



UNIVERSIDAD
COMPLUTENSE
DE MADRID



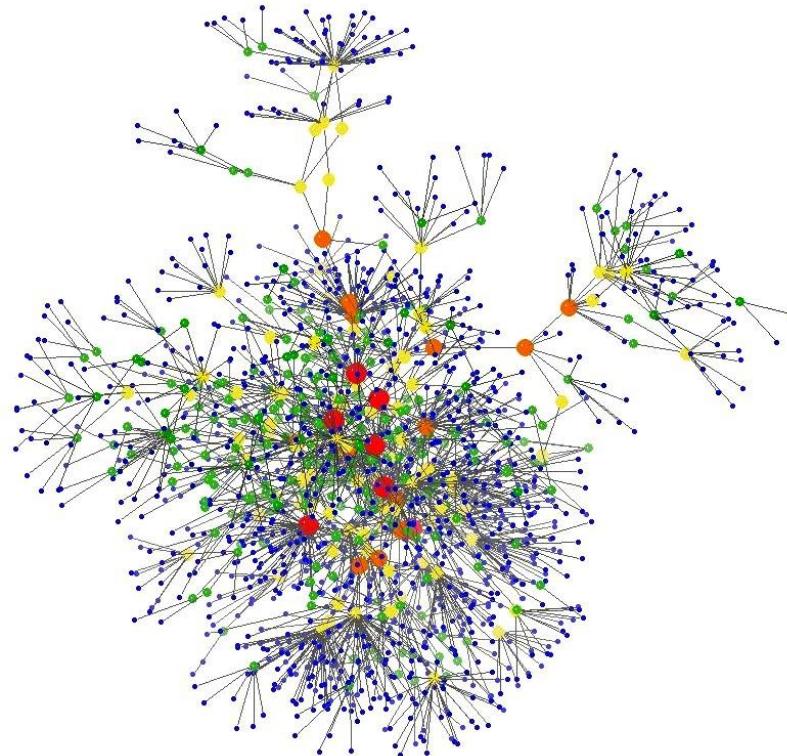
Introducción al Análisis de redes sociales.

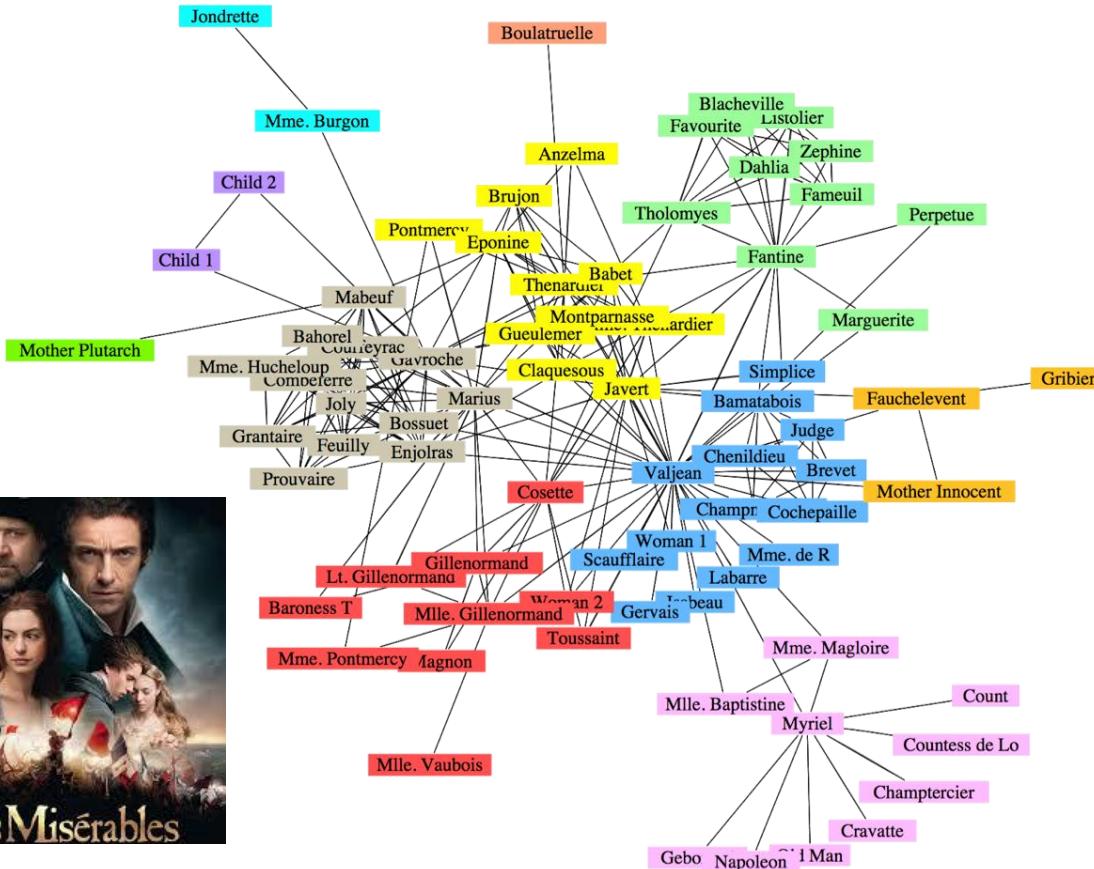
Javier Castro Cantalejo



1. Preliminares

¿Qué es una red Social?





Applications of SNA

▪ Terrorism and crimes

- Social Network analysis is an important part of a conspiracy investigation and is used as an investigative tool. Group structure may be important to investigations of racketeering enterprises, narcotics operations, illegal gambling, and business frauds.

▪ Medicine – epidemiology

- valuable epidemiological tool for understanding the progression of the spread of an infectious disease.

▪ Marketing

- Emarketer projected that Social Network Marketing spending in the USA will reach approximately \$1.3 billion in 2009.
http://www.emarketer.com/Reports/All/Emarketer_2000541.aspx

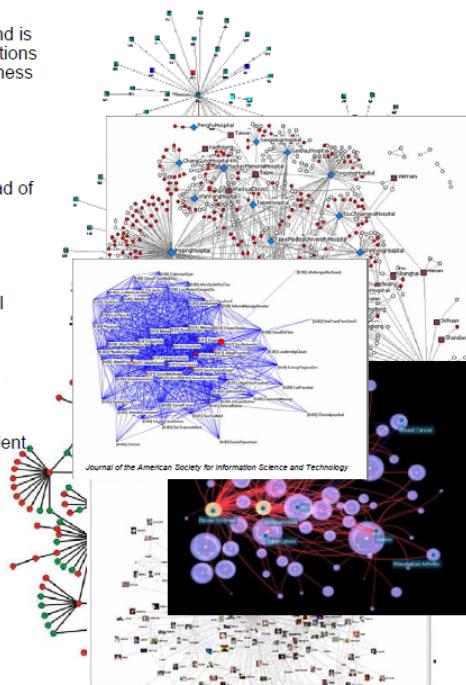
▪ Product Recommendation

- Current recommendation models assume all users' opinions to be independent. Use of SNA relaxes the iid assumption.

▪ Bio-informatics (protein interaction)

▪ Relevance Ranking

▪ Information and Library Science

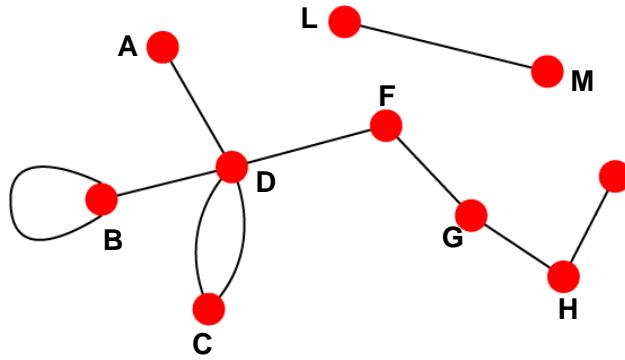


UNDIRECTED VS. DIRECTED NETWORKS

Undirected

Links: undirected (*symmetrical*)

Graph:



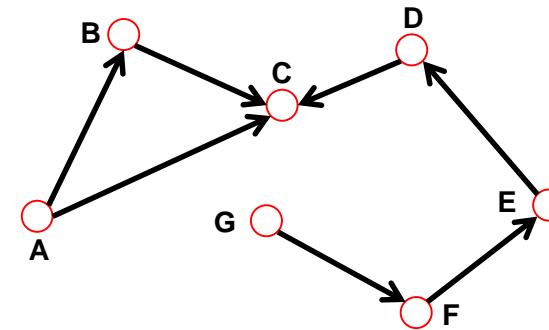
Undirected links :

coauthorship links
Actor network
protein interactions

Directed

Links: directed (*arcs*).

Digraph = directed graph:



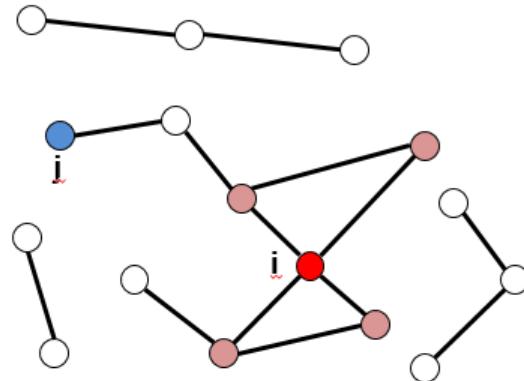
An undirected link is the superposition of two opposite directed links.

Directed links :

URLs on the www
phone calls
metabolic reactions

AVERAGE DEGREE

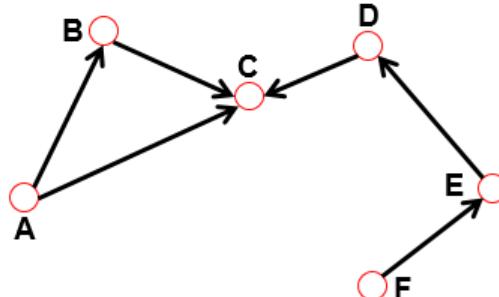
Undirected



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

N – the number of nodes in the graph

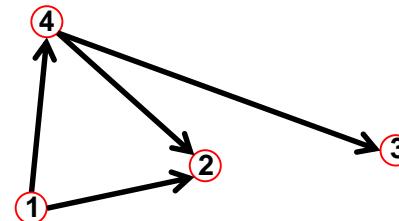
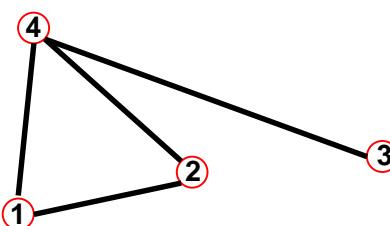
Directed



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

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

ADJACENCY MATRIX



$A_{ij}=1$ if there is a link between node i and j

$A_{ij}=0$ if nodes i and j are not connected to each other.

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

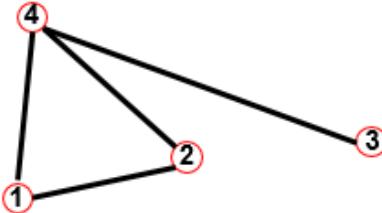
Note that for a directed graph (right) the matrix is not symmetric.

$A_{ij} = 1$ if there is a link pointing from node j and i

$A_{ij} = 0$ if there is no link pointing from j to i .

ADJACENCY MATRIX AND NODE DEGREES

Undirected



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

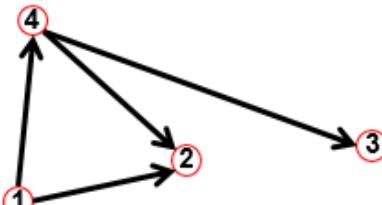
$$A_{ij} = A_{ji}$$
$$A_{ii} = 0$$

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

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

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

Directed



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

$$A_{ij} \neq A_{ji}$$
$$A_{ii} = 0$$

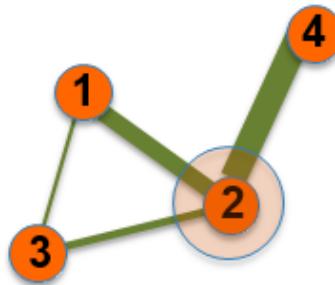
$$k_i^{in} = \sum_{j=1}^N A_{ij}$$

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

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

a_{ij} and w_{ij}

In the literature we often use a double notation: A_{ij} and w_{ij} (somewhat redundant)



Adjacency Matrix (A_{ij})

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

Weight Matrix (W_{ij})

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

Node Strength (weighted degree) s_i :

$$s_i = \sum_{j=1}^N a_{ij} w_{ij}$$

$$s_2 = \sum_{j=1}^N a_{ij} w_{ij} = W_{21} + W_{23} + W_{24}$$

EXAMPLE 1: AIRLINE TRAFFIC

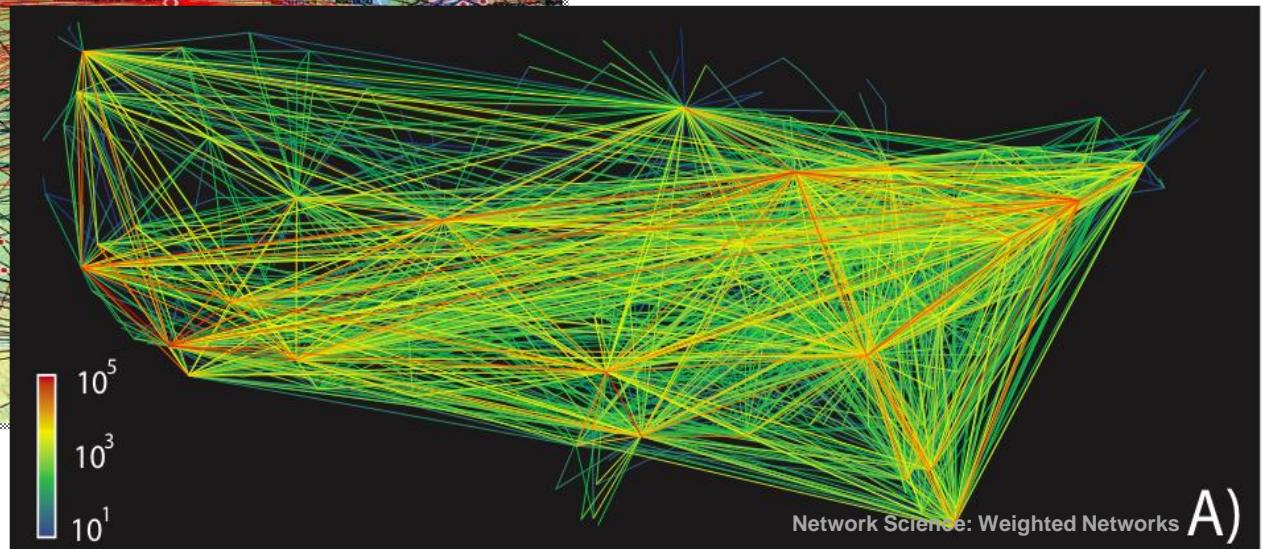
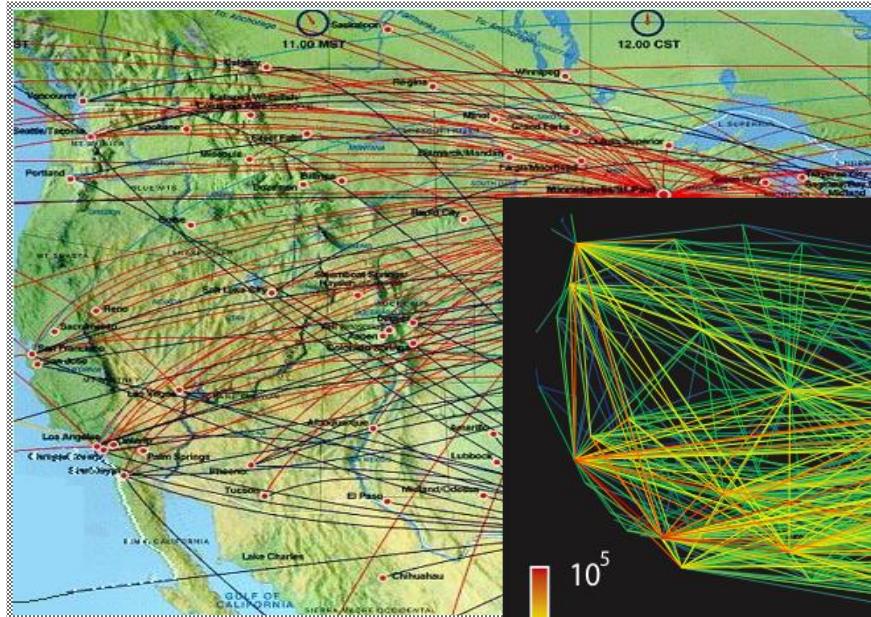
Nodes: airports

Links: direct flights

Weights: Number of seats

- 2002 IATA database
- $V = 3880$ airports
- $E = 18810$ direct flights

- $\langle k \rangle = 10 \quad k_{\max} = 318$
- $\langle l \rangle = 4 \quad l_{\max} = 16$
- $\langle w \rangle = 10^5 \quad w_{\max} = 10^7$
- $N_{\min} = 10^3 \quad N_{\max} = 10^7$



EXAMPLE 2: SCIENCE COLLABORATION NETWORK

- Nodes: scientists
- Links: joint publications
- Weights: number of joint pubs.

