Complex Networks

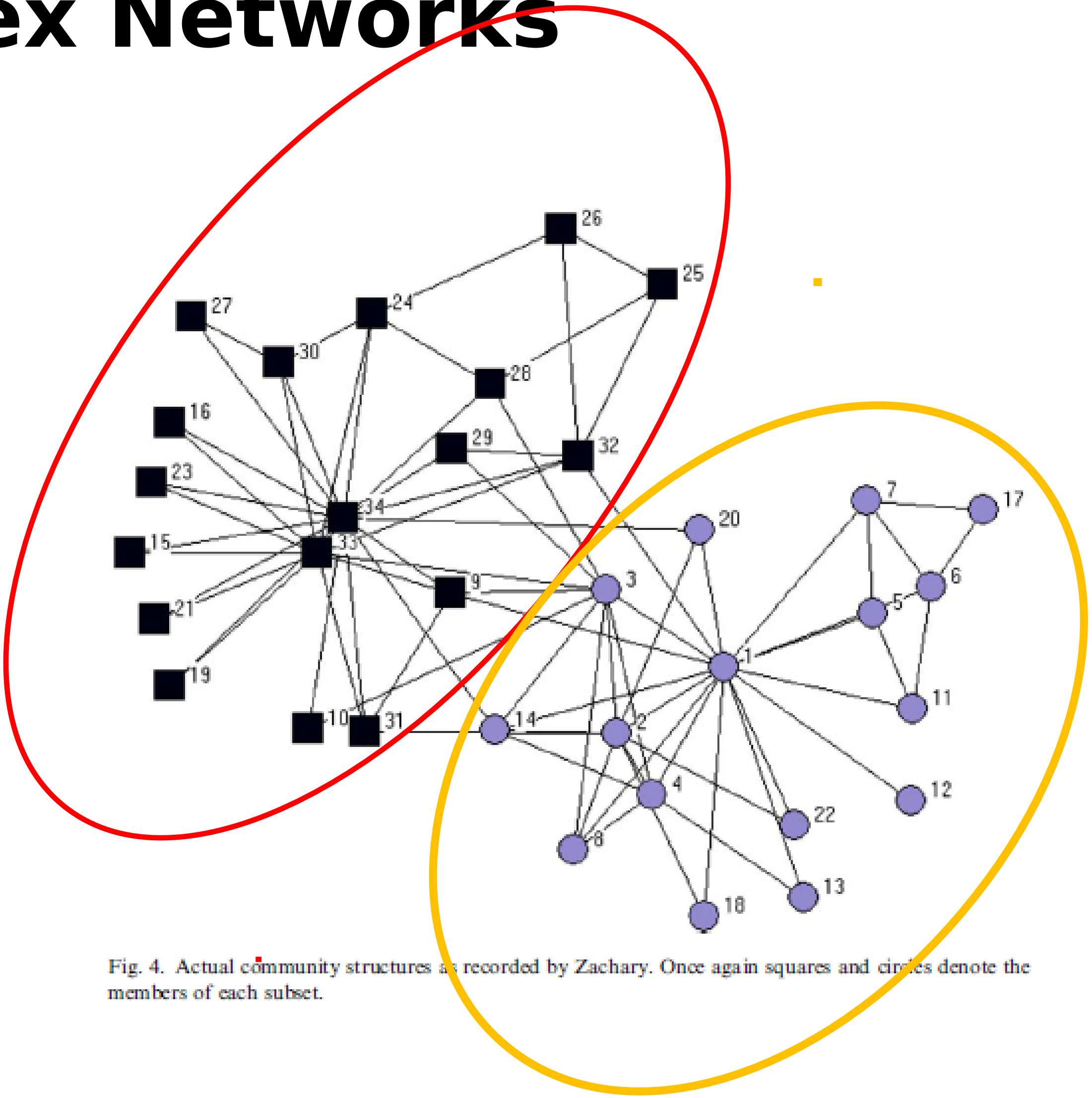


Fig. 4. Actual community structures as recorded by Zachary. Once again squares and circles denote the members of each subset.

REVIEWS OF MODERN PHYSICS, VOLUME 74, JANUARY 2002

Statistical mechanics of complex networks

Réka Albert* and Albert-László Barabási

Department of Physics, University of Notre Dame, Notre Dame, Indiana 46556

(Published 30 January 2002)

Complex networks describe a wide range of systems in nature and society. Frequently cited examples include the cell, a network of chemicals linked by chemical reactions, and the Internet, a network of routers and computers connected by physical links. While traditionally these systems have been modeled as random graphs, it is increasingly recognized that the topology and evolution of real networks are governed by robust organizing principles. This article reviews the recent advances in the field of complex networks, focusing on the statistical mechanics of network topology and dynamics. After reviewing the empirical data that motivated the recent interest in networks, the authors discuss the main models and analytical tools, covering random graphs, small-world and scale-free networks, the emerging theory of evolving networks, and the interplay between topology and the network's robustness against failures and attacks.

ELSEVIER

Available online at www.sciencedirect.com

SCIENCE DIRECT

Physics Reports ■■■ (■■■■) ■■■–■■■

PHYSICS REPORTS

www.elsevier.com/locate/physrep

Complex networks: Structure and dynamics

S. Boccaletti[a,*], V. Latora[b,c], Y. Moreno[d,e], M. Chavez[f], D.-U. Hwang[a]

[a] *CNR-Istituto dei Sistemi Complessi, Largo E. Fermi, 6, 50125 Florence, Italy*
[b] *Dipartimento di Fisica e Astronomia, Universitá di Catania, Via S. Sofia, 64, 95123 Catania, Italy*
[c] *Istituto Nazionale di Fisica Nucleare, Sezione di Catania, Via S. Sofia, 64, 95123 Catania, Italy*
[d] *Instituto de Biocomputación y Física de Sistemas Complejos, Universidad de Zaragoza, Zaragoza 50009, Spain*
[e] *Departamento de Fisica Teórica, Universidad de Zaragoza, Zaragoza 50009, Spain*
[f] *Laboratoire de Neurosciences Cognitives et Imagerie Cérébrale (LENA) CNRS UPR-640, Hôpital de la Salpêtrière. 47 Bd. de l'Hôpital, 75651 Paris CEDEX 13, France*

Accepted 27 October 2005

editor : I. Procaccia

Networks, Dynamics, and the Small-World Phenomenon

Duncan J. Watts

American Journal of Sociology, Volume 105, Issue 2 (Sep., 1999), 493-527.

