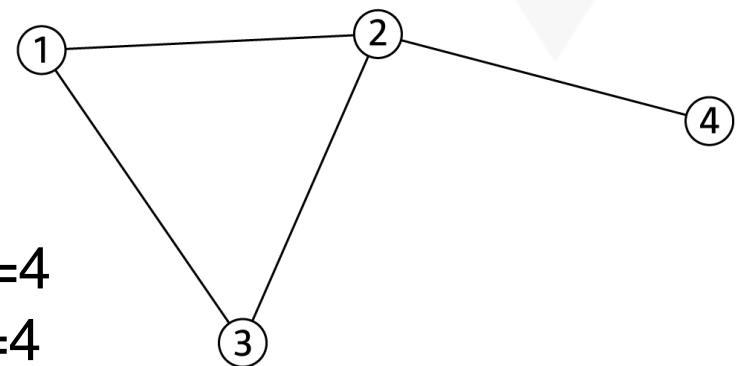
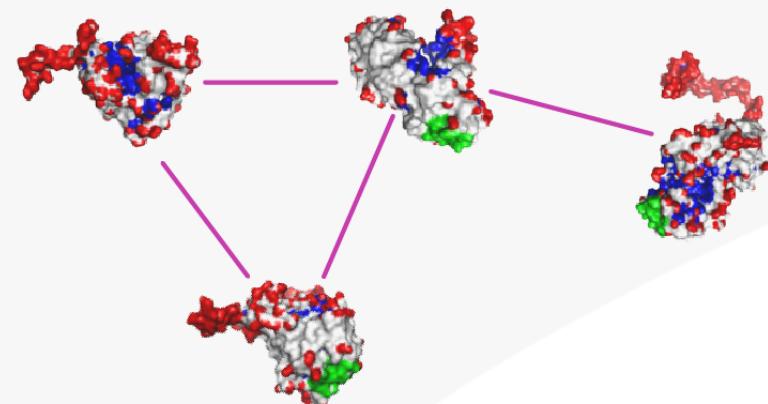
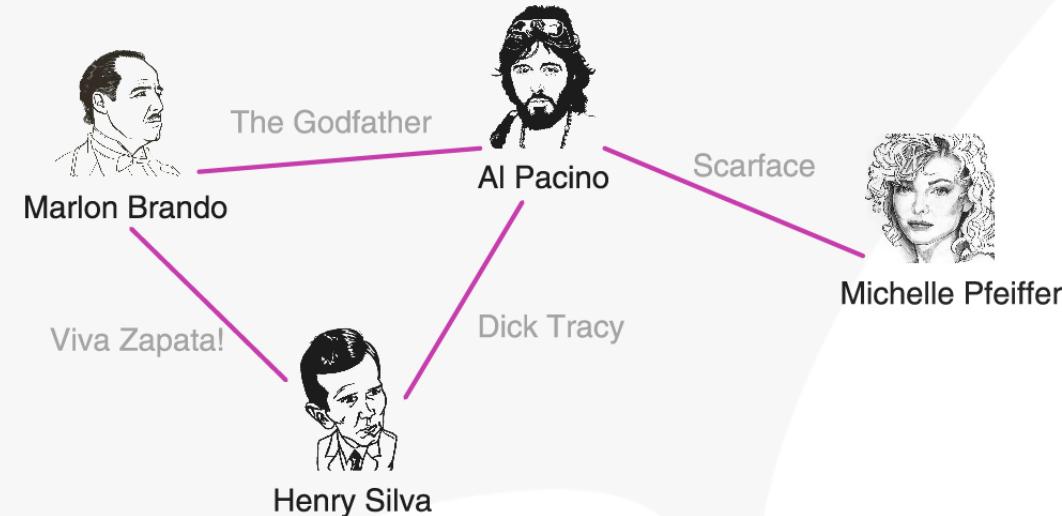
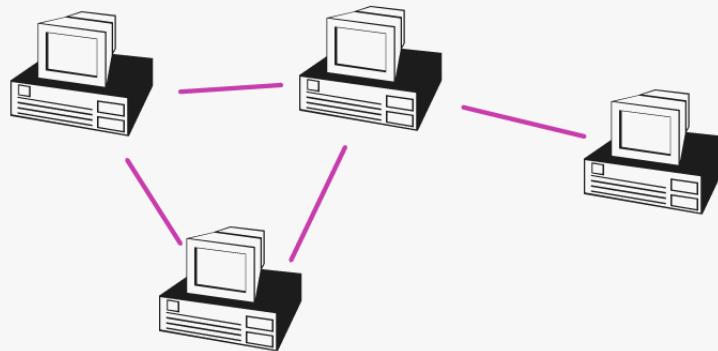
**grafo**: representación matemática de una red

- 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



## CHOOSING A PROPER REPRESENTATION

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

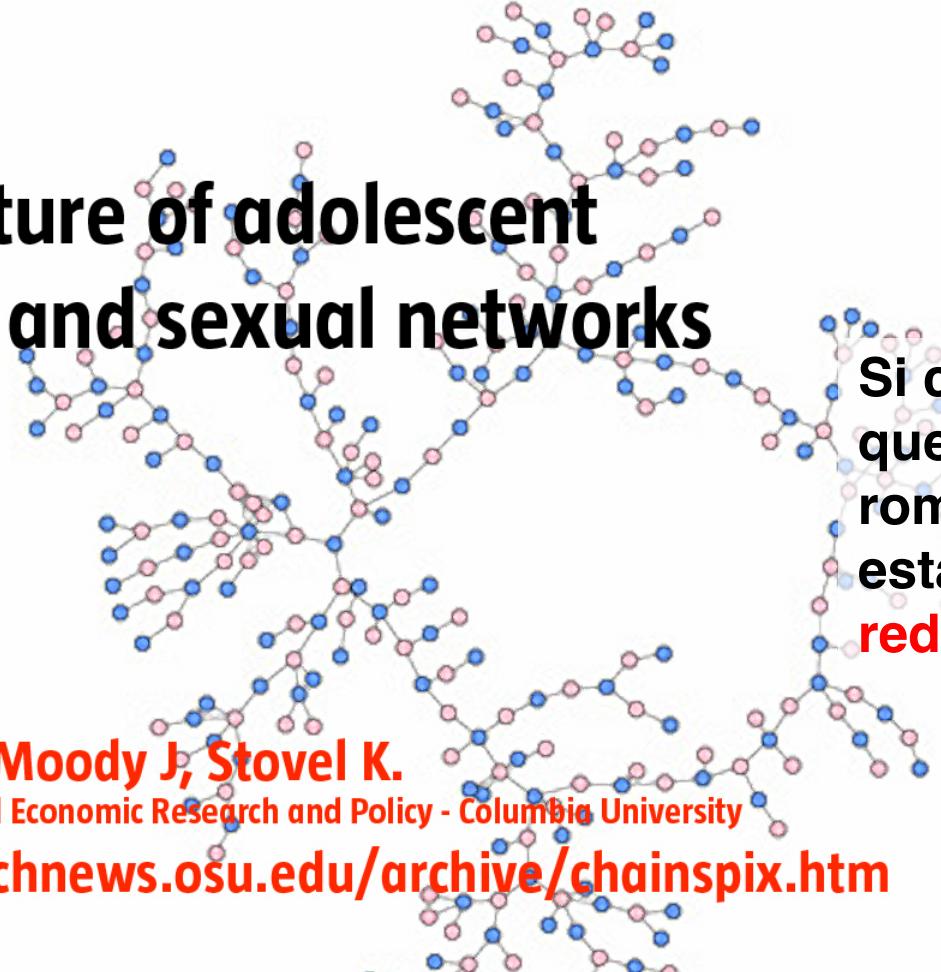
En algunos casos hay una representación única e inequívoca.  
En otros casos, la representación no es de ninguna manera  
única.

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



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

Sin embargo, es una red.

# Redes dirigidas vs no dirigidas

# Enlaces/Bordes



No-dirigida

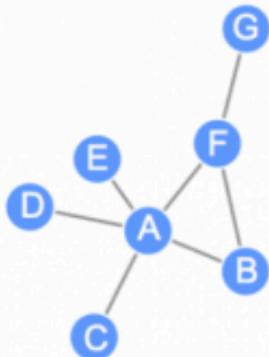


Dirigida



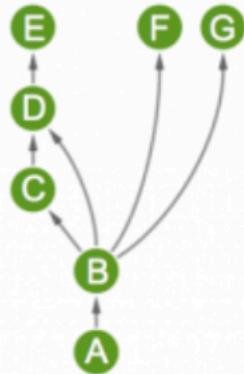
Con pesos

# Tipos de Redes



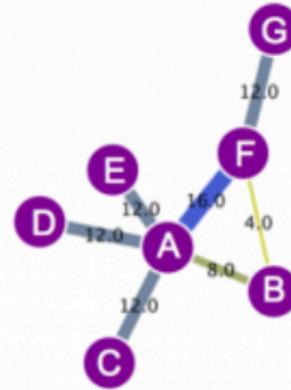
**Friendships, Influence**

**No-dirigida**



**Parenthood,  
Dependences**

**Dirigida**



**Similarity, Financial Ties**

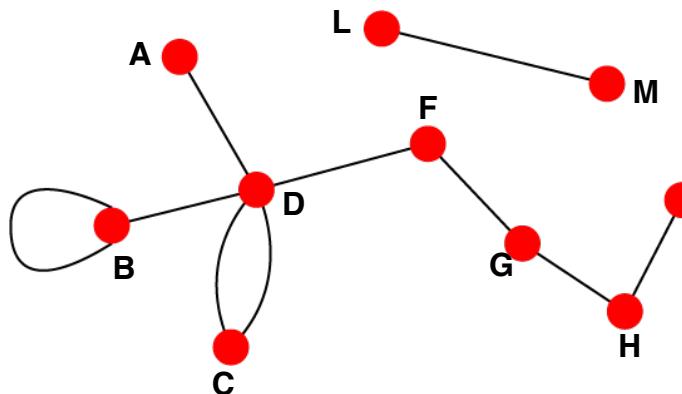
**Con pesos**

# REDES DIRIGIDAS VS REDES NO-DIRIGIDAS

## No-dirigidas

Links: no-dirigidos (*simétricos*)

Grafo:



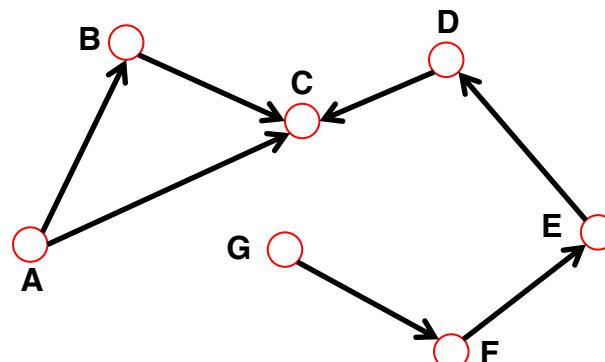
### Links no-dirigidos:

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  
Llamadas de teléfono  
Reacciones metabólicas

## Reference Networks

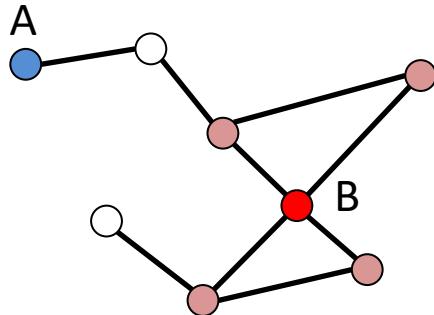
### Section 2.2

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

# Grado, Grado promedio y Distribución de grado

# GRADOS DE NODOS

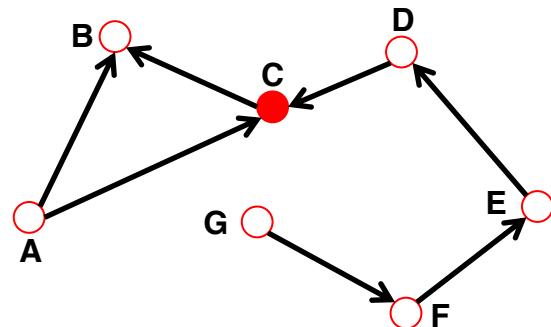
## No-dirigido



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

$$k_A = 1 \quad k_B = 4$$

## Dirigido



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

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

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  $x_1, \dots, x_N$ :

### Promedio:

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

### El n-esimo momento:

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

## Desviación estándar

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

## Distribución de $x$

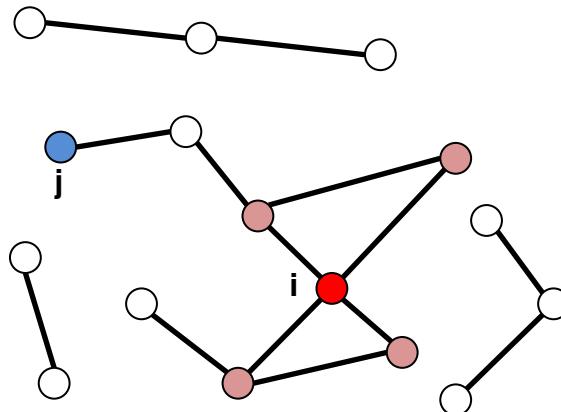
$$p_x = \frac{1}{N} \sum_i \delta_{x,x_i}$$

## Donde $x$ sigue:

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

# GRADO PROMEDIO

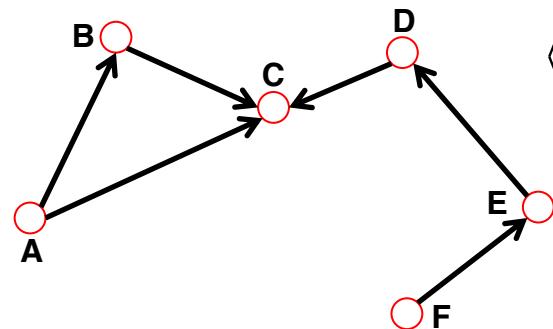
No-dirigido



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

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

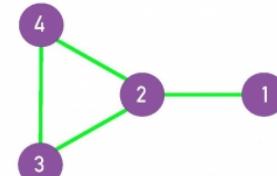
NETWORK	NODES	LINKS	DIRECTED UNDIRECTED	N	L	$\langle k \rangle$
Internet	Routers	Internet connections	Undirected	192,244	609,066	6.33
WWW	Webpages	Links	Directed	325,729	1,497,134	4.60
Power Grid	Power plants, transformers	Cables	Undirected	4,941	6,594	2.67
Mobile Phone Calls	Subscribers	Calls	Directed	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

# DISTRIBUCION DE GRADO

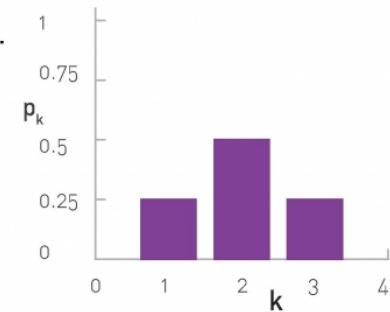
## Distribución de grado

$P(k)$ : probabilidad de que un nodo elegido aleatoriamente tenga grado  $k$

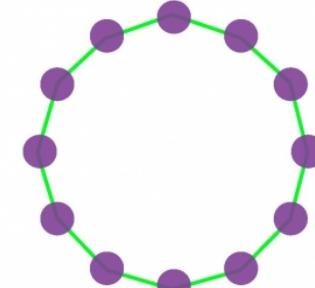
a.



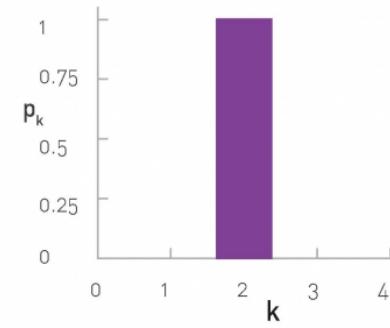
b.



c.



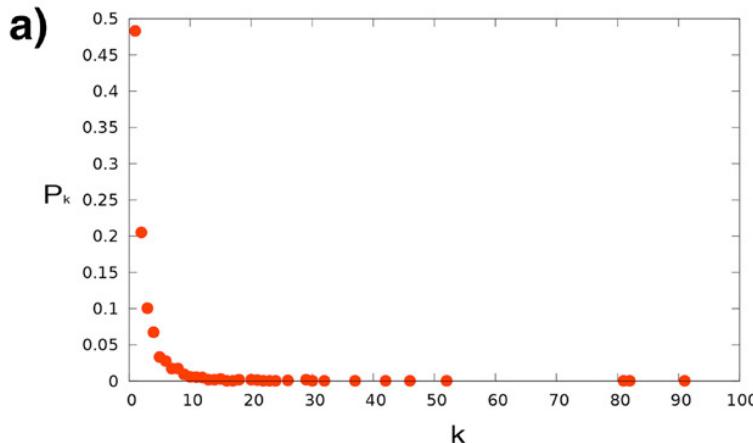
d.



$N_k = \# \text{ nodos con grado } k$

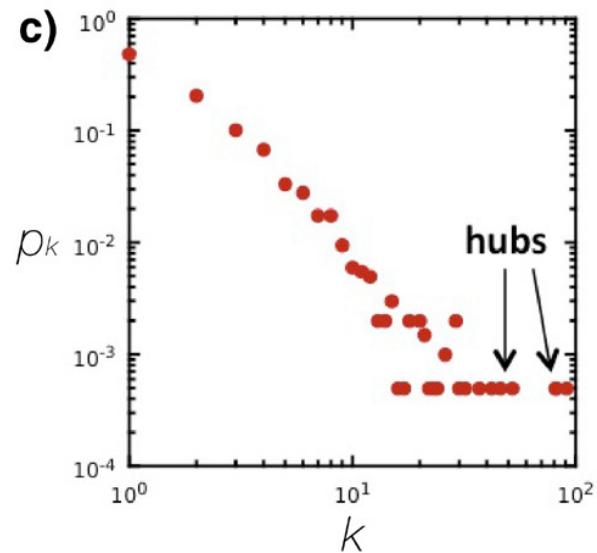
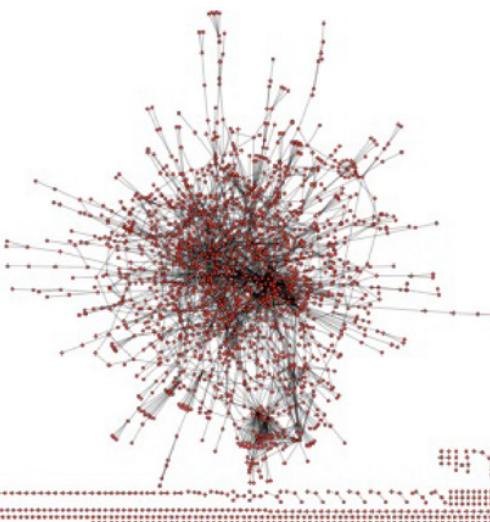
$P(k) = N_k / N \rightarrow \text{plot}$

# 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,  $p_1 = 0,48$ ), 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.

b)



# DISTRIBUCION DE GRADO

**Representación discreta:**  $p_k$  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  $k_1$  y  $k_2$ .

**Condición de normalización:**

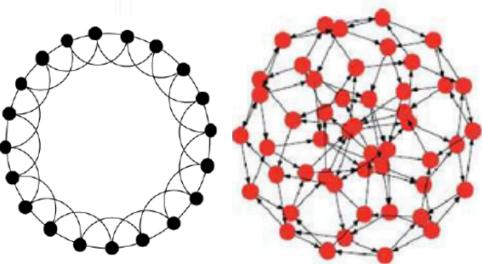
$$\sum_0^{\infty} p_k = 1$$

$$\int_{K_{\min}}^{\infty} p(k) dk = 1$$

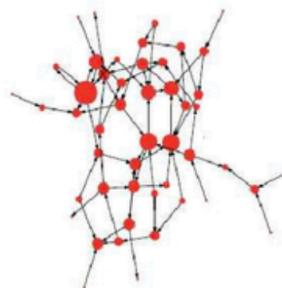
donde  $K_{\min}$  es el mínimo grado de la red.