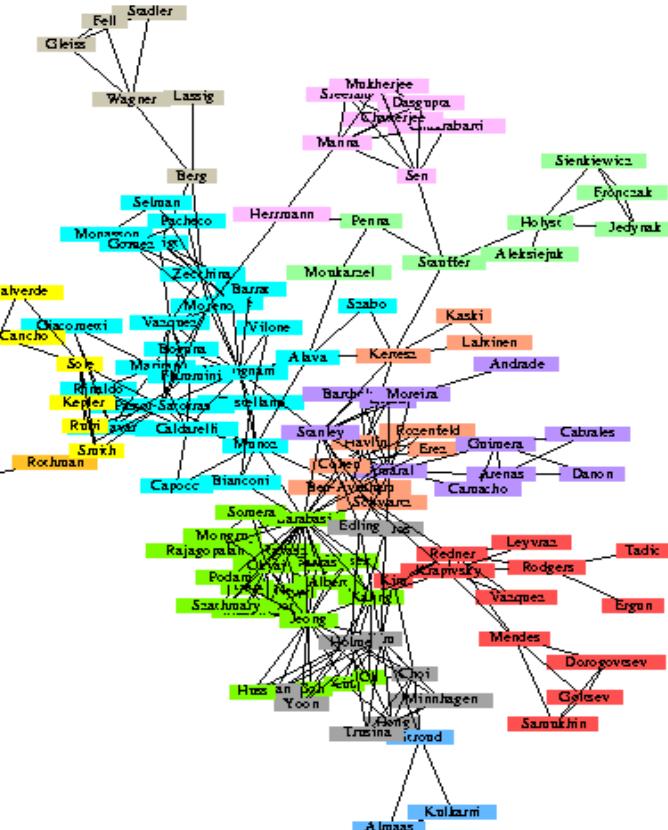
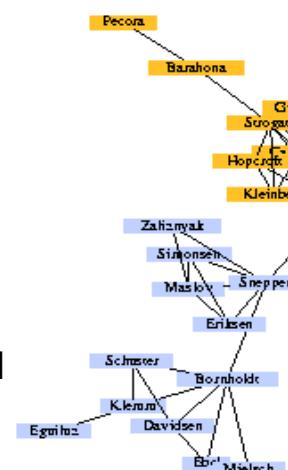
$$w_{ij} = \sum_k \frac{\delta_i^k \delta_j^k}{n_k - 1}$$

i, j: authors

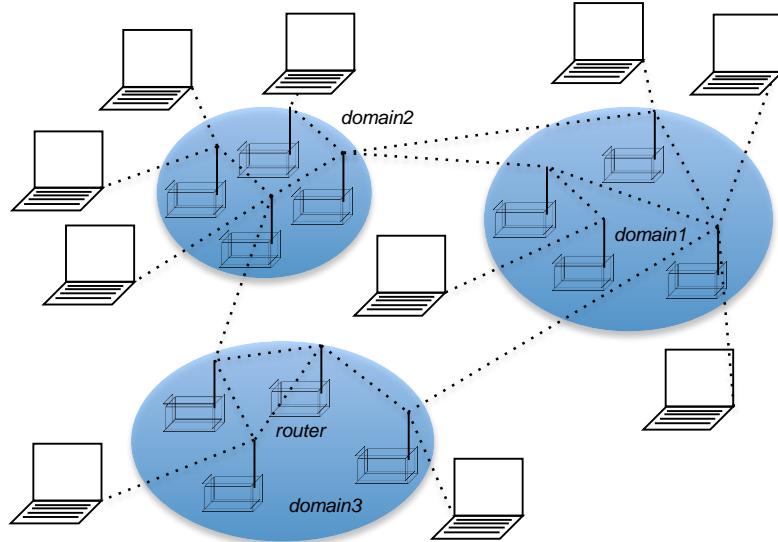
k: paper

n_k : number of authors

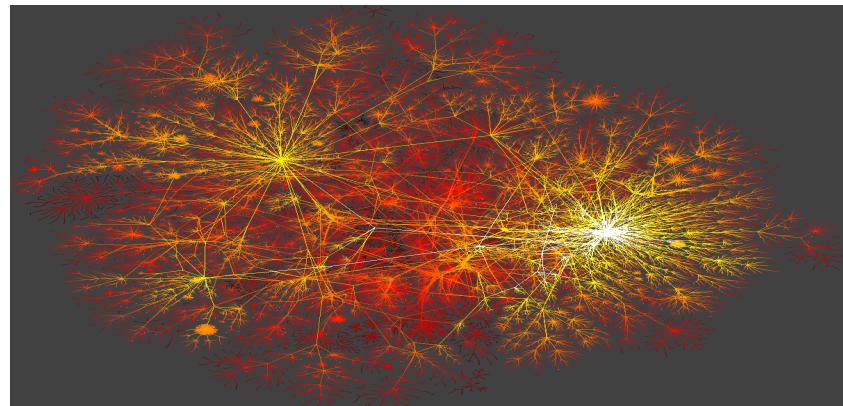
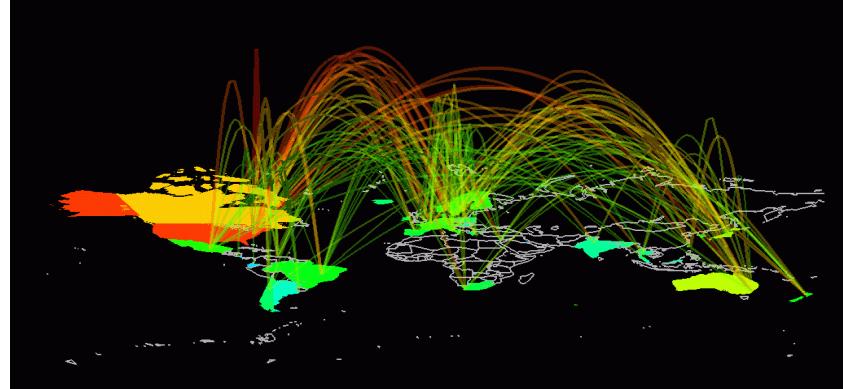
$\delta_i^k = 1$ if author i contributed to paper k



EXAMPLE 3: INTERNET

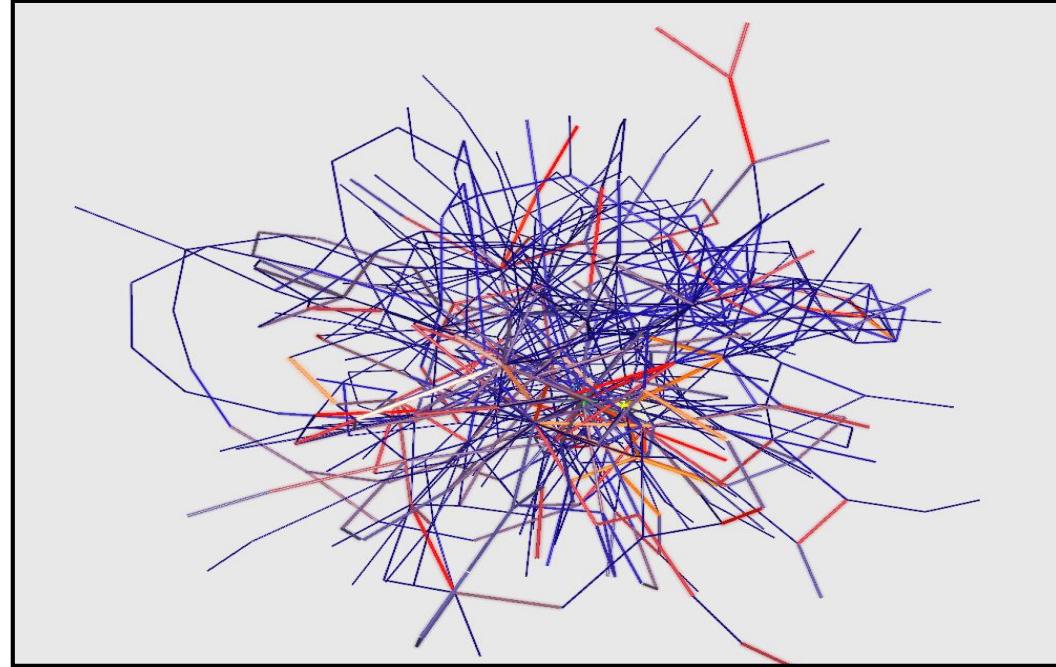


- Nodes: routers
- Links: physical lines
- Link Weights: bandwidth or traffic

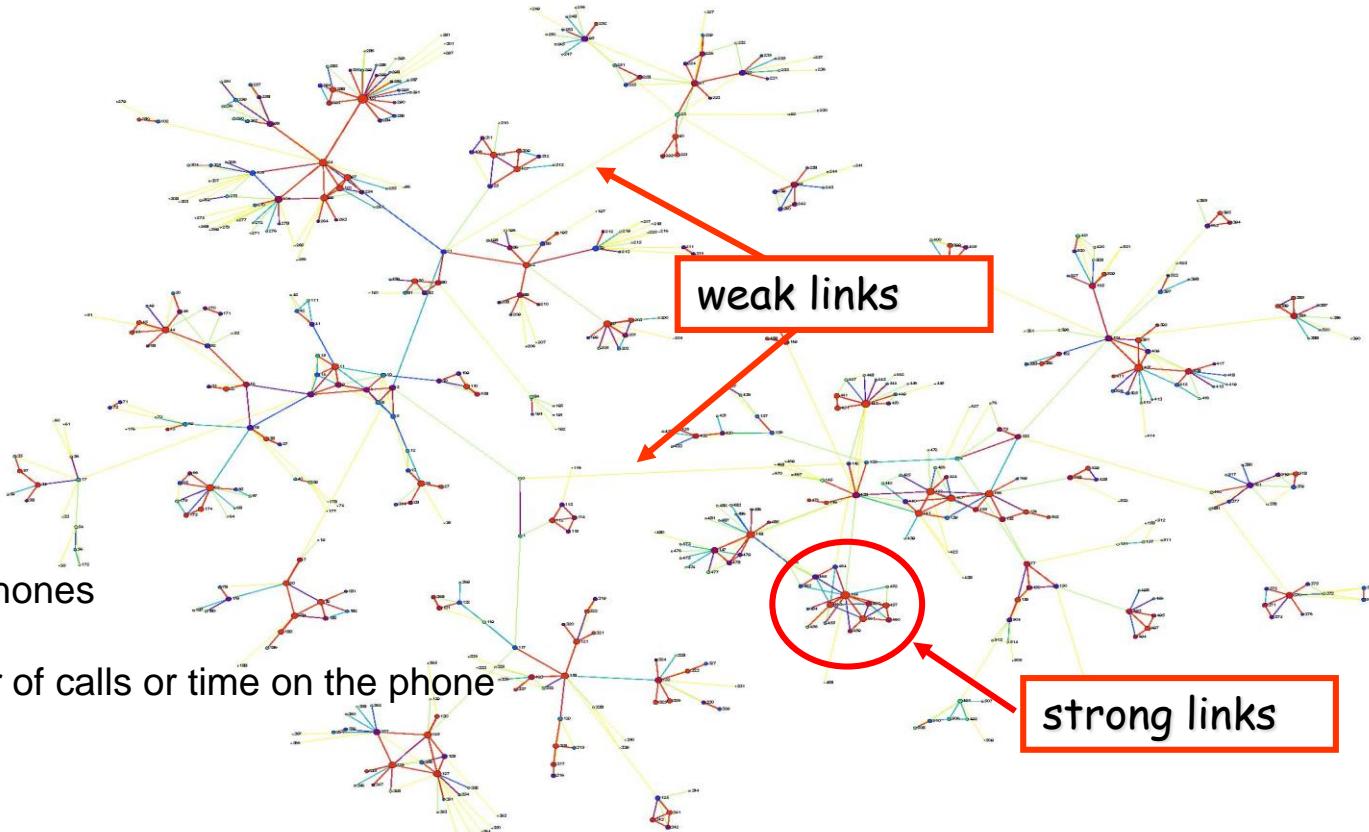


EXAMPLE 4: METABOLIC NETWORK

- Nodes: metabolites
- Links: reactions
- Link Weights: reaction flux



EXAMPLE 5: MOBILE CALL GRAPH



- Nodes: mobile phones
- Links: calls
- Weights: number of calls or time on the phone

2. ¿Cómo es una red Social? ¿Cómo cambia con el tiempo?

- ¿Es posible simular redes que sigan un comportamiento real?
- ¿De esta manera podríamos entender como evoluciona la red?
- Dada una red real, ¿es posible identificar como se ha llegado hasta allí?
- Y quizás lo mas importante, ¿como evolucionará?
 - ✓ Modelos de redes aleatorias.
 - ✓ Simulación de redes.
 - ✓ Análisis de la distribución del grado,
 - ✓ Análisis de las Distancias entre nodos en una red.



MODELO DE RED ALEATORIA

Pál Erdős
(1913-1996)

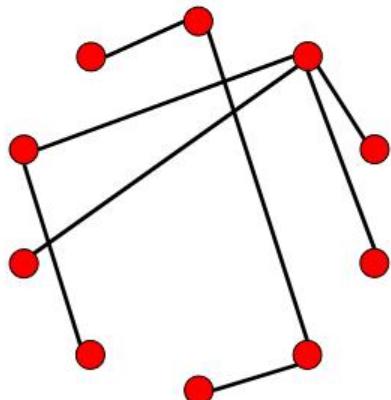


Alfréd Rényi
(1921-1970)



$N=10 \quad p=1/6$

$\langle k \rangle \sim 1.5$



Modelo Erdős-Rényi (1960):

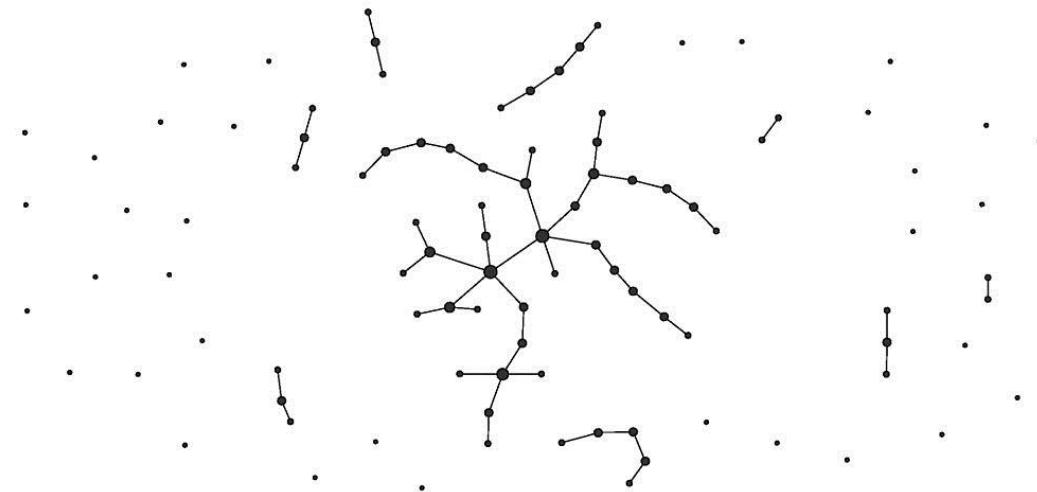
Un **grafo aleatorio** $G(N,p)$ es un grafo no dirigido con N nodos donde cada par de nodos está conectado aleatoriamente con una probabilidad prefijada p

A primera vista, muchas redes reales parecen aleatorias

Este primer modelo de red compleja se basa en modelar esa aparente aleatoriedad mediante redes realmente aleatorias

Modelo Erdos-Renyi

En teoría de grafos el modelo Erdős–Rényi, nombrado así por ser un estudio que realizaron los matemáticos Paul Erdős y Alfréd Rényi, es uno de los métodos empleados en la generación de grafos aleatorios. En este modelo se tiene que un nuevo nodo se enlaza con igual probabilidad con el resto de la red, es decir posee una independencia estadística con el resto de nodos de la red. Hoy en día se emplea como base teórica para la generación de otras redes



Modelo Erdos-Renyi (Algoritmo 1)

Si consideramos N nodos de una red sin conectar y distribuidos de forma aleatoria, podemos imaginar que en un instante inicial enlazamos dos cualesquiera, de esta forma en pasos sucesivos vamos enlazando aleatoriamente de dos a dos nodos. Los nodos que se encuentren enlazados se descarta.

Si repetimos el proceso M veces eligiendo un par de nodos en cada turno al final habremos establecido como máximo M enlaces entre parejas de nodos. Si M es un valor pequeño con respecto al valor total de nodos muchos de los nodos estarán desconectados, mientras que por el contrario otros nodos estarán formando pequeñas islas.

Por el contrario si M es grande en comparación con N el número total de nodos, es muy posible que casi todos los nodos estén enlazados entre sí. Cuando se enlazan los nodos de esta forma aparecen propiedades específicas en la distribución de grado $P(k)$ ya que posee propiedades de distribución de Poisson. Durante muchas décadas a partir de los años 1950 se pensó que las redes con esta característica eran las más adecuadas para describir ciertas redes complejas y pronto se vio que no era del todo cierto.

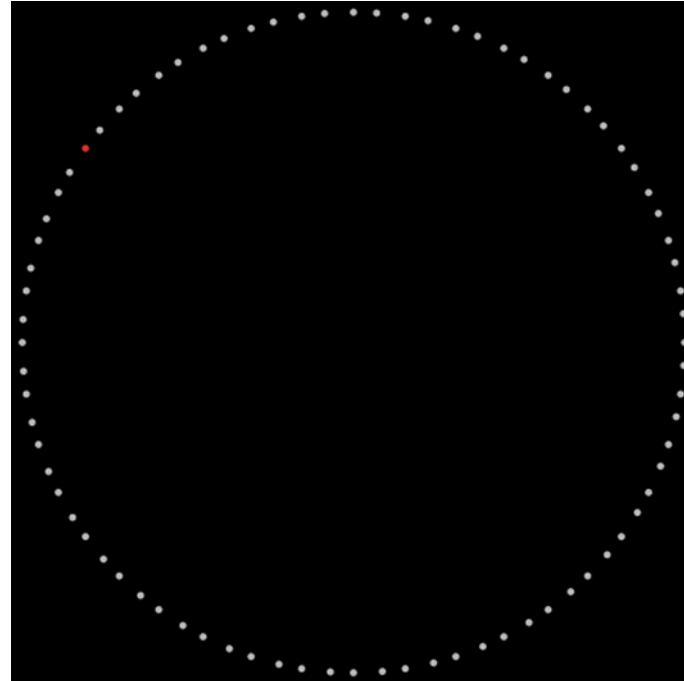
Modelo Erdos-Renyi (Algoritmo 2)

- Elija una probabilidad p para la creación de aristas. El valor de p determinará cuán densamente conectada está la gráfica.
- Comience con un conjunto de nodos (que se muestran distribuidos uniformemente alrededor del círculo en la primera imagen, aunque el diseño no es importante) pero sin conexiones entre ellos.
- Elija dos nodos al azar y agregue una arista entre ellos con probabilidad p .