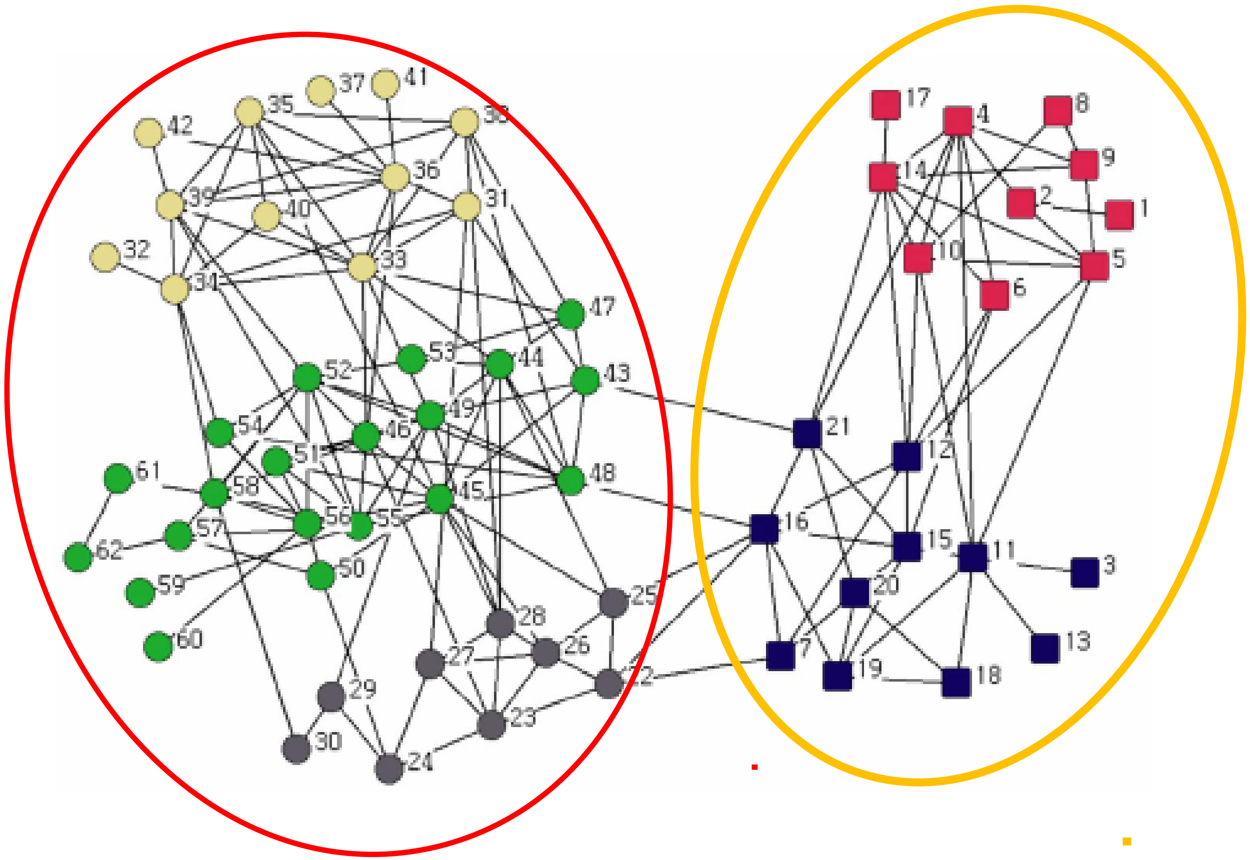


FIG. 4. (Color online) Bottlenose dolphin network. This network has a size of 62 nodes and it is known from direct observation that it has two communities. In this figure squares and circles denote the communities detected by our strong community approach and the colors (shades of gray or colors online) show the results of the weak community approach. Notice that the optimization according to Q_W merely subdivides the communities obtained through Q_S optimization [17].

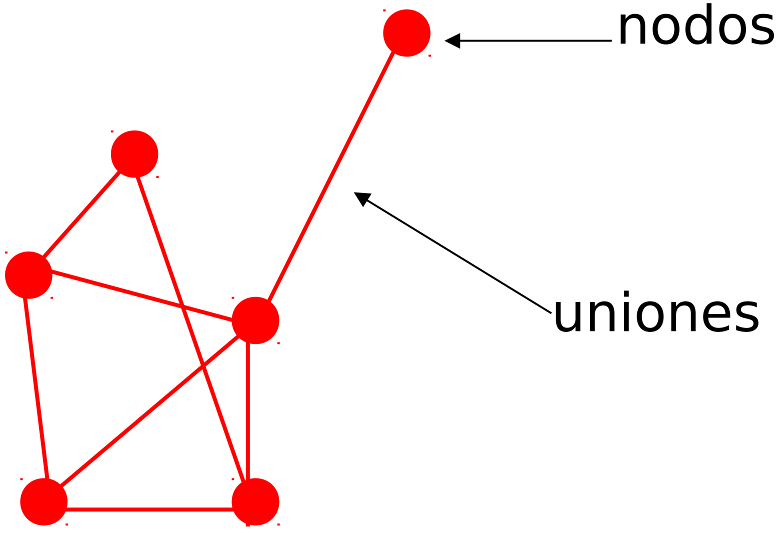
PHYSICAL REVIEW E **79**, 066111 (2009)

Alternative approach to community detection in networks

A. D. Medus[*] and C. O. Dorso[†]
Departamento de Física, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Pabellón 1, Ciudad Universitaria, Ciudad Autónoma de Buenos Aires 1428, Argentina
(Received 28 July 2008; revised manuscript received 21 May 2009; published 23 June 2009)

Que es una red?

Una red esta compuesta por



Entonces, una red es un conjunto de n nodos unidos por M uniones

$$M << \frac{n(n-1)}{2}$$

(diluido)

Los nodos :

a) en sistemas sociales suelen representar personas

b) en sistemas tecnológicos como el de transmisión de electricidad representan estaciones, subestaciones, etc.

c) en sistemas de información como el WWW representan paginas

Las uniones :

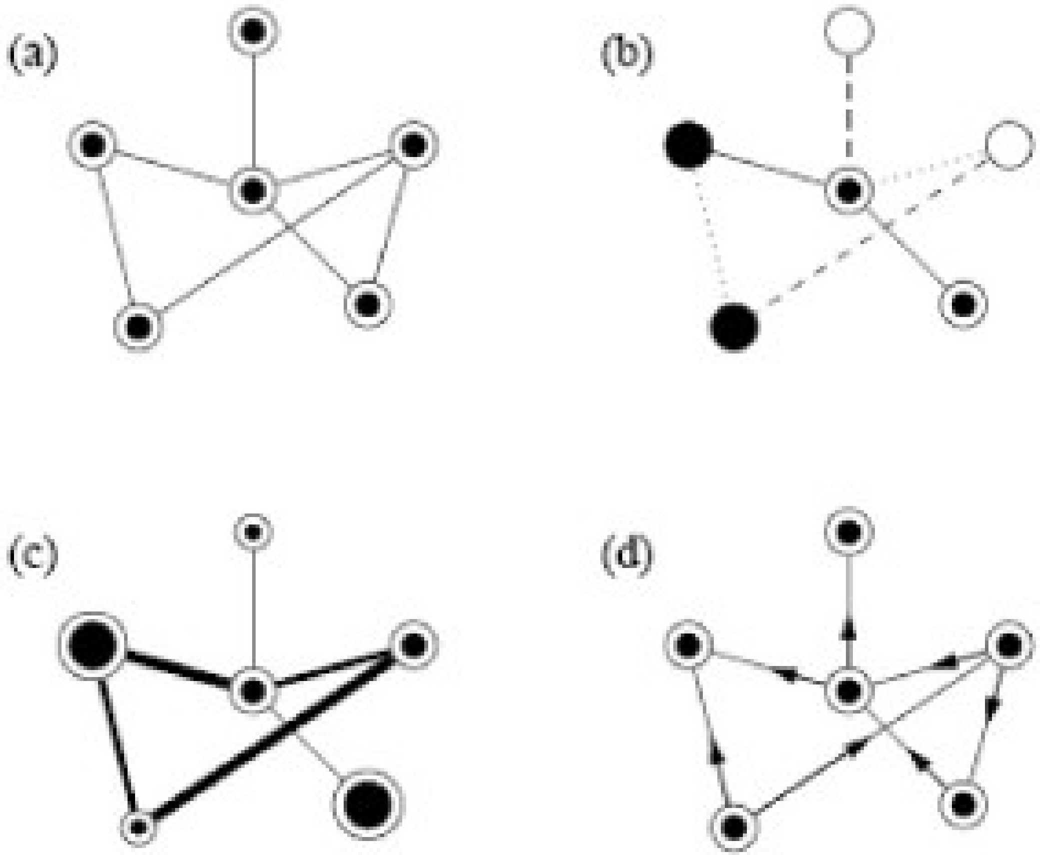
En a) pueden representar amistad, enemistad, lazos familiares, etc.

En b) representan las líneas de transmisión entre nodos

En c) los links

Tipos de Networks

homogéneo



heterogéneo

pesado

dirigido

Elementos de un Network

Vertex (pl. vertices): The fundamental unit of a network, also called a site (physics), a node (computer science), or an actor (sociology).

Edge: The line connecting two vertices. Also called a bond (physics), a link (computer science), or a tie (sociology).

Directed/undirected: An edge is directed if it runs in only one direction (such as a one-way road between two points), and undirected if it runs in both directions. Directed edges, which are sometimes called *arcs*, can be thought of as sporting arrows indicating their orientation. A graph is directed if all of its edges are directed. An undirected graph can be represented by a directed one having two edges between each pair of connected vertices, one in each direction.

Degree: The number of edges connected to a vertex. Note that the degree is not necessarily equal to the number of vertices adjacent to a vertex, since there may be more than one edge between any two vertices. In a few recent articles, the degree is referred to as the "connectivity" of a vertex, but we avoid this usage because the word connectivity already has another meaning in graph theory. A directed graph has both an in-degree and an out-degree for each vertex, which are the numbers of in-coming and out-going edges respectively.

Component: The component to which a vertex belongs is that set of vertices that can be reached from it by paths running along edges of the graph. In a directed graph a vertex has both an in-component and an out-component, which are the sets of vertices from which the vertex can be reached and which can be reached from it.

Geodesic path: A geodesic path is the shortest path through the network from one vertex to another. Note that there may be and often is more than one geodesic path between two vertices.

Diameter: The diameter of a network is the length (in number of edges) of the longest geodesic path between any two vertices. A few authors have also used this term to mean the *average* geodesic distance in a graph, although strictly the two quantities are quite distinct.