# DISTRIBUCION DE GRADO

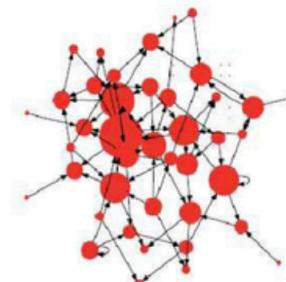
**Regular**



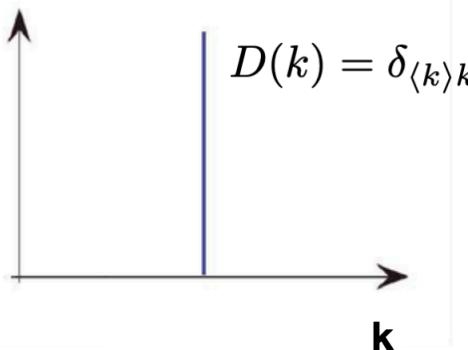
**Random**



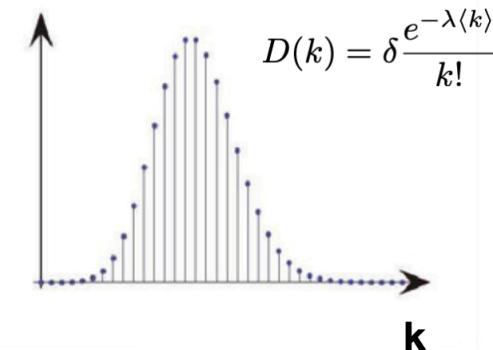
**Scale Free**



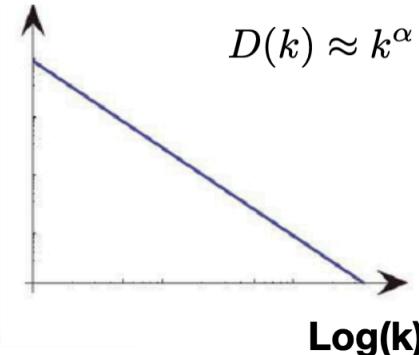
$D(k)$



$D(k)$

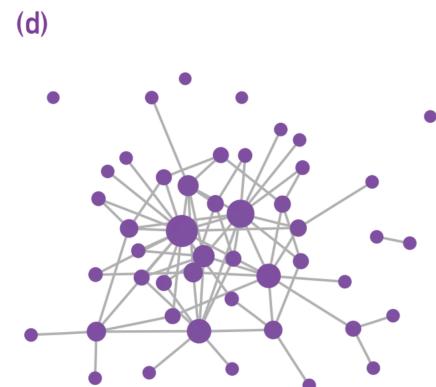
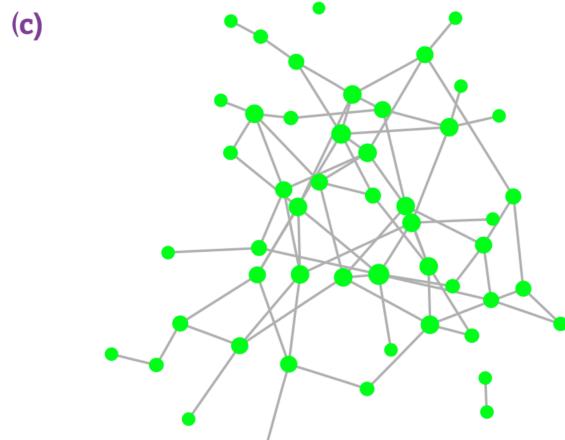
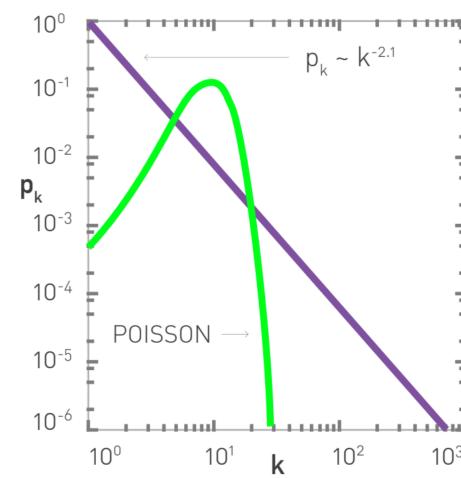
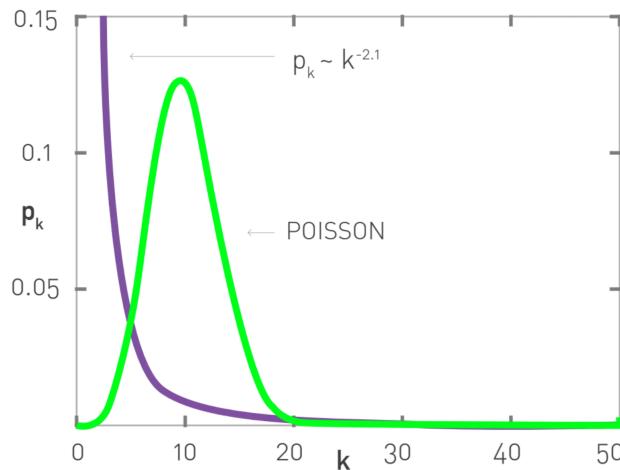


$\text{Log}(D(k))$



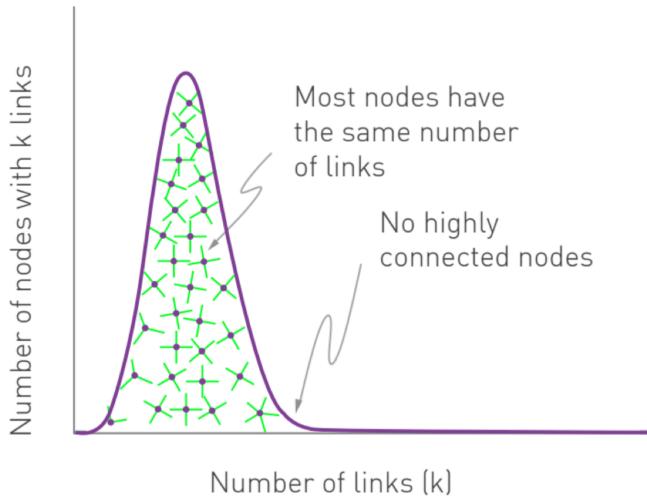
Piensen en los histogramas de fercuencias de grado  
¿Cómo se comparan los promedios con los valores máximos?

# Diferencia entre red aleatoria y libre de escala

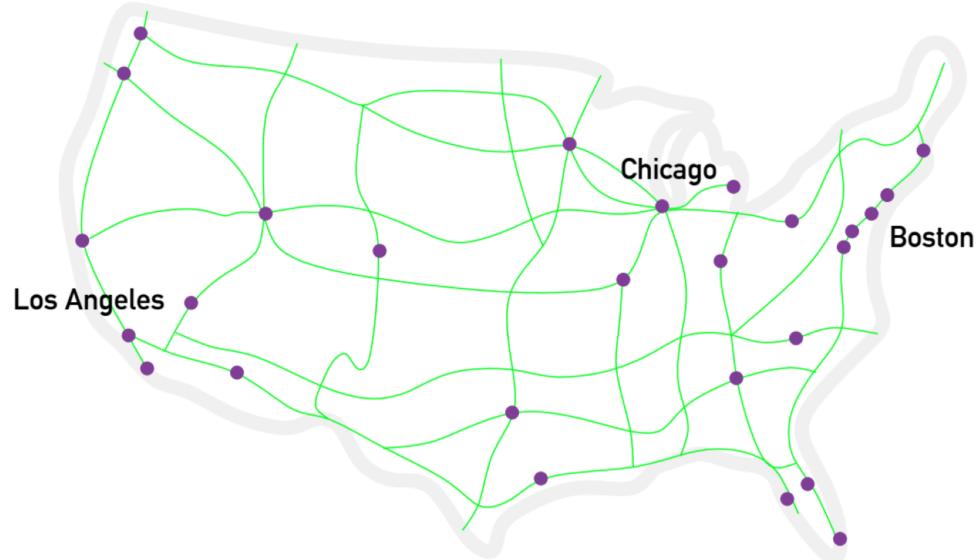


# Diferencia entre red aleatoria y libre de escala

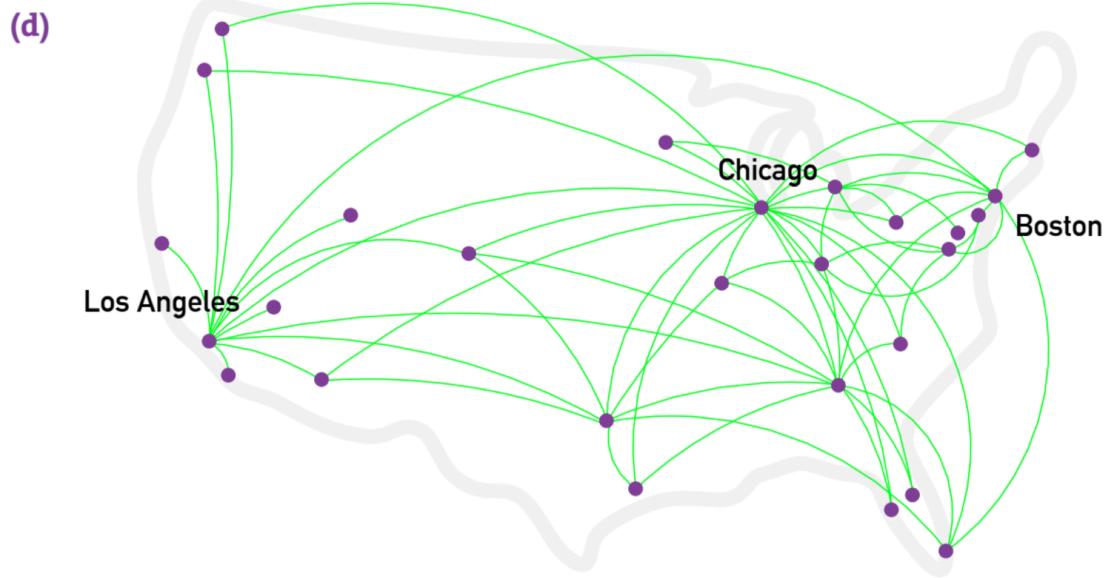
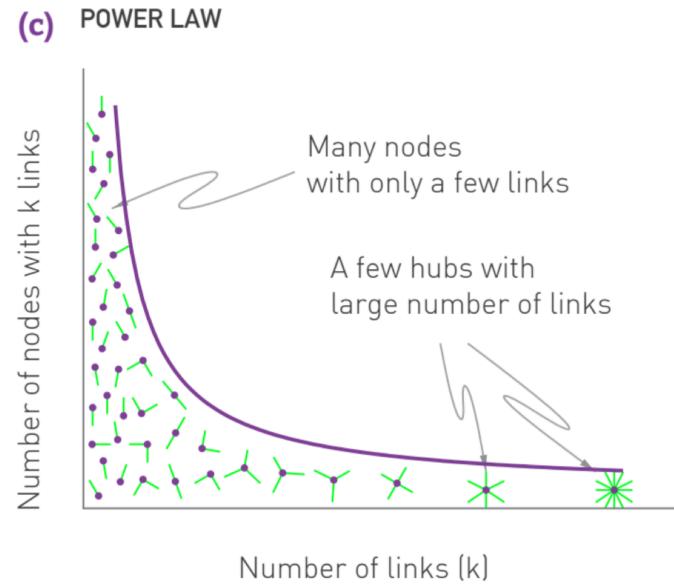
(a) POISSON



(b)



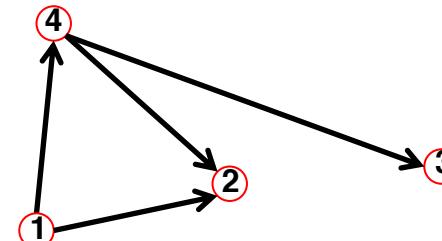
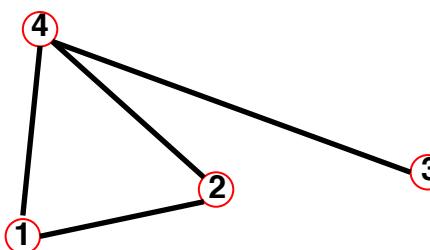
# Diferencia entre red aleatoria y libre de escala



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

# Matriz de adyacencia

# MATRIZ DE ADYACENCIA



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

$A_{ij}=0$  si los nodos  $i$  y  $j$  no estan conectados entre si

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

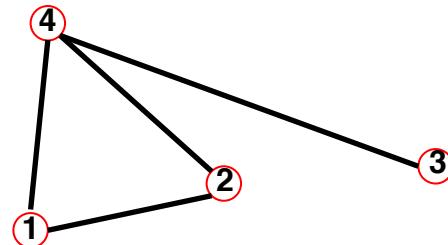
Tenga en cuenta que para un grafo dirigido (derecha) la matriz no es simétrica.

$A_{ij} = 1$  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 Y EL GRADO LOS NODOS

No-dirigido



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

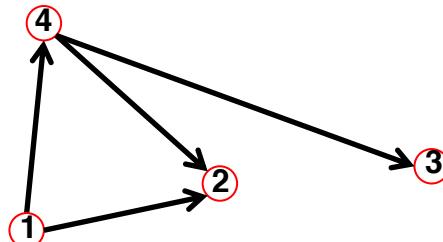
$$\begin{aligned} A_{ij} &= A_{ji} \\ A_{ii} &= 0 \end{aligned}$$

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

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

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

Dirigido



$$A_{ij} = \left( \begin{array}{ccc|c} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{array} \right)$$

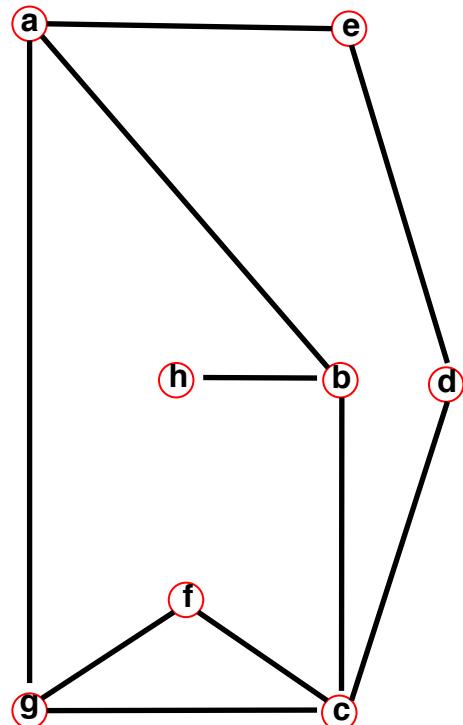
$$\begin{aligned} A_{ij} &\neq A_{ji} \\ A_{ii} &= 0 \end{aligned}$$

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

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

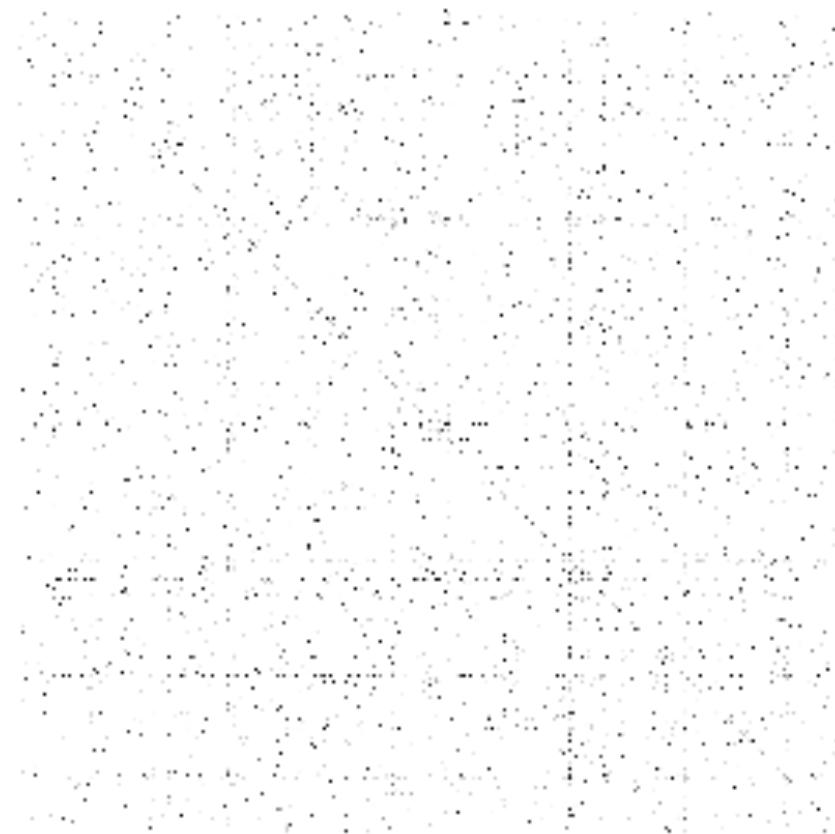
$$L = \sum_{i=1}^N k_i^{in} = \sum_{j=1}^N k_j^{out} = \sum_{i,j} A_{ij}$$

# MATRIZ DE ADYACENCIA

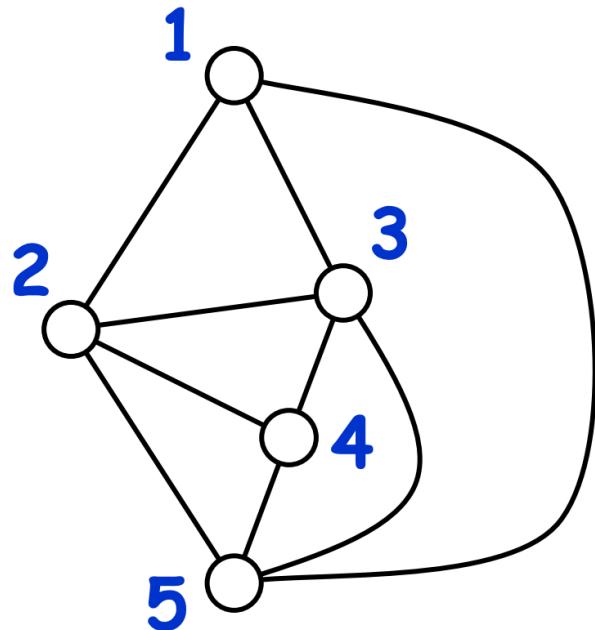


	a	b	c	d	e	f	g	h
a	0	1	0	0	1	0	1	0
b	1	0	1	0	0	0	0	1
c	0	1	0	1	0	1	1	0
d	0	0	1	0	1	0	0	0
e	1	0	0	1	0	0	0	0
f	0	0	1	0	0	0	1	0
g	1	0	1	0	0	0	0	0
h	0	1	0	0	0	0	0	0

# LAS MARICES DE ADYACENCIA SON "SPARSE"



# Ejercicio



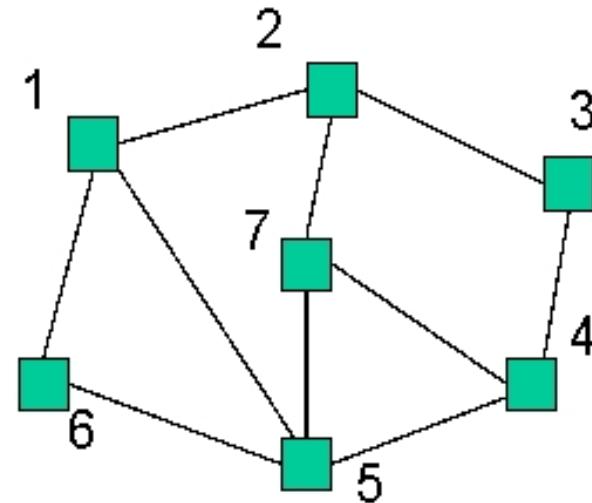
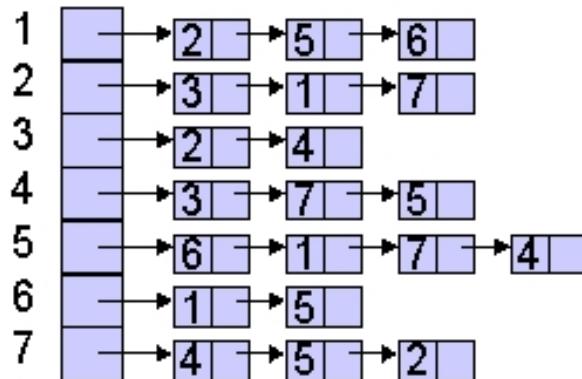
Representa el grafo como una matriz de adyacencia

# Lista de enlaces y lista de adyacencia

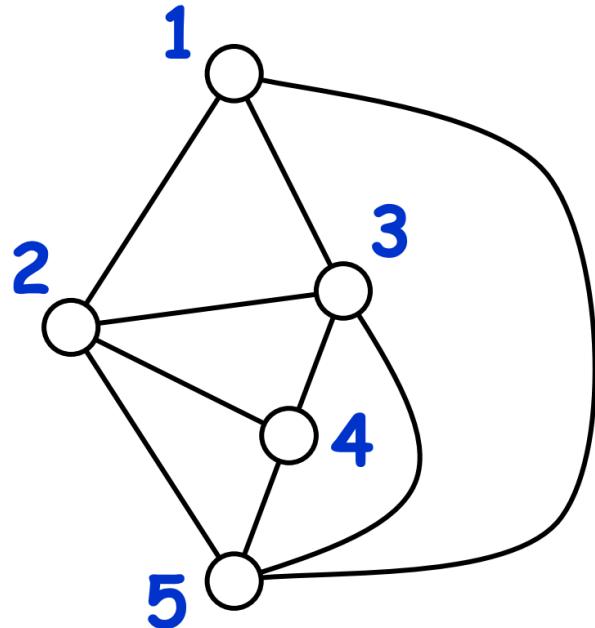
- List of edges

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

- Adjacency lists



# Ejercicio



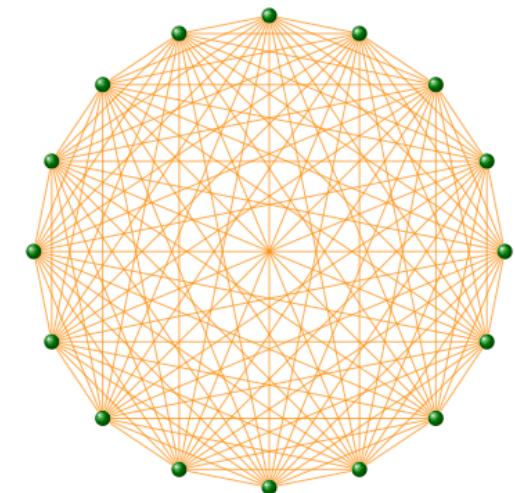
Representa el grafo como una lista de adyacencia  
Representa el grafo como una lista de enlaces