Este procedimiento es similar al anterior

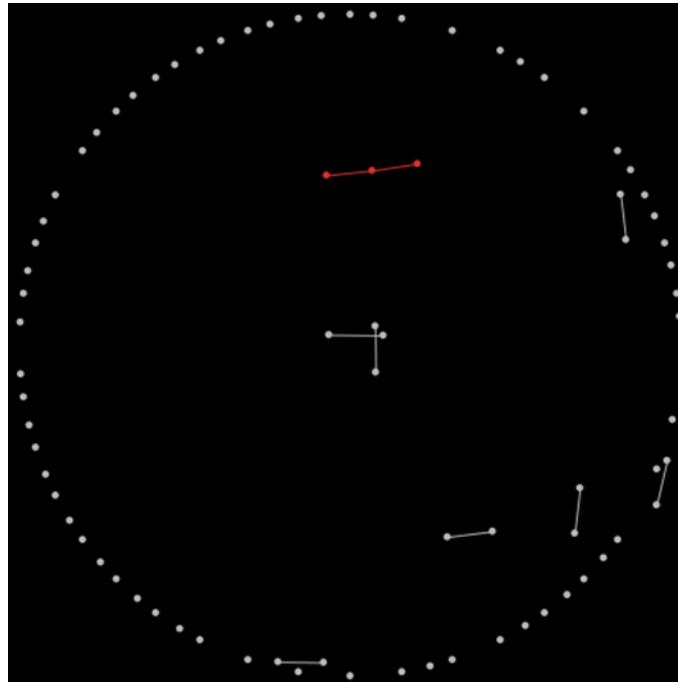
Modelo Erdos-Renyi

Situación inicial: Determinación del numero de nodos de la red



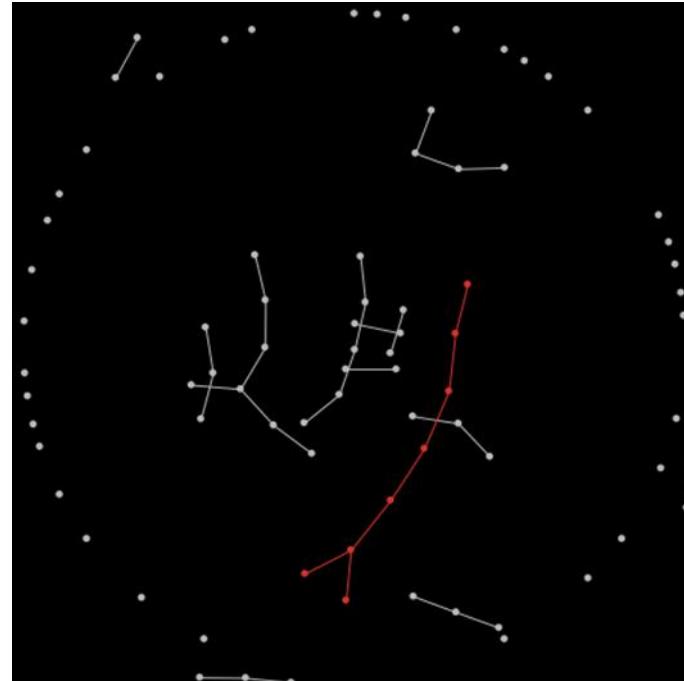
Modelo Erdos-Renyi

Situación intermedia: Primeras conexiones



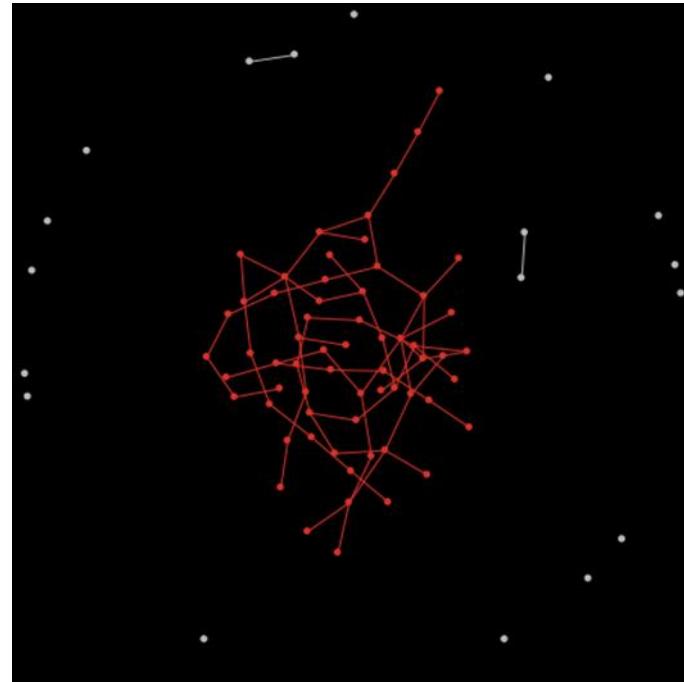
Modelo Erdos-Renyi

Situación intermedia: Emergen conexiones más largas

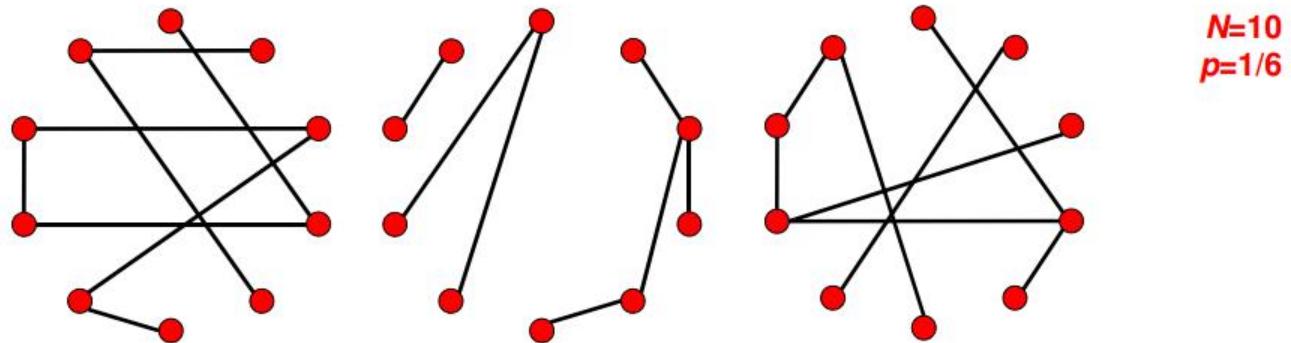


Modelo Erdos-Renyi

Situación final: Se termina generando una gran componente conexa



N y *p* no definen únicamente la red—puede haber muchas instancias, diferentes en el número de enlaces *L* y en los enlaces concretos. ¿Cuántas?



La probabilidad de generar un grafo aleatorio *concreto* $G(N,p)$ con *L* enlaces es:

$$P(G(N,p)) = p^L (1-p)^{\frac{N(N-1)}{2} - L}$$

Es decir, cada grafo $\mathbf{G}(N,p)$ aparece con probabilidad $P(G(N,p))$

P(L): probabilidad de obtener una red con exactamente L enlaces ([distribución binomial](#)):

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

- El número esperado de enlaces $\langle L \rangle$ de un grafo aleatorio $G(N,p)$ se calcula como el producto de la probabilidad de conexión y el número de enlaces ([media de la distribución](#)):

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

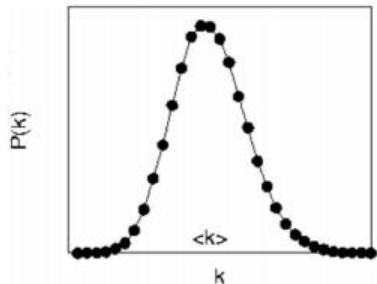
- **El grado medio es:**
- La varianza es:

$$\langle k \rangle = 2L/N = p(N-1)$$

$$\sigma^2 = p(1-p) \frac{N(N-1)}{2}$$

MODELO DE RED ALEATORIA:

Distribución (binomial) del grado



Selección de k nodos de los $N-1$ a los que el nodo actual puede conectarse

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

Probabilidad de que k enlaces estén presentes

Probabilidad de que los otros $N-1-k$ enlaces no estén presentes

$$\langle k \rangle = p(N-1)$$

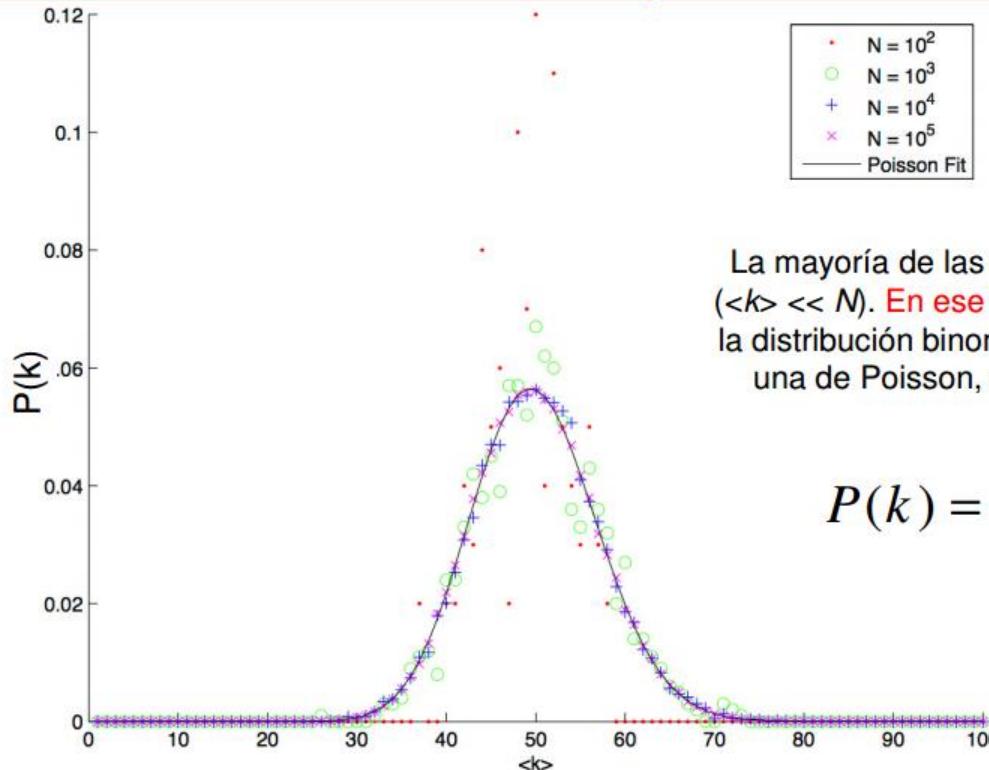
$$\sigma_k^2 = p(1-p)(N-1)$$

$$\frac{\sigma_k}{\langle k \rangle} = \left[\frac{1-p}{p} \frac{1}{(N-1)} \right]^{1/2} \approx \frac{1}{(N-1)^{1/2}}$$

Conforme aumenta el tamaño de la red, la distribución se va volviendo más estrecha, es decir, tenemos una seguridad creciente de que el grado de un nodo está en la vecindad de $\langle k \rangle$

MODELO DE RED ALEATORIA:

Distribución (Poisson) del grado



La mayoría de las redes reales son dispersas ($\langle k \rangle \ll N$). En ese caso y cuando N es grande, la distribución binomial se puede aproximar por una de Poisson, más cómoda de manejar:

$$P(k) = e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!}$$

MODELO DE RED ALEATORIA:

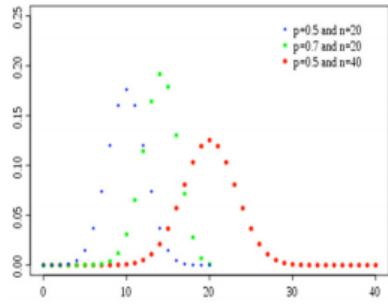
Comparativa de distribuciones del grado

Resultado Exacto

-distribución binomial-

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

Función de Distribución de Probabilidad (PDF)



$$\langle k \rangle = (N-1)p$$

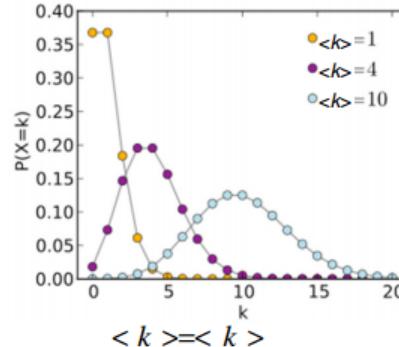
$$\langle k^2 \rangle = p(1-p)(N-1) + p^2(N-1)^2$$

$$\sigma_k = (\langle k^2 \rangle - \langle k \rangle^2)^{1/2} = [p(1-p)(N-1)]^{1/2}$$

Límite con $\langle k \rangle \ll N$ y N grande

-distribución de Poisson-

$$P(k) = e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!}$$



$$\langle k \rangle = \langle k^2 \rangle$$

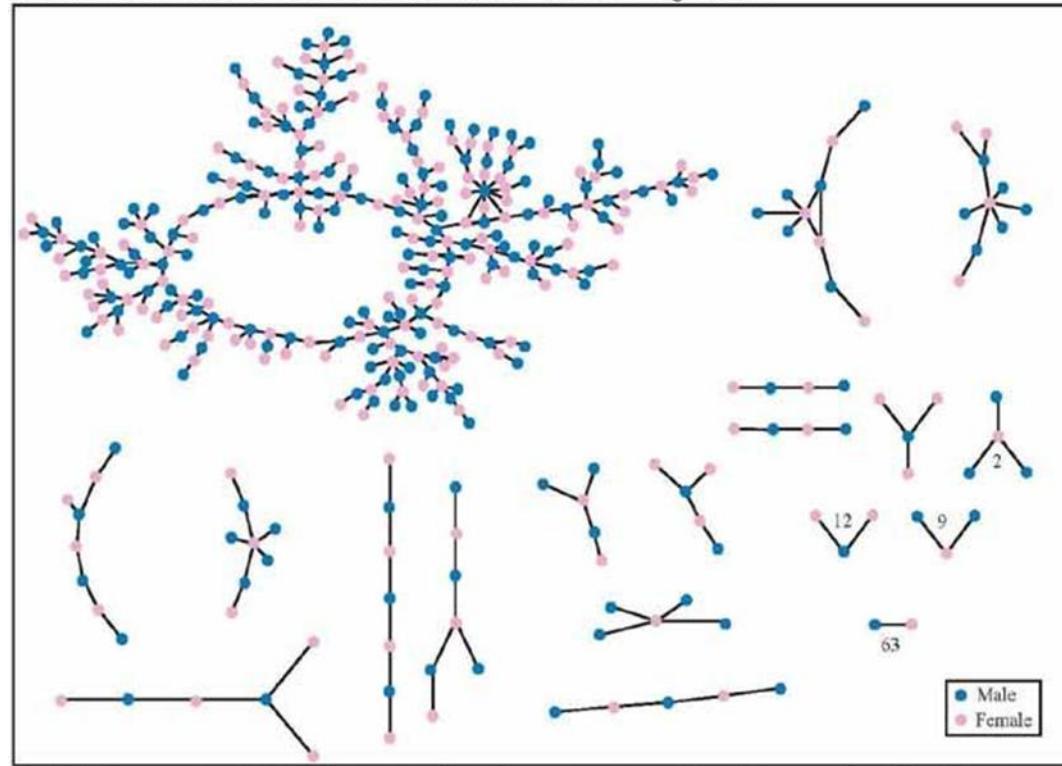
$$\langle k^2 \rangle = \langle k \rangle (1 + \langle k \rangle)$$

$$\sigma_k = (\langle k^2 \rangle - \langle k \rangle^2)^{1/2} = \langle k \rangle^{1/2}$$

Poisson: Todo depende de un solo parámetro $\langle k \rangle$ y no depende explícitamente del número de nodos N . Predice que la distribución del grado es la misma en redes de distintos tamaños en tanto en cuanto tengan el mismo grado medio

¿Hay redes aleatorias? ¿El amor es aleatorio?

The Structure of Romantic and Sexual Relations at "Jefferson High School"



Each circle represents a student and lines connecting students represent romantic relations occurring within the 6 months preceding the interview. Numbers under the figure count the number of times that pattern was observed (i.e. we found 63 pairs unconnected to anyone else).

Y si es erróneo e irrelevante, ¿por qué se estudia?

A pesar de todo, es un modelo de referencia. Posibilita calcular muchas medidas que pueden confrontarse con los datos reales, permitiendo entender hasta qué punto una propiedad particular es consecuencia de un proceso aleatorio

Principios organizativos: patrones que se producen en redes reales, observados en un gran número de ellas, y que se desvían de las predicciones del modelo de red aleatoria

Para identificarlos, necesitamos comprender cómo sería una propiedad particular si estuviese guiada por un proceso puramente aleatorio

Si la propiedad observada está ausente en las redes aleatorias, merece la pena estudiarla porque puede representar algún tipo de orden

¡Aunque ERRÓNEO e IRRELEVANTE, el modelo resulta ser extremadamente ÚTIL!

Modelo Erdos-Renyi

Ejercicio 1.

1. Generar una red no-dirigida aleatoria de Erdos de 50 nodos fijando el número de aristas en 60.
2. Visualizar la red bruta.
3. Aplicar algoritmo de visualización para minimizar el numero de aristas que se cortan.
4. Generar un fichero Excel sobre el grado.
5. Calcular un histograma para la variable grado.
6. Visualizar la red dando mayor peso a los nodos con mayor grado.

Ejercicio 2.

1. Generar una red no-dirigida aleatoria de Erdos de 50 nodos fijando la probabilidad de conexión equivalente a la del ejercicio 1.
2. Visualizar la red bruta.
3. Aplicar algoritmo de visualización para minimizar el numero de aristas que se cortan.
4. Generar un fichero Excel sobre el grado.
5. Calcular un histograma para la variable grado.
6. Visualizar la red dando mayor peso a los nodos con mayor grado.

Modelo Barabasi-Albert

En teoría de redes se denomina Modelo de Barabási–Albert como un algoritmo empleado para generar redes aleatorias complejas libres de escala empleando una regla o mecanismo denominado conexión preferencial.

Las redes generadas por este algoritmo poseen una distribución de grado de tipo potencial y que se denominan: redes libres de escalas. Las redes de este tipo son muy frecuentes en los sistemas elaborados por el ser humano así como en la naturaleza.

Ejemplos de sistemas de este tipo son Internet, el world wide web, redes de citas, y algunas redes sociales, redes eléctricas.

El modelo toma el nombre de Albert-László Barabási y Réka Albert autores que lo popularizaron en 1999.

Modelo Barabasi-Albert

La red comienza con un conjunto de m_0 nodos conectados aleatoriamente.

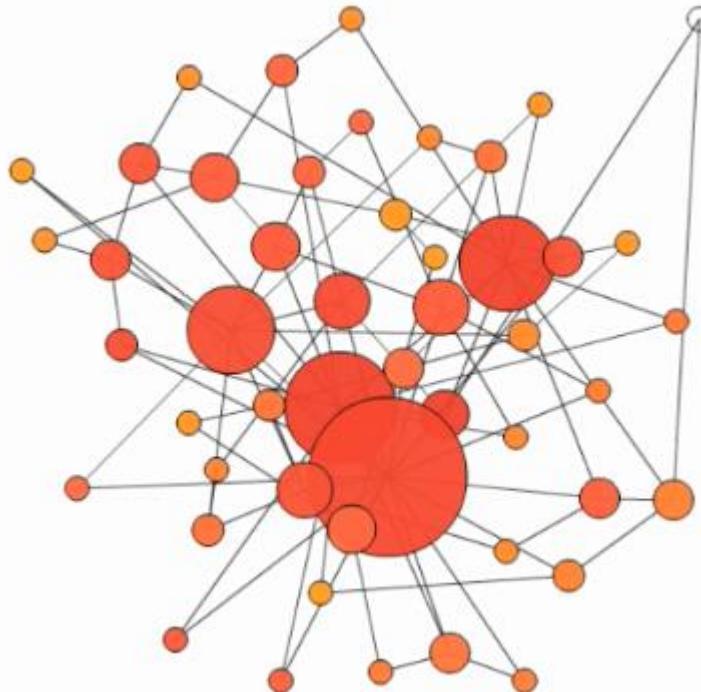
Debe notarse que $m_0 \geq 2$ y el grado de cada nodo en la red inicial debe ser al menos 1, de otra forma la evolución de la red, a medida que se van añadiendo nodos, haría que éstos permanecieran desconectados completamente de la red.

