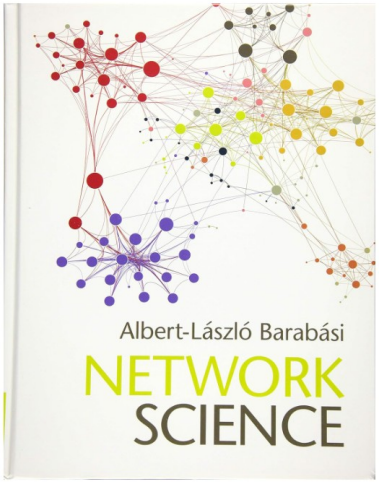
IIC-3641 Aprendizaje Automático basado en Grafos

https://github.com/marcelomendoza/IIC3641

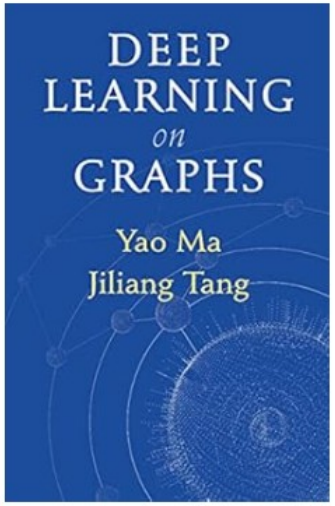
- PRELIMINARES -

Referencias bibliográficas



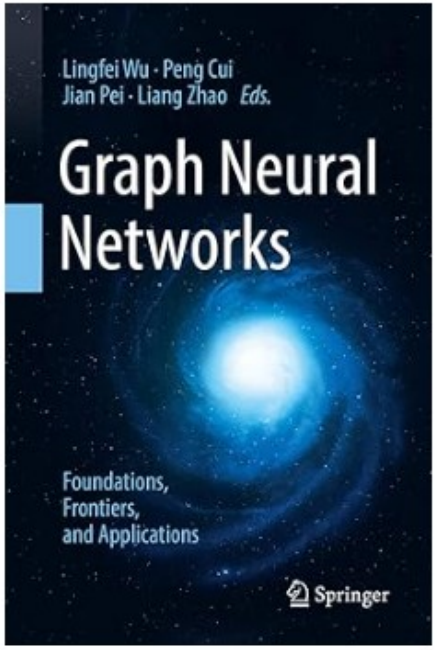
Unidad 1: Modelos de grafos

Barabási, Albert-Lásló. Network Science. Cambridge University Press, 2016.



Unidad 2: Aprendizaje de representaciones en grafos

Ma, Yao & Tang, Jiliang. Deep learning on graphs. Cambridge University Press, 2021



Unidad 3: Redes neuronales artificiales con grafos

Wu, Lingfei; Cui, Peng; Pei, Jian & Zhao, Liang. Graph Neural

Networks: Foundations, Frontiers, and Applications. Springer, 2022.

Grado, grado promedio y distribución de grado

Se denota con la variable k_i el grado del i-th nodo de un grafo:

Se denota con la variable L el número total de enlaces del grafo. En una red no dirigida se calcula según:

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

Evita contar dos veces cada enlace

Grado, grado promedio y distribución de grado

Se denota con la variable <*k*> el grado promedio del grafo. En una red no dirigida se calcula según:

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

En redes dirigidas se distingue entre grado de entrada k_i^{in} y grado de salida k_i^{out}, los que representan el número de enlaces que apuntan al nodo *i* y que salen del nodo *i*, respectivamente. En estos grafos, el grado total del nodo *i* está dado por:

$$k_i = k_i^{\text{in}} + k_i^{\text{out}}.$$

Notemos que el número total de enlaces en un grafo dirigido es:

$$L = \sum_{i=1}^{N} k_i^{in} = \sum_{i=1}^{N} k_i^{out} \quad .$$

De esta misma forma, el grado promedio en un grafo dirigido está dado por:

$$\langle k^{\text{in}} \rangle = \frac{1}{N}\sum_{i=1}^{N} k_i^{\text{in}} = \langle k^{\text{out}} \rangle = \frac{1}{N}\sum_{i=1}^{N} k_i^{\text{out}} = \frac{L}{N}$$

Grado, grado promedio y distribución de grado

Se denota con la variable p_k la distribución de grado. Esta variable indica la probabilidad de que al muestrear un nodo al azar del grafo su grado sea k. Dado que p_k es una probabilidad, se cumple que:

$$\sum_{k=1}^{\infty} p_k = 1 \, .$$

La distribución de grado se calcula en base a la frecuencia relativa:

$$p_k = \frac{N_k}{N} \, ,$$

donde N_k es el número de nodos del grafo que tienen grado k.