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  $\langle k \rangle = N-1$

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

# REDES REALES SON "SPARSE"

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

$$L \ll L_{\max}$$

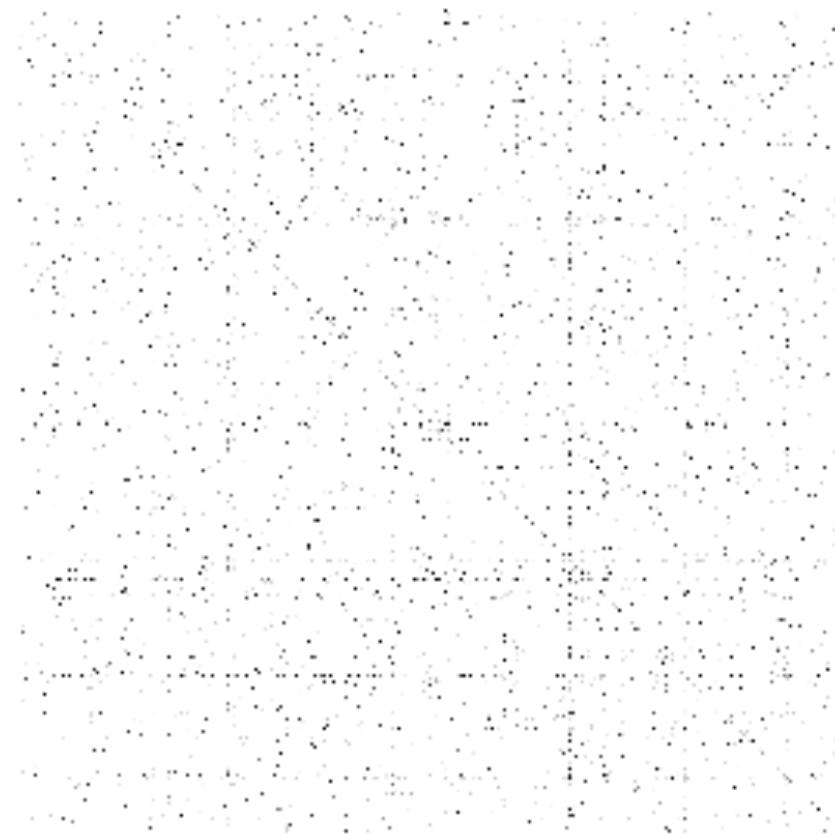
o

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

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

(Source: Albert, Barabasi, RMP2002)

# LAS MARICES DE ADYACENCIA SON "SPARSE"

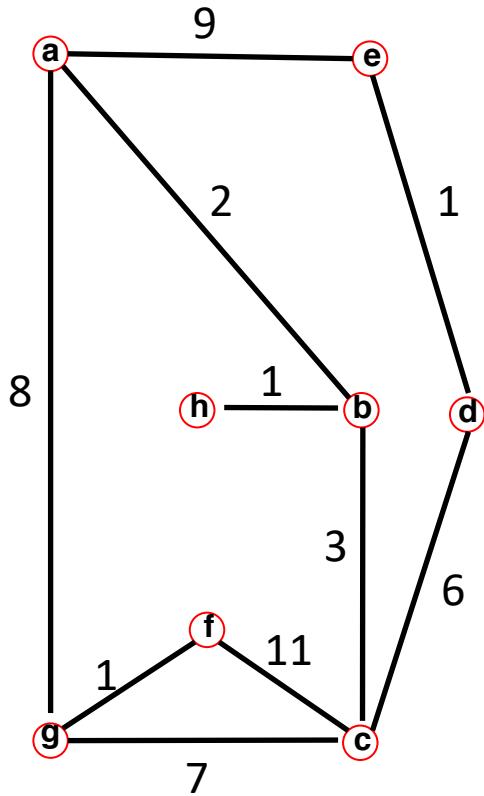


# REDES CON Y SIN PESOS

## REDES CON Y SIN PESOS

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

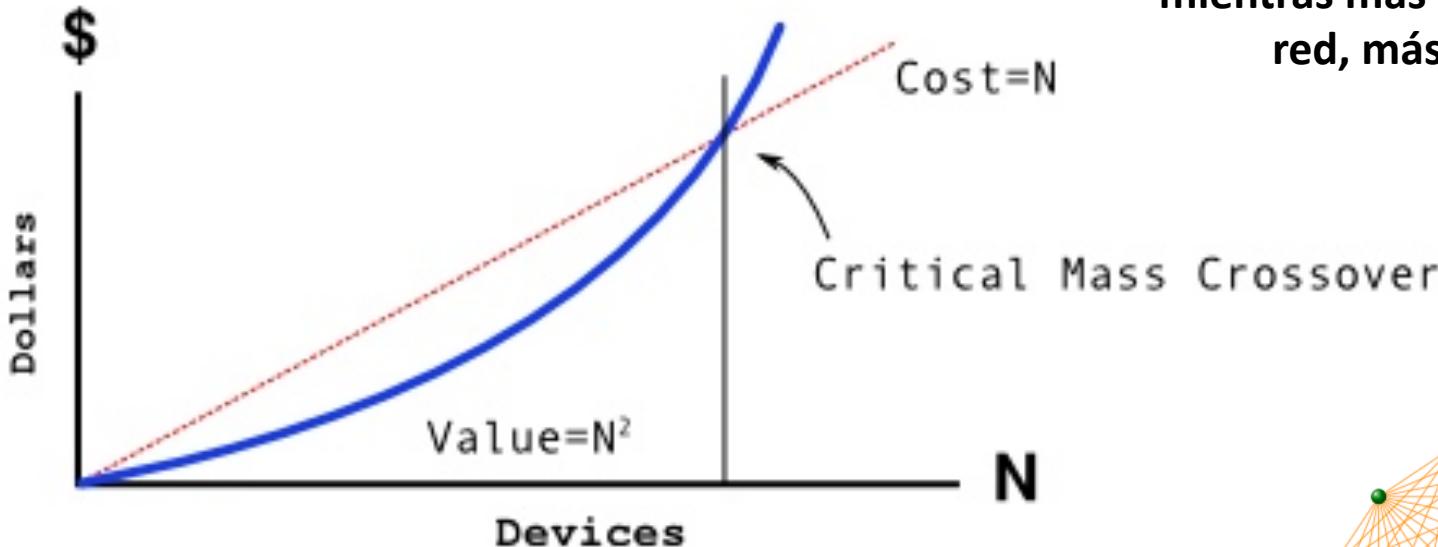
# MATRIZ DE ADYACENCIA CON PESOS



	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

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

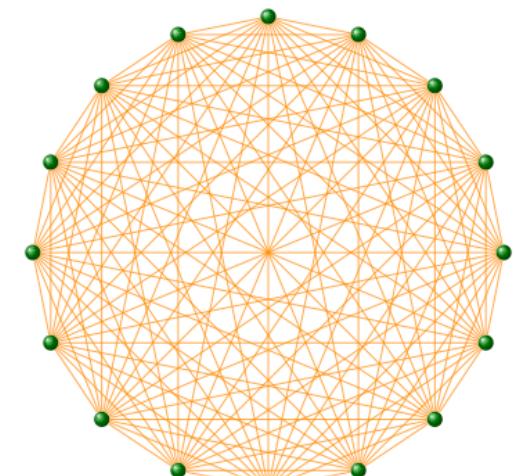
# METCALFE'S LAW



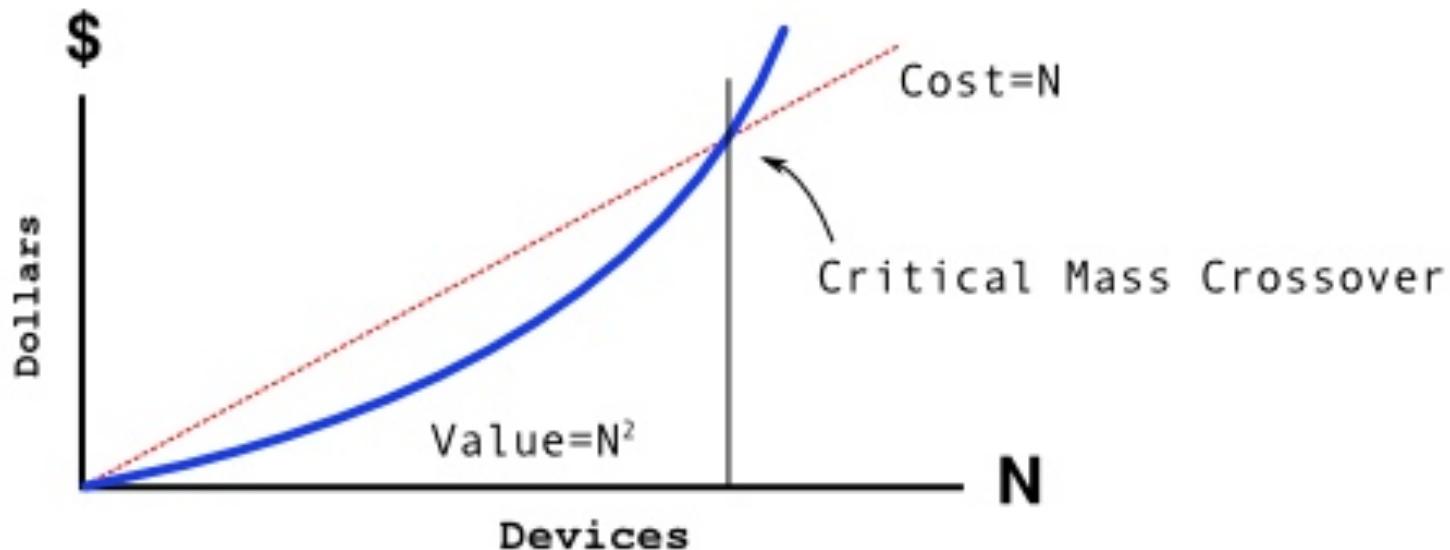
“mientras más personas usan una red, más valiosa se vuelve”

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.

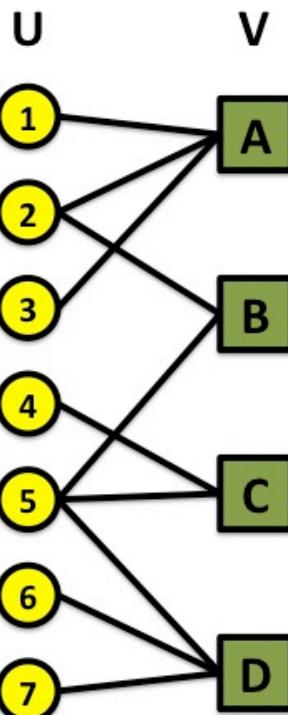
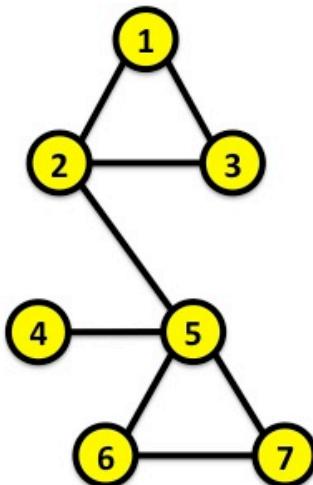
No todos los enlaces son de igual valor (pesos). Algunos enlaces se usan mucho mientras que la mayoría de los enlaces son 'débiles', es decir, rara vez se utilizan.

# REDES BIPARTITAS

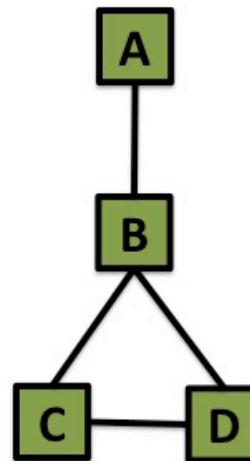
# GRAFOS BIPARTITOS

Un **grafo bipartito** (o bigrafo) es un gráfo cuyos nodos se pueden dividir en dos **conjuntos separados** U y V, de manera que cada enlace conecta un nodo en U con uno en V; es decir, U y V son conjuntos **independientes**.

Projection U



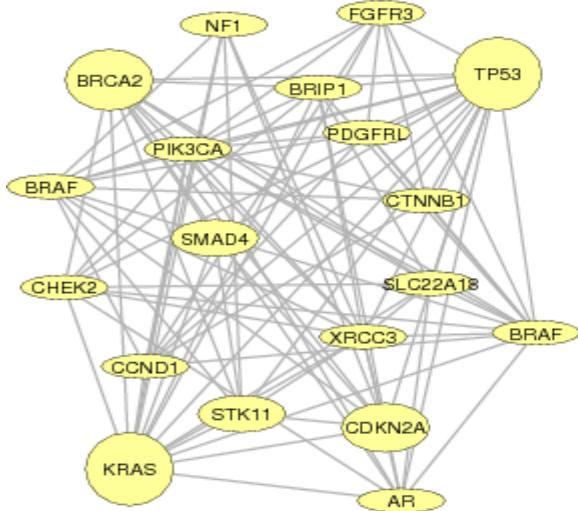
Projection V



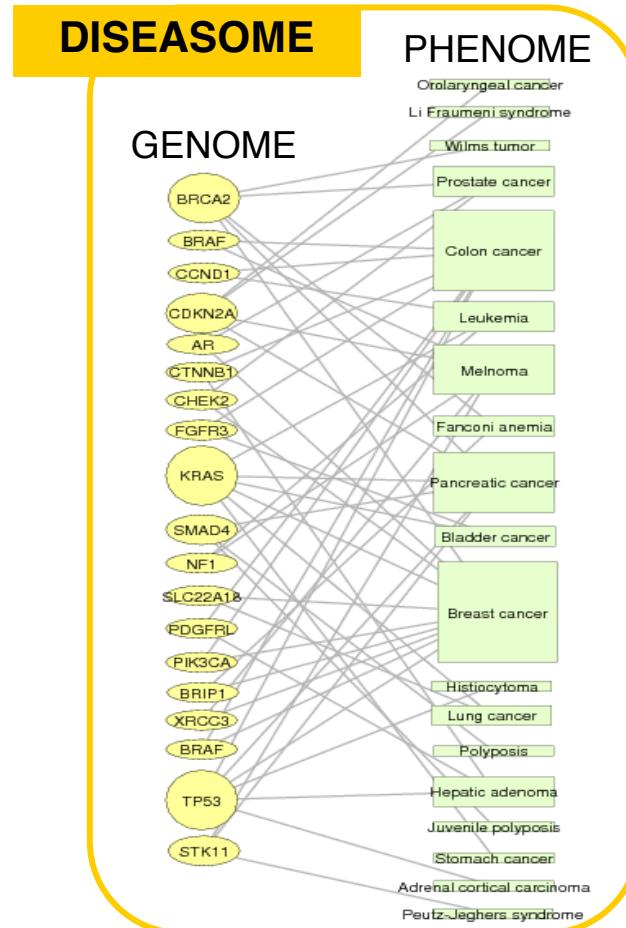
## Ejemplos:

- Red de actores de Hollywood
- Redes de colaboración
- Red de enfermedades (diseasome)

# Red de genes– Red de enfermedades



Gene network



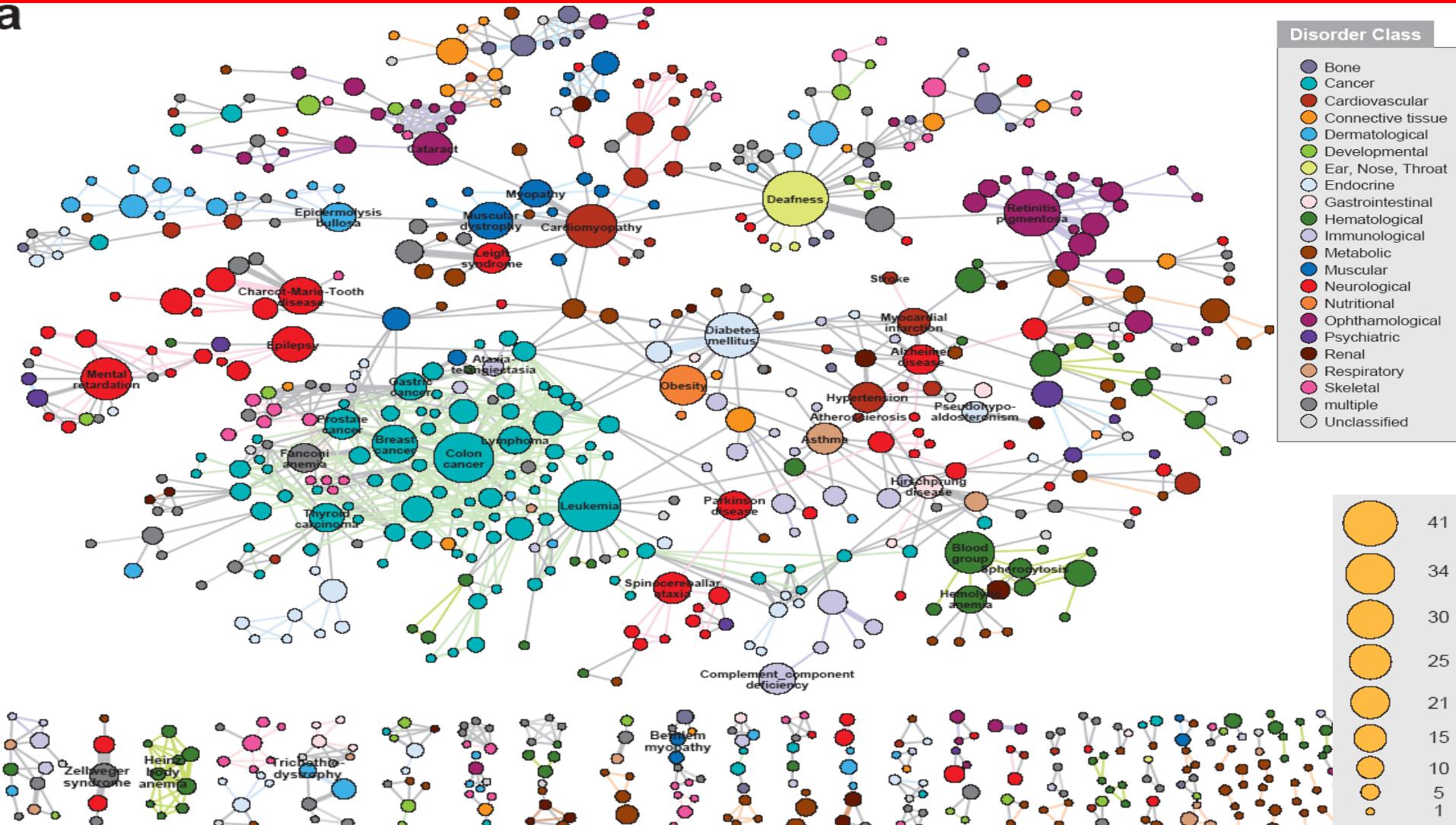
Goh, Cusick, Valle, Childs, Vidal & Barabási, PNAS (2007)