Caminos y conectividad

En una red conexa → todo nodo es “alcanzable” desde cualesquiera nodo

Podemos introducir el Camino P_{i_0,i_n}

Para un grafo $G=(V,E)$

P_{i_0,i_n} Corresponde a un conjunto ordenado de ***n*** nodos y ***(n-1)*** links

$$V_P=\{i_0,\ldots,i_n\} \qquad E_P=\{(i_0,i_1),\ldots,(i_{n-1},i_n)\}$$

Mundo pequeño

(el mundo es un pañuelo)

Experimento de Milgram (1967):

Envió paquetes a personas elegidas al azar en el medio oeste
De USA y les pidió que se los hiciesen llegar a personas en Boston
Pero debían hacerlo a través de personas que conociesen por el
Nombre (debían llegar por amigos de amigos de amigos…).

Como resultado de este experimento es que en valor medio
Eran suficientes “6 amigos”,para alcanzar el éxito.

Esto fue el inicio de una enorme serie de trabajos en el tema.

Herramientas para estudiar el efecto de mundo pequeño

Distancia mínima promedio en una red :

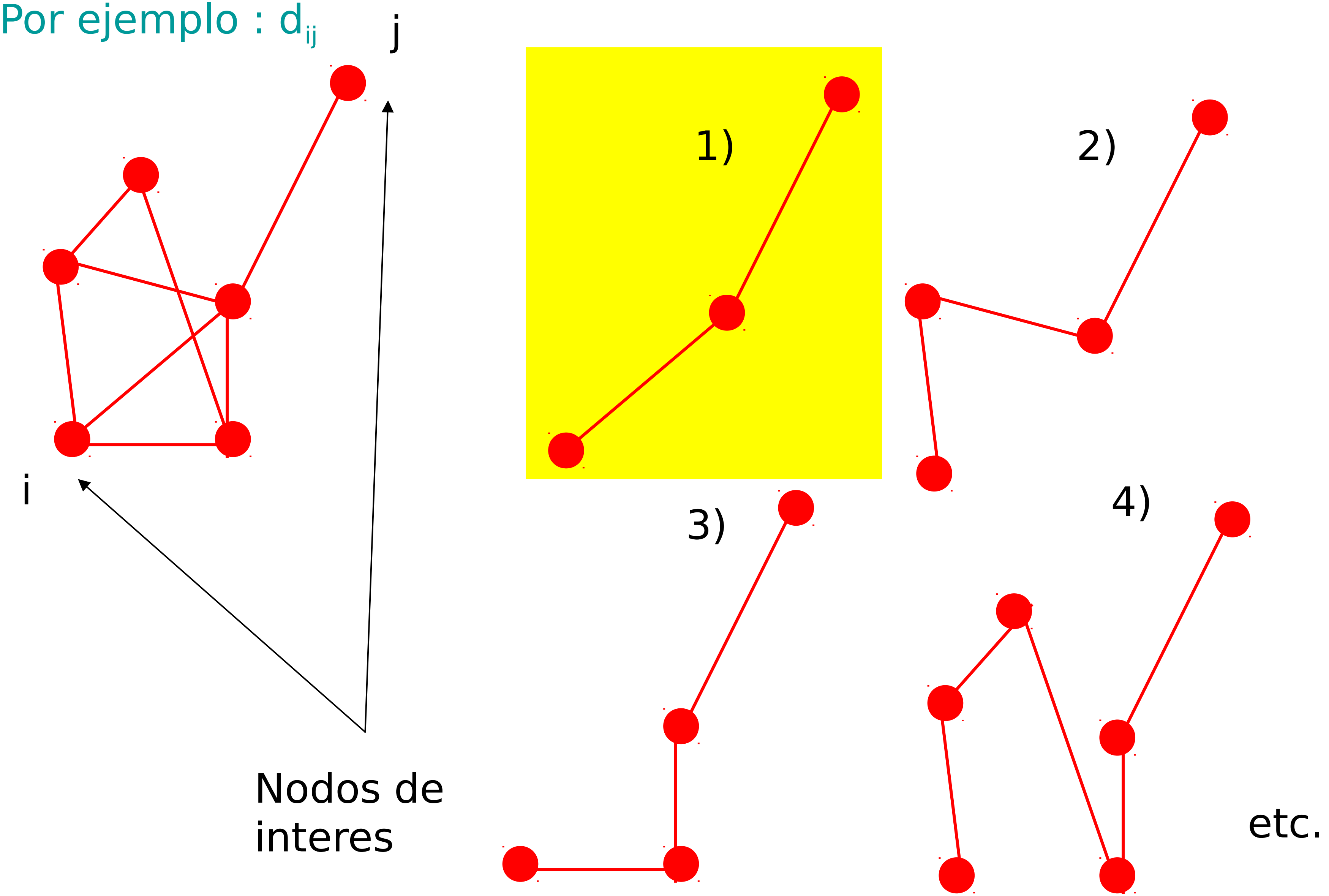
Suma sobre todos
Los pared de nodos

$$l(G) = \frac{1}{\left[n(n-1)\right]/2} \sum_{i \leq j \in G} d_{ij}$$

d_{ij} es el camino mínimo entre los nodos i y j

[n(n-1)]/2 es el numero pares de nodos para n nodos

Por ejemplo : d_{ij}



$$l = \frac{1}{\frac{1}{2}n(n+1)} \sum_{i \geq j} d_{ij}$$

Donde d_{ij} es la distancia minima (en terminos del numero de pasos) entre los nodos i y j. O sea la geodesica entre esos dos puntos.(ojo se incluye la distancia a si mismo que es 0)

Esto esta bien definido si trabajamos con networks conexos, pero si no hacemos un analisis preliminar de clusters podrian aparecer infinitos

Quizas entonces es mejor usar

$$l^{-1} = \frac{1}{\frac{1}{2}n(n+1)} \sum_{i \geq j} d_{ij}^{-1}$$

Clusterizacion

$$C(G) = \frac{1}{n}\sum_i C_i$$

Donde C_i responde la siguiente pregunta

Que fracción de mis amigos son amigos entre si?

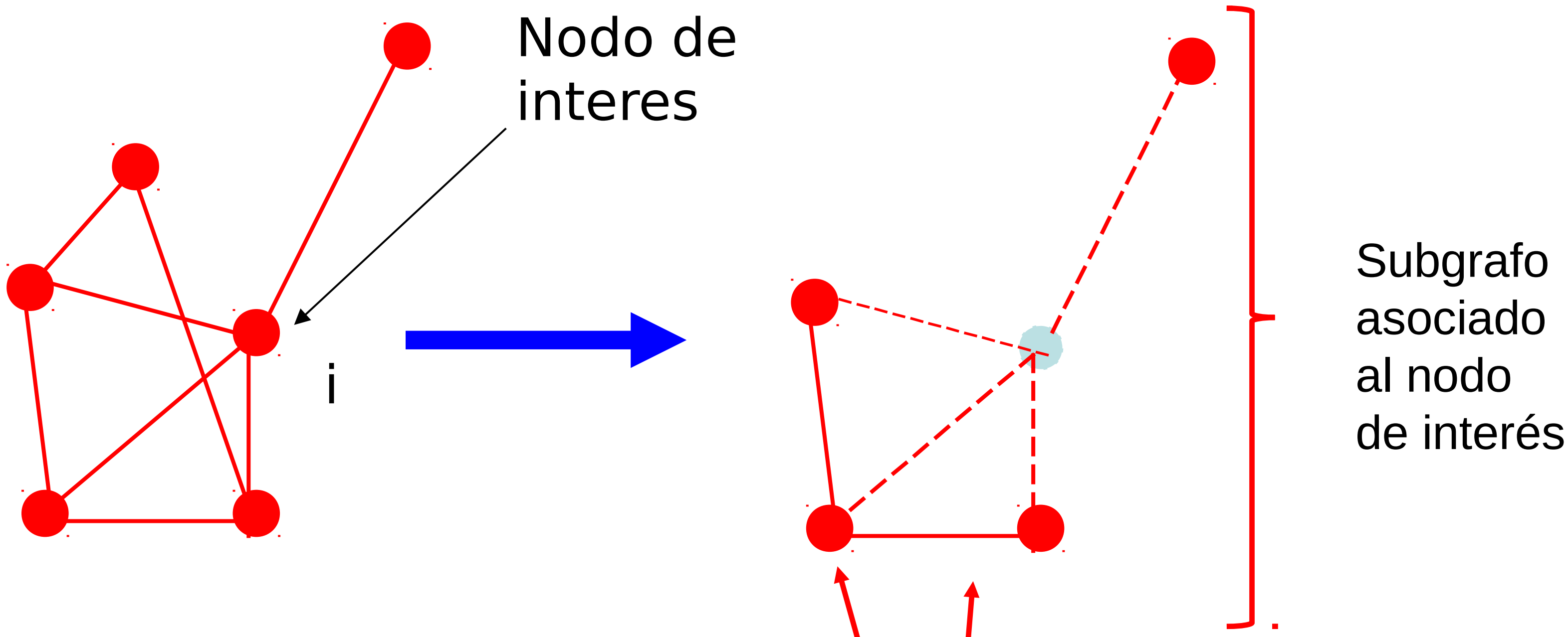
$$C_i = \frac{E_i}{k_i(k_i - 1)/2}$$

E_i es el numero de uniones entre los k_i vecinos de i

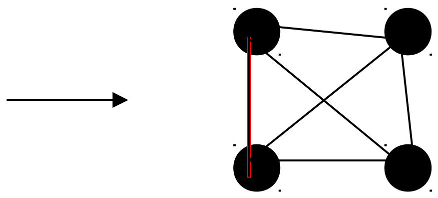
Numero de pares de nodos con ki el numero de nodos vecinos de i