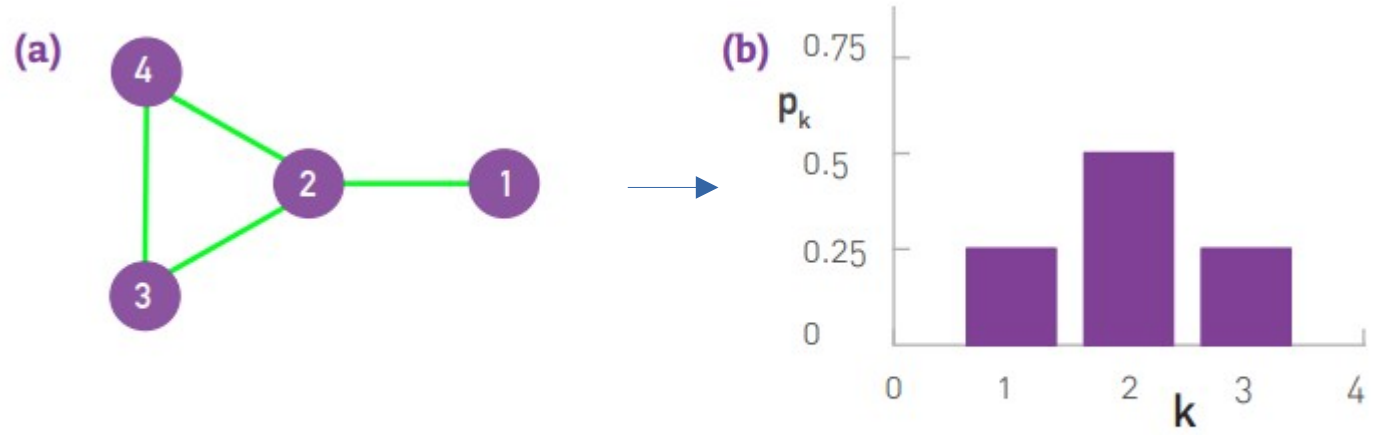
Notemos que el grado promedio se puede calcular sobre la base de la distribución de grado:

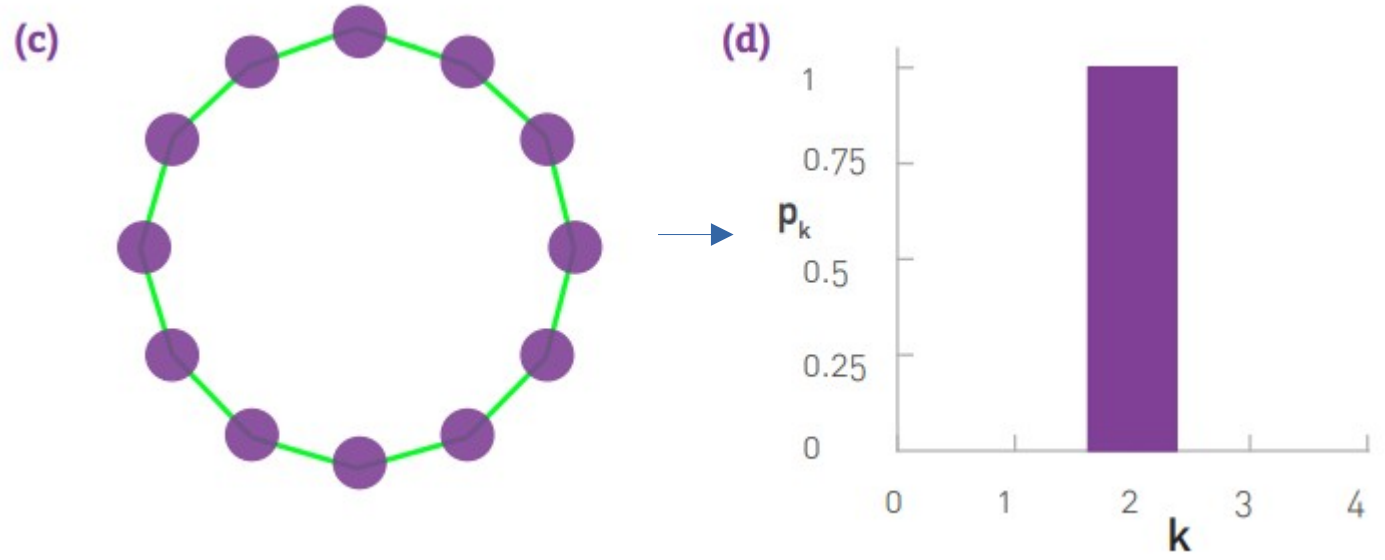
$$\langle k \rangle = \sum_{k=0}^{\infty} k p_k$$

La distribución de grado es muy importante ya que determina varios fenómenos, como la robustez de la red y la viralidad de la estructura de la red.