Disease network

# RED DE ENFERMEDADES HUMANAS

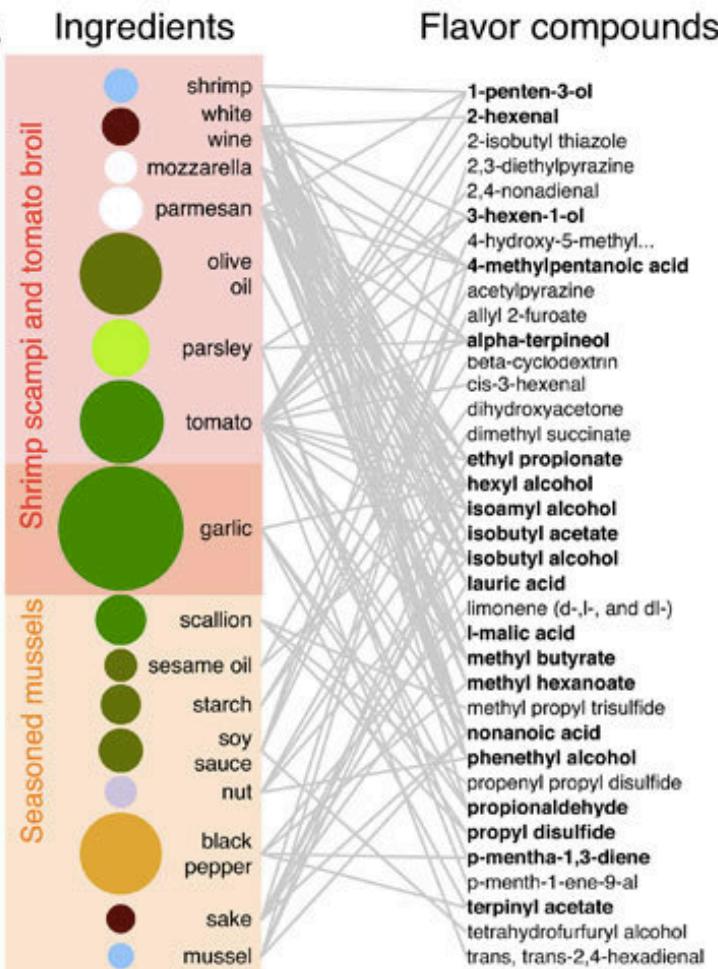
<https://www.nytimes.com/2008/05/06/health/research/06dise.html>

a

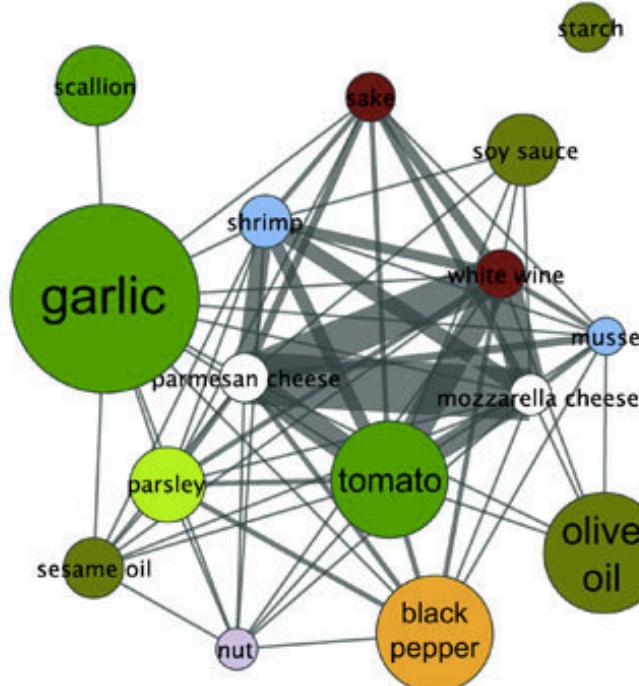


# Red bipartita ingredient-sabor

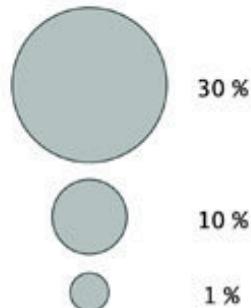
A



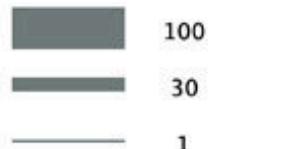
B Flavor network



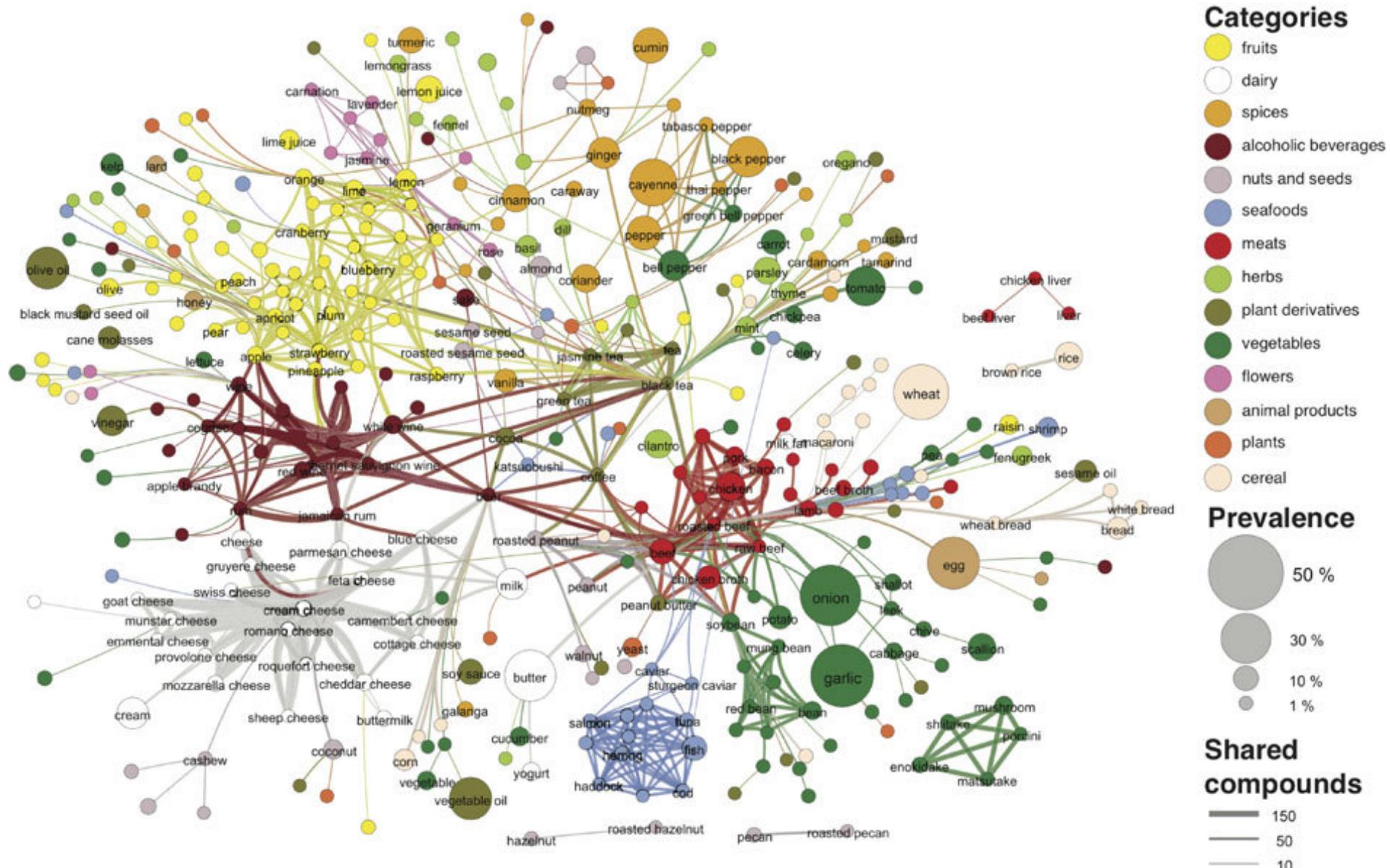
Prevalence



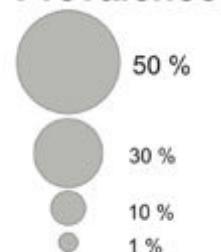
Shared compounds



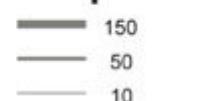
## Categories



## Prevalence

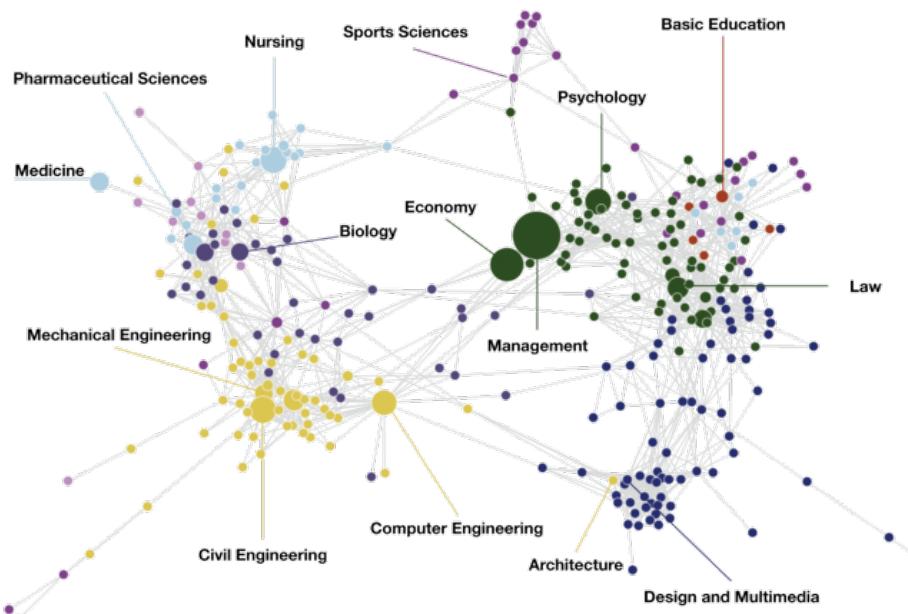


## Shared compounds

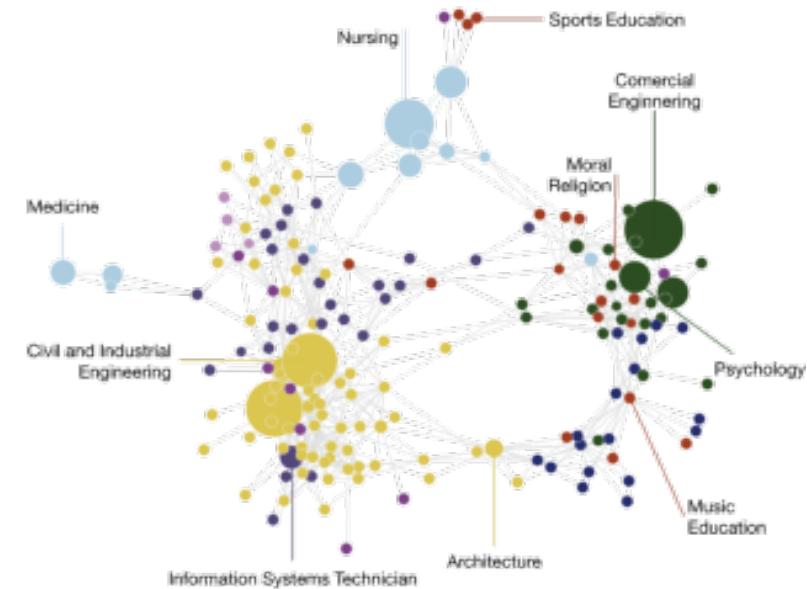


# Red bipartita carreras-postulantes

a) Portuguese Higher Education System [2008-2015]

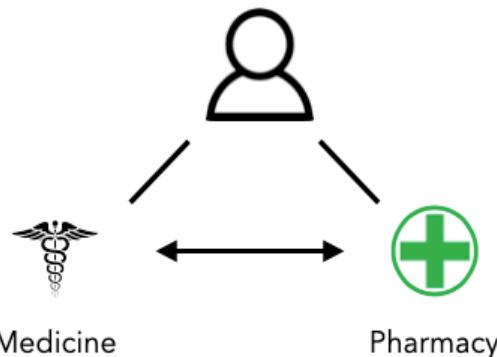


c) Chilean Higher Education System [2012-2017]



# The Higher Education Space

Connecting Degree Programs



## Lista de Preferencias

1. Medicina
2. Odontología
3. Tecnología Médica
4. Odontología
5. Cs. Físicas y Astronómicas



## Pares de Carreras

Medicina	●	Odontología	●	Odontología	●	Odontología
Medicina	●	Tecnología Médica	●	Odontología	●	Cs. Físicas y Astronómicas
Medicina	●	Odonotología	●	Tecnología Médica	●	Odontología
Medicina	●	Cs. Físicas y Astronómicas	●	Tecnología Médica	●	Cs. Físicas y Astronómicas
Odontología	●	Tecnología Médica	●	Odontología	●	Cs. Físicas y Astronómicas

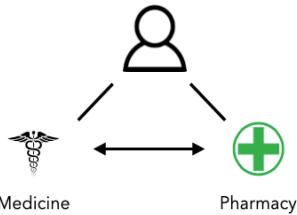
Consideramos todas las preferencias de cada postulante.  
Las preferencias repetidas indican una postulación a dos  
instituciones de educación superior distintas.

Creamos todos los pares de carreras posibles. Luego, descartamos los que contienen la misma carrera en ambos extremos (color gris).

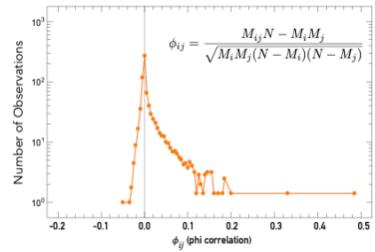
# Red bipartita carreras-postulantes

# The Higher Education Space

Connecting Degree Programs

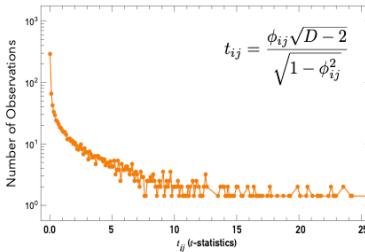


Finding Correlations Between Degrees

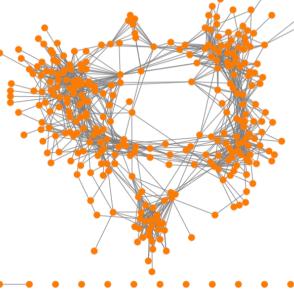


Discard all negative correlations

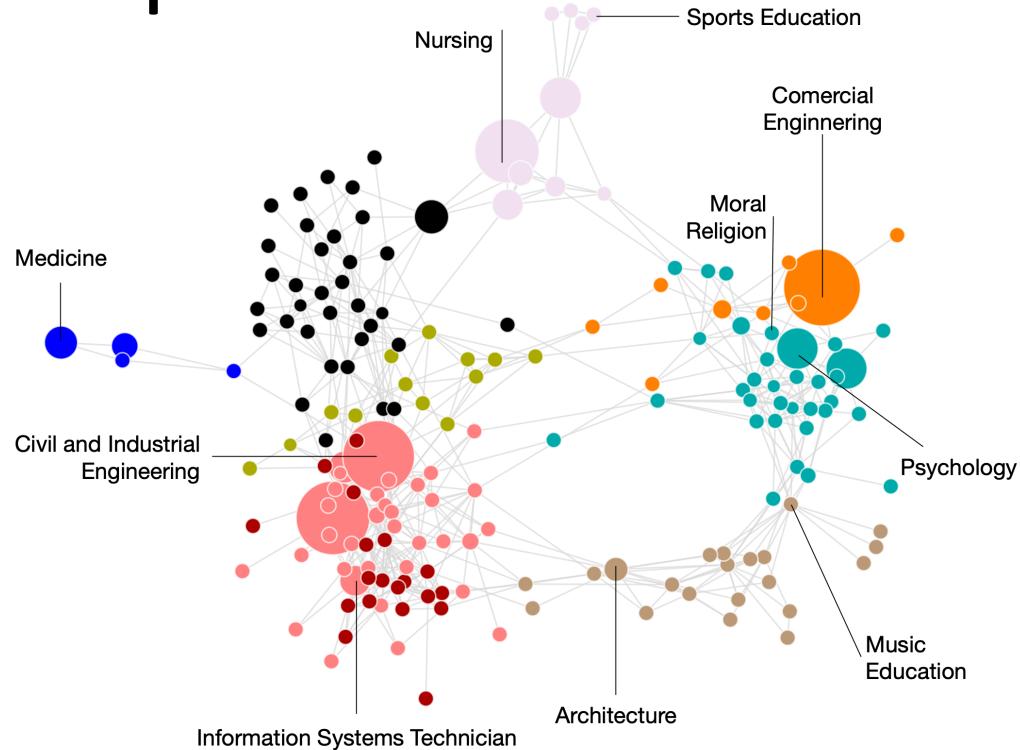
Finding Statistical Significance of Correlations



Discard all non-significant links



Discard all loose Nodes



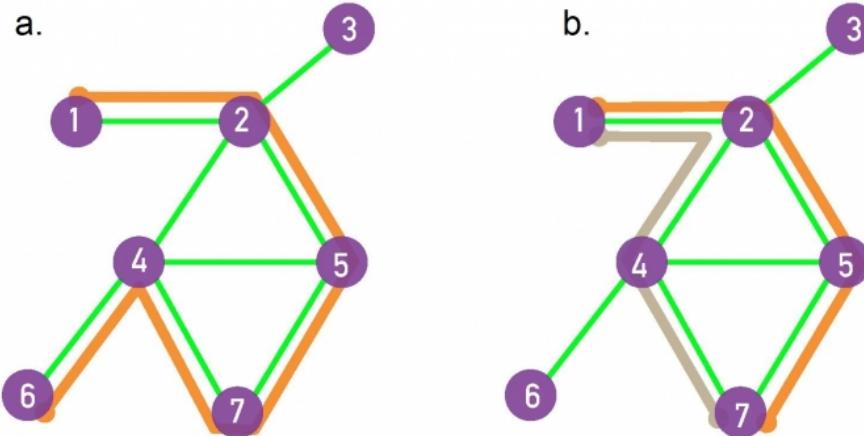
# “CAMINOLOGIA” (PATHOLOGY)

# CAMINOS

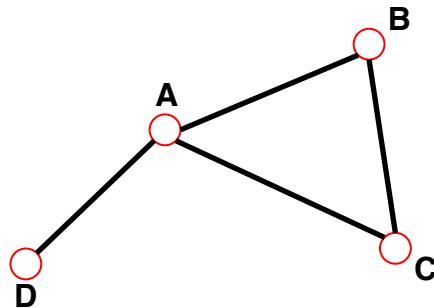
Un *camino* (path) es una secuencia de nodos en los que cada nodo es adyacente al siguiente

$P_{i_0, i_n}$  de longitud  $n$  entre los nodos  $i_0$  y  $i_n$  es una colección ordenada de  $n+1$  nodos y  $n$  links

$$P_n = \{i_0, i_1, i_2, \dots, i_n\} \quad P_n = \{(i_0, i_1), (i_1, i_2), (i_2, i_3), \dots, (i_{n-1}, i_n)\}$$

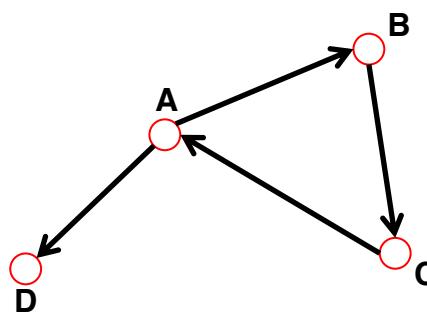


- En una red dirigida, la ruta solo puede seguir la dirección de una flecha.



**La distancia (ruta más corta, ruta geodésica)** entre dos nodos se define como el número de enlaces a lo largo de la ruta más corta que los conecta.

\* Si los dos nodos están desconectados, la distancia es infinita.



En los grafos dirigidos, cada ruta debe seguir la dirección de las flechas.

Así, en un grafo, la distancia desde el nodo A hasta B (en una ruta AB) es generalmente diferente de la distancia desde el nodo B hasta A (en una ruta BCA).

**N<sub>ij</sub>**, número de caminos entre dos nodos cualesquiera *i* y *j*:

Longitud n=1: Si existe un link entre *i* y *j*, entonces A<sub>ij</sub>=1 y A<sub>ij</sub>=0 en otro caso.

Longitud n=2: Si existe un camino de longitud dos entre *i* y *j*, entonces A<sub>ik</sub>A<sub>kj</sub>=1, y A<sub>ik</sub>A<sub>kj</sub>=0 en otro caso.

El número de caminos de longitud 2 es:

$$N_{ij}^{(2)} = \sum_{k=1}^N A_{ik}A_{kj} = [A^2]_{ij}$$

Longitud n: en general, si existe un camino de longitud *n* entre *i* y *j*, entonces A<sub>ik</sub>...A<sub>lj</sub>=1 y A<sub>ik</sub>...A<sub>lj</sub>=0 en otro caso.

El número de caminos de longitud *n* entre *i* y *j* es\*

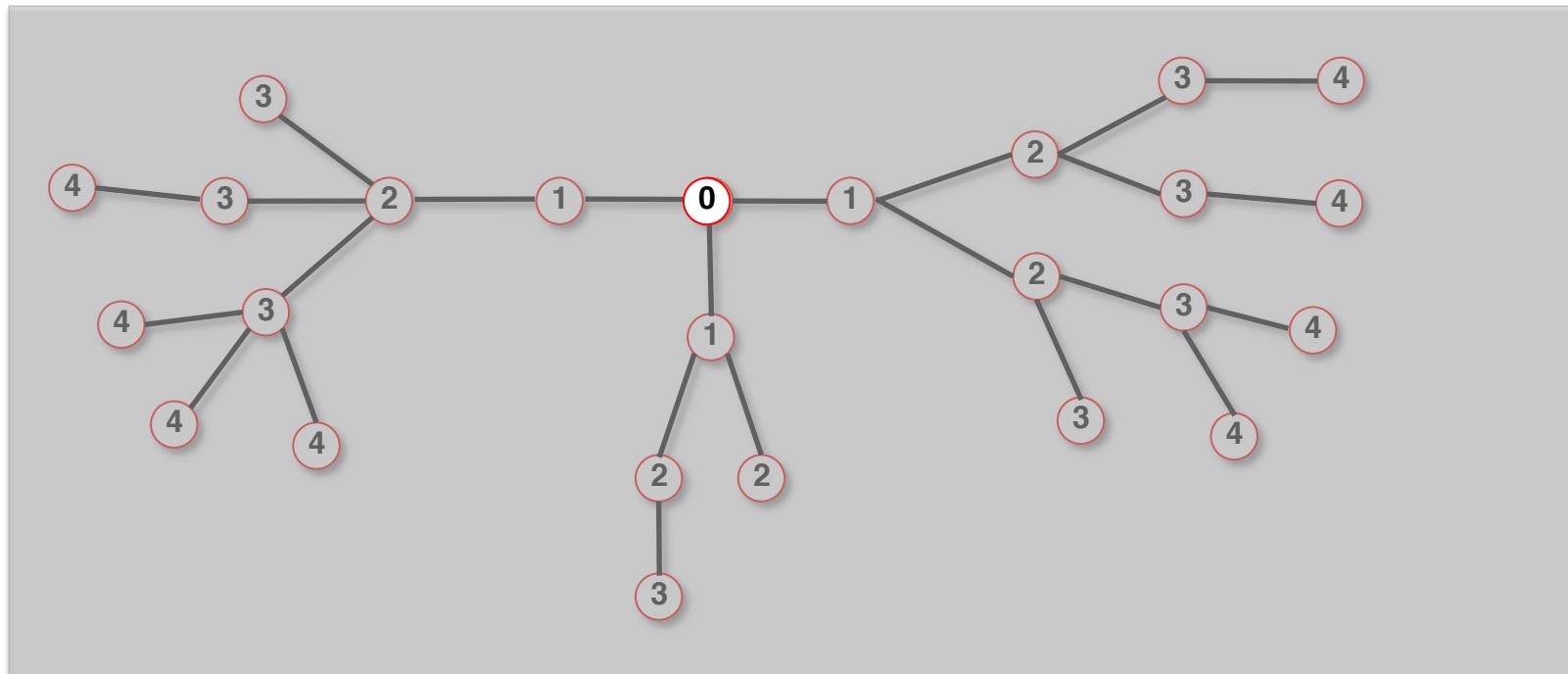
$$N_{ij}^{(n)} = [A^n]_{ij}$$

\* Se aplica tanto a redes dirigidas como no dirigidas.