Si todos mis amigos son amigos entre si entonces C_i=1

Por ejemplo, para calcular C_i



Numero de nodos vecinos al de interés : 4
Numero máximo de uniones de 4 nodos : 6
Uniones presentes : 2



$C_i = 2/6$

Coeficiente de Clusterizacion

Para el grafo “global”

$$\frac{\#\ \text{triangles} \times 3}{\#\ \text{connected triples}}$$

Cada triangulo X 3

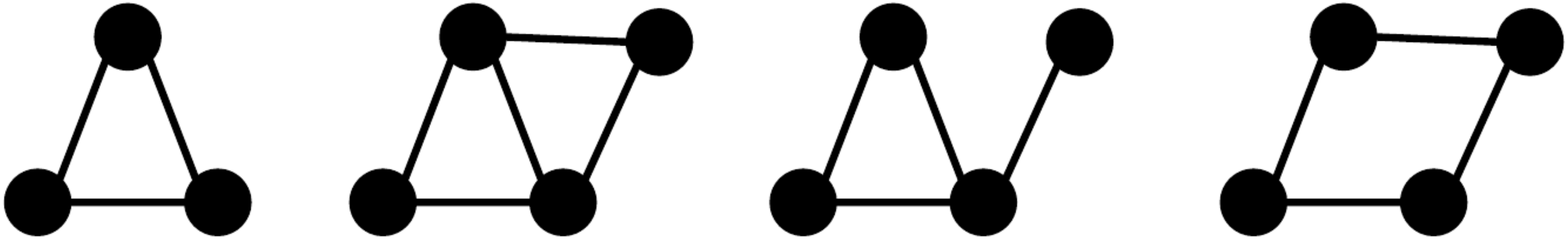


TABLE I. The general characteristics of several real networks. For each network we have indicated the number of nodes, the average degree $\langle k \rangle$, the average path length ℓ, and the clustering coefficient C. For a comparison we have included the average path length ℓ_{rand} and clustering coefficient C_{rand} of a random graph of the same size and average degree. The numbers in the last column are keyed to the symbols in Figs. 8 and 9.

Network	Size	$\langle k \rangle$	ℓ	ℓ_{rand}	C	C_{rand}	Reference	Nr.
WWW, site level, undir.	153 127	35.21	3.1	3.35	0.1078	0.00023	Adamic, 1999	1
Internet, domain level	3015–6209	3.52–4.11	3.7–3.76	6.36–6.18	0.18–0.3	0.001	Yook *et al.*, 2001a, Pastor-Satorras *et al.*, 2001	2
Movie actors	225 226	61	3.65	2.99	0.79	0.00027	Watts and Strogatz, 1998	3
LANL co-authorship	52 909	9.7	5.9	4.79	0.43	1.8×10^{-4}	Newman, 2001a, 2001b, 2001c	4
MEDLINE co-authorship	1 520 251	18.1	4.6	4.91	0.066	1.1×10^{-5}	Newman, 2001a, 2001b, 2001c	5
SPIRES co-authorship	56 627	173	4.0	2.12	0.726	0.003	Newman, 2001a, 2001b, 2001c	6
NCSTRL co-authorship	11 994	3.59	9.7	7.34	0.496	3×10^{-4}	Newman, 2001a, 2001b, 2001c	7
Math. co-authorship	70 975	3.9	9.5	8.2	0.59	5.4×10^{-5}	Barabási *et al.*, 2001	8
Neurosci. co-authorship	209 293	11.5	6	5.01	0.76	5.5×10^{-5}	Barabási *et al.*, 2001	9
E. coli, substrate graph	282	7.35	2.9	3.04	0.32	0.026	Wagner and Fell, 2000	10
E. coli, reaction graph	315	28.3	2.62	1.98	0.59	0.09	Wagner and Fell, 2000	11
Ythan estuary food web	134	8.7	2.43	2.26	0.22	0.06	Montoya and Solé, 2000	12
Silwood Park food web	154	4.75	3.40	3.23	0.15	0.03	Montoya and Solé, 2000	13
Words, co-occurrence	460.902	70.13	2.67	3.03	0.437	0.0001	Ferrer i Cancho and Solé, 2001	14
Words, synonyms	22 311	13.48	4.5	3.84	0.7	0.0006	Yook *et al.*, 2001b	15
Power grid	4941	2.67	18.7	12.4	0.08	0.005	Watts and Strogatz, 1998	16
C. Elegans	282	14	2.65	2.25	0.28	0.05	Watts and Strogatz, 1998	17

Observamos que

a) los caminos minimos medios de los networks reales son del orden de los random

b) la clusterizacion de los reales es mucho mayor que las de los random.

Networks que les interesan

- Facebook's data team released two papers in Nov. 2011
  - 721 million users with 69 billion friendship links
  - Average distance of 4.74
- Twitter studies
  - Sysomos reports the average distance is 4.67 (2010)
    - 50% of people are 4 steps apart, nearly everyone is 5 steps or less
  - Bakhshandeh et al. (2011) report an average distance of 3.435 among 1,500 random Twitter users

e) Estructura de Comunidades

definicion intuitiva de comunidad → un subconjunto de nodos (subgrafo) tal que los nodos en el subgrafo estan mas unidos entre si que con el resto