Grado, grado promedio y distribución de grado

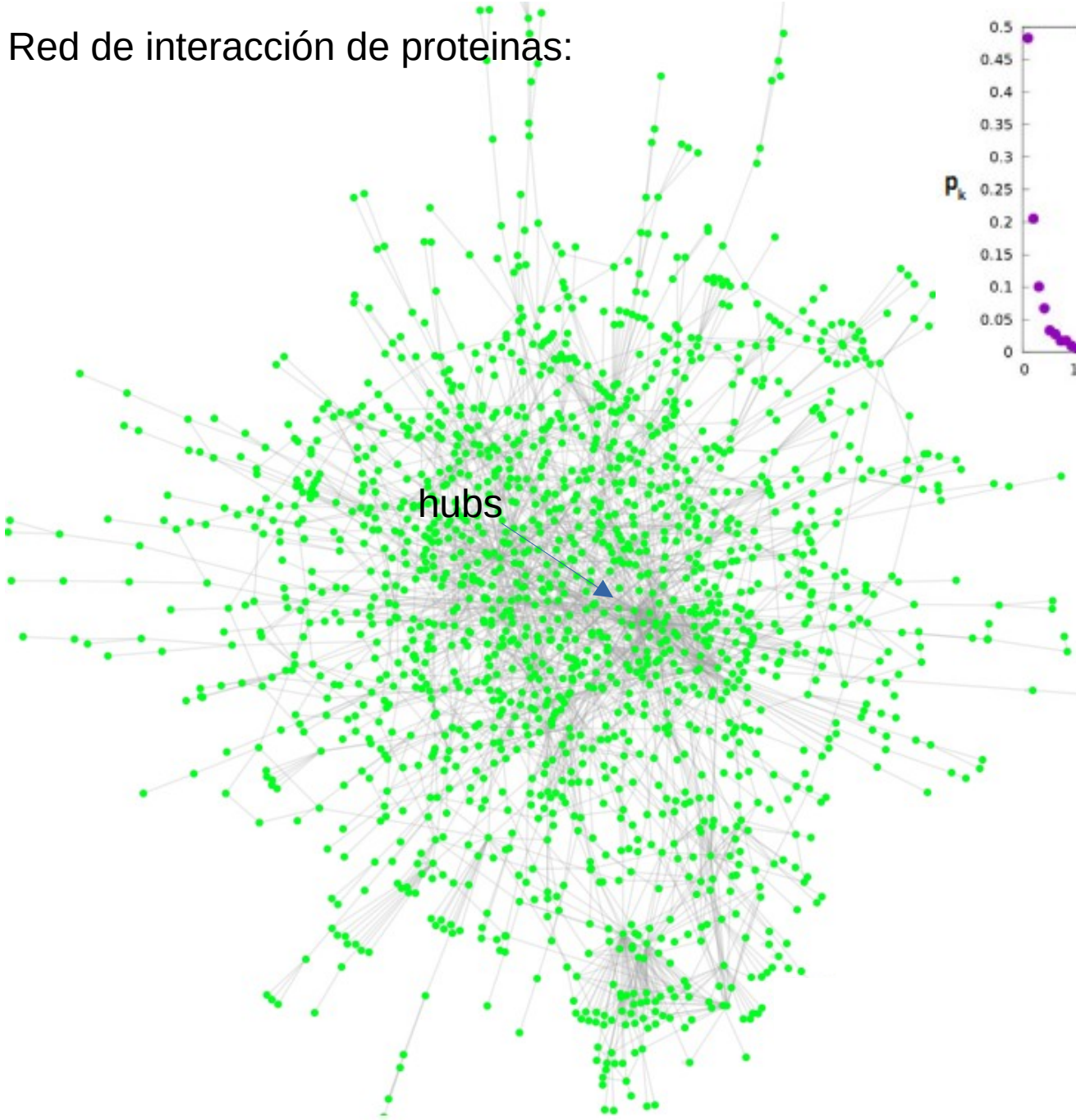
Algunos ejemplos de distribución de grado:

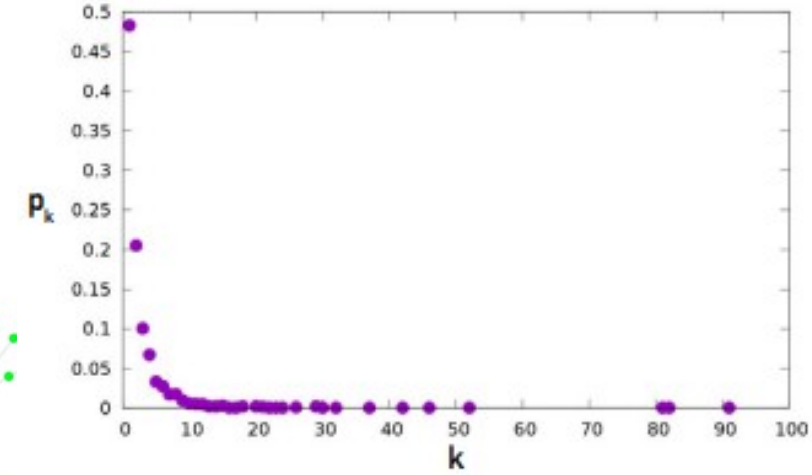




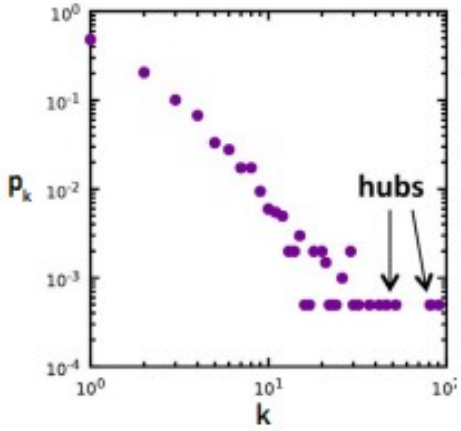
Grado, grado promedio y distribución de grado