# ECONTRANDO DISTANCIAS: “BREADTH FIRST SEARCH”

La distancia entre el nodo 0 y nodo 4:

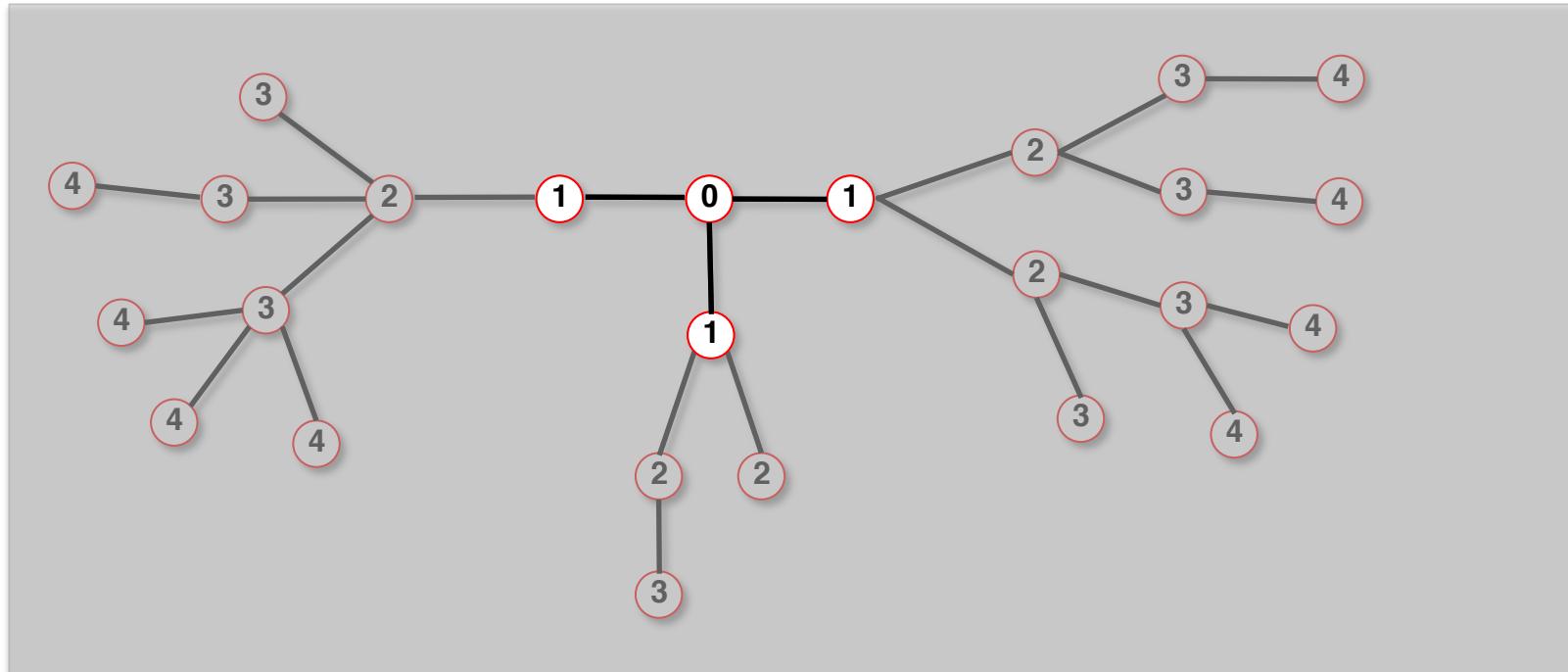
1. Comienza en 0.



# ECONTRANDO DISTANCIAS: “BREADTH FIRST SEARCH”

La distancia entre el nodo 0 y nodo 4:

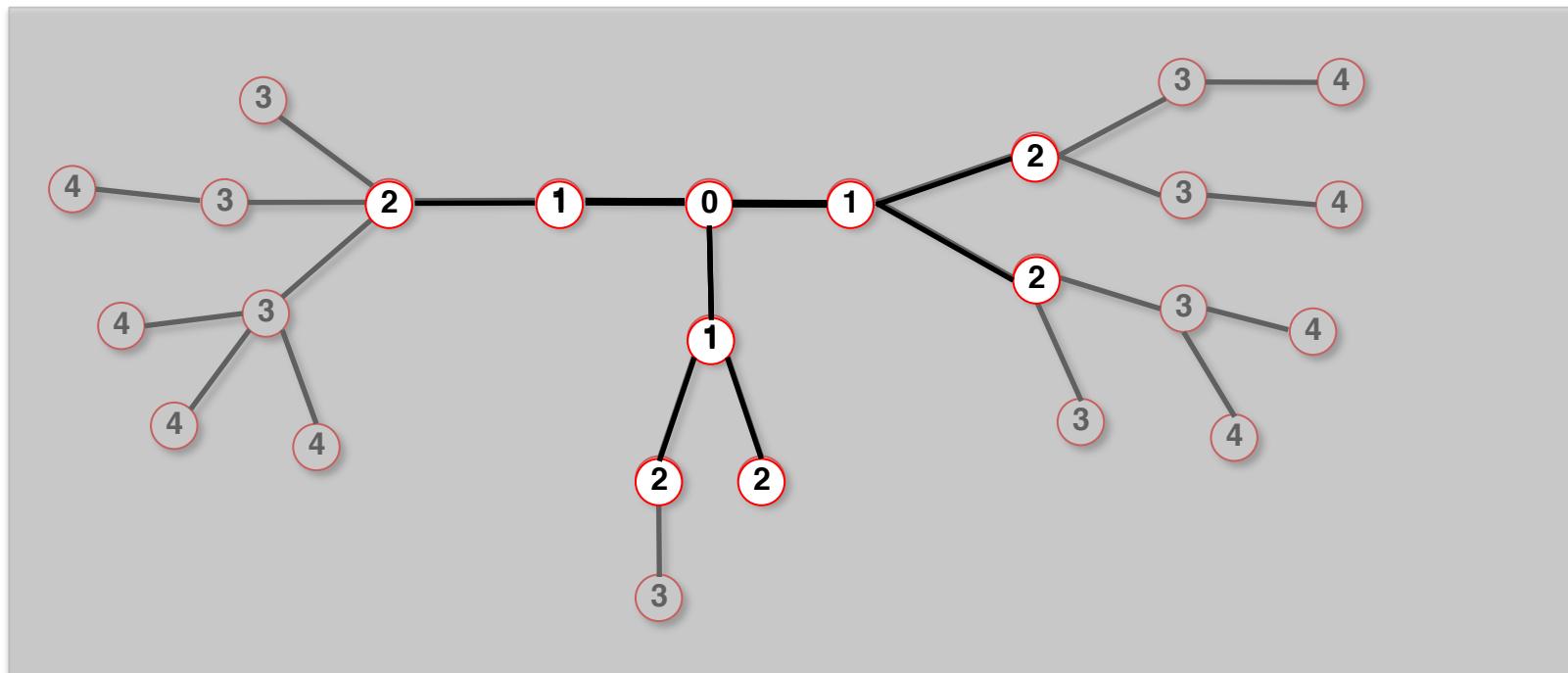
1. Comienza en 0.
2. Encuentra los nodos adyacentes a 0. Márcalos con la etiqueta 1. Ponlos en una fila.



# ECONTRANDO DISTANCIAS: “BREADTH FIRST SEARCH”

**La distancia entre el nodo 0 y nodo 4:**

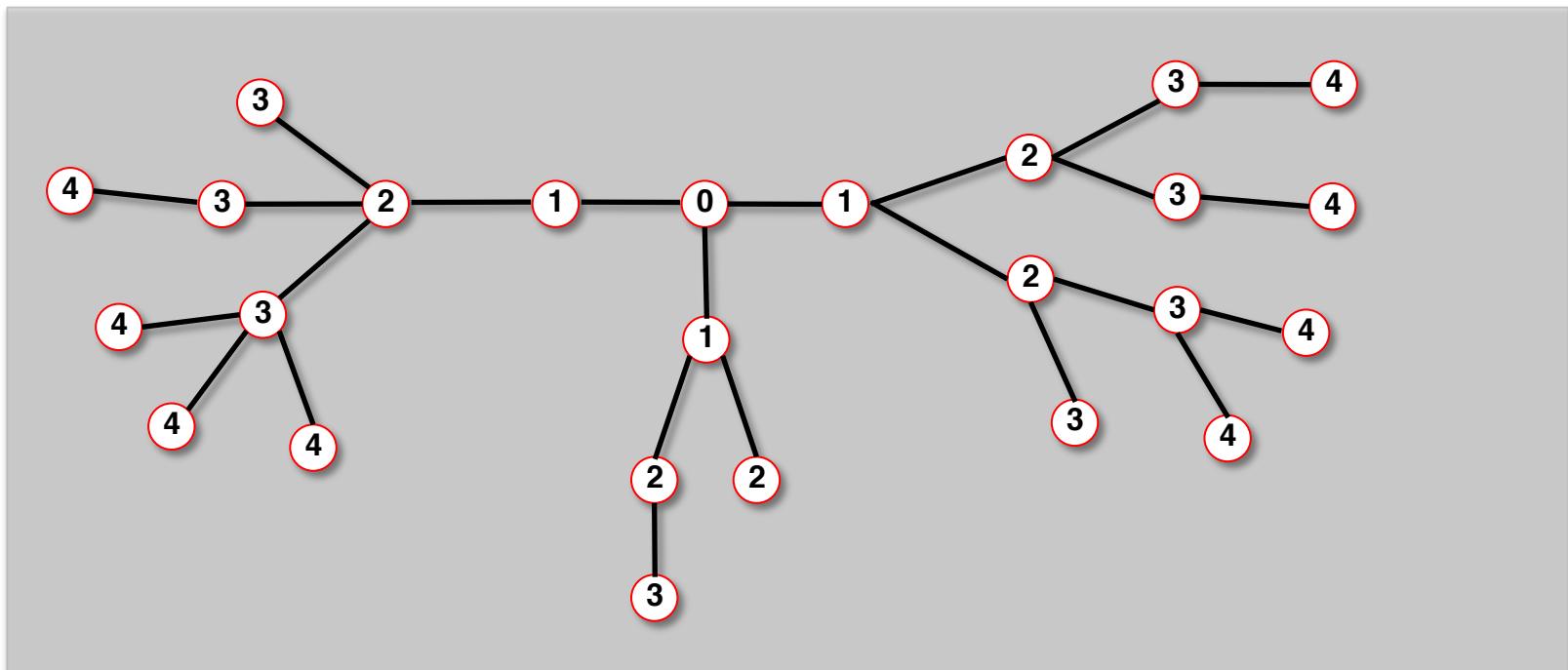
1. Comienza en 0.
2. Encuentra los nodos adyacentes a 1. Márcalos con la etiqueta 1. Pon los en una fila.
3. Toma el primer nodo de la fila. Encuentra los nodos no-marcados adyacentes en el grafo. Márcalos con la etiqueta 2. Ponlos en la fila.



# ECONTRANDO DISTANCIAS: “BREADTH FIRST SEARCH”

La distancia entre el nodo 0 y nodo 4:

1. Repite hasta que encuentres el nodo 4 o no hayan mas nodos en la fila.
2. La distancia entre 0 y 4 es la etiqueta 4 o, si 4 no tiene una etiqueta, infinito.



# DIÁMETRO DE RED Y DISTANCIA MEDIA

*Diámetro:  $d_{max}$*  La distancia máxima entre cualquier par de nodos en el grafo.

*Longitud / distancia media del camino,  $\langle d \rangle$ , para un **grafo conectado**:*

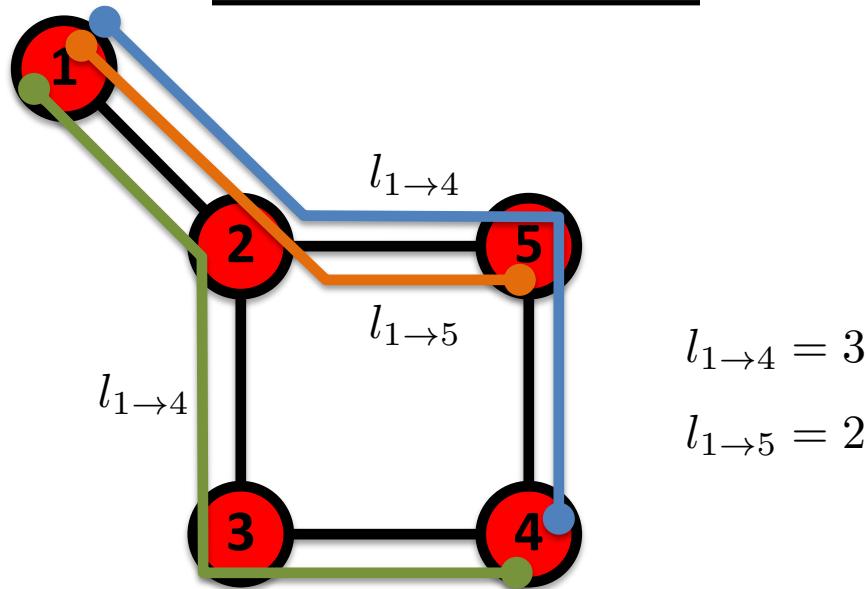
dónde  $d_{ij}$  es la distancia desde el nodo  $i$  al nodo  $j$

$$\langle d \rangle \equiv \frac{1}{2L_{\max}} \sum_{i,j \neq i} d_{ij}$$

En un grafo no-dirigido  $d_{ij} = d_{ji}$ , solo necesitamos contarlos una sola vez:

$$\langle d \rangle \equiv \frac{1}{L_{\max}} \sum_{i,j > i} d_{ij}$$

Camino más corto:

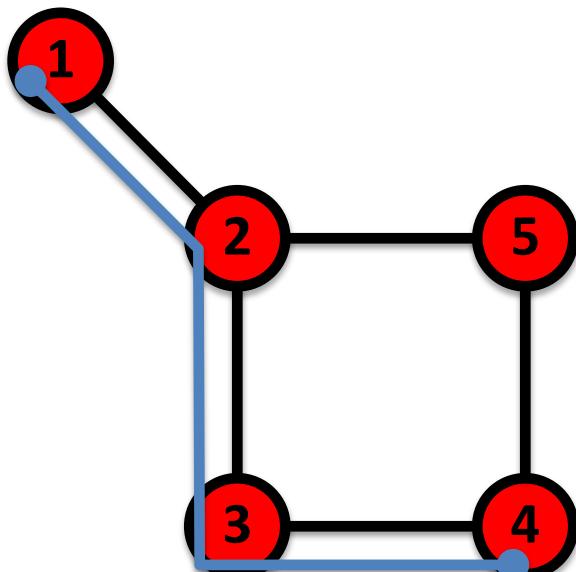


$$l_{1 \rightarrow 4} = 3$$

$$l_{1 \rightarrow 5} = 2$$

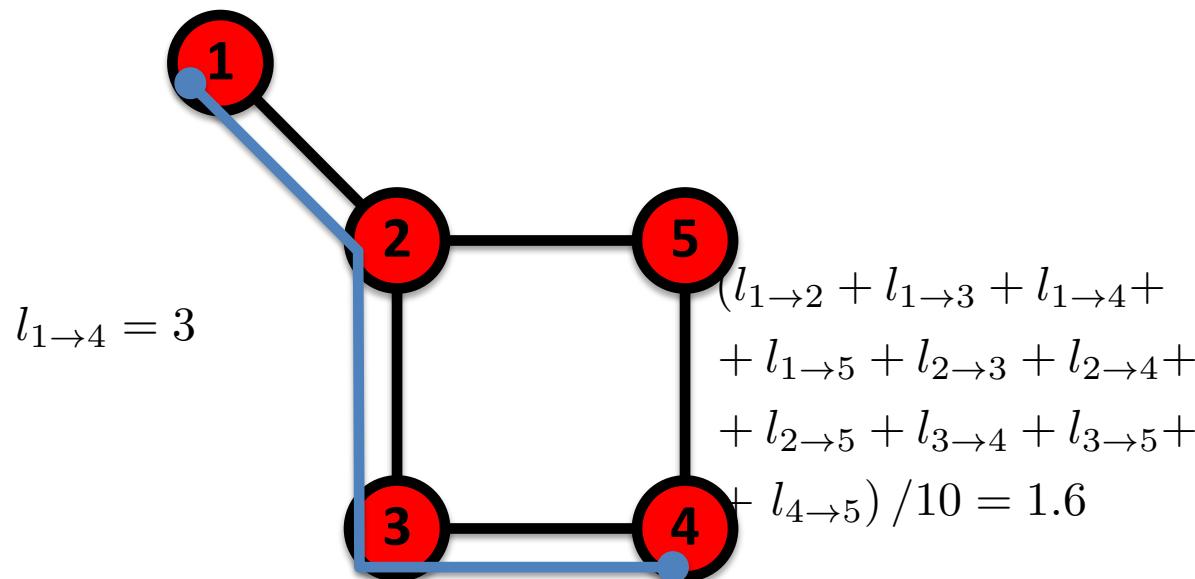
El camino con la longitud más corta entre dos nodos (distancia).

## Diámetro



La distancia más larga del grafo

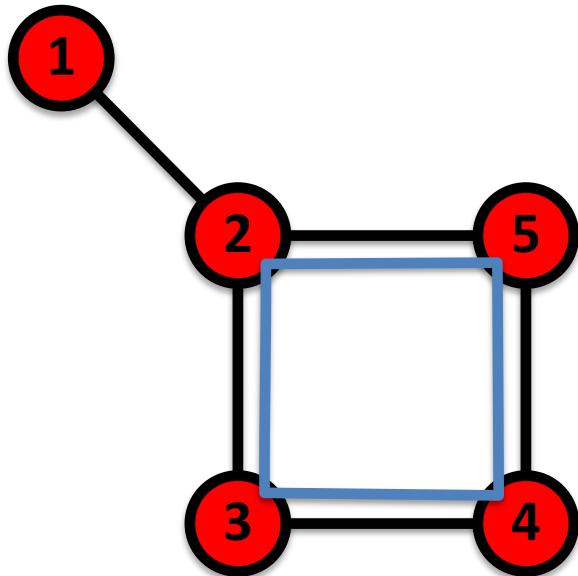
## Longitud de camino promedio



El promedio de los caminos más cortos para todos los pares de nodos

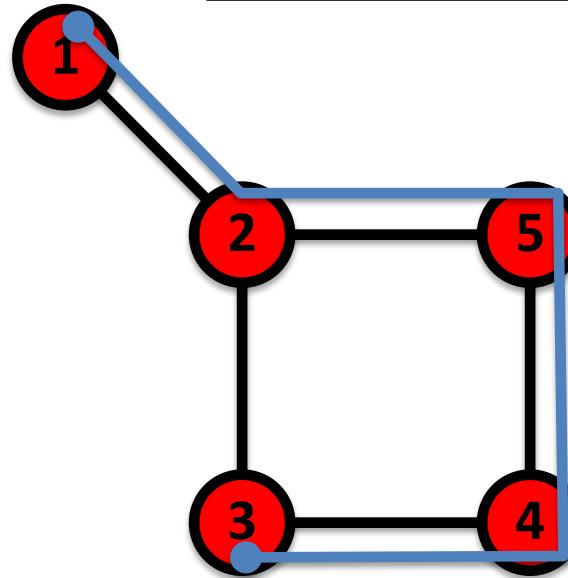
# CAMINOLOGIA: resumen

Círculo



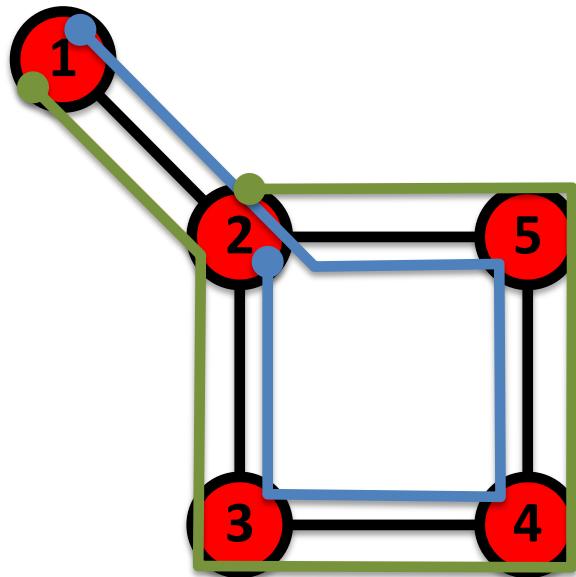
Una ruta con el mismo nodo de inicio y final.