Los nuevos nodos se añaden a la red de uno a uno. Cada nodo es conectado a m nodos de la red con una probabilidad que es proporcional al número de enlaces que posee los nodos de la red, es decir, los nuevos nodos se enlazan preferiblemente con los nodos más conectados. Formalmente la probabilidad p_i de que un nuevo nodo se conecte con j es:

$$p_i = \frac{k_i}{\sum_j k_j}$$

donde k_i es el grado del nodo i . Los nodos con gran cantidad de conexiones ("hubs") tienden a acumular rápidamente más enlaces, mientras que los que poseen pocos enlaces rara vez son el origen de nuevos enlaces. Los nuevos nodos según este algoritmo se dice que poseen una "preferencia" a ser enlazados con los nodos más solicitados. Este algoritmo se fundamenta en el concepto de "conexión prefencial" de los nuevos nodos que se incorporan a la red.

Modelo Barabasi-Albert ($m=2$ empezando con un triangulo)



Modelo Barabasi-Albert

Ejercicio 3.

1. Suponiendo que se empieza con un triangulo y que en cada iteración se añaden 2 aristas. Y que el tamaño final es 23 determinar el numero de aristas y el grado medio.

Solución.

Empezamos con 3 nodos y 3 aristas. Para el resto de 20 nodos añadimos 2 aristas, por lo que el numero total de aristas será de $20 \cdot 2 = 40$ aristas nuevas a las 3 que ya teníamos.

Nodos=N=23

Aristas=L=43

$\langle k \rangle = 1.86$

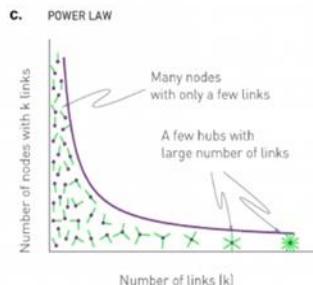
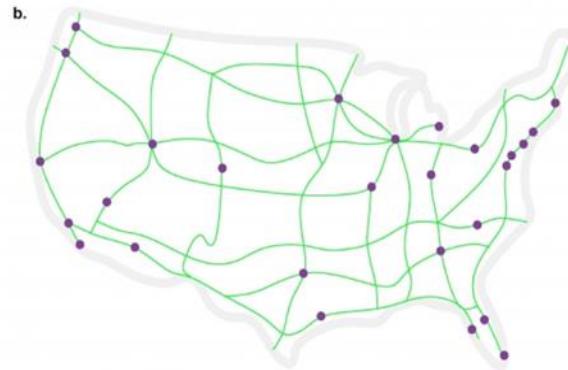
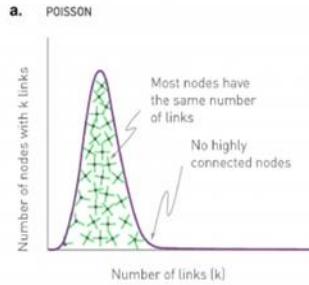
Para un m genérico y un N genérico tenemos. Suponiendo $No < N$, $Lo < L$, tenemos

Nodos=N

Empezamos con No nodos y Lo aristas. Para el resto de los $N-No$ nodos añadimos m aristas, por lo que el numero total de aristas será $L=Lo+(N-No)*m$.

$\langle k \rangle = Lo + (N-No)*m/N$.

Modelo Erdos VS Barabasi-Albert



SUPONIENDO QUE EL TAMAÑO Y EL NUMERO DE ARISTAS ES EL MISMO

¿CUALES SON LAS DIFERENCIAS ENTRE REDES ALEATORIAS Y REDES LIBRES DE ESCALA?

Red de libre escala

$$k_i(t) = m \left(\frac{t}{t_i} \right)^b \quad b = \frac{1}{2}$$

$$P(k) = \frac{2m^2 t}{m_o + t} \frac{1}{k^3} \sim k^{-\gamma}$$

$$\boxed{\gamma = 3}$$

- (i) El exponente del grado es independiente de m.
- (ii) A medida que la ley de potencia describe sistemas de duraciones y tamaños bastante diferentes, se espera que un modelo correcto proporcione una distribución del grado independiente del tiempo. De hecho, asintóticamente, la distribución en grados del modelo BA es independiente del tiempo (y del tamaño del sistema N) → La red alcanza un estado estacionario sin escala.
- (iii) El coeficiente de la distribución de la ley de potencia es proporcional a m^2 .

Scale-free network

From Wikipedia, the free encyclopedia

A **scale-free network** is a network whose degree distribution follows a [power law](#), at least asymptotically. That is, the fraction $P(k)$ of nodes in the network having k connections to other nodes goes for large values of k as

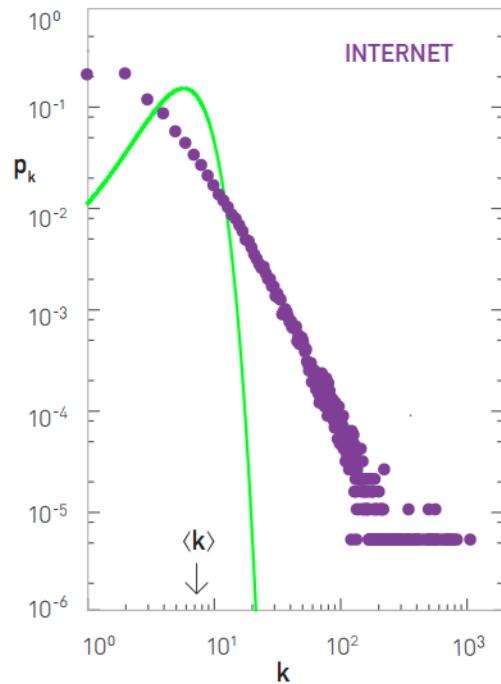
$$P(k) \sim k^{-\gamma}$$

where γ is a parameter whose value is typically in the range $2 < \gamma < 3$, although occasionally it may lie outside these bounds.^{[1][2]}

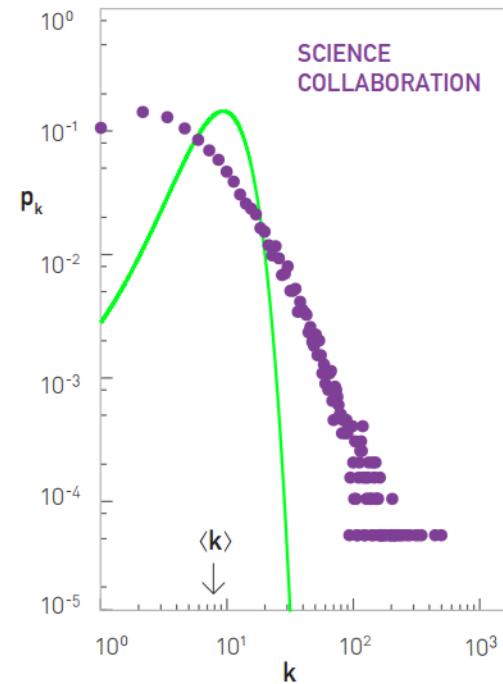
Many networks have been reported to be scale-free, although statistical analysis has refuted many of these claims and seriously questioned others.^[3] Preferential attachment and the [fitness model](#) have been proposed as mechanisms to explain conjectured power law degree distributions in real networks.

Red de libre escala

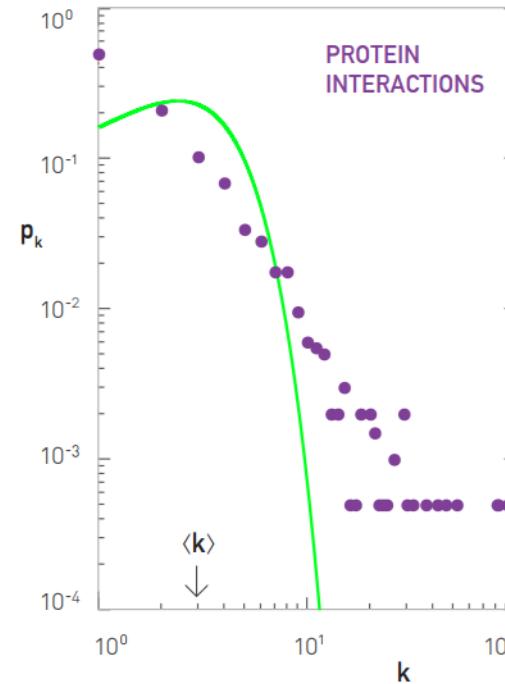
(a)



(b)



(c)



Red de libre escala

Las redes sociales se suelen clasificar según la distribución que sigue el grado. A partir de ahí, se buscan diferentes algoritmos para simular como se ha llegado hasta esa distribución. Scale-Free network. El grado sigue una distribución de “power law”.

- a) The Barabási–Albert model
- b) Two-level network model
- c) Mediation-driven attachment (MDA) model
- d) Non-linear preferential attachment
- e) Hierarchical network model
- f) Fitness model
- g) Hyperbolic geometric graphs
- h) Edge dual transformation to generate scale free graphs with desired properties.

Estas redes son de distribución altamente heterogénea, que sigue una “ley de poder” (Barabasi & Albert 1999). Se denominan sin escala, ya que amplían cualquier parte de la distribución no cambia su forma: hay pocos, pero un número significativo de nodos con mucho de conexiones y hay una cola final de nodos con muy pocas conexiones en cada nivel de ampliación. Dichas redes son típicas de la estructura de la red mundial, mapas semánticos, circuitos electrónicos.

Red de libre escala

Ejercicio 4.

1. Generar una red no-dirigida aleatoria Scale-Free con 100 nodos. Variando los parámetros alfa desde 0 hasta 0.5 de 0.1 en 0.1. Con grado medio 2 y 3. Analizar la distribución del grado y extraer conclusiones.

Red de pequeño mundo

Las redes sociales se suelen clasificar según la distribución que sigue el grado. A partir de ahí, se buscan diferentes algoritmos para simular como se ha llegado hasta esa distribución.

Sin embargo, la mayoría de las redes del mundo real, especialmente las redes sociales, no tienen sistemas homogéneos en la distribución del grado y las conexiones de cada nodo. Se colocan en algún lugar entre las redes regulares y aleatorias. De hecho, Watts y Strogatz (1998) propusieron un modelo donde las conexiones entre los nodos en un gráfico regular fueron reconfigurados con cierta probabilidad. Los gráficos resultantes fueron entre los regulares y los aleatorios en su estructura y se conocen como redes de mundo pequeño (SW).

Las redes de SW están muy cerca estructuralmente de muchas redes sociales en el sentido de que tienen una agrupación más alta y casi la misma ruta promedio que las redes aleatorias con el mismo número de nodos y aristas. Las redes SW suelen tener una alta modularidad (grupos de nodos que son más densos conectados entre sí que al resto de la red).

Red de pequeño mundo

Una red de mundo pequeño es un tipo de grafo para el que la mayoría de los nodos no son vecinos entre sí, y sin embargo la mayoría de los nodos pueden ser alcanzados desde cualquier nodo origen a través de un número relativamente corto de saltos entre ellos.

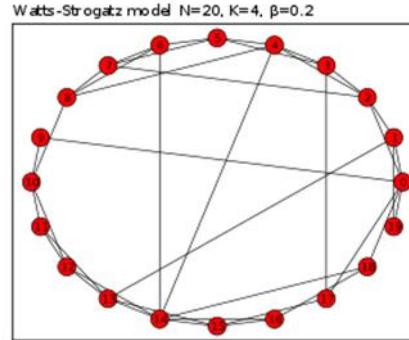
En el caso de una red social, donde los nodos son personas y los enlaces son el conocimiento/relación entre ellos se puede decir que captura muchos de los fenómenos de las redes de mundo pequeño.

Pronto se empezaría a ver que las redes de mundo pequeño son más frecuentes de lo que se presupone y pronto aparecieron otras redes bajo esta categoría: un ejemplo muy claro es la topología de Internet. Este fenómeno ha dado la posibilidad de aplicación de este tipo de redes en diferentes áreas de la ciencia como puede ser el modelado de redes sociales, física, biología, epidemiología, etc.

Este tipo de redes tienen como ejemplo “los seis grados de separación” que conectan el mundo (Stanley Milgram (1967)).

Red de pequeño mundo (algoritmo de watts y strogatz)

El algoritmo de construcción propuesto por watts y strogatz para las redes de mundo pequeño es el siguiente: se establece una red inicial unidimensional con N nodos, estos nodos se pueden disponer en forma de anillo de tal forma que cada uno de los vértices (o nodos) se une con $2k$ vecinos. La probabilidad de conectar un nodo con otro cualquiera es de p . Para un grafo con $p=0$ se puede ver que la conectividad es la misma y de valor $2k$. por otro lado un valor no nulo de p introduce desorden en la red de tal forma que la conectividad no es uniforme, manteniendo todavía de media un valor de $2k$.



Red de 20 nodos construida según el modelo Watts y Strogatz ($N=20$, $K=4$, $\beta=0.2$).

Red de pequeño mundo (algoritmo de watts y strogatz)

Algorithm [\[edit\]](#)

Given the desired number of nodes N , the mean degree K (assumed to be an even integer), and a special parameter β , satisfying $0 \leq \beta \leq 1$ and $N \gg K \gg \ln N \gg 1$, the model constructs an undirected graph with N nodes and $\frac{NK}{2}$ edges in the following way:

1. Construct a regular ring lattice, a graph with N nodes each connected to K neighbors, $K/2$ on each side. That is, if the nodes are labeled $n_0 \dots n_{N-1}$, there is an edge (n_i, n_j) if and only if $0 < |i - j| \pmod{N-1} - \frac{K}{2} \leq \frac{K}{2}$.
2. For every node $n_i = n_0, \dots, n_{N-1}$ take every edge (n_i, n_j) with $i < j$, and rewire it with probability β . Rewiring is done by replacing (n_i, n_j) with (n_i, n_k) where k is chosen with uniform probability from all possible values that avoid self-loops ($k \neq i$) and link duplication (there is no edge $(n_i, n_{k'})$ with $k' = k$ at this point in the algorithm).

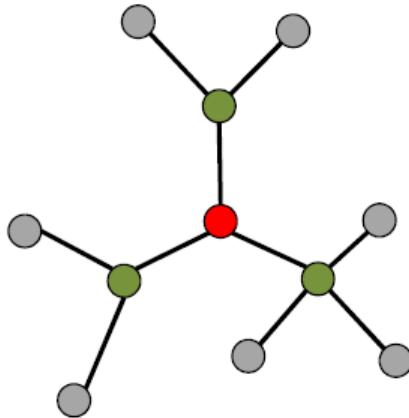
Properties [\[edit\]](#)

The underlying lattice structure of the model produces a locally clustered network, and the random links dramatically reduce the average path lengths. The algorithm introduces about $\beta \frac{NK}{2}$ non-lattice edges. Varying β makes it possible to interpolate between a regular lattice ($\beta = 0$) and a random graph ($\beta = 1$)

approaching the Erdős–Rényi random graph $G(n, p)$ with $n = N$ and $p = \frac{NK}{2 \binom{N}{2}}$.

The three properties of interest are the average path length, the clustering coefficient, and the degree distribution.

Los grafos aleatorios tienden a tener una topología en forma de árbol con nodos de grado casi constante:



- no. de vecinos a distancia 1: $N_1 \cong \langle k \rangle$
- no. de vecinos a distancia 2: $N_2 \cong \langle k \rangle^2$
- no. de vecinos a distancia d : $N_d \cong \langle k \rangle^d$
- Estimación de la distancia máxima
(fórmula de los mundos pequeños):

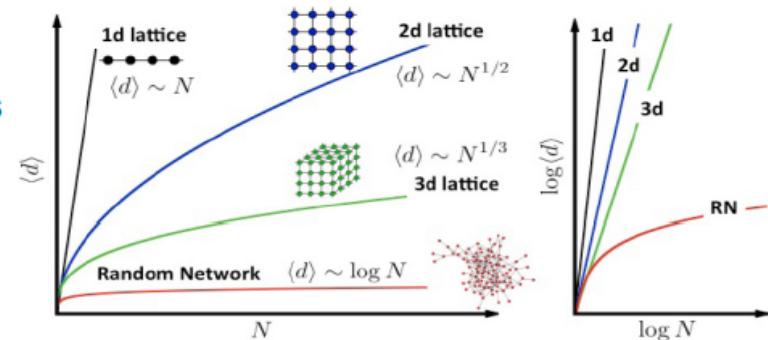
$$N = 1 + \langle k \rangle + \langle k \rangle^2 + \dots + \langle k \rangle^d = \frac{\langle k \rangle^{d+1} - 1}{\langle k \rangle - 1} \approx \langle k \rangle^d \quad \rightarrow \quad d_{\max} = \frac{\log N}{\log \langle k \rangle}$$

Comentarios:

- En la mayoría de los casos, la fórmula aproxima mejor $\langle d \rangle$ que d_{\max} . Esto se debe a que en la práctica d_{\max} está dominada por unos pocos caminos de longitud extrema mientras que $\langle d \rangle$ está promediada entre todos los pares de nodos

$$\langle d \rangle = \frac{\log N}{\log \langle k \rangle}$$

- En general, $\log N \ll N$. Por tanto, el hecho de que $\langle d \rangle$ dependa de $\log N$ implica que las distancias en una red aleatoria son varios órdenes de magnitud menores que el tamaño de la red



- El término $1/\log \langle k \rangle$ implica que, cuanto más densa es la red, menor es la distancia entre sus nodos

MODELO DE RED ALEATORIA:

Distancias en redes aleatorias (3)

En la práctica, la fórmula aproxima mejor $\langle d \rangle$ que d_{\max}

Validación sobre redes reales:

$$\langle d \rangle = \frac{\log N}{\log \langle k \rangle}$$

<i>Network Name</i>	<i>N</i>	<i>L</i>	$\langle k \rangle$	$\langle d \rangle$	d_{\max}	$\frac{\log N}{\log \langle k \rangle}$
Internet	192,244	609,066	6.34	6.98	26	6.59
WWW	325,729	1,497,134	4.60	11.27	93	8.32
Power Grid	4,941	6,594	2.67	18.99	46	8.66
Mobile Phone Calls	36,595	91,826	2.51	11.72	39	11.42
Email	57,194	103,731	1.81	5.88	18	18.4
Science Collaboration	23,133	186,936	8.08	5.35	15	4.81
Actor Network	212,250	3,054,278	28.78	-	-	-
Citation Network	449,673	4,707,958	10.47	11.21	42	5.55
E Coli Metabolism	1,039	5,802	5.84	2.98	8	4.04
Yeast Protein Interactions	2,018	2,930	2.90	5.61	14	7.14

Dadas las grandes diferencias en temática, tamaño y grado medio, el ajuste es bastante bueno

Red de pequeño mundo

Ejercicio 5.