Red de interacción de proteinas:





log-log



Grado, grado promedio y distribución de grado

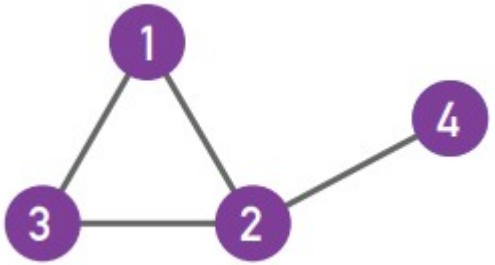
Muestras de redes:

NETWORK	NODES	LINKS	DIRECTED UNDIRECTED	N	L	⟨k⟩
Internet	Routers	Internet connections	Undirected	192,244	609,066	6.34
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

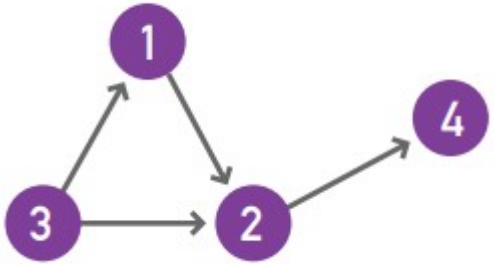
Matriz de adyacencia

A_{ij} = *1* si existe un enlace entre el nodo *i* y el *j*, 0 si no existe.

$$A_{ij} = \begin{pmatrix} A_{11} & A_{12} & A_{13} & A_{14} \\ A_{21} & A_{22} & A_{23} & A_{24} \\ A_{31} & A_{32} & A_{33} & A_{34} \\ A_{41} & A_{42} & A_{43} & A_{44} \end{pmatrix}$$



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



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

Notar que en redes dirigidas se cumple que: $k_i^{\text{in}} = \sum_{j=1}^{N} A_{ij}$, $k_i^{\text{out}} = \sum_{j=1}^{N} A_{ji}$.

y en no dirigidas: $k_i = \sum_{j=1}^{N} A_{ji} = \sum_{i=1}^{N} A_{ji}$. $2L = \sum_{i=1}^{N} k_i^{\text{in}} = \sum_{i=1}^{N} k_i^{\text{out}} = \sum_{ij} A_{ij}$.

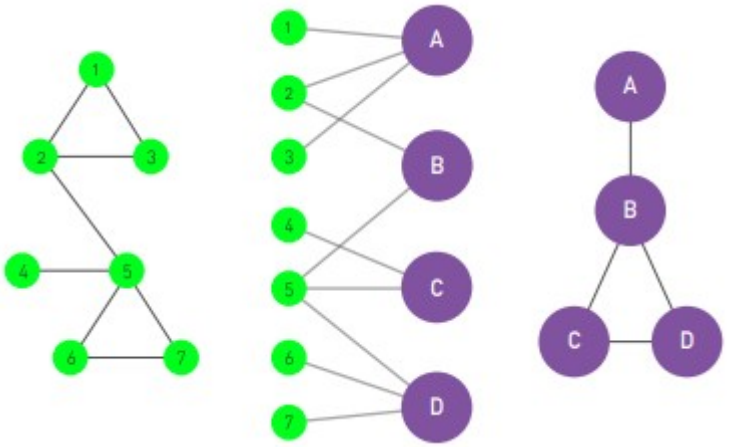
Grafos con pesos

En algunas redes nos interesará definir un peso para cada enlace, es decir, $A_{ij} = w_{ij}$ (power grids, comms). Típicamente son grafos dirigidos.

Grafos bipartitos

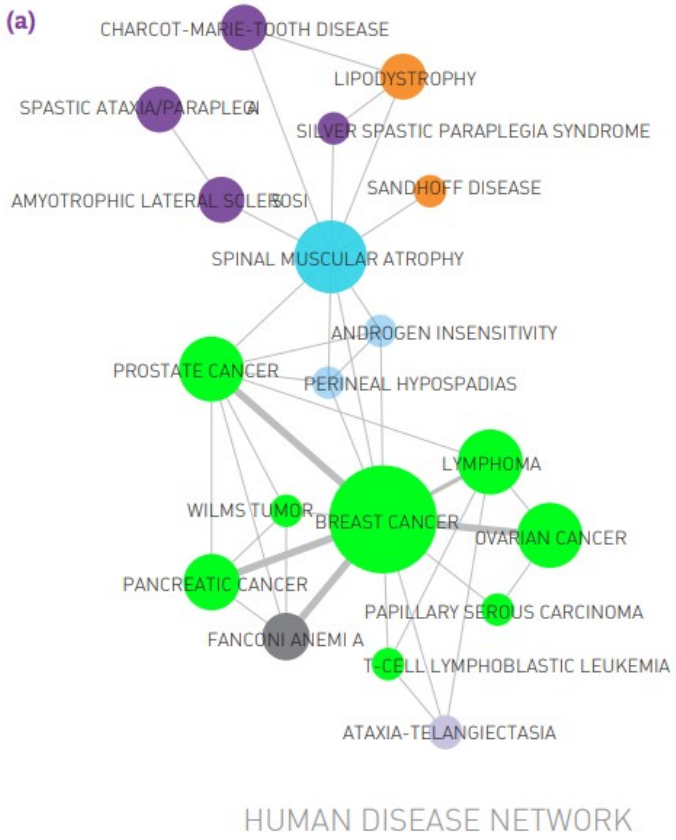
Una grafo bipartito es un grafo en el cual sus nodos se subdividen en dos subconjuntos disjuntos *U* y *V*. Todo enlace del grafo conecta nodos de *U* y *V*, es decir, no existen enlaces entre pares de nodos de *U* o *V*.