Camino de auto-evasión



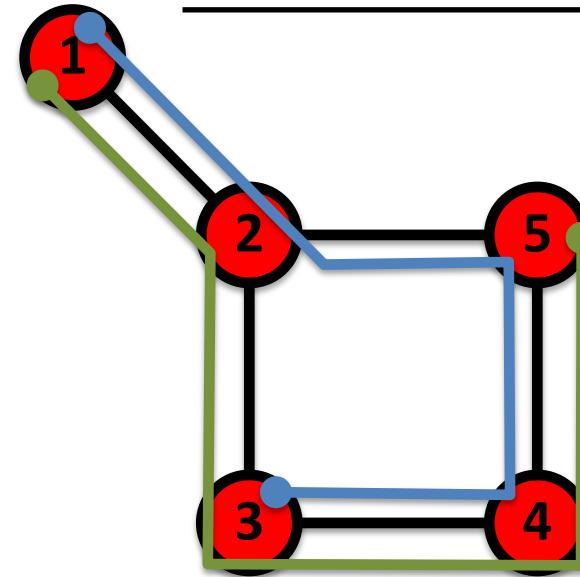
Un camino que no se intersecta.

## Camino Euleriano



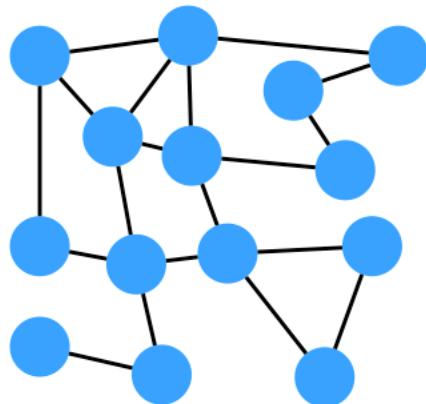
Un camino que atraviesa cada enlace exactamente una vez.

## Camino Hamiltoniano



Una ruta que visita cada nodo exactamente una vez.

## CAMINOLOGIA: resumen



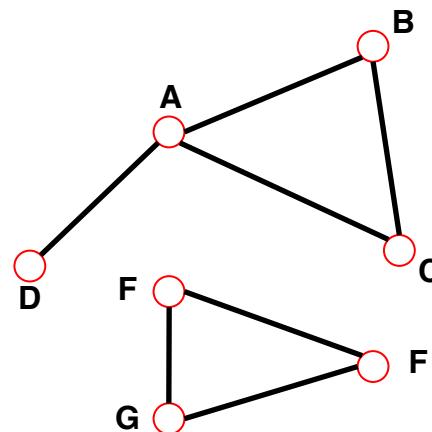
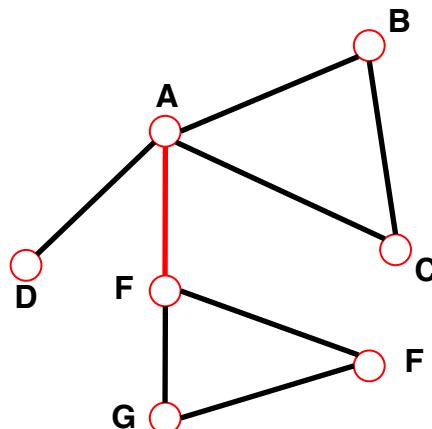
[**Importante**] La red proporciona una métrica natural para el sistema. Es decir, nos permite cuantificar distancias entre elementos.

Esto es particularmente importante ya que trataremos con sistemas en los que las distancias / similitudes de cálculo no son naturales y deben realizarse a través de un proxy (la red).

# CONECTIVIDAD

# CONECTIVIDAD DE GRAFOS NO-DIRIGIDOS

Gráfico conectado (no dirigido): cualquiera de los dos vértices se puede unir por una ruta. Un gráfico desconectado está compuesto por dos o más componentes conectados.



Componente más grande:  
**Componente gigante**

El resto: **Aíslados**

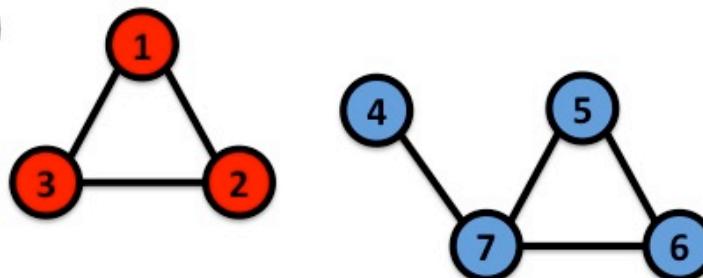
Puente: si lo borramos, la gráfica se desconecta.

# CONECTIVIDAD DE GRAFOS NO-DIRIGIDOS.

## Matriz de adyacencia

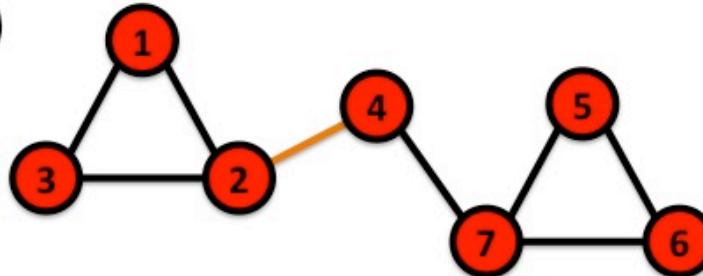
La matriz de adyacencia de una red con varios componentes se puede escribir en forma de diagonal de bloques, de modo que los elementos distintos de cero se limiten a los cuadrados, y todos los demás elementos sean cero:

(a)



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

(b)

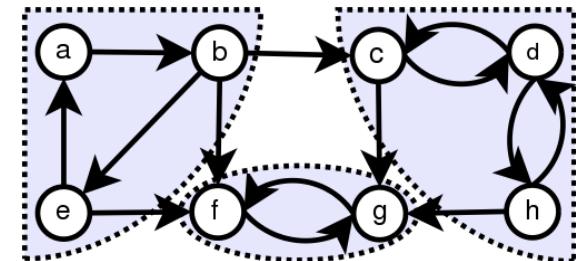
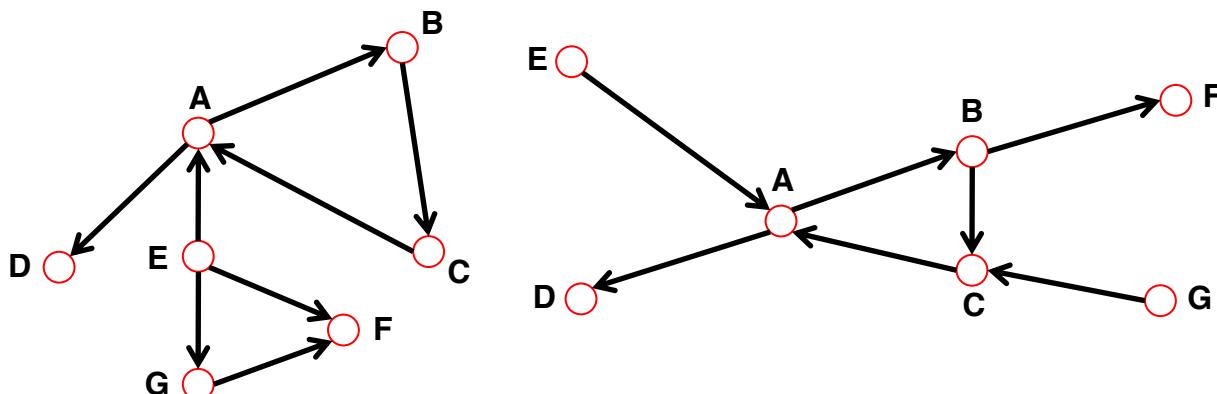


$$\begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 \end{pmatrix}$$

# CONECTIVIDAD DE GRAFOS DIRIGIDOS

Grafo dirigido fuertemente conectado: tiene una ruta desde cada nodo a todos los demás nodos **y viceversa** (por ejemplo, ruta AB y ruta BA).

Grafo dirigido débilmente conectada: está conectada si ignoramos el direcciones de borde.



**In-component:** nodos que pueden alcanzar el scc (strongly connected component),  
**Out-component:** Nodos a los que se puede acceder desde el scc.

# Coeficiente de Clustering

# COEFICIENTE DE CLUSTERING

## \* Coeficiente de clustering:

¿Qué fracción de tus vecinos están conectados?

\* Nodo  $i$  con grado  $k_i$

\*  $C_i$  entre  $[0,1]$

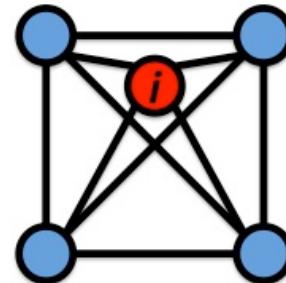
\*  $e_i$  (links entre los vecinos del nodo  $i$ )

Dirigido

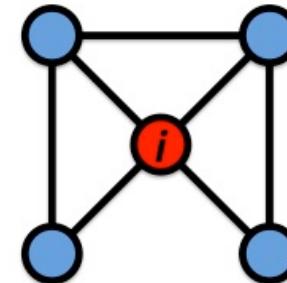
$$C_i = \frac{|\{e_{jk} : v_j, v_k \in N_i, e_{jk} \in E\}|}{k_i(k_i - 1)}.$$

No-Dirigido

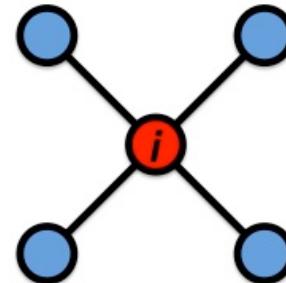
$$C_i = \frac{2e_i}{k_i(k_i - 1)}$$



$$C_i = 1$$



$$C_i = 1/2$$



$$C_i = 0$$

# COEFICIENTE DE CLUSTERING

## \* Coeficiente de clustering:

¿Qué fracción de tus vecinos están conectados?

\* Nodo  $i$  con grado  $k_i$

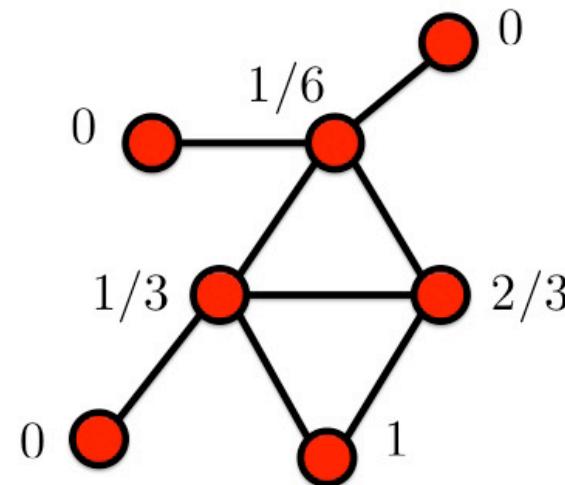
\*  $C_i$  entre  $[0,1]$

\*  $\epsilon_i$  (links entre los vecinos del nodo  $i$ )

$$C_i = \frac{1}{k_i(k_i - 1)} \sum_{j,k} A_{ij} A_{jk} A_{ki}$$

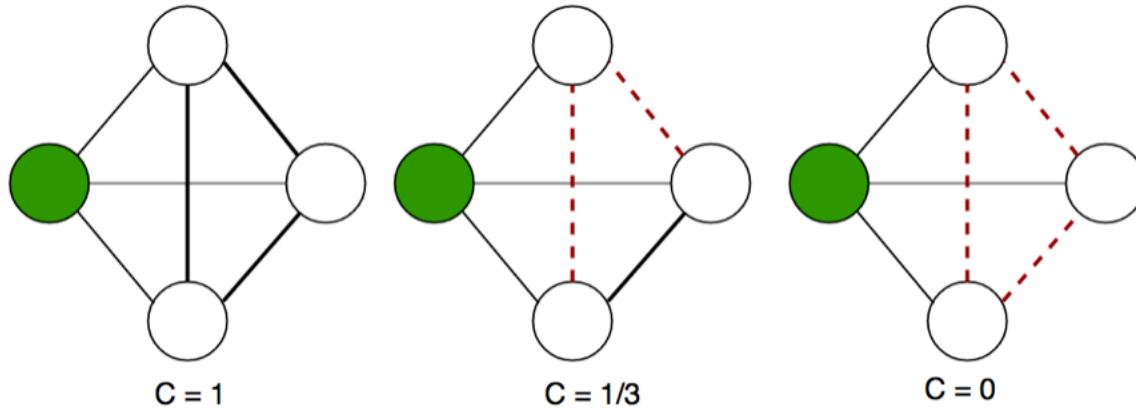
$$k_i = \sum_j A_{ij}$$

$$\bar{C} = \frac{1}{n} \sum_{i=1}^n C_i.$$



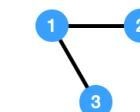
$$\langle C \rangle = \frac{13}{42} \approx 0.310$$

# COEFICIENTE DE CLUSTERING

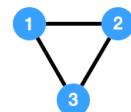


$$C = \frac{\text{number of closed triplets}}{\text{number of all triplets (open and closed)}}.$$

Medición de la densidad de los enlaces  
Enlaces Fuertes entre nodos.

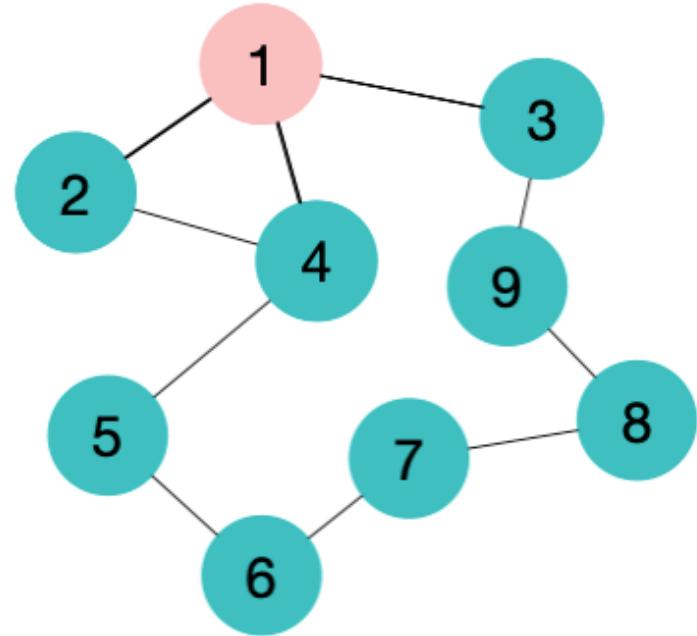


Open Triangle  
(triplet)



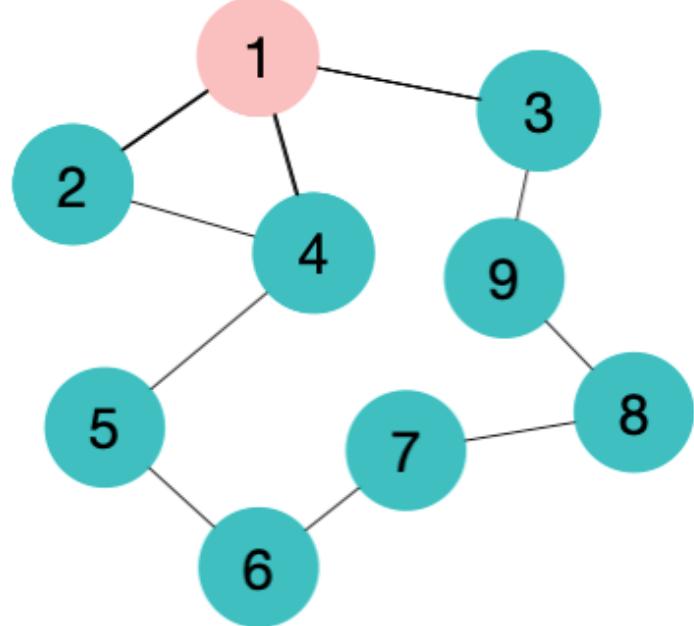
Closed Triangle  
(closed triplet)

# COEFICIENTE DE CLUSTERING

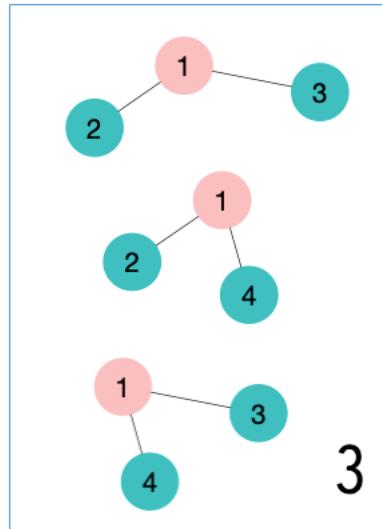


¿Cuál es el CC del nodo 1?

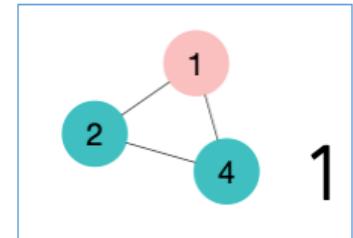
# COEFICIENTE DE CLUSTERING



Tripletes



Triángulos Cerrados



Coeficiente de Clustering:

$$1/3=0.33$$

# Resumen y Conceptos Importantes

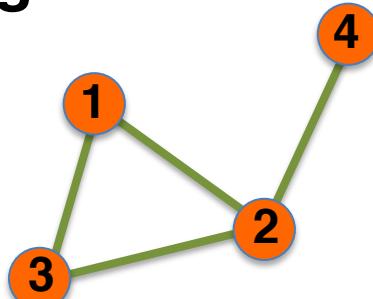
# TRES PRINCIPALES CANTIDADES EN NETWORK SCIENCE

**Distribución de grado:**  $P(k)$

**Longitud de camino:**  $\langle d \rangle$

**Coeficiente de clustering:**  $C_i = \frac{2e_i}{k_i(k_i - 1)}$

## No-dirigido



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

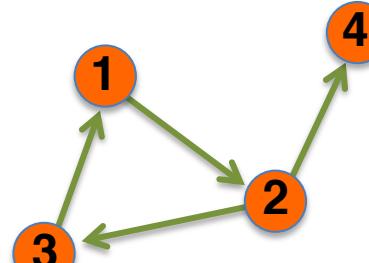
$$A_{ii} = 0$$

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

$$L = \frac{1}{2} \sum_{i,j=1}^N A_{ij} \quad \langle k \rangle = \frac{2L}{N}$$

*Red de actores, interacciones proteína-proteína*

## Dirigido



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

$$A_{ii} = 0$$

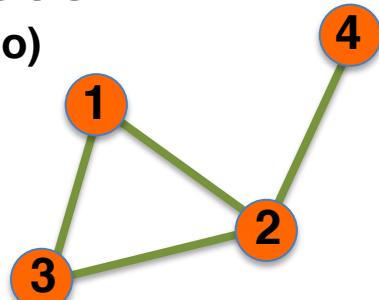
$$A_{ij} \neq A_{ji}$$

$$L = \sum_{i,j=1}^N A_{ij} \quad \langle k \rangle = \frac{L}{N}$$

*WWW, red de citas*

In=columnas  
Out=filas

## Sin pesos (no-dirigido)



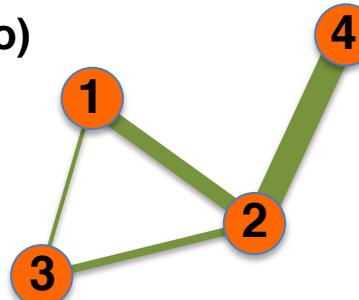
$$A_{ij} = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

$$A_{ii} = 0$$

$$L = \frac{1}{2} \sum_{i,j=1}^N A_{ij} \quad \langle k \rangle = \frac{2L}{N}$$

Interacciones proteína-proteína, www

## Con pesos (no-dirigido)



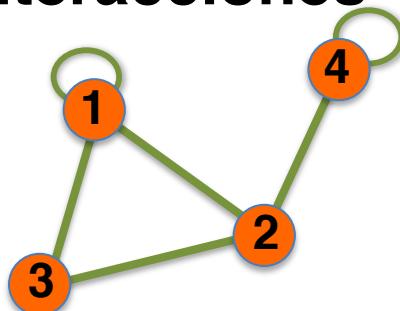
$$A_{ij} = \begin{pmatrix} 0 & 2 & 0.5 & 0 \\ 2 & 0 & 1 & 4 \\ 0.5 & 1 & 0 & 0 \\ 0 & 4 & 0 & 0 \end{pmatrix}$$

$$A_{ii} = 0$$

$$L = \frac{1}{2} \sum_{i,j=1}^N \text{nonzero}(A_{ij}) \quad \langle k \rangle = \frac{2L}{N}$$