1. Generar una red no-dirigida aleatoria Small world network con 10 nodos. Y únicamente 2 vecinos conectados. P muy baja.
2. Visualizar la red bruta.
3. Analizar las distancias.
4. Aumentar p y ver como se genera una red de pequeño mundo.

Ejercicio 6.

1. Generar una red no-dirigida aleatoria Small world network con 100 nodos. Variando el número de vecinos conectados y el valor de P.
2. Visualizar la red bruta.
3. Aplicar algoritmo de visualización para minimizar el numero de aristas que se cortan.
4. Generar un fichero Excel sobre el grado.
5. Calcular un histograma para la variable grado.
6. Visualizar la red dando mayor peso a los nodos con mayor grado.

Red de pequeño mundo

Ejercicio 7.

Análisis y ajuste completo de las siguientes redes reales:

Red de Football (Mundial1998).

Red de Karate club network.

Red del glosario de palabras de redes.

Red de los delfines.

¿Son una red aleatoria?

¿Son una red libre de escala?

¿Son una red de pequeño mundo?

2. ¿Cuál es su capacidad de difusión?

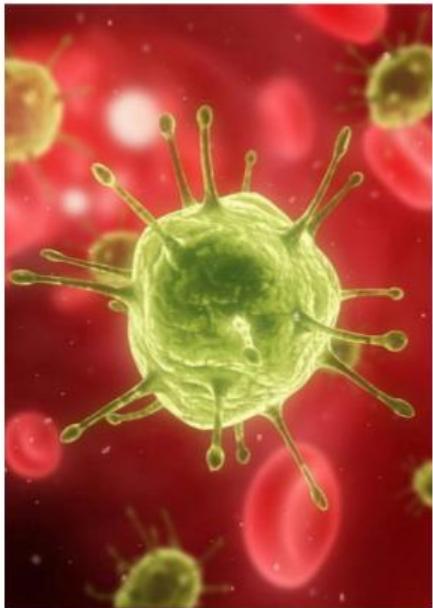
Existen diferentes modelos para estudiar como a partir de unos nodos infectados la información, virus, .. Se propaga en la red, en función de como se produce la imunidad y el contagio.

Estos modelos son complejos y en muchas ocasiones se realizan mediante simulación.

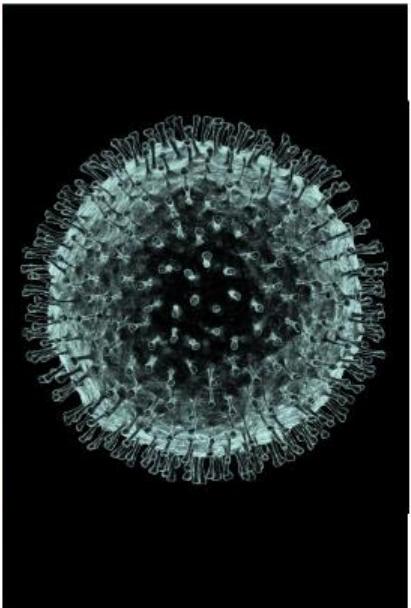
No obstante, uno de los problemas mas relevantes son a quien debo infectar/proteger para que la idea/virus se propague con mayor velocidad o para proteger a la población de una epidemia.

Este problema es un problema np-duro y se resuelve mediante distintas heurísticas.

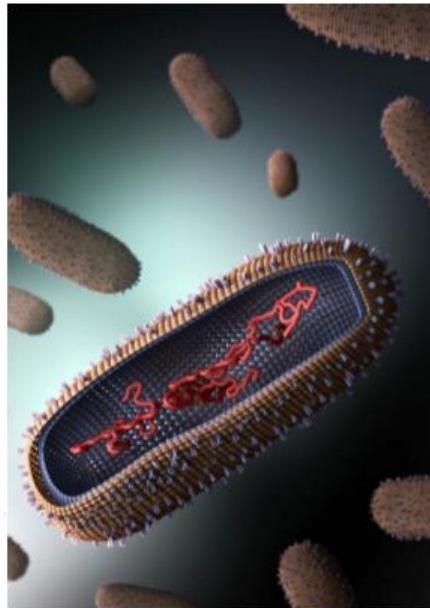




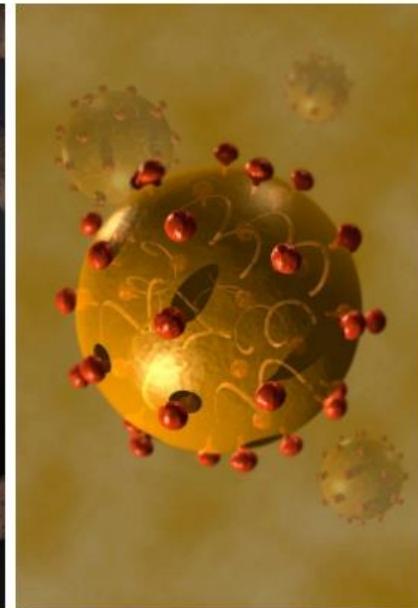
HIV



SARS

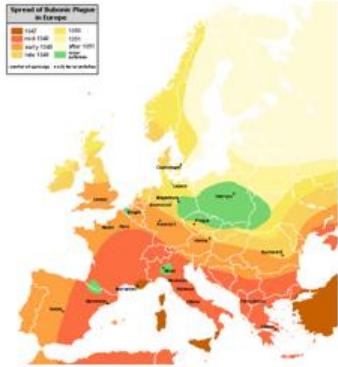


influenza



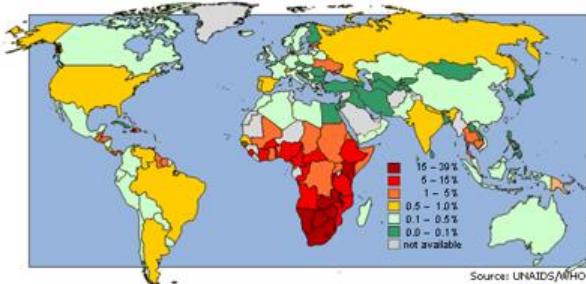
Hepatitis C

The Great Plague



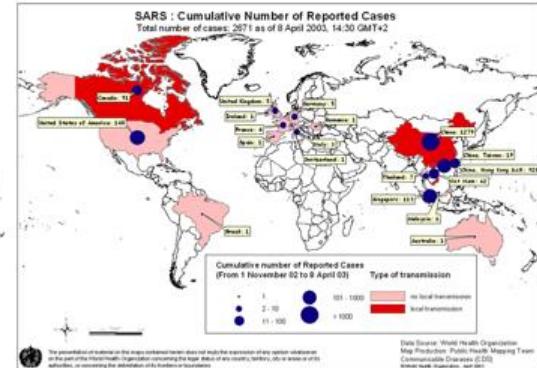
HIV

HIV prevalence in adults, end 2001



Note: This map does not reflect a position by UNICEF on the legal status of any country or territory or the delimitation of any frontier.

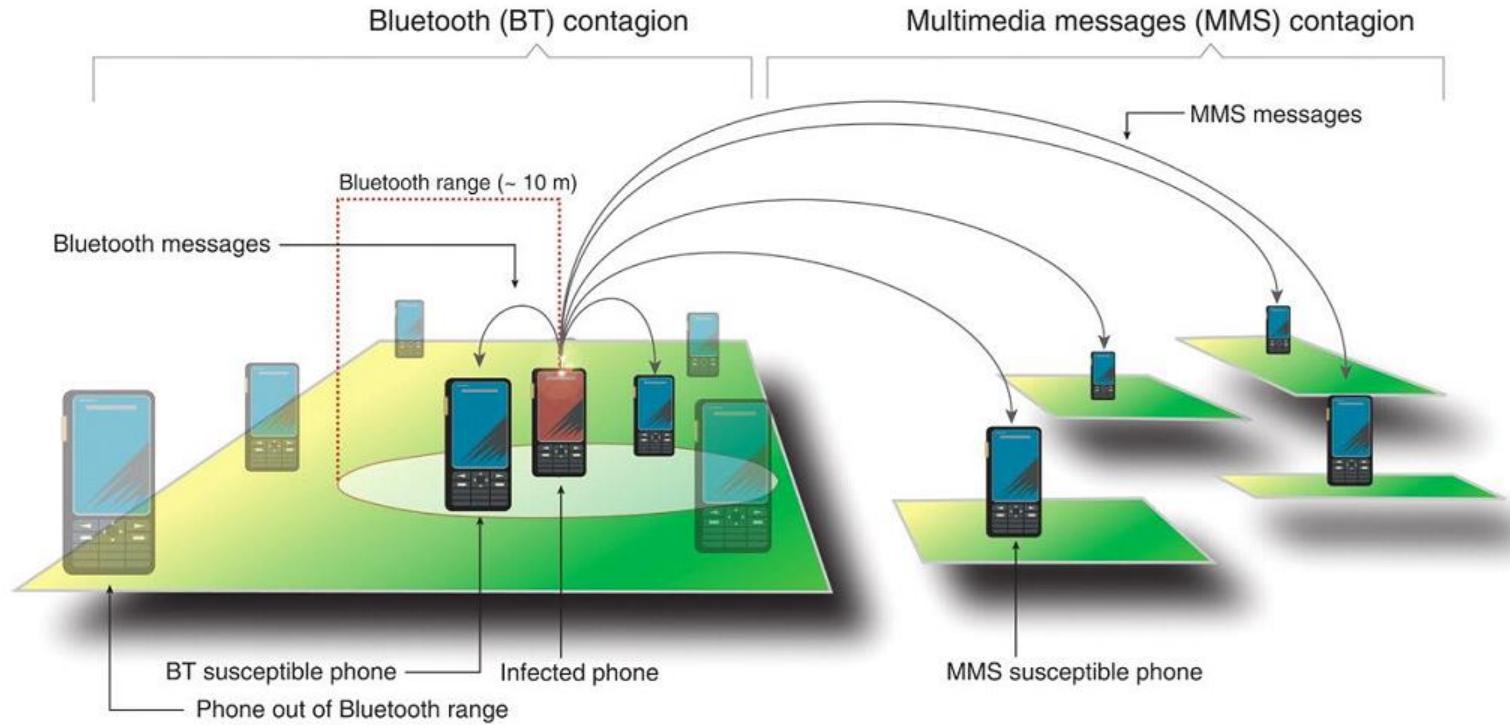
SARS



1918 Spanish flu



H1N1 flu



PHENOMENA	AGENT	NETWORK
Venereal Disease	Pathogens	Sexual Network
Rumor Spreading	Information, Memes	Communication Network
Diffusion of Innovations	Ideas, Knowledge	Communication Network
Computer Viruses	Malwares, Digital viruses	Internet
Mobile Phone Virus	Mobile Viruses	Social Network/Proximity Network
Bedbugs	Bedbugs	Hotel - Traveler Network
Malaria	Plasmodium	Mosquito - Human network

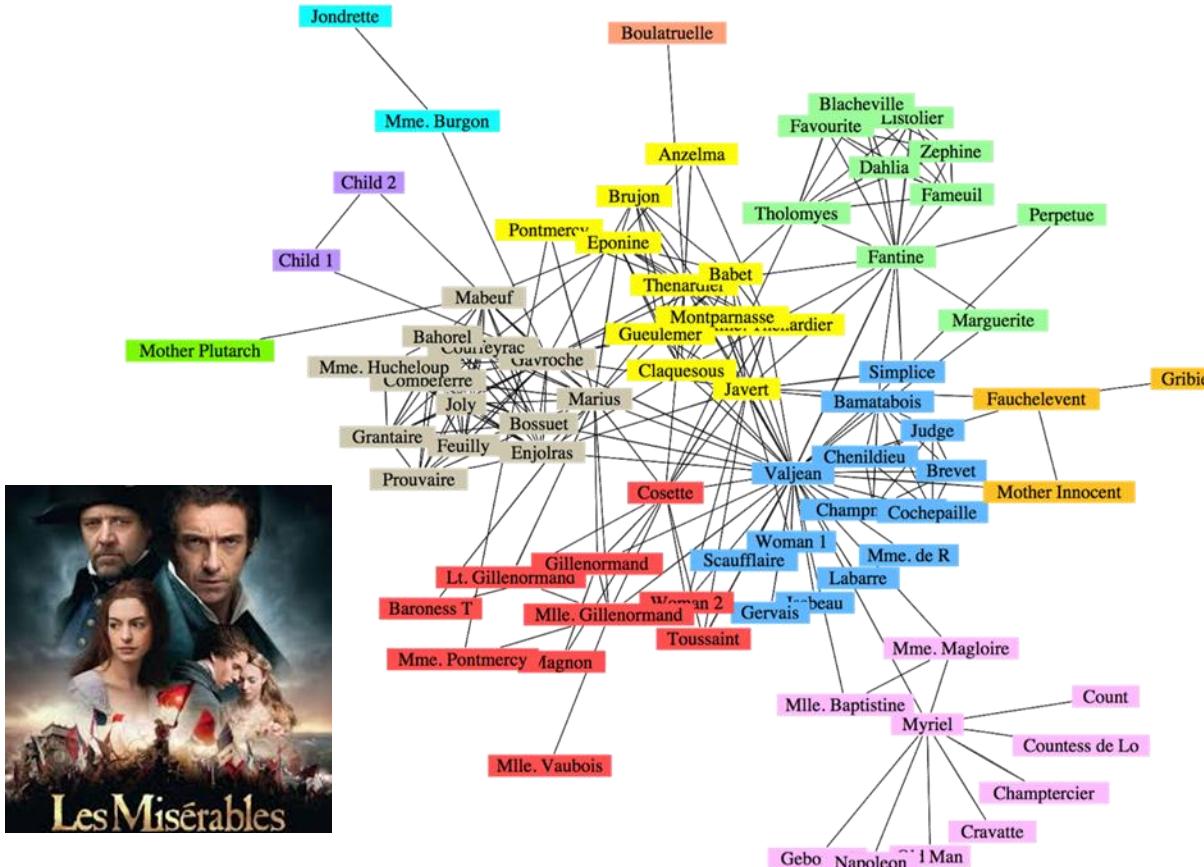
3. ¿Cuál es el poder e influencia de cada individuo-nodo en la red?



¿ Dónde medir la importancia-relevancia de la posición que ocupa un Individuo-nodo en una red ?

- Identificación de genes relevantes en para diabetes, alzheimer, enfermedades mentales, crecimiento.
- Control de epidemias (tanto a nivel global como local).
- Contagiar la red con información relevante (marketing viral?)
- Proteger una red frente ataques. Terrorismo.
- Problemas de transporte. Localización.
- Determinacion de lideres en una estructura. Blogeros mas influyentes en Twiter, Instagram, Facebook,....
- Conocer la influencia de unos nodos sobre otros (administracion, politica, empresa).
- Internet Google Page Rank.
- Agricultura, Forestales,-----

¿ Dónde medir la importancia-relevancia de la posición que ocupa un Individuo-nodo en una red ?



¿ Dónde medir la importancia-relevancia de la posición que ocupa un Individuo-nodo en una red ?

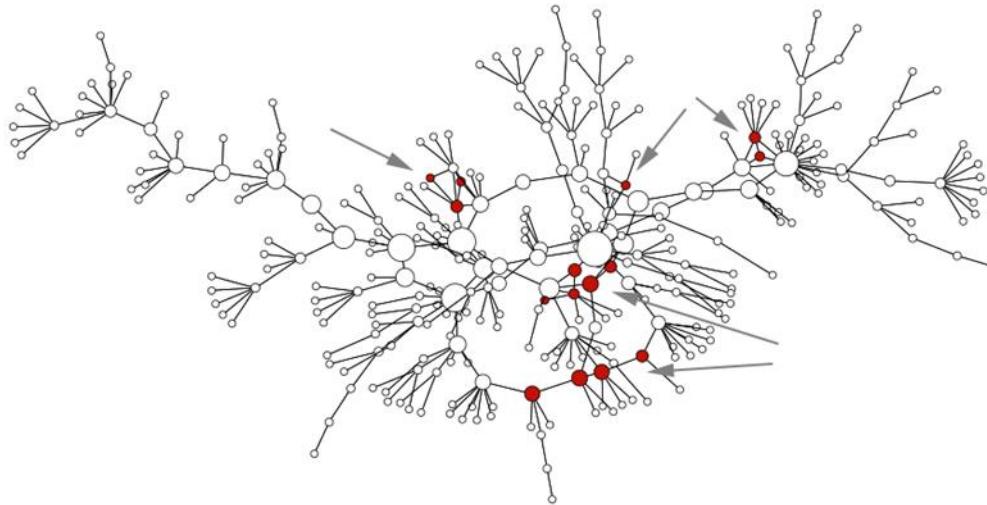
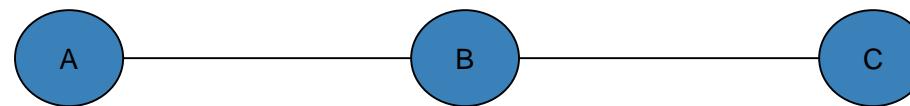


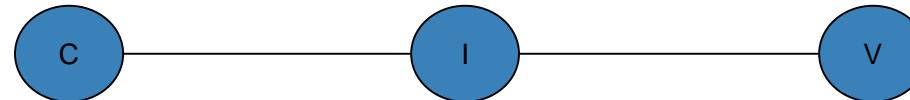
Figure 5: The largest component of a network of sexual contacts between high-risk actors in the city of Colorado Springs, CO, as reconstructed by Potterat *et al.* (2002). The size of the vertices increases linearly with their random-walk betweenness, as defined in this paper. The highlighted vertices (also indicated by the arrows) are those for which the random-walk betweenness is substantially greater than shortest-path betweenness (a factor of two or more).