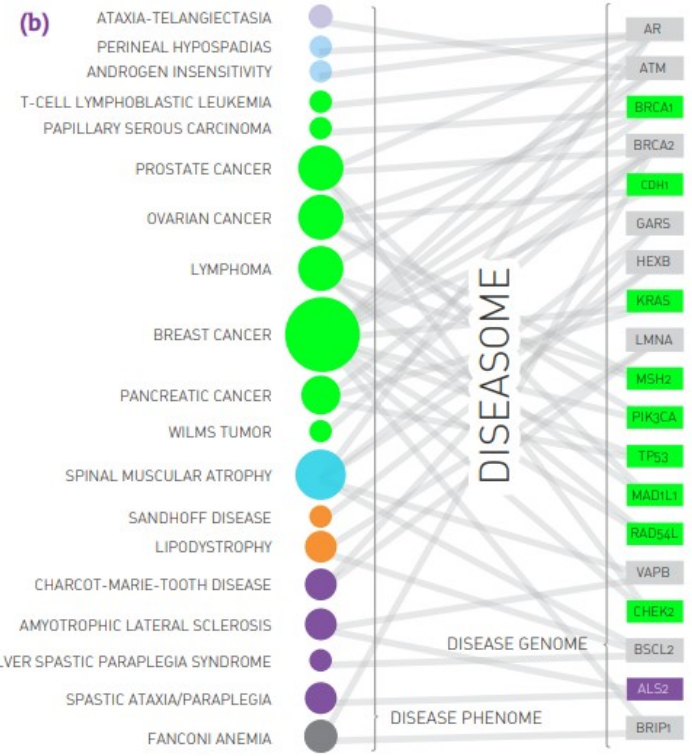
Se pueden proyectar los subconjuntos U y V si el mismo nodo del subconjunto contrario los conecta.

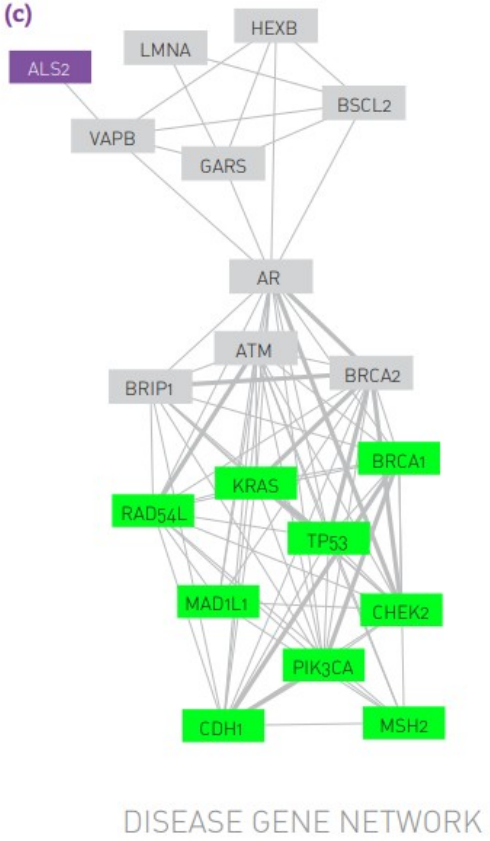


Proj U *U* *V* *Proj V*

Grafo bipartito (human disease)







Proj U *U* *V* *Proj V*

Caminos y distancias

Un camino en un grafo entre los nodos *i* y *j* es una lista ordenada de enlaces que los conecta. El largo del camino es el número de enlaces que lo conforman.

El camino más corto entre los nodos *i* y *j* es el camino que tiene el menor número de enlaces. También se le comoce como la distancia entre *i* y *j*.

El diámetro de un grafo es la distancia entre el par de nodos más lejanos. Se denota por d_{max}.