Grafo de llamadas, red metabólica

## Auto-interacciones



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

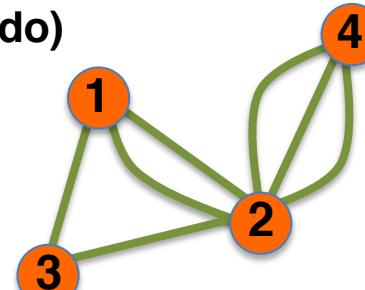
$$A_{ii} \neq 0$$

$$L = \frac{1}{2} \sum_{i,j=1, i \neq j}^N A_{ij} + \sum_{i=1}^N A_{ii}$$



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

## Multigrafo (no-dirigido)

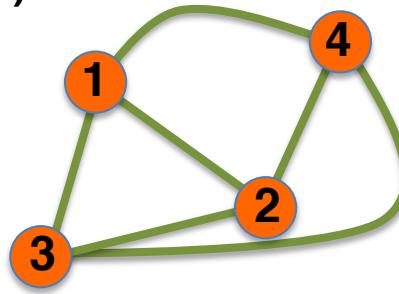


$$A_{ij} = \begin{pmatrix} 0 & 2 & 1 & 0 \\ 2 & 0 & 1 & 3 \\ 1 & 1 & 0 & 0 \\ 0 & 3 & 0 & 0 \end{pmatrix}$$

$$A_{ii} = 0$$

$$L = \frac{1}{2} \sum_{i,j=1}^N \text{nonzero}(A_{ij}) \quad \langle k \rangle = \frac{2L}{N}$$

# Grafo completo (no-dirigido)

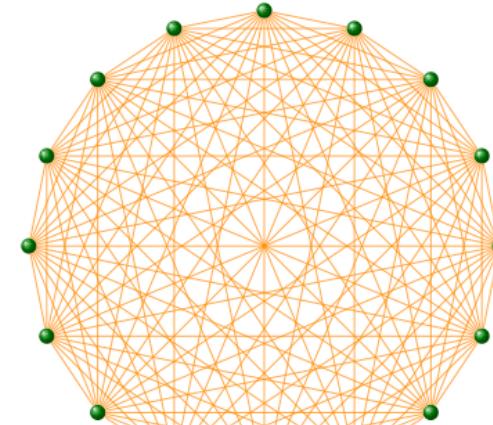


$$A_{ij} = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$

$$A_{ii} = 0 \qquad \qquad A_{i \neq j} = 1$$

$$L = L_{\max} = \frac{N(N-1)}{2} \qquad \langle k \rangle = N - 1$$

*Red de actores, Interacciones proteína-proteína*



# GRAPHOLOGY: Las redes reales pueden tener multiples características

WWW > Multigrafo dirigido con auto-interacciones

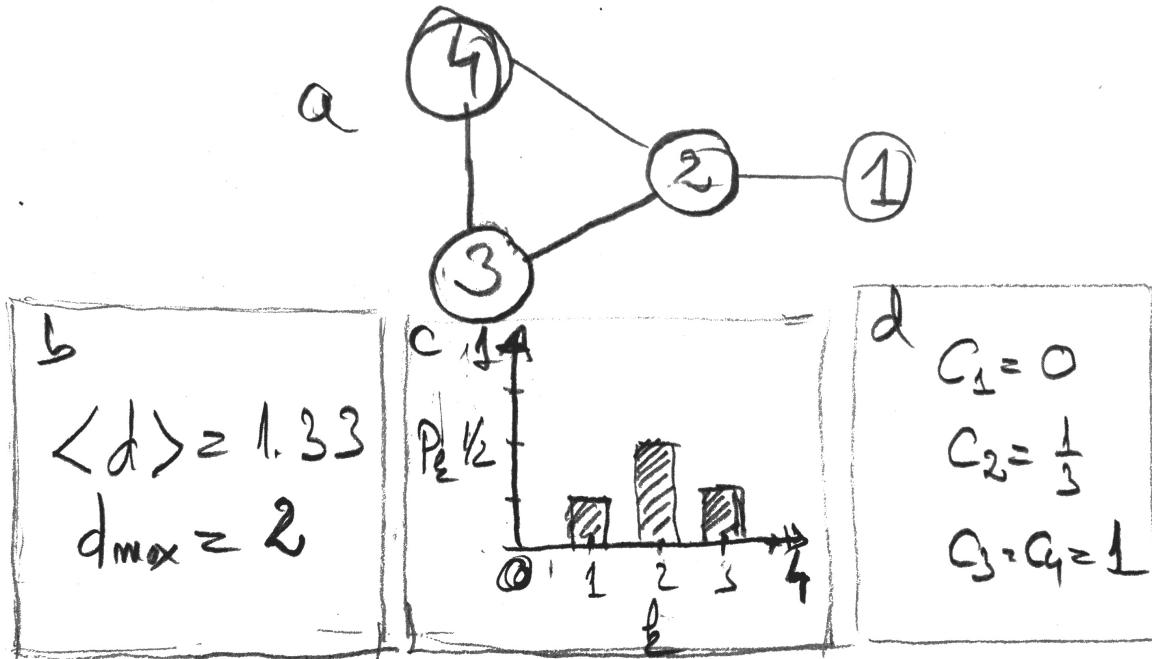
Protein Interactions > no-dirigido, sin pesos, con auto-interacciones

Collaboration network > no-dirigido, multigrafo or con pesos

Mobile phone calls > dirigido, con pesos.

Facebook Friendship links > no-dirigido, sin pesos.

# TRES CANTIDADES CENTRALES EN NETWORK SCIENCE



A. Distribución de grado:

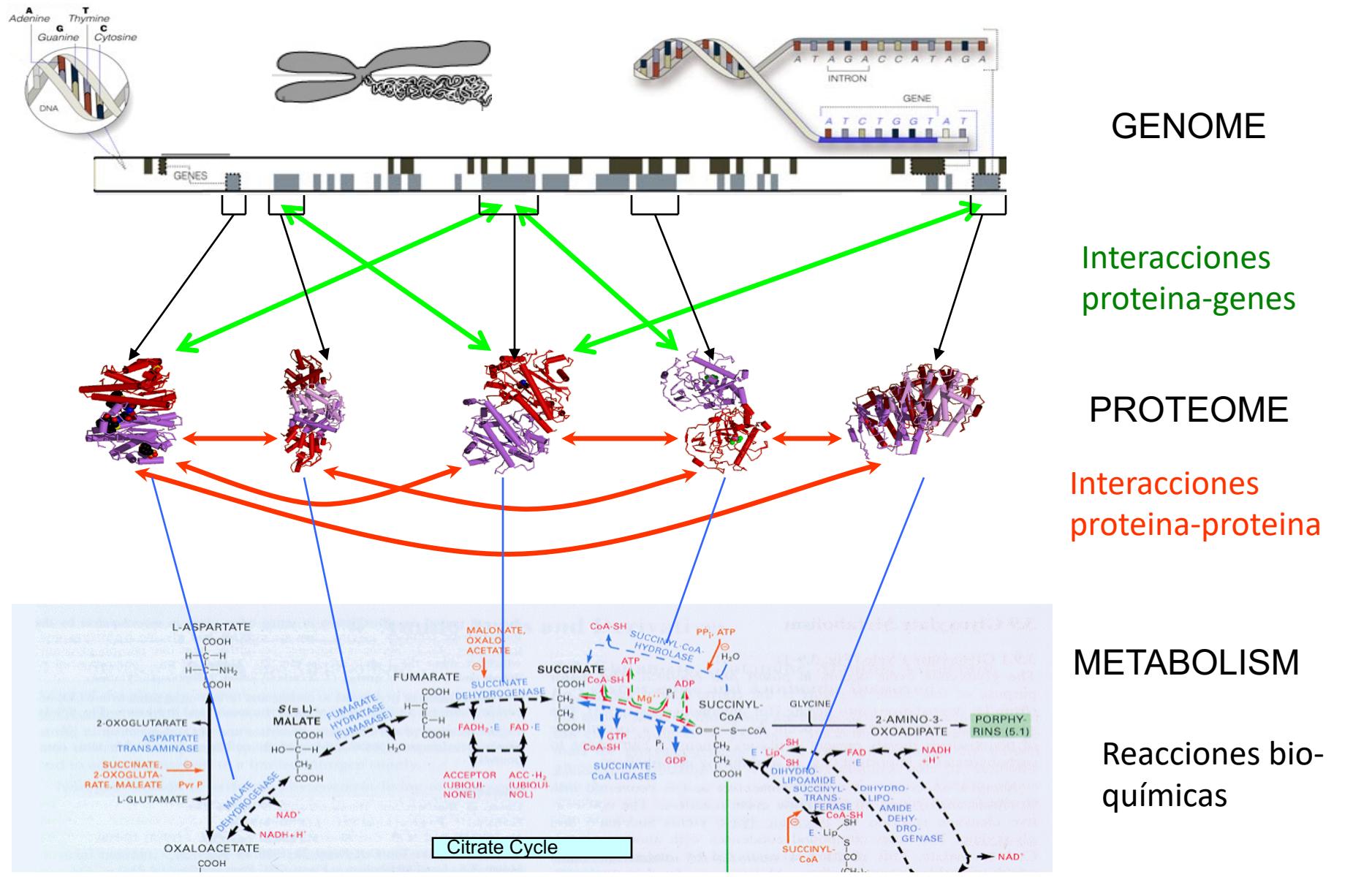
$$p_k$$

B. Longitud de camino:

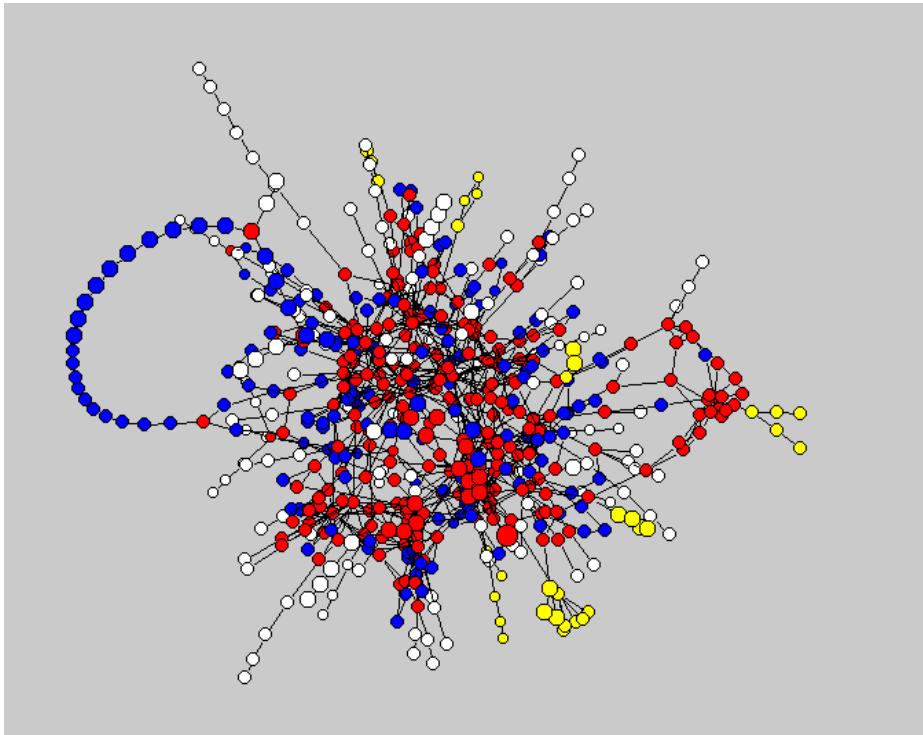
$$\langle d \rangle$$

C. Coeficiente de clustering:

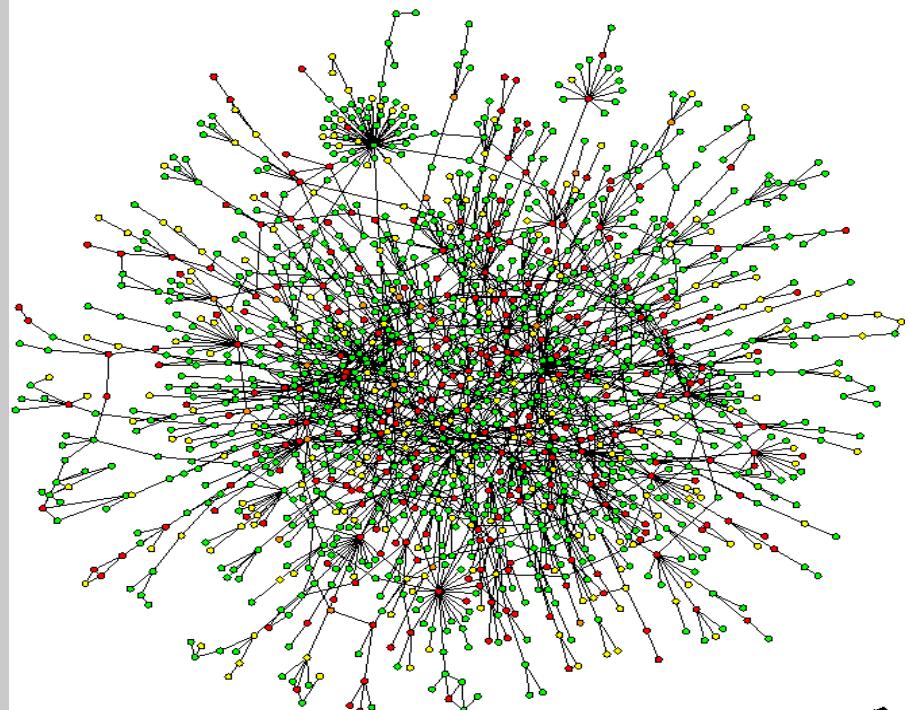
$$C_i = \frac{2e_i}{k_i(k_i - 1)}$$



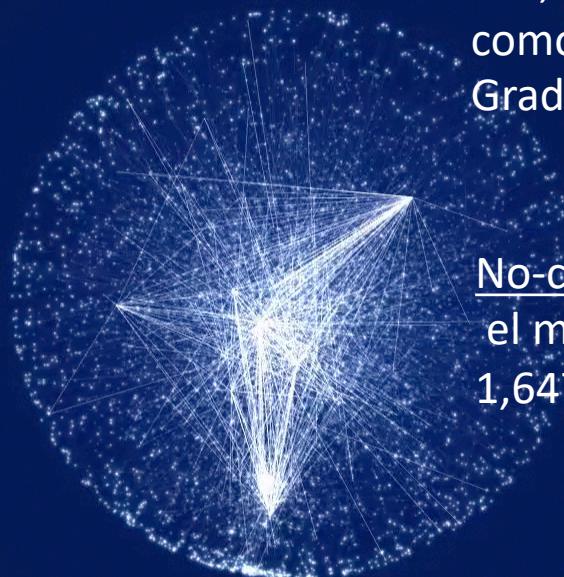
# Red metabólica



# Interacciones entre Proteinas



# A CASE STUDY: PROTEIN-PROTEIN INTERACTION NETWORK



Red no-dirigida

N=2,018 proteinas como nodos

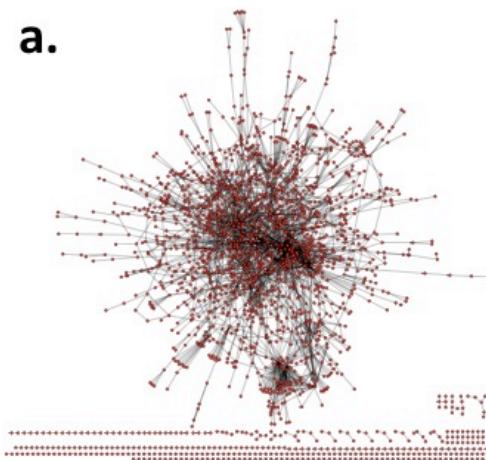
L=2,930 interacciones vinculantes  
como links.

Grado promedio  $\langle k \rangle = 2.90$ .

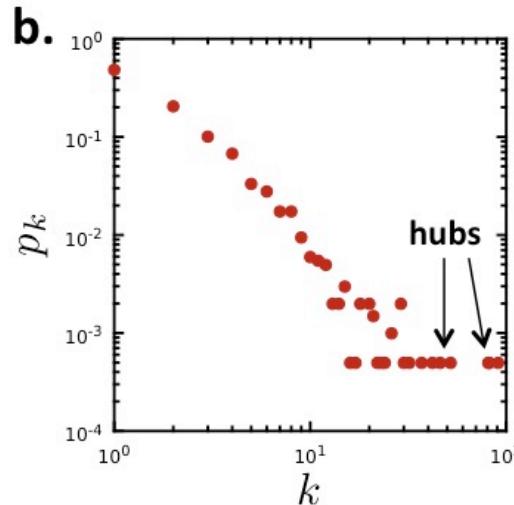
No-conectado: 185 componentes  
el más grande (componente gigante)  
1,647/2018 nodos

# UN CASO DE ESTUDIO: INTERACCION PROTEINA-PROTEINA

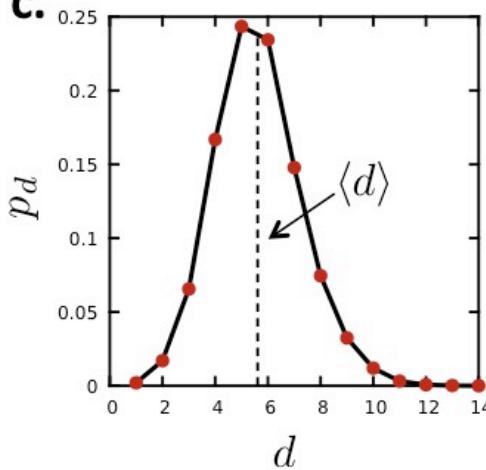
a.



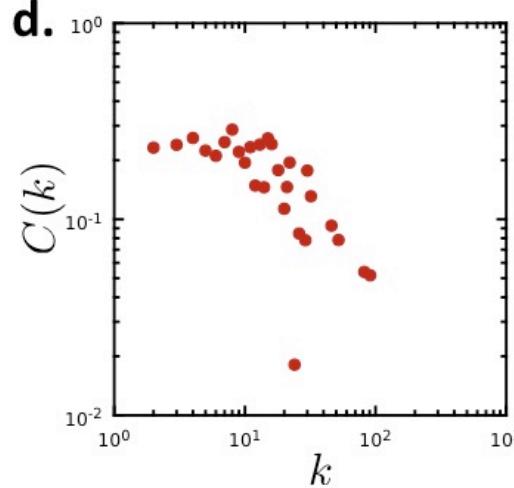
b.



c.



d.



Red no-dirigida

N=2,018 proteínas como nodos

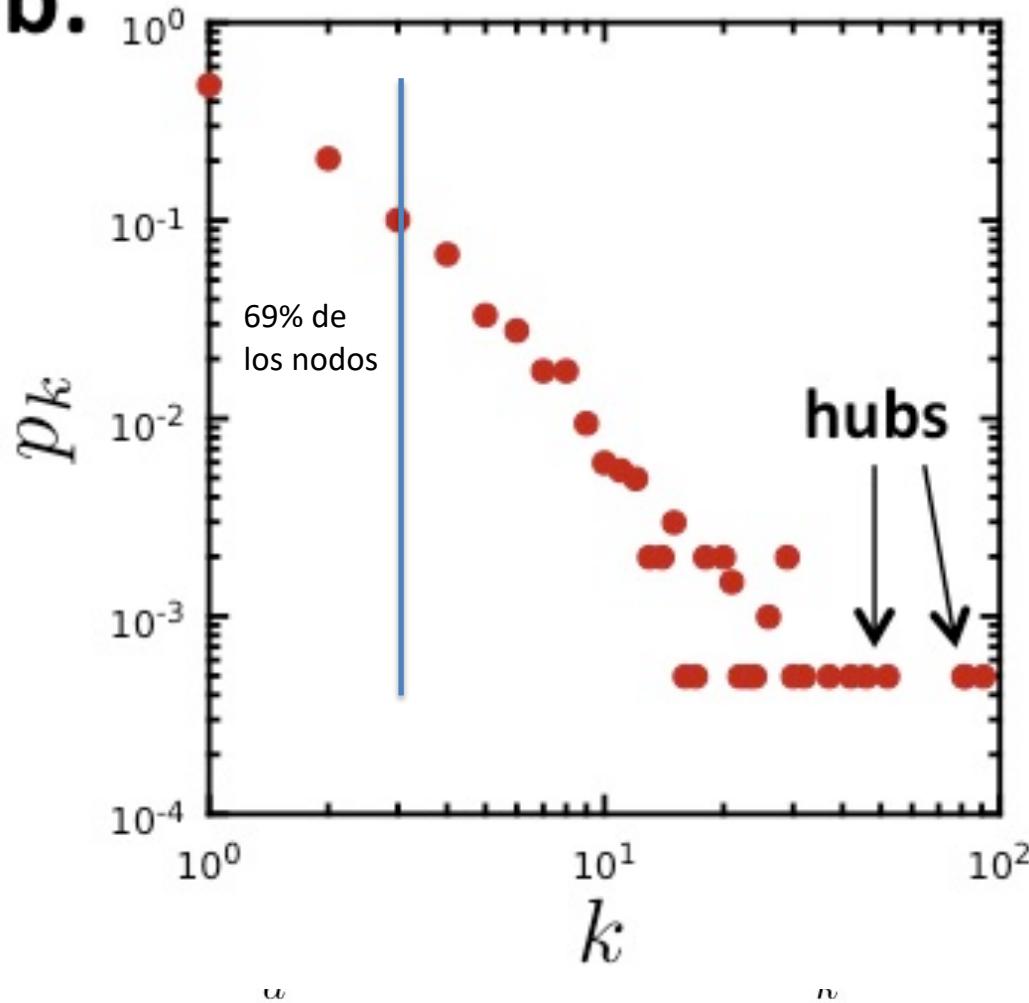
L=2,930 interacciones vinculantes como links.