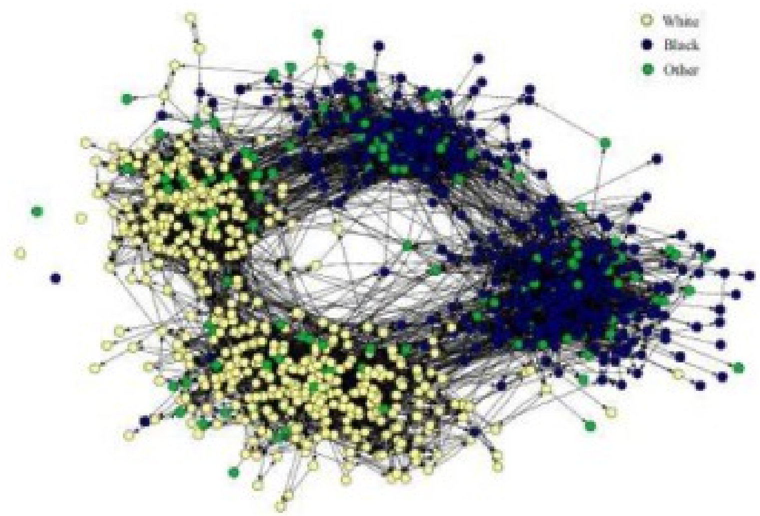


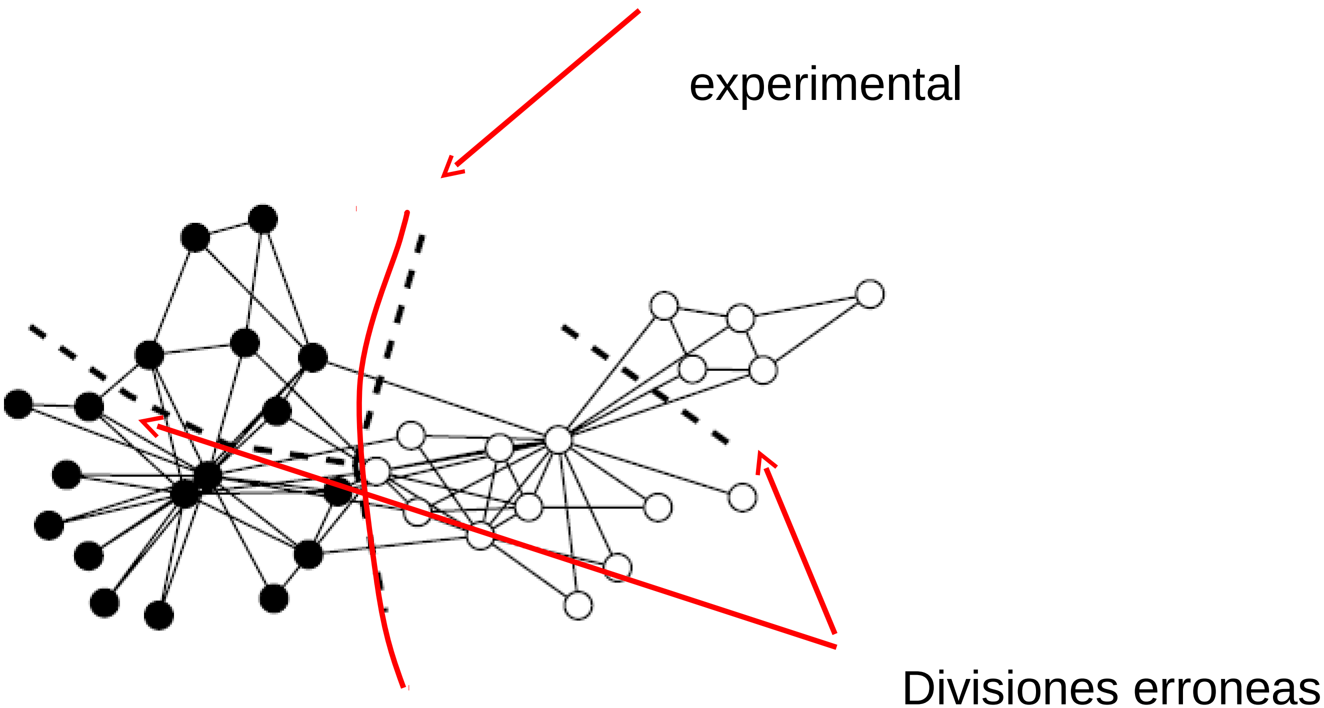
FIG. 8 Friendship network of children in a US school. Friendships are determined by asking the participants, and hence are directed, since A may say that B is their friend but not vice versa. Vertices are color coded according to race, as marked, and the split from left to right in the figure is clearly primarily along lines of race. The split from top to bottom is between middle school and high school, i.e., between younger and older children. Picture courtesy of James Moody.

Comunidad

Definición de Comunidad (intuitiva)

Dado un conjunto de nodos y links que constituyen un grafo conexo

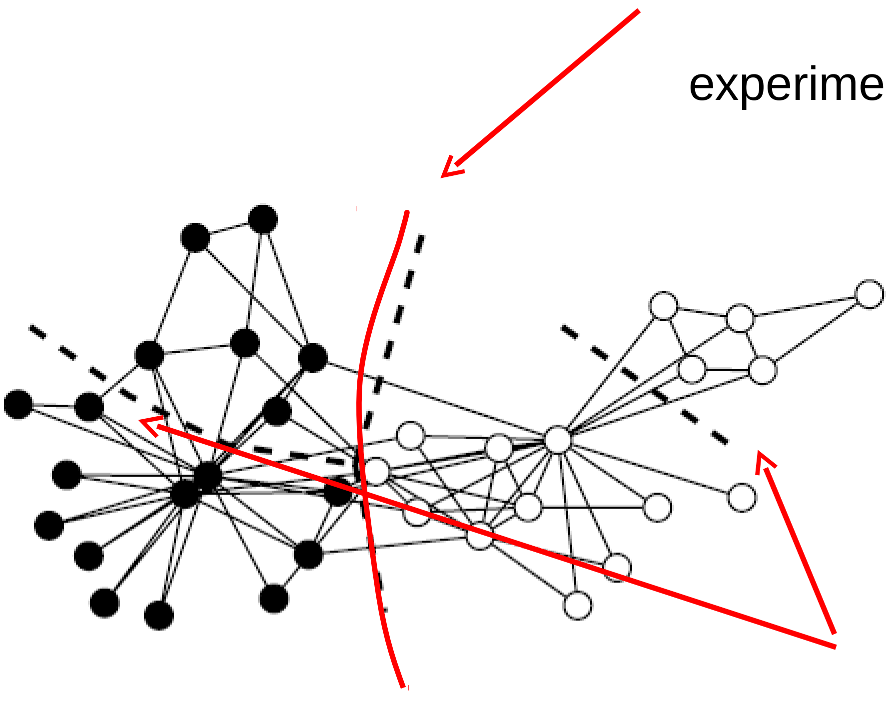
Descomponerlo en comunidades es encontrar aquellos subconjuntos de nodos mas unidos entre si que con respecto al resto del grafo)



Comunidad

Resolver un problema de comunas es :
dar definicion de comuna
dar algoritmo para resolverlo

experimental



Divisiones erroneas

Respecto de las comunidades

El analisis de Newman

Sea

$$A_{v\omega} = \begin{cases} 1 \text{ si los vertices } v \text{ y } \omega \text{ estan conectados} \\ 0 \text{ en otro caso} \end{cases}$$

La misma comuna

$$\sum_{\mu\nu} A_{\mu\nu} \overbrace{\delta(c_\mu, c_\nu)} \propto$$

Numero de links que pertenecen a la misma comunidad

Links, edges

La fraccion de lados que pertenecen a las mismas comunidades es

$$\frac{\Sigma_{v\omega} A_{v\omega}\delta(c_v, c_\omega)}{\Sigma_{v\omega} A_{v\omega}} = \frac{1}{2m}\sum_{v\omega} A_{v\omega}\delta(c_v, c_\omega)$$

Por otro lado

$$k_v = \sum_{\omega} A_{v\omega}$$

Grado del nodo v

son los links que salen de el nodo v (tomar en cuenta que $A_{vv} = 0$)

Define ahora Q como

$$Q = \frac{1}{2m} \sum_{v\omega} \left[A_{v\omega} - \frac{k_v k_\omega}{2m} \right] \delta(c_v, c_\omega)$$

o tambien

$$Q = \sum_{v\omega} \left[\frac{A_{v\omega}}{2m} - \frac{k_v \cdot k_\omega}{2m \cdot 2m} \right] \delta(c_v, c_\omega)$$

donde el termino $\frac{k_v}{2m}$ es la fraccion de nodos que salen del nodo v o sea

$$\frac{k_v}{2m} = \frac{\sum_\omega A_{v\omega}}{\Sigma_{v\omega} A_{v\omega}}$$

Esto deberia ser la probabilidad de que un link al azar salga del nodo v entonces la proba de que un Link salga de v y llegue a ω es

$$P_{v\omega} = \frac{k_v}{m} \cdot \frac{k_\omega}{m}$$

Escribe ahora $\delta(c_v, c_\omega) = \sum_i \delta(c_v, i)\delta(c_\omega, i)$

$$Q = \frac{1}{2m} \sum_{v\omega} \left[A_{v\omega} - \frac{k_v k_\omega}{2m} \right] \sum_i \delta(c_v, i)\delta(c_\omega, i)$$

operando con esto es

$$Q = \sum_i \left\{ \frac{1}{2m} \sum_{v\omega} A_{v\omega}\delta(c_v, i)\delta(c_\omega, i) - \sum_{v\omega} \frac{k_v}{2m} \frac{k_\omega}{2m} \delta(c_v, i)\delta(c_\omega, i) \right\}$$

$$Q = \sum_i \left\{ \frac{1}{2m} \sum_{v\omega} A_{v\omega}\delta(c_v, i)\delta(c_\omega, i) - \sum_v \frac{k_v}{2m}\delta(c_v, i) \sum_\omega \frac{k_\omega}{2m}\delta(c_\omega, i) \right\}$$

Atención , estas dos expresiones son iguales

$$Q = \sum_i \left\{ \frac{1}{2m} \sum_{v\omega} A_{v\omega}\delta(c_v, i)\delta(c_\omega, i) - \sum_v \frac{k_v}{2m}\delta(c_v, i) \sum_\omega \frac{k_\omega}{2m}\delta(c_\omega, i) \right\}$$