¿ Dónde medir la importancia-relevancia de la posición que ocupa un Individuo-nodo en una red ?

Esquema político



Comprador-Intermediario-Vendedor



Medidas de Centralidad

Centralidad. Noción sociológica que no ha sido rigorosamente definida. Se suele definir de forma indirecta. Diremos que un nodo es central si:

- Puede comunicarse directamente con muchos nodos ó
 - Es cercano a muchos nodos ó
 - Existen muchos pares de nodos que le necesitan como intermediario en sus comunicaciones.

Medidas de Centralidad

- Degree centrality (Shaw, 1954, and Nieminen, 1974)
- Closeness centrality (Beauchamp, 1965 and Sabidussi, 1966)
- Betweenness centrality (Bavelas, 1948 and Freeman, 1977)
- Stephenson and Zelen (1989)
- Katz centrality
- Bonacich (1972) (autovalor).
- Flow centrality (Freeman)
- Page Rank (Google)
- Hubs and Authorities
- Game theory centrality measure (Gomez et al. 2003)
- Multicriteria centrality measure (Gomez et al. 2014)

Medidas de Centralidad (Closeness Centrality)

Closeness centrality [\[edit\]](#)

In connected [graphs](#) there is a natural distance metric between all pairs of nodes, defined by the length of their [shortest paths](#). The **farness** of a node x is defined as the sum of its distances from all other nodes, and its closeness was defined by Bavelas as the reciprocal of the farness,[\[9\]](#)[\[10\]](#) that is:

$$C(x) = \frac{1}{\sum_y d(y, x)}.$$

Thus, the more central a node is the lower its total distance from all other nodes. Note that taking distances from or to all other nodes is irrelevant in undirected graphs, whereas in directed graphs distances to a node are considered a more meaningful measure of centrality, as in general (e.g., in the web) a node has little control over its incoming links.

When a graph is not [strongly connected](#), a widespread idea is that of using the sum of reciprocal of distances, instead of the reciprocal of the sum of distances, with the convention $1/\infty = 0$:

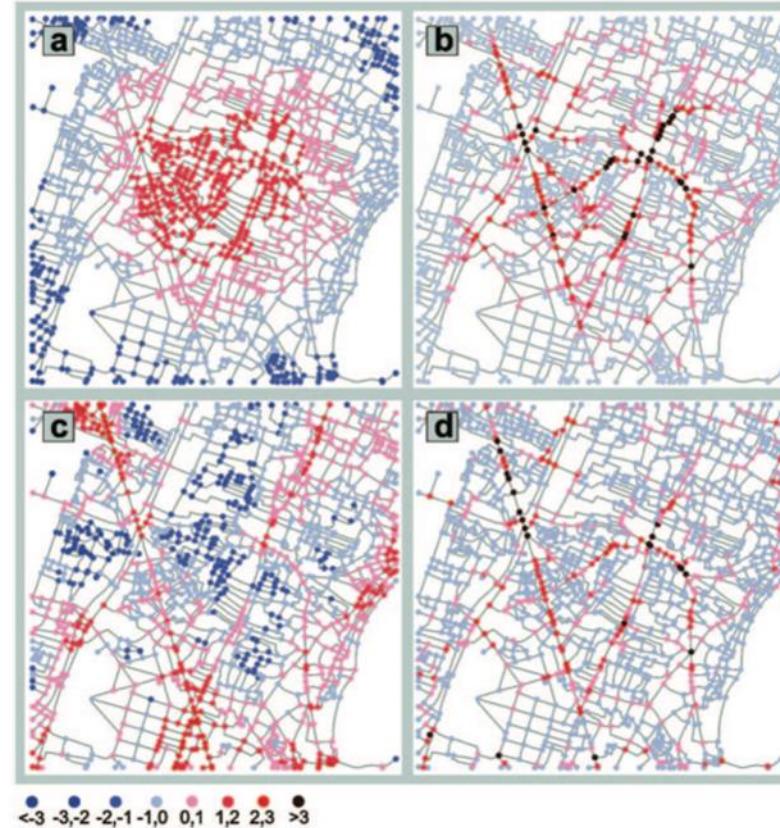
$$H(x) = \sum_{y \neq x} \frac{1}{d(y, x)}.$$

Medidas de Centralidad

Street Centralities - Centrality in a network of self-organised streets in Cairo, showing (a) closeness, (b) betweenness, (c) straightness, (d) information centralities.

Crucitti, Latora, Porta (2006)

[[source](#)]

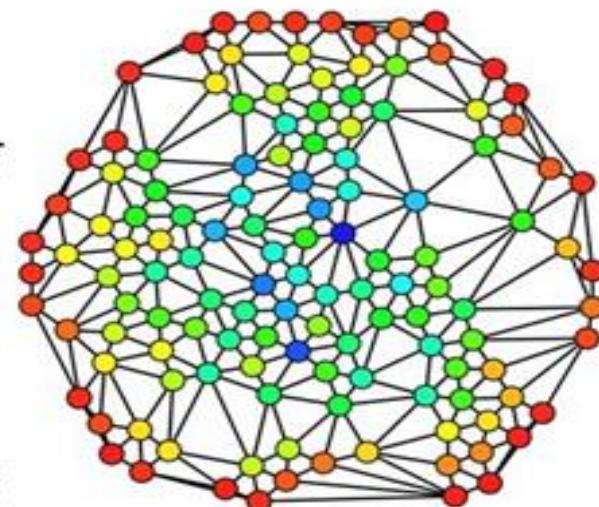


Medidas de Centralidad (Betweenness Centrality)

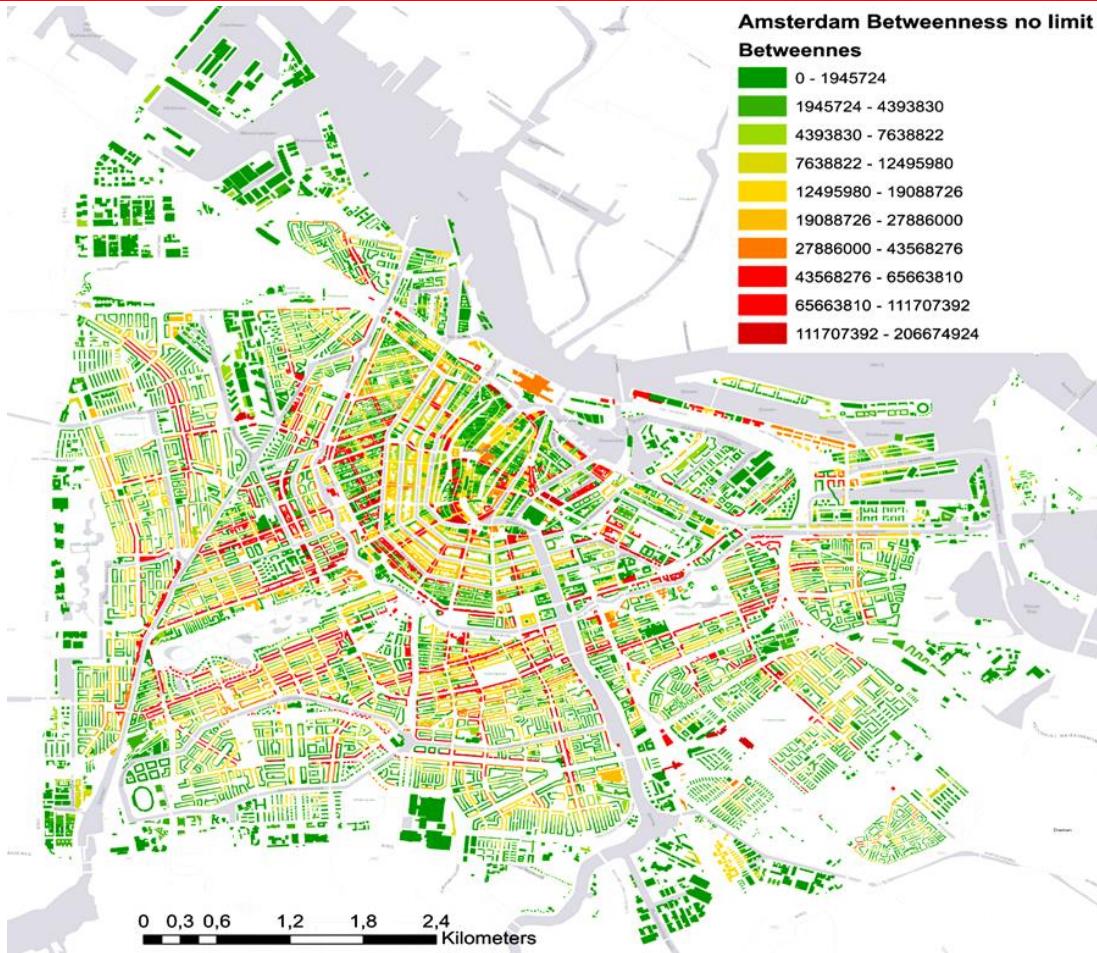
- Betweenness centrality
 - a centrality measure of a vertex within a graph
 - Vertices that occur on many shortest paths between other vertices have higher betweenness than those that do not
 - Act as “broker” or “bridge”
 - $O(V^3)$ complexity
 - $O(V^2 \log V + VE)$ for sparse network

$$C_B(v) = \sum_{\substack{s=v=t \in V \\ s \neq t}} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

σ_{st} is the geodesic path between s and t. $\sigma_{st}(v)$ is the geodesic path between s and t passing through v.



Medidas de Centralidad



Medidas de Centralidad (Eigenvector Centrality)

Eigenvector centrality [edit]

Eigenvector centrality is a measure of the influence of a [node](#) in a [network](#). It assigns relative scores to all nodes in the network based on the concept that connections to high-scoring nodes contribute more to the score of the node in question than equal connections to low-scoring nodes. [Google's PageRank](#) is a variant of the eigenvector centrality measure.^[24] Another closely related centrality measure is [Katz centrality](#).

Using the adjacency matrix to find eigenvector centrality [edit]

For a given graph $G := (V, E)$ with $|V|$ number of vertices let $A = (a_{v,t})$ be the [adjacency matrix](#), i.e. $a_{v,t} = 1$ if vertex v is linked to vertex t , and $a_{v,t} = 0$ otherwise. The centrality score of vertex v can be defined as:

$$x_v = \frac{1}{\lambda} \sum_{t \in M(v)} x_t = \frac{1}{\lambda} \sum_{t \in G} a_{v,t} x_t$$

where $M(v)$ is a set of the neighbors of v and λ is a constant. With a small rearrangement this can be rewritten in vector notation as the [eigenvector](#) equation

$$\mathbf{Ax} = \lambda \mathbf{x}$$

In general, there will be many different [eigenvalues](#) λ for which an eigenvector solution exists. However, the additional requirement that all the entries in the eigenvector be positive implies (by the [Perron–Frobenius theorem](#)) that only the greatest eigenvalue results in the desired centrality measure.^[25] The v^{th} component of the related eigenvector then gives the centrality score of the vertex v in the network. [Power iteration](#) is one of many [eigenvalue algorithms](#) that may be used to find this dominant eigenvector.^[24] Furthermore, this can be generalized so that the entries in A can be real numbers representing connection strengths, as in a [stochastic matrix](#).

Medidas de Centralidad (Katz and PageRank Centrality)

Katz centrality and PageRank [\[edit\]](#)

Main article: [Katz centrality](#)

Katz centrality^[26] is a generalization of degree centrality. Degree centrality measures the number of direct neighbors, and Katz centrality measures the number of all nodes that can be connected through a path, while the contributions of distant nodes are penalized. Mathematically, it is defined as $x_i = \sum_{k=1}^{\infty} \sum_{j=1}^N \alpha^k (A^k)_{ji}$ where α is an attenuation factor in $(0, 1)$.

Katz centrality can be viewed as a variant of eigenvector centrality. Another form of Katz centrality is $x_i = \alpha \sum_{j=1}^N a_{ij}(x_j + 1)$. Compared to the expression of eigenvector centrality, x_j is replaced by $x_j + 1$.

It is shown that^[27] the principal eigenvector (associated with the largest eigenvalue of A , the adjacency matrix) is the limit of Katz centrality as α approaches $1/\lambda$ from below.

PageRank satisfies the following equation $x_i = \alpha \sum_j a_{ji} \frac{x_j}{L(j)} + \frac{1-\alpha}{N}$, where $L(j) = \sum_j a_{ij}$ is the number of neighbors of node j (or number of outbound links in a directed graph). Compared to eigenvector centrality and Katz centrality, one major difference is the scaling factor $L(j)$. Another difference between PageRank and eigenvector centrality is that the PageRank vector is a left hand eigenvector (note the factor a_{ji} has indices reversed).^[28]

Medidas de Centralidad (Original PageRank Centrality)

Algoritmo [editar]

El algoritmo inicial del PageRank lo podemos encontrar en el documento original donde sus creadores presentaron el prototipo de Google: "The Anatomy of a Large-Scale Hypertextual Web Search Engine".²

$$\text{PR}(A) = (1 - d) + d \sum_{i=1}^n \frac{\text{PR}(i)}{C(i)}$$

Donde:

- $\text{PR}(A)$ es el PageRank de la página A.
- d es un factor de amortiguación que tiene un valor entre 0 y 1.
- $\text{PR}(i)$ son los valores de PageRank que tienen cada una de las páginas i que enlazan a A.
- $C(i)$ es el número total de enlaces salientes de la página i (sean o no hacia A).

Algunos expertos aseguran que el valor de la variable d suele ser 0,85. Representa la probabilidad de que un navegante continúe pulsando links al navegar por Internet en vez de escribir una url directamente en la barra de direcciones o pulsar uno de sus marcadores y es un valor establecido por Google. Por lo tanto, la probabilidad de que el usuario deje de pulsar links y navegue directamente a otra web aleatoria es $1-d$.³ La introducción del factor de amortiguación en la fórmula resta algo de peso a todas las páginas de Internet y consigue que las páginas que no tienen enlaces a ninguna otra página no salgan especialmente beneficiadas. Si un usuario aterriza en una página sin enlaces, lo que hará será navegar a cualquier otra página aleatoriamente, lo que equivale a suponer que *una página sin enlaces salientes tiene enlaces a todas las páginas de Internet*.

La calidad de la página y el número de posiciones que ascienda se determina por una "votación" entre todas las demás páginas de la World Wide Web acerca del nivel de importancia que tiene esa página. Un hiperenlace a una página cuenta como un voto de apoyo. El PageRank de una página se define **recursivamente** y depende del número y PageRank de todas las páginas que la enlazan. Una página que está enlazada por muchas páginas con un PageRank alto consigue también un PageRank alto. Si no hay enlaces a una página web, no hay apoyo a esa página específica. El PageRank de la barra de Google va de 0 a 10. Diez es el máximo PageRank posible y son muy pocos los sitios que gozan de esta calificación, 1 es la calificación mínima que recibe un sitio normal, y cero significa que el sitio ha sido penalizado o aún no ha recibido una calificación de PageRank. Parece ser una escala logarítmica. Los detalles exactos de esta escala son desconocidos. En los últimos tiempos Google está tratando de mantener un poco "privado" su PageRank para evitar manipulaciones, pero existen sitios donde se puede comprobar el PageRank.⁴

Una alternativa al algoritmo PageRank propuesto por Jon Kleinberg, es el algoritmo HITS.

Medidas de Centralidad (hubs and authorities Centrality)

Para comenzar el **ranking**, $\forall p$, $\text{auth}(p) = 1$ y $\text{hub}(p) = 1$. Consideramos dos tipos de actualizaciones: Regla de actualización de autoridad y Regla de actualización de concentrador. Para calcular las puntuaciones de concentrador / autoridad de cada nodo, se aplican iteraciones repetidas de la Regla de Actualización de Autoridades y la Regla de Actualización de Concentrador. Una aplicación en k-pasos del algoritmo HITS implica solicitar k veces primero la Regla de Actualización de Autoridades y luego la Regla de Actualización de Concentrador.

Regla de actualización de autoridad [\[editar\]](#)

$\forall p$, actualizamos $\text{auth}(p)$ para que sea la sumatoria:

$$\text{auth}(p) = \sum_{i=1}^n \text{hub}(i)$$

donde n es el número total de páginas conectadas a p e i es una página conectada a p. Es decir, la puntuación de la Autoridad de una página es la suma de todas las puntuaciones de concentrador de las páginas que apuntan a ella.

Regla de actualización de concentrador (Hub) [\[editar\]](#)

$\forall p$, actualizamos $\text{hub}(p)$ para que sea la sumatoria:

$$\text{hub}(p) = \sum_{i=1}^n \text{auth}(i)$$

donde n es el número total de páginas enlazadas desde p e i es una página enlazada desde p. Así, la puntuación de una página en el Concentrador es la suma de las puntuaciones de la Autoridad de todas sus páginas de enlace.

Medidas de Centralidad (Flow Centrality)

$$C_F(x_i) = \sum_{j < k}^n \sum_{j < k}^n m_{jk}(x_i). \quad (1')$$

If we divide the flow that passes through x_i by the total flow between all pairs of points where x_i is neither a source nor a sink, we can determine the proportion of the flow that depends on x_i

$$C'_F(p_i) = \frac{\sum_{j < k}^n m_{jk}(x_i)}{\sum_{j < k}^n \sum_{j < k}^n m_{jk}}. \quad (2')$$

Medidas de Centralidad

Ejercicio 8. Analizar y discutir diferentes medidas de centralidad en las siguientes redes ($N=50$):

Red simulada aleatoria Erdos-Renyi

Red simulada Scale-Free

Red simulada Small-world networks

Red de Football (Mundial1998).

Red de Karate club network.

Red del glosario de palabras de redes.