Grado promedio  $\langle k \rangle = 2.90$ .

No conectado: 185 componentes  
las más grande largest (componente gigante) 1,647 nodos

# UN CASO DE ESTUDIO: INTERACCION PROTEINA-PROTEINA

b.



$p_k$  es la probabilidad de que un nodo tenga grado  $k$ .

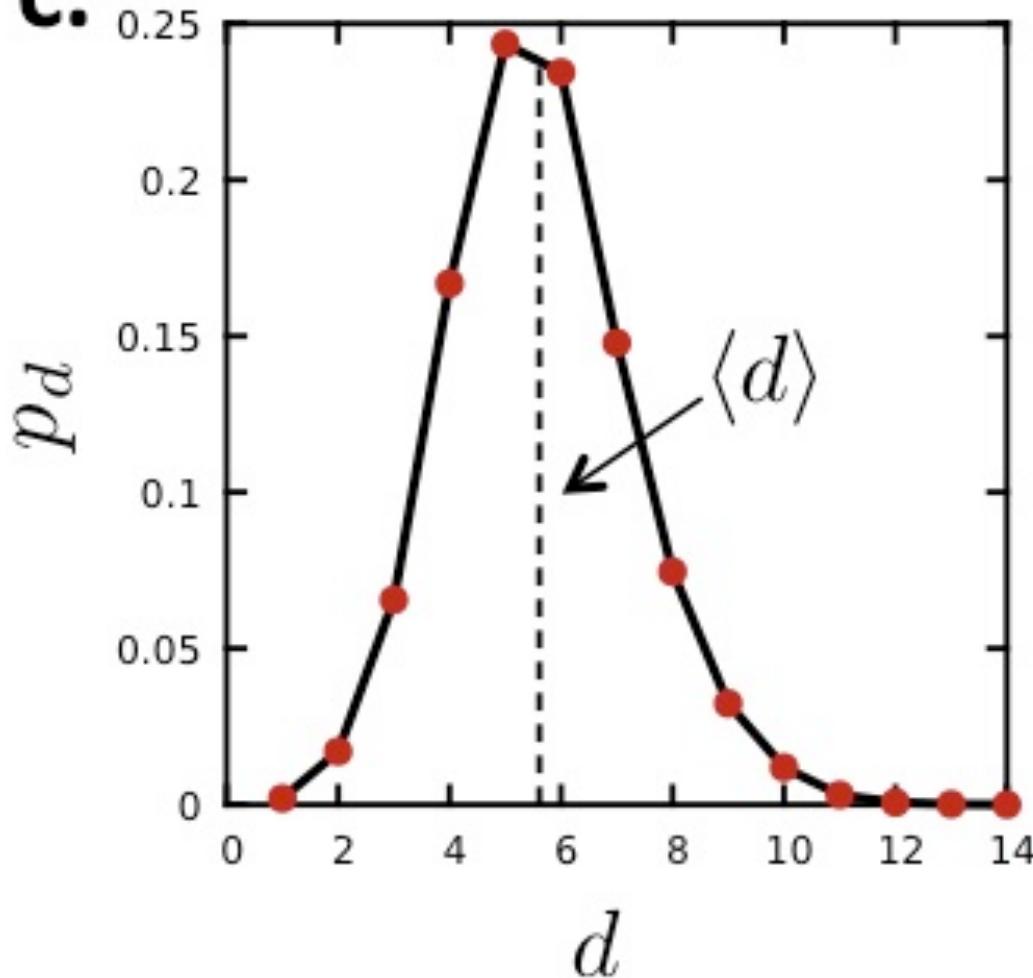
$N_k = \# \text{ nodos con grado } k$

$$p_k = N_k / N$$

Propiedad libre de escala (scale free)

# UN CASO DE ESTUDIO: INTERACCION PROTEINA-PROTEINA

C.



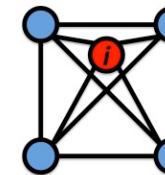
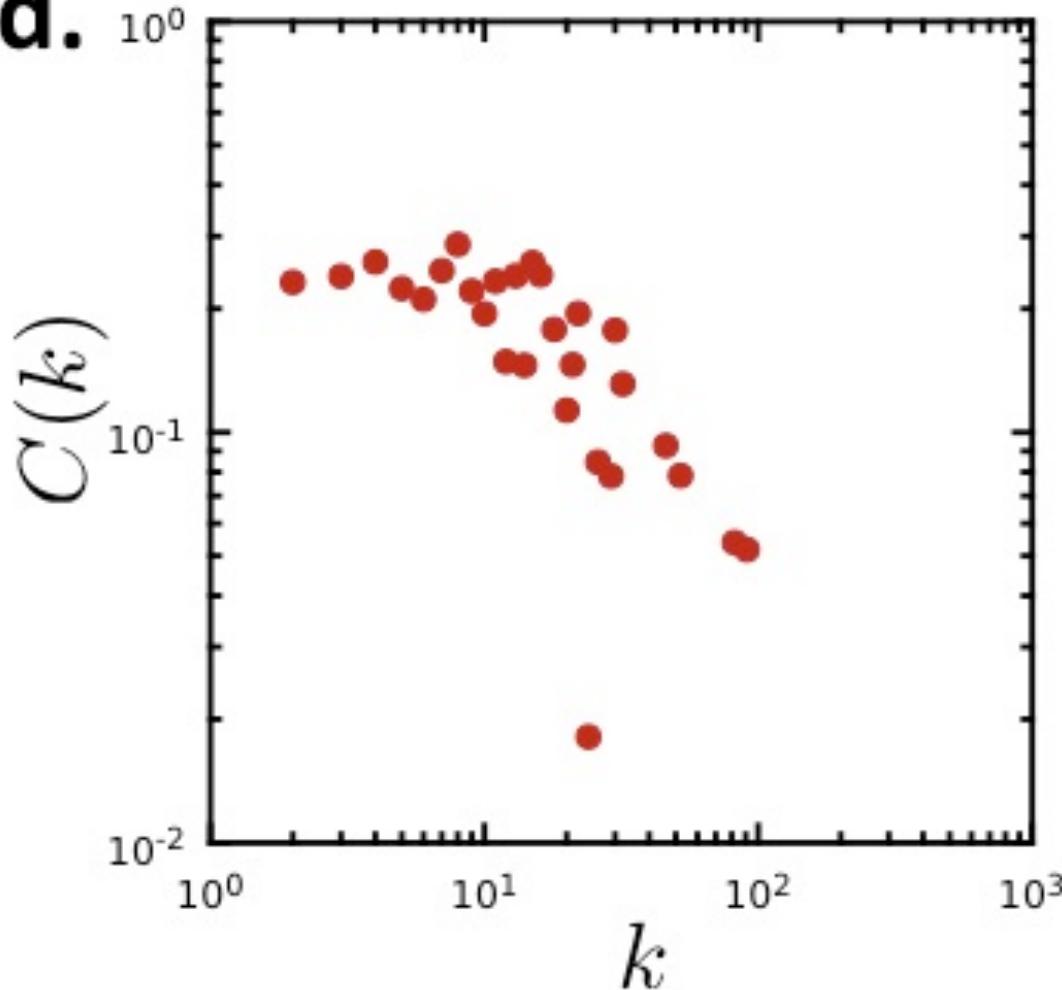
$$d_{\max} = 14$$

$$\langle d \rangle = 5.61$$

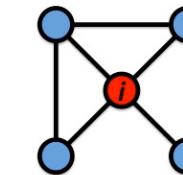
Propiedad de mundo pequeño

# UN CASO DE ESTUDIO: INTERACCION PROTEINA-PROTEINA

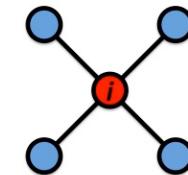
d.



$$C_i = 1$$



$$C_i = 1/2$$



$$C_i = 0$$

$$C_i = \frac{2e_i}{k_i(k_i - 1)}$$

$$\langle C \rangle = 0.12$$

Propiedad de jerarquía

# PROYECTOS FINALES

# COMPONENTES DEL PROYECTO

## 1. ADQUISICIÓN DE DATOS

Descargar la data y ponerla en un formato usable

## 2. REPRESENTACIÓN DE LA RED

Qué representan los nodos y los links?

## 3. Análisis de redes

Qué preguntas quieres responder con esta red,  
y que herramientas/medidas usarás?

## ADQUISICIÓN DE DATOS

- Muchas fuentes de datos en línea tendrán una API (interfaz de programación de aplicaciones) que permite consultar y descargar los datos de forma específica
  - Ejemplo: ¿Cuáles son todas las películas de 1984-1995 protagonizadas por Kevin Bacon y distribuidas por Paramount Pictures?
  - Esto se hace a través de una interfaz web o de una biblioteca dentro de un lenguaje de programación
- Otras fuentes proporcionarán datos en bruto sin procesar (por ejemplo, hojas de cálculo de Excel) que requieren procesamiento, ya sea manualmente o mediante un programa

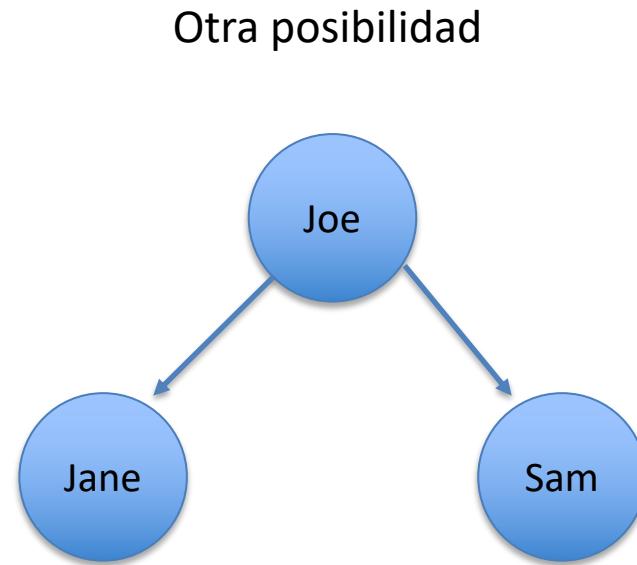
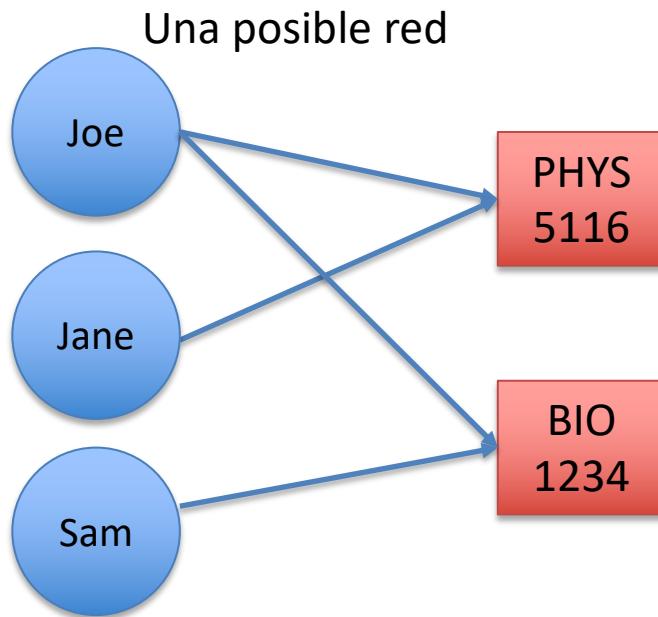
# “GRAPH” ≠ “NETWORK”

- La mayoría de los conjuntos de datos admitirán más de una representación como una red
- Algunas representaciones serán más o menos informativas que otras.
- ¡Descubrir la "red" que está oculta en sus datos es parte de su proyecto!

## RECONSTRUCCION DE RED

“GRAPH”  $\neq$  “NETWORK”

Supongamos que tiene una lista de estudiantes y los cursos para los que están registrados



# EJEMPLOS

goodreads

Meet your next  
favorite book.



# LIBROS

- Como IMDB para libros (contiene libros, calificaciones, reseñas, recomendaciones, etc.)
- API disponible en  
<https://www.goodreads.com/api>
- Áreas potenciales de investigación:
  - Red de similitud de libros.
  - Detección de comunidades (descubrir géneros)

# Comics

Global comics database

<http://www.comics.org/>



## Comics

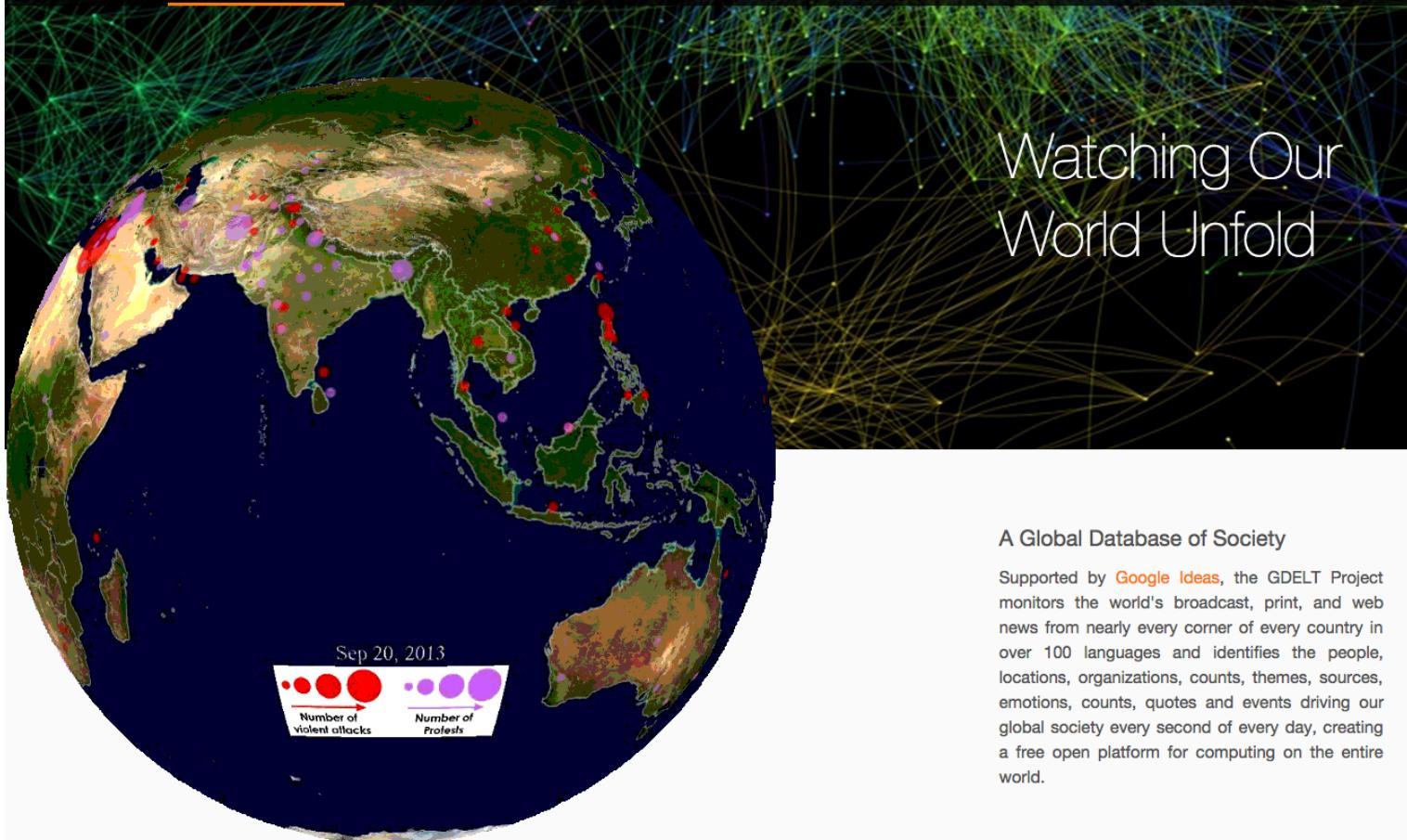
- Muchos datos diferentes sobre cada cómic, por ejemplo:
  - Editor
  - Quien escribió guión / lápiz / tinta
  - Fecha de publicación
- Wiki y la interfaz de búsqueda avanzada disponibles
- Áreas potenciales de investigación:
  - Comics vinculados por personajes comunes.
  - Red de colaboración entre artistas.



<http://www.mendeley.com/>

- Gran base de datos de publicaciones científicas / red social para investigadores.
- API disponible ([dev.mendeley.com](http://dev.mendeley.com))
- Idea: utilizar lectores para asignar crédito de autoría
  - Los datos consisten en perfiles de usuario + documentos que el usuario ha leído.
  - Las publicaciones (nodos) están vinculadas si ambas están presentes en una o más listas de usuarios
  - Use técnicas desarrolladas recientemente para inferir el crédito de autoría basado en la percepción del usuario:  
[\(http://www.pnas.org/content/111/34/12325.abstract\)](http://www.pnas.org/content/111/34/12325.abstract)

## The GDELT Project

[Blog](#)[Data](#)[Solutions](#)[About](#)[Intro](#)[Watching](#)[Computing](#)[Downloading](#)[Blogging](#)[Starting](#)

### A Global Database of Society

Supported by [Google Ideas](#), the GDELT Project monitors the world's broadcast, print, and web news from nearly every corner of every country in over 100 languages and identifies the people, locations, organizations, counts, themes, sources, emotions, counts, quotes and events driving our global society every second of every day, creating a free open platform for computing on the entire world.

## The GDEL Project

[Blog](#)[Data](#)[Solutions](#)[About](#)

- Es un conjunto de datos de monitoreo de noticias (transmisión, impresión y web) desde 1979 hasta hoy en todo el mundo. Identifica nombres, lugares, organizaciones, emociones, cuentas.
- Ofrecen archivos de datos en bruto y / o posibilidad de consultar una base de datos.
- Proyectos: (i) estudiar la red individuo - individuo (dos personas están conectadas si aparecen en las mismas noticias) a lo largo del tiempo, vea cómo emergen los líderes. (ii) estudiar la red de ubicaciones, con dos ubicaciones conectadas si se informa la misma noticia. ¿Cómo viajan las noticias sobre el espacio?
- El conjunto de datos se puede utilizar para muchos más proyectos!



A Global Database of Society

At its core, the GDEL Project monitors the world's broadcast, print, and web news from nearly every corner of every country in over 100 languages and identifies the people, locations, organizations, counts, themes, sources, emotions, counts, quotes and events driving our global society every second of every day, creating a free open platform for computing on the entire world.

# Baseball



<http://seanlahman.com/baseball-archive/statistics/>

# Baseball

- Amplia base de datos de estadísticas, a nivel de jugador (estadísticas individuales) y a nivel de equipo (composiciones de equipo, salón de la fama, gerentes, etc.)
- No obstante, posibles direcciones de investigación.
  - ¿Hay características de la red que distinguen a los miembros del Salón de la Fama?
  - Movilidad de jugadores / directivos entre equipos.

# Lineamientos finales del proyecto

Medida: N (t), L (t) [t-tiempo si tiene un sistema dependiente del tiempo); P (k) (distribución en grados);  $\langle l \rangle$  longitud de camino promedio; C (coeficiente de agrupamiento), C\_rand, C (k); Visualización / comunidades; P(w) si tiene una red ponderada; robustez de la red (si corresponde); propagación (si es apropiado).

No es suficiente medir las cosas, es necesario discutir las ideas que ellas ofrecen (significado):

¿Qué aprendiste de cada cantidad que mediste?

¿Cuáles fueron tus expectativas?

¿Cómo se comparan los resultados con tus expectativas?

La restricción de tiempo serán estrictas. Aproximadamente 10min + 3 min preguntas;

Es necesario escribir un informe y también se entrega la presentación (google slides).

Deben enviar un email con nombres / títulos / nombre curso con 24 horas antes de la presentación (intermedia y final)

La primera diapositiva debe contener nombres / títulos / programa.

Ven antes y prueba tus diapositivas con el proyector (es tu responsabilidad comunicar la información de la mejor manera)

## Criterio de evaluación:

Uso de herramientas de red (integridad / uso correcto);

Capacidad para extraer información/perspectivas de sus datos utilizando las herramientas de red;

**(data != información)**

Calidad general del proyecto / presentación.