$\sum_v \frac{k_v}{2m}\delta(c_v, i) = \frac{1}{2m}\sum_v k_v\delta(c_v, i)$, se suman los numeros de links que salen de cada nodo que pertenecen a la comunidad i

$$Q = \sum_i \left\{ \frac{1}{2m} \sum_{v\omega} A_{v\omega} \delta(c_v, i)\delta(c_\omega, i) - \sum_v \frac{k_v}{2m} \delta(c_v, i) \sum_\omega \frac{k_\omega}{2m} \delta(c_\omega, i) \right\}$$

$$Q = \sum_i \left\{ \frac{1}{2m} \sum_{v\omega} A_{v\omega} \delta(c_v, i)\delta(c_\omega, i) - \left[\sum_v \frac{k_v}{2m} \delta(c_v, i) \right]^2 \right\}$$

El termino $\sum_{v\omega} \frac{1}{2m} \frac{k_v k_\omega}{2m} \delta(c_v, c_\omega)$ representa la probabilidad de que si yo "coloco" un link al azar preservando el grado de cada nodo este caiga entre nodos de la comunidad. Es como elegir un link "saliente" con proba $\frac{k_\omega}{2m}$ y uno "entrante" con proba $\frac{k_\omega}{2m}$

Entonces tenemos una definición un factor de merito de una comuna lo que Sugiere que la mejor partición es la que maximice

Veremos ahora que pasa para un tipo particular de network

Grafos de mundo pequeño

Los grafos de Watts - Strogatz

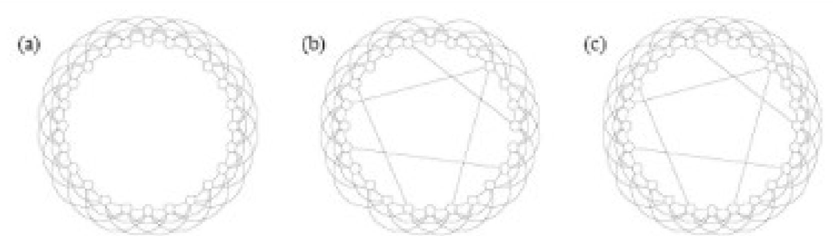


FIG. 11 (a) A one-dimensional lattice with connections between all vertex pairs separated by k or fewer lattice spacing, with $k = 3$ in this case. (b) The small-world model [412, 416] is created by choosing at random a fraction p of the edges in the graph and moving one end of each to a new location, also chosen uniformly at random. (c) A slight variation on the model [289, 324] in which shortcuts are added randomly between vertices, but no edges are removed from the underlying one-dimensional lattice.

Red de Wats Strogatz

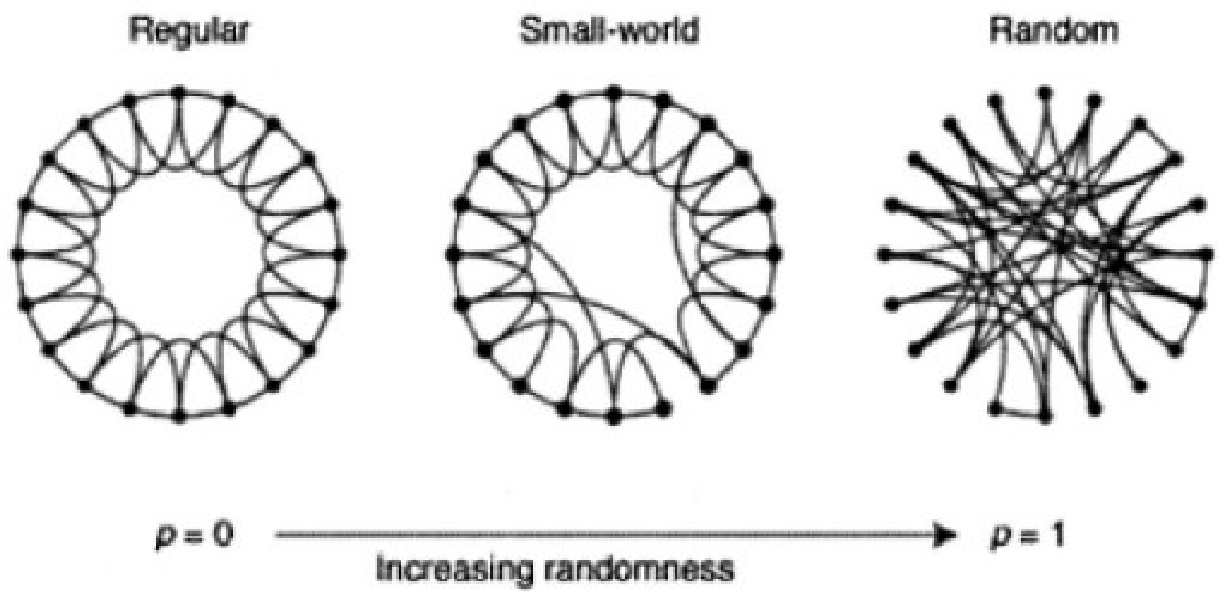


FIG. 15. The random rewiring procedure of the Watts-Strogatz model, which interpolates between a regular ring lattice and a random network without altering the number of nodes or edges. We start with $N=20$ nodes, each connected to its four nearest neighbors. For $p=0$ the original ring is unchanged; as p increases the network becomes increasingly disordered until for $p=1$ all edges are rewired randomly. After Watts and Strogatz, 1998.