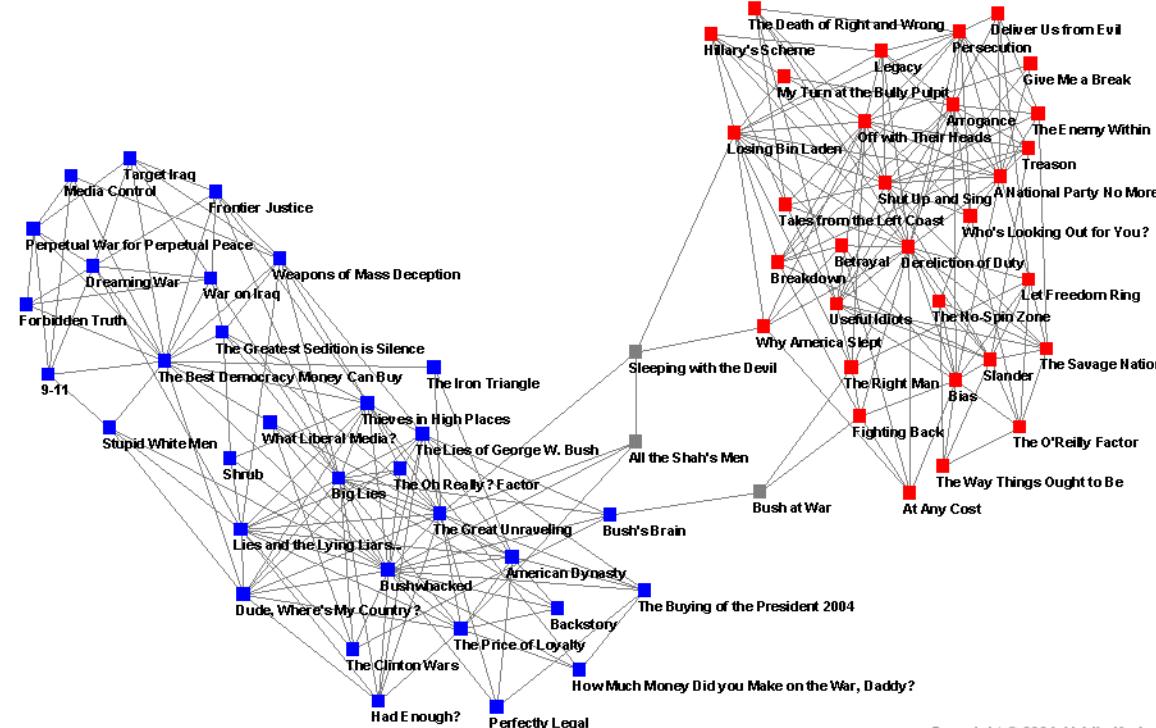
Red de los delfines.

4. ¿Qué individuos-nodos forman comunidades?

- Los problemas típicos de Clustering (sin redes) asumen que los ítem que se agrupan no están relacionados. Si lo están estos algoritmos son poco apropiados.
- En ocasiones la clasificación de un conjunto de objetos no se hace teniendo en cuenta sus características inherentes, si no en base a sus relaciones. Por ejemplo libros de política, votaciones a Juntas de Facultad, relaciones de amistad...

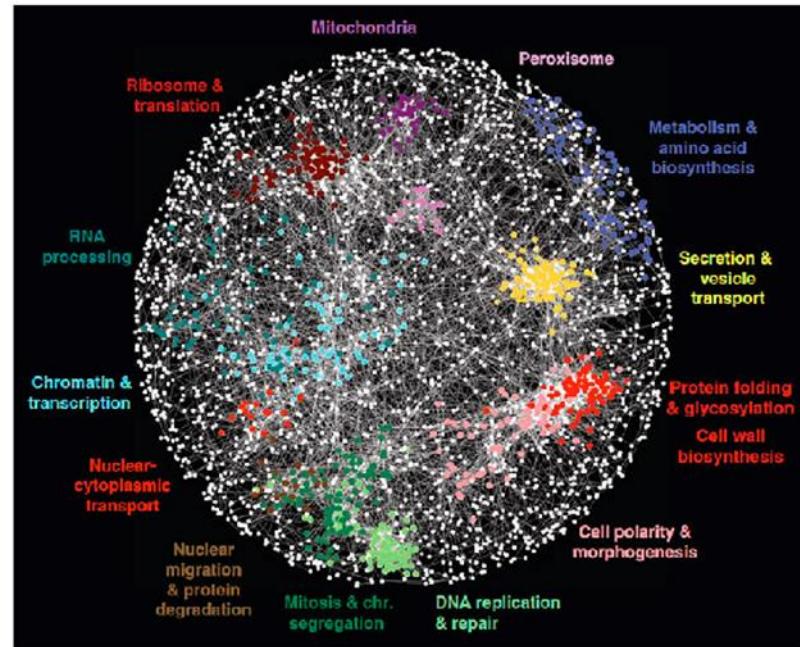
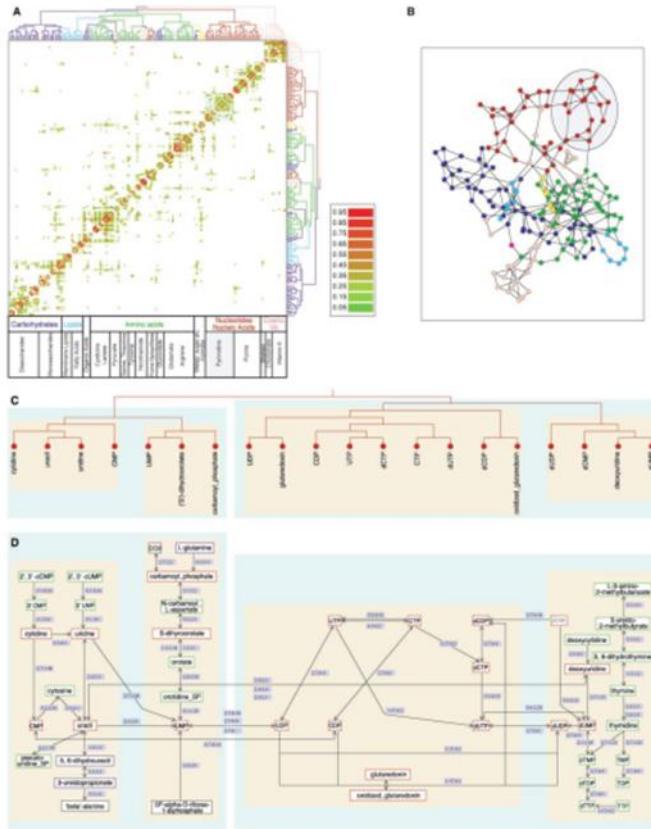


Clustering



Copyright © 2004, Valdis Krebs

Clustering



Clustering

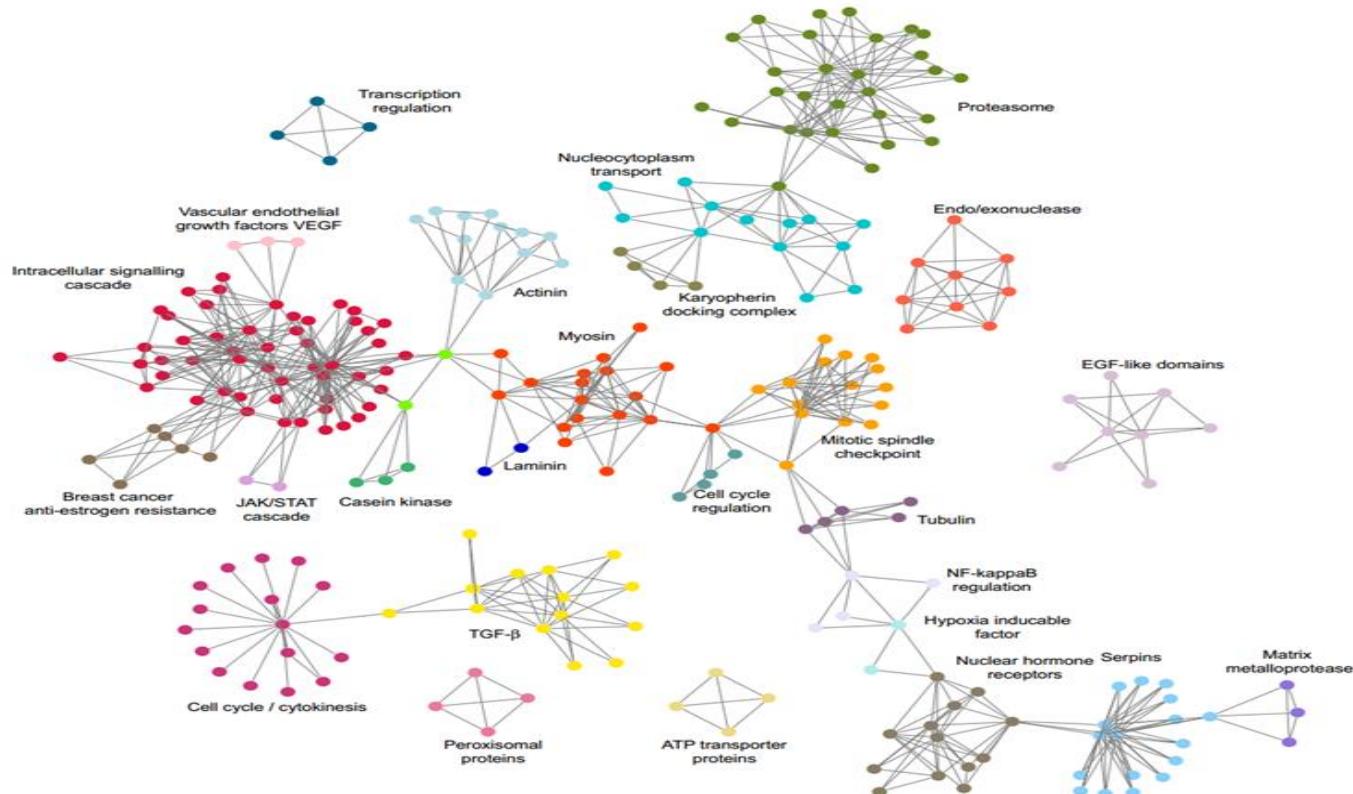
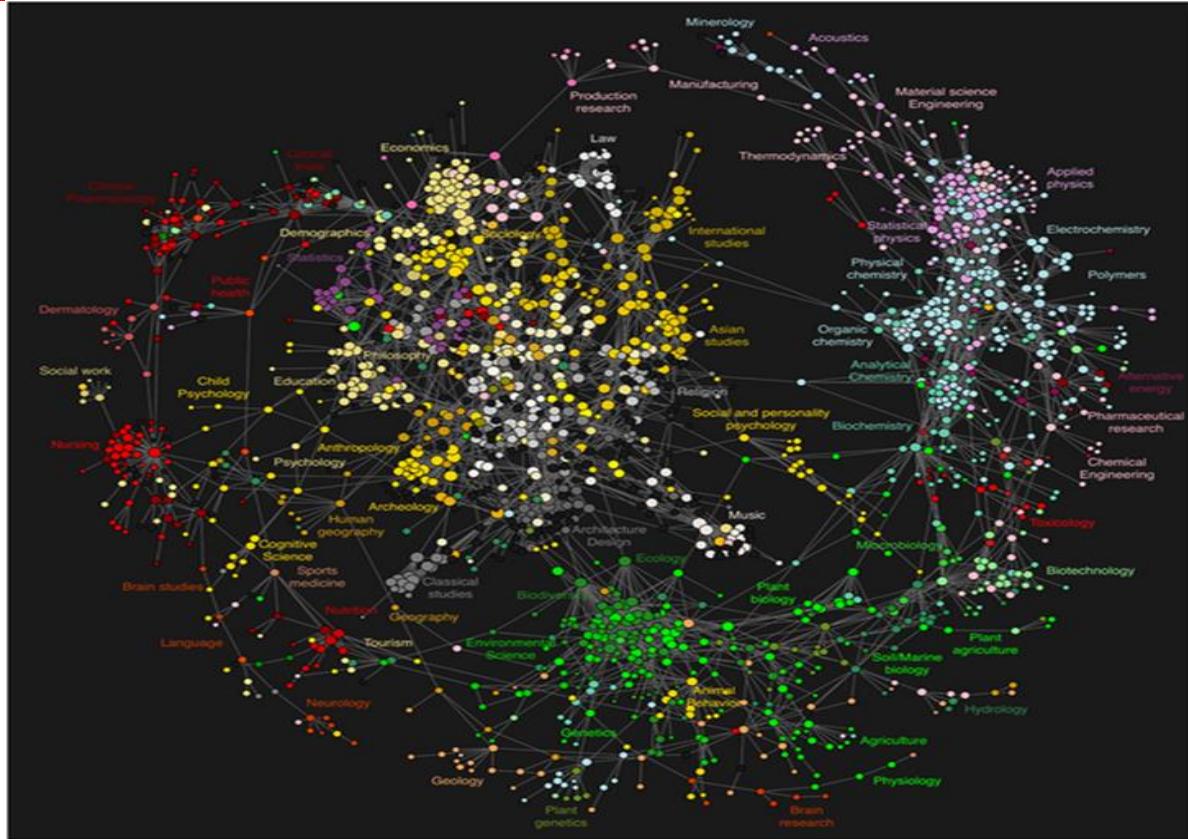


FIG. 3 Community structure in protein-protein interaction networks. The graph pictures the interactions between proteins in cancerous cells of a rat. Communities, labeled by colors, were detected with the Clique Percolation Method by Palla et al. (Section XI.A). Reprinted figure with permission from Ref. (Jonsson *et al.*, 2006). ©2006 by PubMed Central.

Clustering

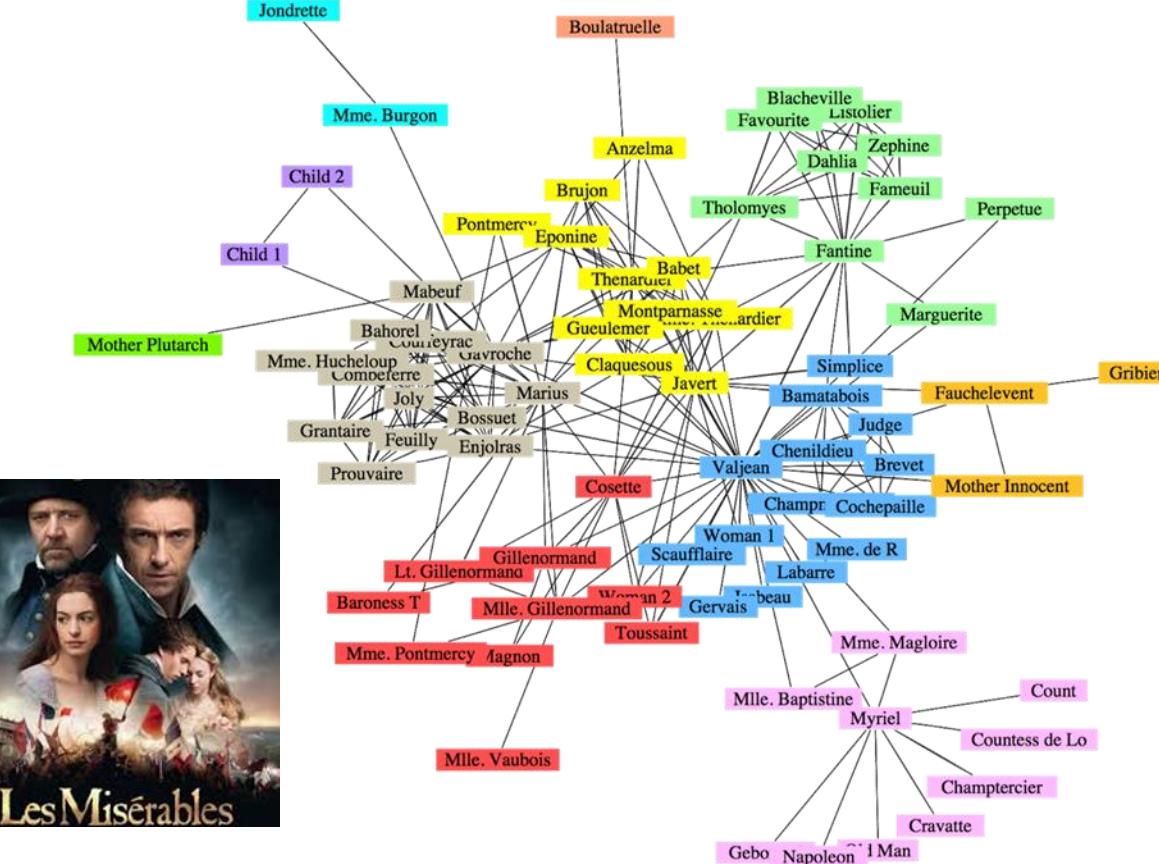


Bollen J, Van de Sompel H, Hagberg A, Bettencourt L, Chute R, et al. (2009) Clickstream Data Yields High-Resolution Maps of Science. PLoS ONE 4(3): e4803. doi:10.1371/journal.pone.0004803
<http://journals.plos.org/plosone/article?id=info:doi/10.1371/journal.pone.0004803>

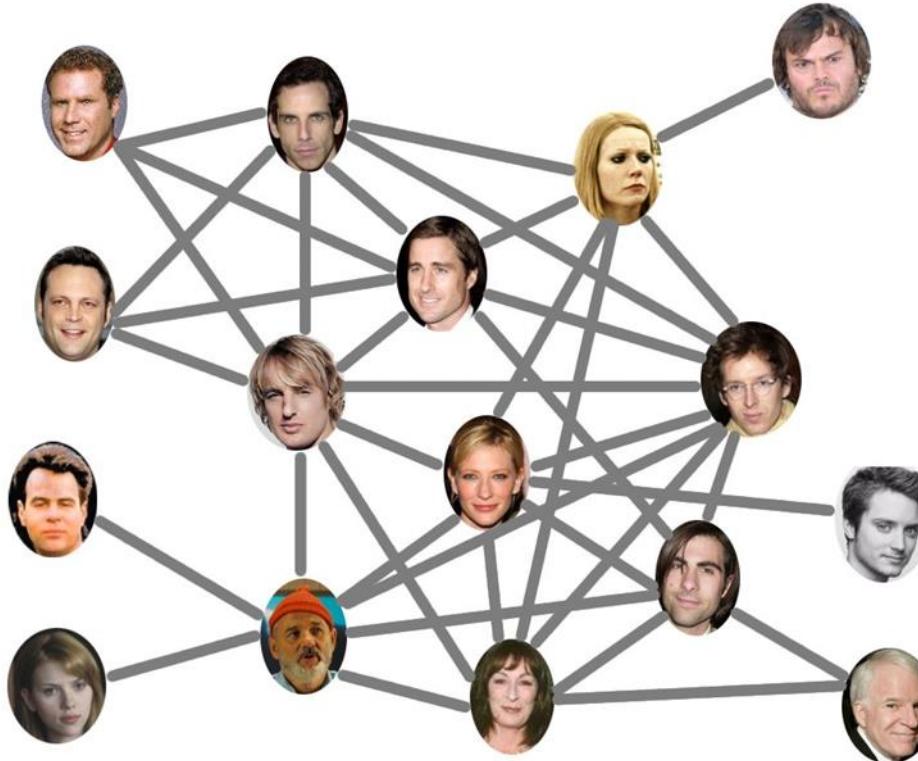
Clustering



Les Misérables



Clustering



Clustering

Los problemas de Clustering en Redes Sociales consiste en encontrar las comunidades, pero existen otros problemas muy relacionados como la predicción de los siguientes arcos o aristas o la clasificación de nodos dudosos