La distancia promedio de un grafo es el promedio de las distancias entre todos los pares de nodos del grafo. Se denota por <*d*> y se calcula según:

$$d = \frac{1}{N(N-1)} \sum_{\substack{i,j=1,N \\ i \neq j}} d_{i,j} \,. \qquad \text{(grafos dirigidos)}$$

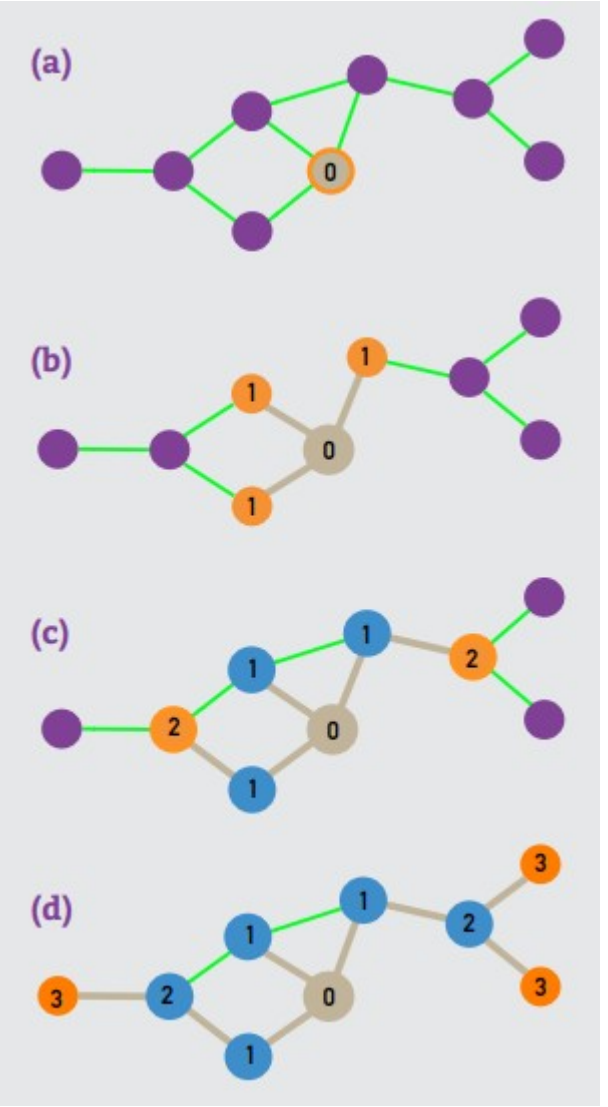
Ciclo: un camino cuyo nodo de inicio y término es el mismo.

Camino Euleriano: Un camino que atraviesa cada enlace de ctamente una vez.

Camino Hamiltoniano: un camino que visita cada nodo del grafo exactamente una vez.

Conectividad

Breadth-First Search (BFS): BFS inicia en un nodo y etiqueta a sus vecinos. Luego se etiqueta a los vecinos de los vecinos, … BFS termina cuando todos los nodos del grafo están etiquetados. Las etiquetas corresponden a la distancia hacia el nodo inicial.