Empezamos con uno unidemensional

1) empezamos con orden

Para el caso con orden tendremos

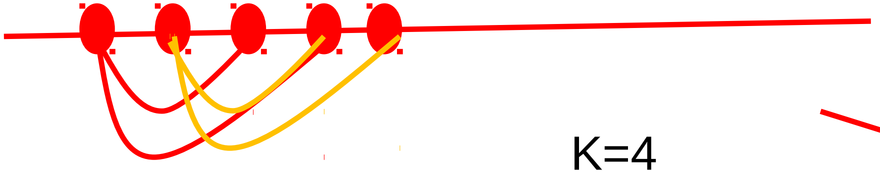
Cada vertice tiene $2k$ vecinos

El numero de links entre estos vecinos es $3k(k-1)/2$

El numero total de links es $2k(2k-1)/2$

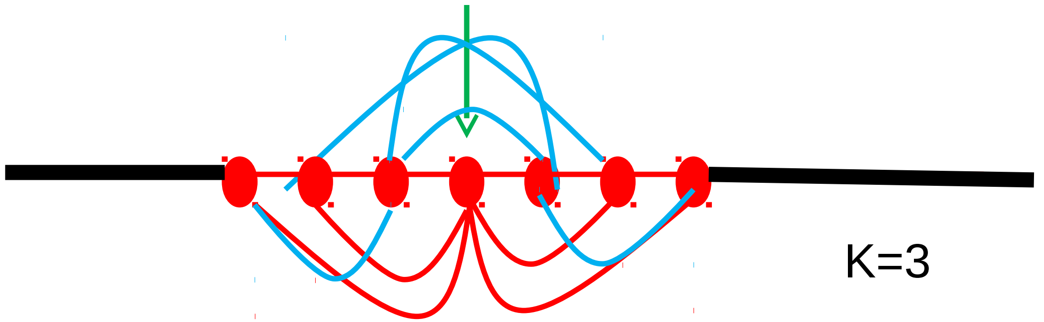
$$C = \frac{3(k-1)}{2(2k-1)}$$

Donde K es el numero de vecinos



K=4

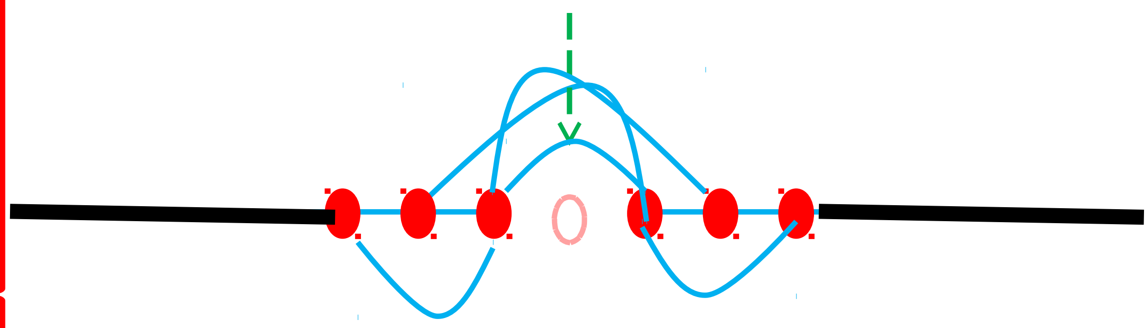
Nos fijamos
En un dado
nodo



K=3

Links totales = 15 = $\dfrac{2k(2k-1)}{2}$

Para el
subgrafo



Links subgrafo = 9 = $\dfrac{3k(k-1)}{2}$

Clustering coefficient

Entonces 2) randomizamos con rewiring

$$C = \frac{3(k-1)}{2(2k-1)}(1-p)^3 \quad = C(p)$$

Que es la proba de que los 3 links sobrevivan

Si k=2 ; C(0)=0.5

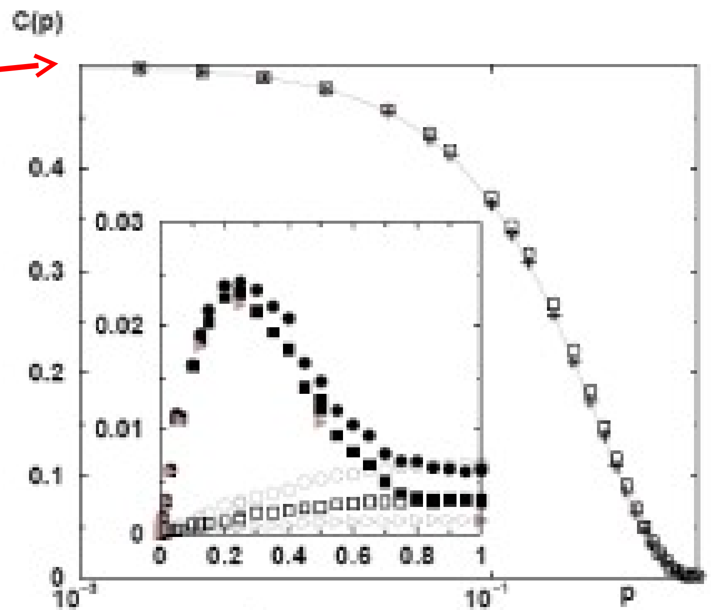


FIG. 9. $C(p)$ and $\tilde{C}(p)$ versus p, for $k = 2$ ($C(0) = \tilde{C}(0) = 0.5$), $N = 1000$, 2000, 5000: open symbols are for $C(p)$, and the crosses are for $\tilde{C}(p)$; the line is $C(0)(1-p)^3$. Inset: corrections $C(p) - C(0)(1-p)^3$ (filled symbols) for $N = 1000$ (circles), $N = 2000$ (squares) and $N = 5000$ (triangles), and $\tilde{C}(p) - C(0)(1-p)^3$ (open symbols) for $N = 1000$ (circles), $N = 2000$ (squares) and $N = 5000$ (triangles). We see that the corrections go to zero as $1/N$ for $\tilde{C}(p)$; the corrections for $C(p)$ are larger, but anyway very small.

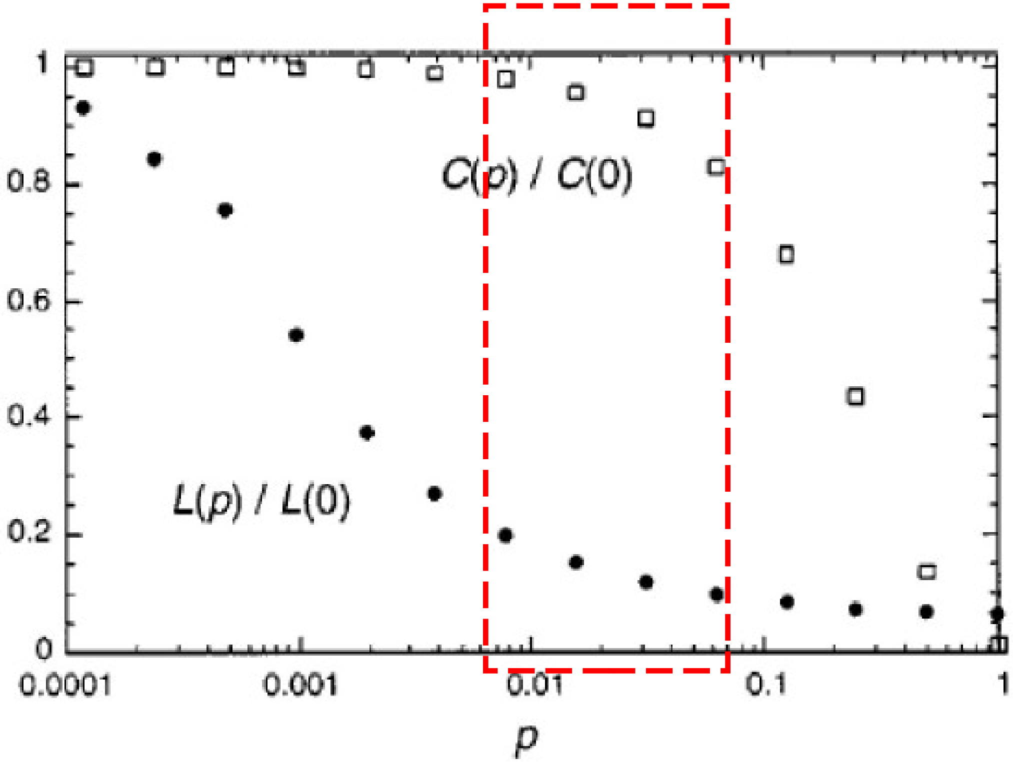


FIG. 16. Characteristic path length $\ell(p)$ and clustering coefficient $C(p)$ for the Watts-Strogatz model. The data are normalized by the values $\ell(0)$ and $C(0)$ for a regular lattice. A logarithmic horizontal scale resolves the rapid drop in $\ell(p)$, corresponding to the onset of the small-world phenomenon. During this drop $C(p)$ remains almost constant, indicating that the transition to a small world is almost undetectable at the local level. After Watts and Strogatz, 1998.

Ha sido conjeturado que el camino minimo medio satisface la siguiente relacion de escala (recuerdos?):

[42] Barthélémy, M. and Amaral, L. A. N., Small-world networks: Evidence for a crossover picture, *Phys. Rev. Lett.* **82**, 3180–3183 (1999).

$$l = \xi g\left(\frac{l}{\xi}\right)$$

con

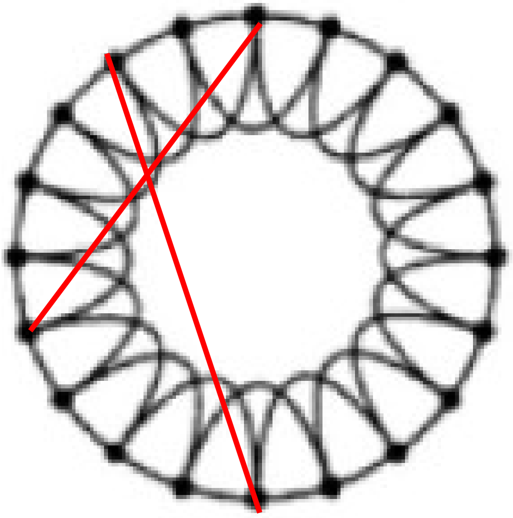
$$g(x) \sim \begin{cases} x & para \quad x \gg 1 \\ \log(x) & para \quad x \ll 1 \end{cases}$$

y ademas ξ diverge cuando $p \to 0$

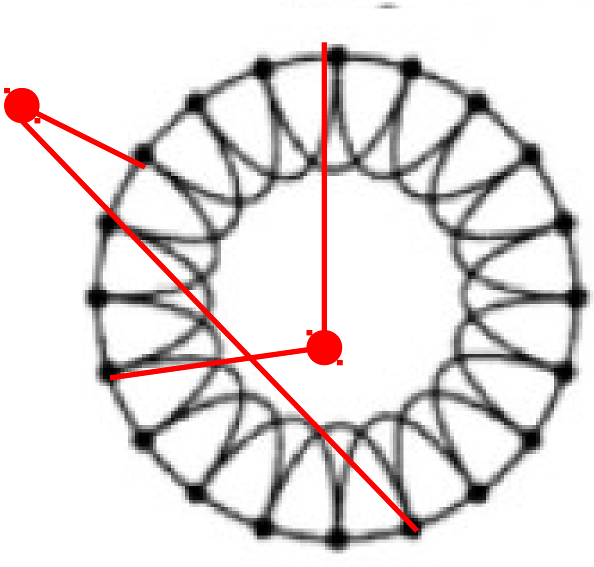
Esto se comprobo "experimentalmente"

Tambien conjeturaron que $\xi \sim p^{-\tau}$ con $\tau = 2/3$ pero se encontro que $\tau = 1$