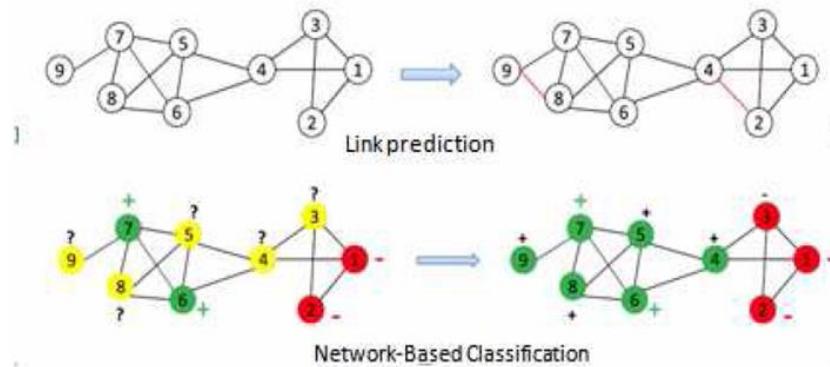


Figure: Link prediction problem and Network-Based Classification

Algoritmos de Clustering

Contents

I. Introduction	2	B. Random walk	45
II. Communities in real-world networks	4	C. Synchronization	47
III. Elements of Community Detection	8	IX. Methods based on statistical inference	48
A. Computational complexity	9	A. Generative models	49
B. Communities	9	B. Blockmodeling, model selection and information theory	52
1. Basics	9	X. Alternative methods	54
2. Local definitions	10	XI. Methods to find overlapping communities	58
3. Global definitions	11	A. Clique percolation	58
4. Definitions based on vertex similarity	12	B. Other techniques	60
C. Partitions	13	XII. Multiresolution methods and cluster hierarchy	62
1. Basics	13	A. Multiresolution methods	63
2. Quality functions: modularity	13	B. Hierarchical methods	65
IV. Traditional methods	16	XIII. Detection of dynamic communities	66
A. Graph partitioning	16	XIV. Significance of clustering	70
B. Hierarchical clustering	19	XV. Testing Algorithms	73
C. Partitional clustering	19	A. Benchmarks	74
D. Spectral clustering	20	B. Comparing partitions: measures	77
V. Divisive algorithms	23	C. Comparing algorithms	79
A. The algorithm of Girvan and Newman	23	XVI. General properties of real clusters	82
B. Other methods	25	XVII. Applications on real-world networks	85
VI. Modularity-based methods	27	A. Biological networks	85
A. Modularity optimization	27	B. Social networks	86
1. Greedy techniques	27	C. Other networks	88
2. Simulated annealing	29	XVIII. Outlook	90
3. Extremal optimization	29	A. Elements of Graph Theory	92
4. Spectral optimization	30	1. Basic Definitions	92
5. Other optimization strategies	33	2. Graph Matrices	94
B. Modifications of modularity	34	3. Model graphs	94
C. Limits of modularity	38	References	96
VII. Spectral Algorithms	41		
VIII. Dynamic Algorithms	43		
A. Spin models	43		

Algoritmos de Clustering (No Jerárquicos)

Su objetivo principal es el de obtener una partición del grafo. Como se ha llegado a esta partición no es problema suyo. La salida es solo la partición final.

Los distintos tipos son:

- ✓ Métodos Tradicionales. Clustering clásico. No funcionan bien.
- ✓ Metodos de Graph partitioning clásicos. Max-Min Flow algorithm.....
- ✓ Spectral Algorithms. Transforman los nodos de un grafo en puntos en el plano y a partir de ahí aplican técnicas clásicas de clustering.
- ✓ Algoritmos basados en la función de modularidad. Optimización.
- ✓ LOUVAIN (También conocido como BLONDEL). Es el más conocido.
- ✓ Algoritmos genéticos.
- ✓ Temple simulado.
- ✓

Algoritmos de Clustering (Jerárquicos)

Muestran la evolución de como los grupos se van rompiendo (divisivos) o formando (aglomerativos) desde la primera fase hasta la ultima. La salida es un dendograma.

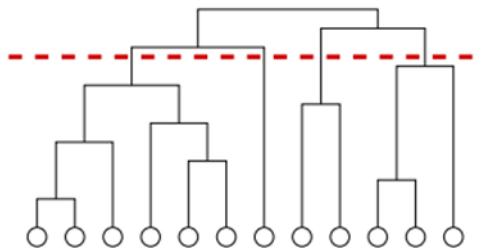


FIG. 8 A dendrogram, or hierarchical tree. Horizontal cuts correspond to partitions of the graph in communities. Reprinted figure with permission from Ref. (Newman and Girvan, 2004). ©2004 by the American Physical Society.

Algoritmos de Clustering (Jerárquicos)

Definition

Dado un grafo (V, E) , *Clustering Network Problem* (CNP) busca una "buena" partición del grafo.

Definition

Dado un grafo (V, E) , $P = \{H_1, \dots, H_r\}$ es una partición del grafo si $V = \cup H_i$, $H_i \cap H_j = \emptyset$ para i distinto de j , y el grafo parcial $(H_i, E|_{H_i})$ es conexo.

- Grafo modeliza una imagen: Aplicacion al campo de la Segmentación de imagenes.
- Grafo modeliza una red social. Aplicacion al campo de los Problemas de detección de comunidades.

Algoritmos de Clustering (Jerárquicos)

Definition

Dado un grafo $G = (V, E)$, the Hierarchical Clustering Network Problem (HCNP) busca una buena partición jerárquica \mathcal{D} del grafo. Una partición jerárquica se suele representar por medio de un dendograma.

Partición mas fina .

Definition

Dadas dos particiones P y Q , diremos que P es mas fina que Q ($P \widetilde{\subseteq} Q$) si para toda $A \in P$, existe $B \in Q$, tal que $A \subseteq B$.

Algoritmos de Clustering (Jerárquicos)

Los distintos tipos son:

1) Divisivos.

- ✓ Girvan-Newman (2003).
- ✓ Gomez & Castro (2018)
- ✓ Divide and Link (2015)

2) Aglomerativos.

- ✓ Clauset Newman and Moore
- ✓ Random Walk.
- ✓ Newman 2015

Random Hypothesis:

Randomly wired networks are not expected to have a community structure.

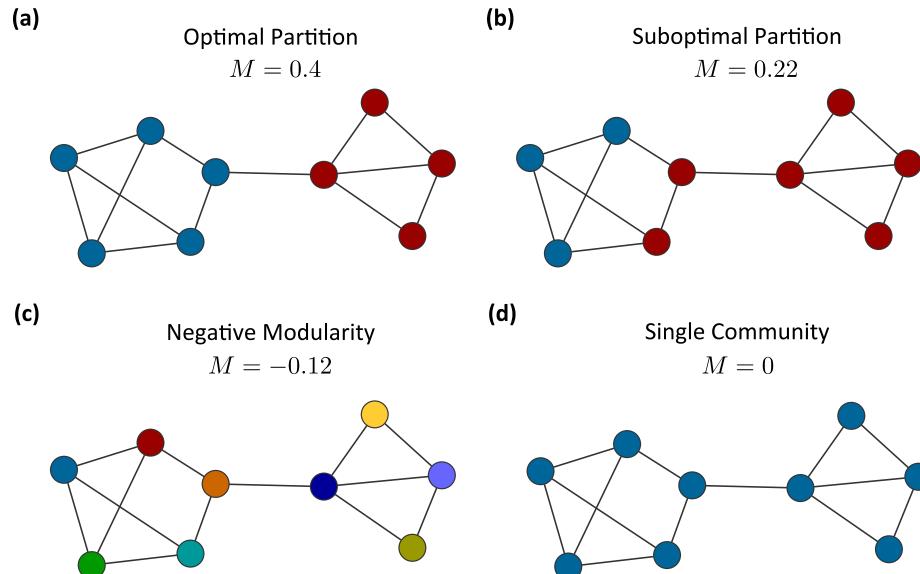
$$\{C_c, c = 1, n_c\}$$

$$M = \frac{1}{2L} \sum_{i,j=1}^N (A_{ij} - P_{ij}) \delta(C_i - C_j)$$

$$P_{ij} = 2L p_i p_j = \frac{k_i k_j}{2L}$$

$$M = \sum_{c=1}^{n_c} \left[\frac{l_c}{L} - \left(\frac{k_c}{2L} \right)^2 \right]$$

$$M = \sum_{c=1}^{n_c} \left[\frac{l_c}{L} - \left(\frac{k_c}{2L} \right)^2 \right]$$



- *Optimal partition*, that maximizes the modularity.
- *Sub-optimal* but positive modularity.
- *Negative Modularity*: If we assign each node to a different community.
- *Zero modularity*: Assigning all nodes to the same community, we obtain , independent of the network structure.
- *Modularity is size dependent*.

Algoritmos de Clustering (Louvain)

The value to be optimized is [modularity](#), defined as a value between -1 and 1 that measures the density of links inside communities compared to links between communities.^[1] For a weighted graph, modularity is defined as:

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j),$$

where

- A_{ij} represents the edge weight between nodes i and j ;
- k_i and k_j are the sum of the weights of the edges attached to nodes i and j , respectively;
- $2m$ is the sum of all of the edge weights in the graph;
- c_i and c_j are the communities of the nodes; and
- δ is a simple [delta function](#).

In order to maximize this value efficiently, the Louvain Method has two phases that are repeated [iteratively](#).

First, each node in the network is assigned to its own community. Then for each node i , the change in modularity is calculated for removing i from its own community and moving it into the community of each neighbor j of i . This value is easily calculated by two steps: (1) removing i from its original community, and (2) inserting i to the community of j . The two equations are quite similar, and the equation for step (2) is:^[1]

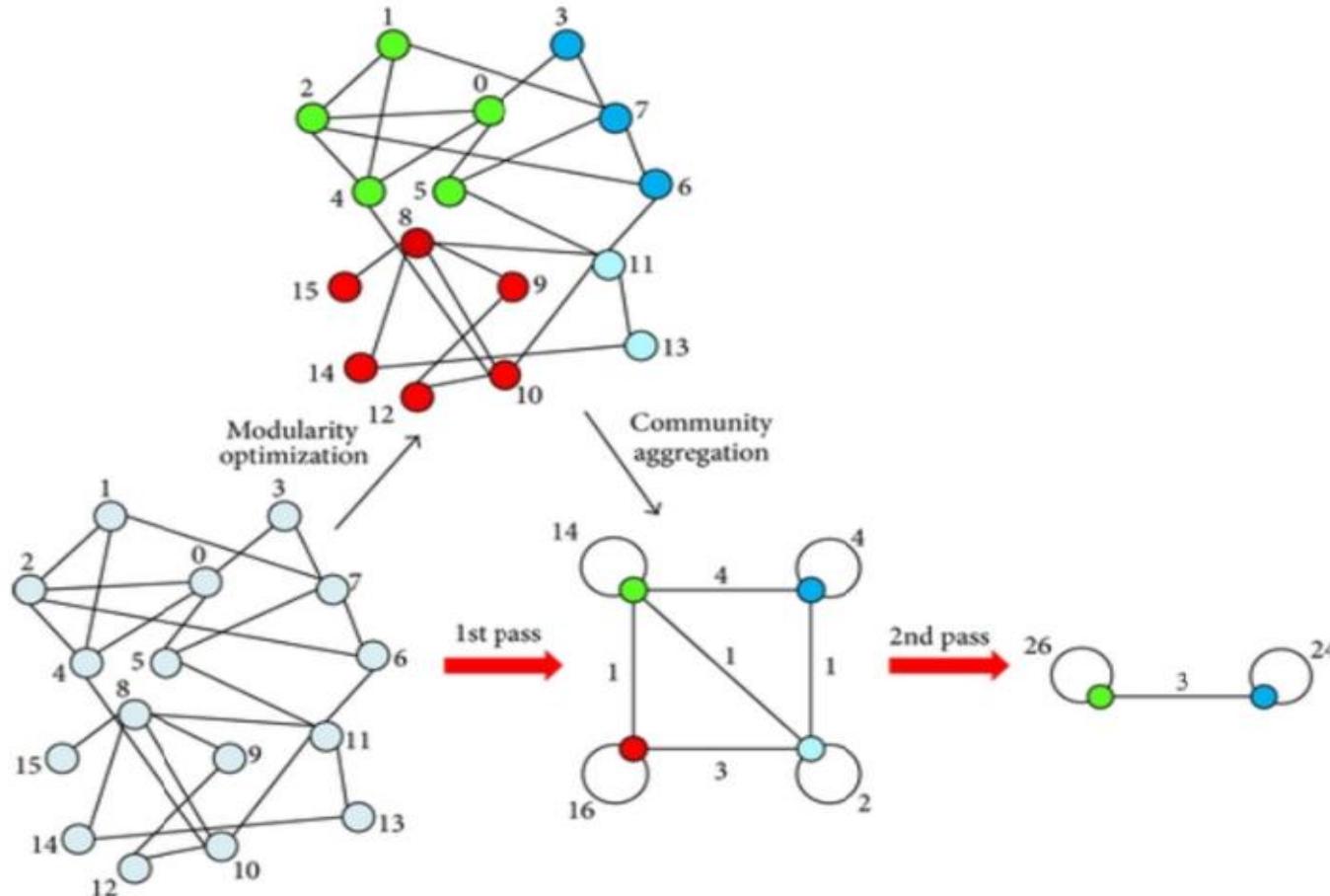
$$\Delta Q = \left[\frac{\Sigma_{in} + 2k_{i,in}}{2m} - \left(\frac{\Sigma_{tot} + k_i}{2m} \right)^2 \right] - \left[\frac{\Sigma_{in}}{2m} - \left(\frac{\Sigma_{tot}}{2m} \right)^2 - \left(\frac{k_i}{2m} \right)^2 \right]$$

Algoritmos de Clustering (Louvain)

Where Σ_{in} is sum of all the weights of the links inside the community i is moving into, Σ_{tot} is the sum of all the weights of the links to nodes in the community i is moving into, k_i is the weighted degree of i , $k_{i,in}$ is the sum of the weights of the links between i and other nodes in the community that i is moving into, and m is the sum of the weights of all links in the network. Then, once this value is calculated for all communities i is connected to, i is placed into the community that resulted in the greatest modularity increase. If no increase is possible, i remains in its original community. This process is applied repeatedly and sequentially to all nodes until no modularity increase can occur. Once this local maximum of modularity is hit, the first phase has ended.

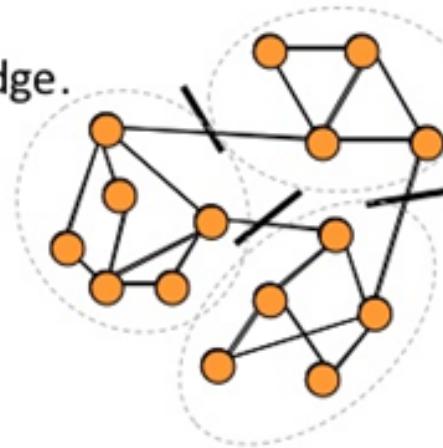
In the second phase of the algorithm, it groups all of the nodes in the same community and builds a new network where nodes are the communities from the previous phase. Any links between nodes of the same community are now represented by self loops on the new community node and links from multiple nodes in the same community to a node in a different community are represented by weighted edges between communities. Once the new network is created, the second phase has ended and the first phase can be re-applied to the new network.

Algoritmos de Clustering (Louvain)



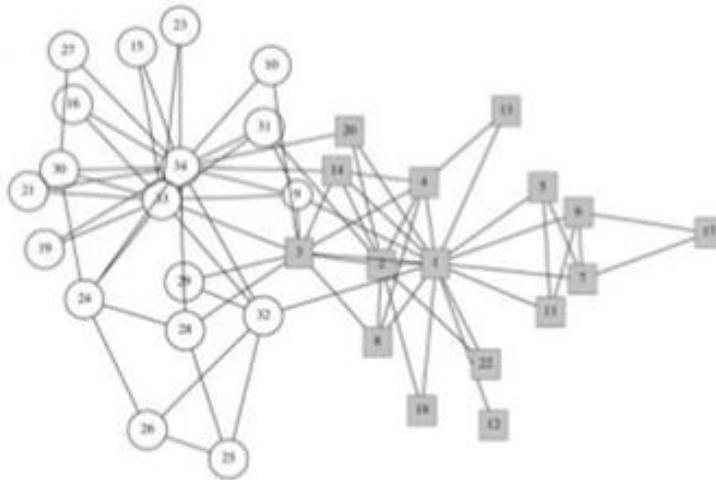
Algoritmos de Clustering (Girvan-Newman)

- GN algorithm is one of the most important algorithms stimulating a whole wave of community detection methods.
- Basic principle:
 - > Compute betweenness centrality for each edge.
 - > Remove edge with highest score.
 - > Re-compute all scores.
 - > Repeat 2nd step.
- Complexity: $O(n^3)$
- Many variations have been presented to improve precision by use of different betweenness measures or reduce complexity, e.g. by sampling or local computations.



Girvan, M., Newman, M.E.J. "Community structure in social and biological networks". In Proceedings of National Academy of Science, U. S. A. 99(12), 7821–7826, 2002

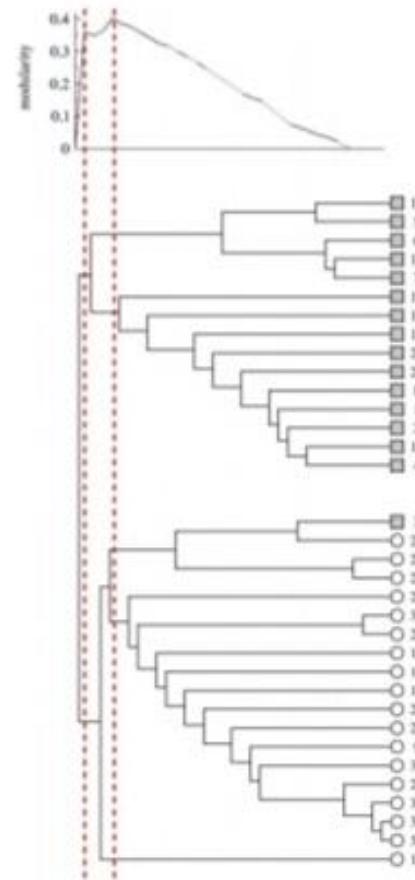
Algoritmos de Clustering (Girvan-Newman)



Optimal community structure for Zachary's karate club.



Modularity without recalculation



Clustering

Ejercicio 9. Analizar y discutir diferentes clusterin en las siguientes redes (N=50):

Red simulada aleatoria Erdos-Renyi

Red simulada Scale-Free

Red simulada Small-world networks

Red de Football (Mundial1998).

Red de Karate club network.

Red de los delfines.

Red del glosario de palabras de redes.