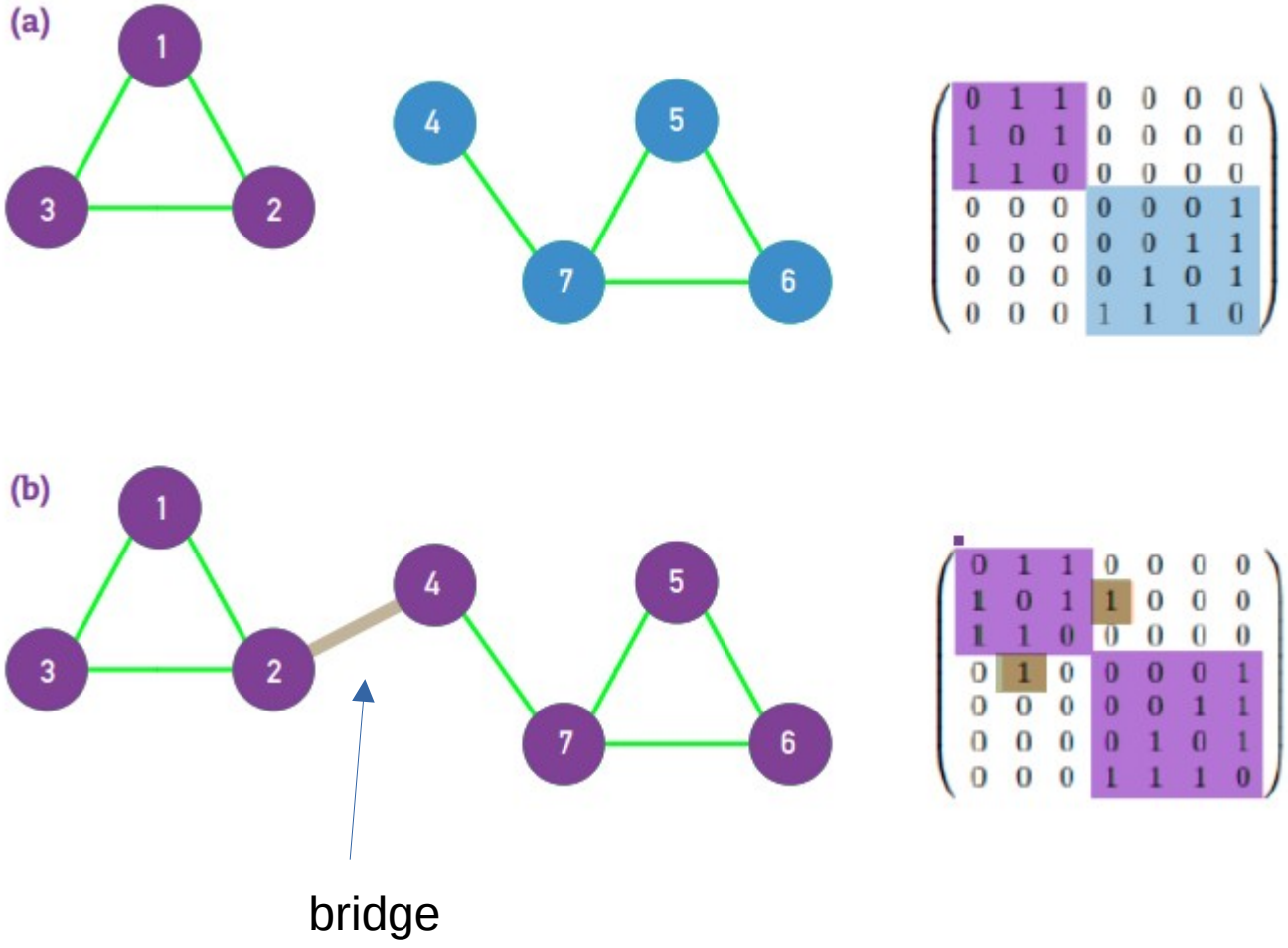


Una grafo es **conexo** si todos los nodos del grafo son etiquetables por BFS.

Si un grafo requiere correr más de un BFS para etiquetar todos sus nodos, es disconexo. Cada componente conexa se puede unir a otra usando un enlace denominado **bridge**.

Conectividad

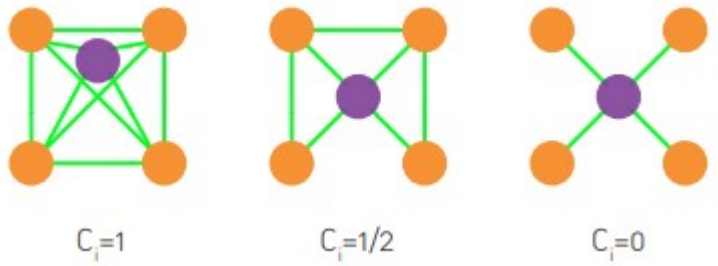
En un grafo disconexo podemos unir sus componentes con enlaces de tipo **bridge**.



Coeficiente de clustering

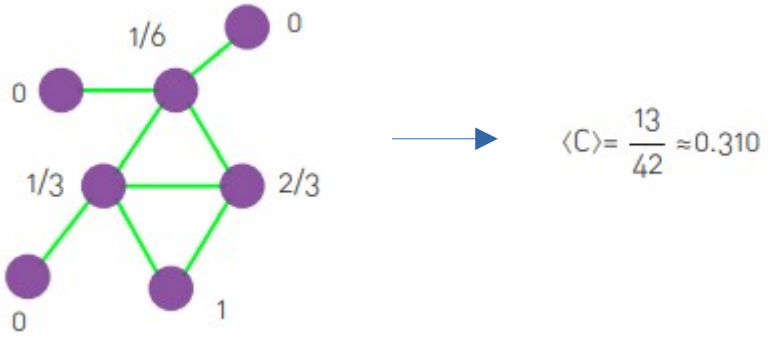
El coeficiente de clustering representa la densidad local de un nodo. Se calcula mediante el coeficiente entre dos veces el número de enlaces entre los vecinos del nodo *i* y la máxima conectividad alcanzable por k_i nodos, donde k_i es el grado del nodo *i*:

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



Se define el coeficiente de clustering medio del grafo, denominado por <*C*>, como el promedio sobre las densidades locales de los nodos del grafo:

$$\langle C \rangle = \frac{1}{N}\sum_{i=1}^{N} C_i .$$



- RANDOM NETWORKS -

Modelos de grafos

Un modelo de grafo permite explicar las reglas mediante las cuales la red se conecta. El modelo será tan cercano a una red real en la medida que sus propiedades sean similares.

Un primer modelo con el que trabajeremos se denomina **Random Network Model**. Estos modelos siguien esta lógica:

1) Comenzamos con N nodos sin enlaces.

2) Seleccionamos un par de nodos al azar. Luego generamos un número al azar entre 0 y 1. Si el número excede a *p* (parámetro del modelo), conectamos los nodos, en otro caso, los dejamos desconectados.

3) Repetimos el paso 2) para todos los restantes pares de nodos (N(N-1)/2 steps en total).

A este tipo de grafos se les conoce como Random Graphs, o también grafos Erdös-Rényi.

Grafos Erdös-Rényi

Observemos lo siguiente:

1) La probabilidad de que L de los intentos conecten pares de nodos es p^L.

2) La probabilidad de que los otros pares no queden conectados es $(1-p)^{N(N-1)/2-L}$.

3) El número total de forma en las cuales podemos colocar L enlaces en un grafo con N nodos está dado por el coeficiente combinatorial:

$$\binom{\frac{N(N-1)}{2}}{L}$$

Luego, la probabilidad de una realización particular de G(N,P) con exactamente L enlaces es:

$$p_L = \binom{\frac{N(N-1)}{2}}{L} p^L (1-p)^{\frac{N(N-1)}{2}-L}$$

Como p_L sigue una distribución binomial, el número esperado de enlaces en un grafo Erdös-Rényi es:

$$\langle L \rangle = \sum_{L=0}^{\frac{N(N-1)}{2}} L p_L = p\frac{N(N-1)}{2}.$$

Notamos que podemos deducir el grado promedio: $\langle k \rangle = \frac{2\langle L \rangle}{N} = p(N-1).$

Grafos Erdös-Rényi (distribución de grado)

La probabilidad de que un nodo *i* tenga exactamente *k* enlaces depende de:

1) La probabilidad de que *i* tenga *k* enlaces, es decir, p^k.

2) La probabilidad de que los remanentes *(N-1-k)* no existan, es decir, $(1-p)^{N-1-k}$.

3) El número de formas que podemos seleccionar *k* enlaces desde los potenciales *N-1* enlaces posibles:

$$\binom{N-1}{k}$$

Luego, la distribución de grado de una random network sigue la binomial:

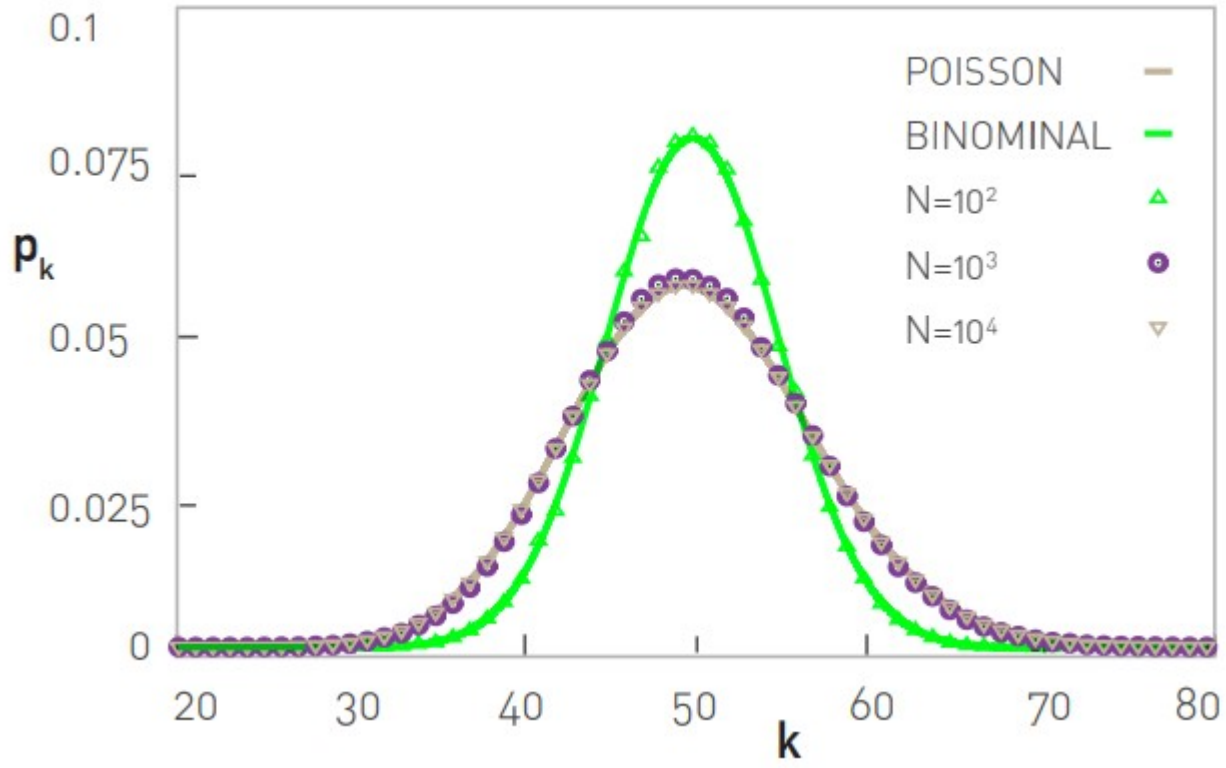
$$p_k = \binom{N-1}{k} p^k (1-p)^{N-1-k}.$$

Si el grafo es sparse, es decir $\langle k \rangle \ll N$, la binomial se aproxima por una Poisson, es decir:

$$p_k = e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!}$$

Grafos Erdös-Rényi (distribución de grado)

La Poisson es menos concentrada en torno de <*k*> que la binomial. Esto ocurre en redes random sparse:



Podemos observar que la distribución de grado es independiente del tamaño de la red (*N*).