Variaciones



Links agregados



Links+nodos agregados

$$Q = \sum_i \left\{ \frac{1}{2m} \sum_{v\omega} A_{v\omega} \delta(c_v, i) \delta(c_\omega, i) - \left[\sum_v \frac{k_v}{2m} \delta(c_v, i) \right]^2 \right\}$$

Estudiemos el Q de newman para el siguiente caso

Sea un WS con

N nodos

kN links $\Rightarrow$ salen $2kN$ links

luego el grado de cada nodo es $2k$

Calculamos Q para el grafo completo

l_s es el numero de links internos

L es el numero total de links

El primer termino de Q resulta ser $= \frac{l_s}{L} = \frac{kN}{KN} = 1$

$$Q = \sum_i \left\{ \frac{1}{2m} \sum_{v\omega} A_{v\omega}\delta(c_v, i)\delta(c_\omega, i) - \left[\sum_v \frac{k_v}{2m}\delta(c_v, i) \right]^2 \right\}$$

El segundo termino de Q es el grado total dividido por dos veces el numero de links = $\frac{2kN}{2kN} = 1$

$$Q = 1 - 1 = 0$$

Ahora lo partimos en 2 (supongo N par), para cada subgrafo

$$Q_i = \left\{ \frac{\frac{kN}{2} - 2k}{kN} \right\} - \left\{ \frac{2kN/2}{2kN} \right\}^2$$

multiplicando por 2

$$Q = 1 - \frac{1}{2} - \frac{4}{N}$$

Con N muy grande

$$Q \simeq 0.5$$

Luego Q crece mal

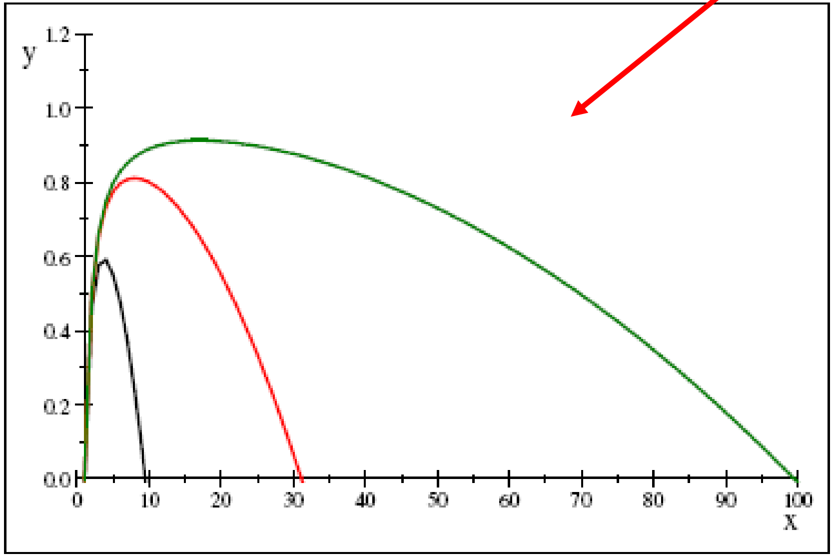
Rompe un grafo simétrico

Entonces si lo parto en n subgrafos

$$Q = 1 - \frac{1}{n} - \frac{n^2}{N}$$

!!!

$$1 - \frac{1}{n} - \frac{n^2}{100}$$



En negro $N = 100$, en rojo $N = 1000$, en verde $N = 10000$

Le gustan los grafos rotos!!!!!!!!!!!!!!!!!

Lo cual invalida los metodos que maximizan Q

Sin embargo al cortar los links los estoy aun contando, asi que pensamos que esos links no existen mas

$$Q_i = \left\{ \frac{\frac{kN}{2} - 2k}{kN} \right\} - \left\{ \frac{2kN/2 - 2k}{2kN} \right\}^2$$

$$= \frac{1}{2} - \frac{2}{N} - \left[\frac{1}{2} - \frac{1}{N} \right]^2 = \frac{1}{2} - \frac{2}{N} - \frac{1}{4} + \frac{1}{N} - \frac{1}{N^2} = \frac{1}{4} - \frac{1}{N} - \frac{1}{N^2}$$

Entonces Q da

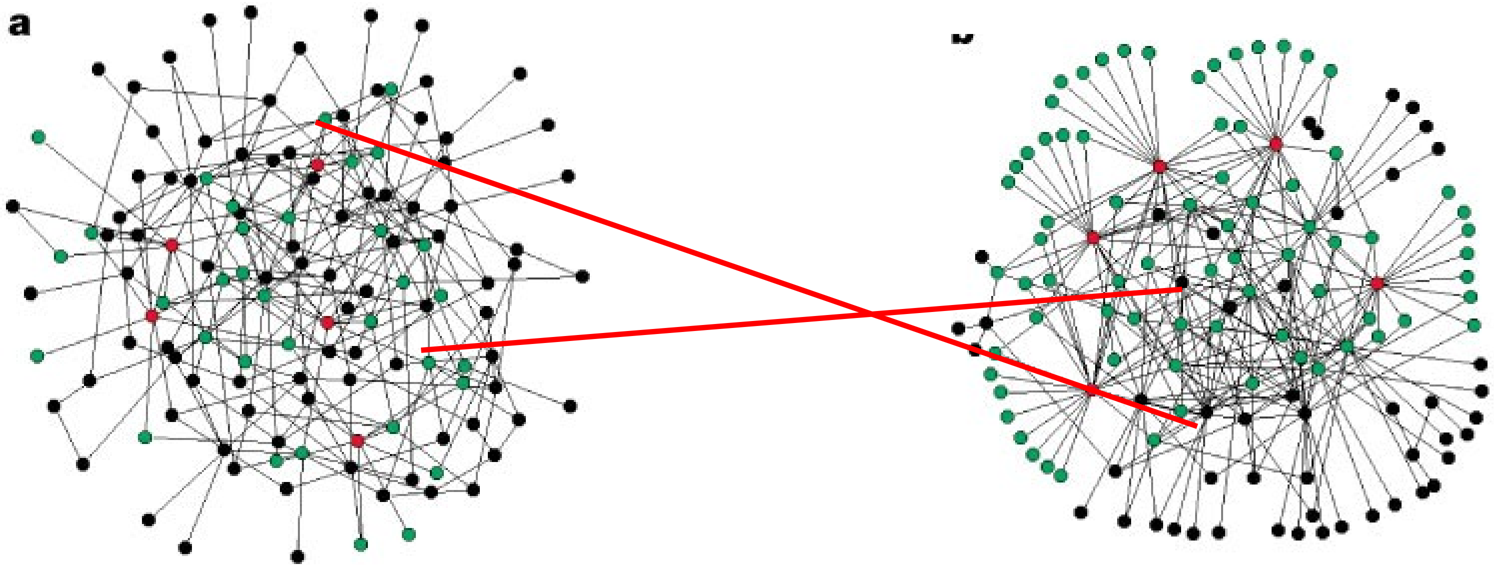
$$Q = \frac{1}{2} - \frac{2}{N} - \frac{2}{N^2}$$

Pero no se arregla

Tenemos un factor de merito que intenta cuantificar el valor de una dada Particion

Para un grafo razonable, explorar todas las posibles particiones es Impracticable (NP complete)

Como generamos la configuraciones?



Recordemos a la Betweenness

$$b_{ij} = \sum_{paths} \alpha_{no}^{-1} \sum_{l_{km} \varepsilon path_{no}} \delta(l_{ij} - l_{km})$$

Physica A 358 (2005) 593–604

Detection of community structures in networks via global optimization ☆

...s, G. Acuña, C.O. Dorso*

Se propone entonces

In a recent work, Newman and Girvan [3] have proposed to study the structure of the network by analyzing the effect of the removal of links with highest betweenness. The betweenness b_{ij} of a given link l_{ij} is

$$b_{ij} = \sum_{paths} \alpha_{no}^{-1} \sum_{l_{km} \varepsilon path_{no}} \delta(l_{ij} - l_{km}) \qquad (2)$$

with $\sum_{paths}$ the sum over all the path joining the n_n nodes. α_{no} is the degeneracy of the path between nodes n and o, and $\sum_{l_{km} \varepsilon path_{no}}$ is the sum over all the links l_{km} that form the path under consideration. In this way the link with highest betweenness is the one that appears most often when we study all the components of all the minimum paths between all the pairs of nodes.

According to this prescription:

(i) One calculates the betweenness of all the links in the network. (ii) The one with the highest betweenness is removed.

The process is continued until a disjoint cluster is obtained. Afterwards, it is applied to each of the resulting subgraphs.

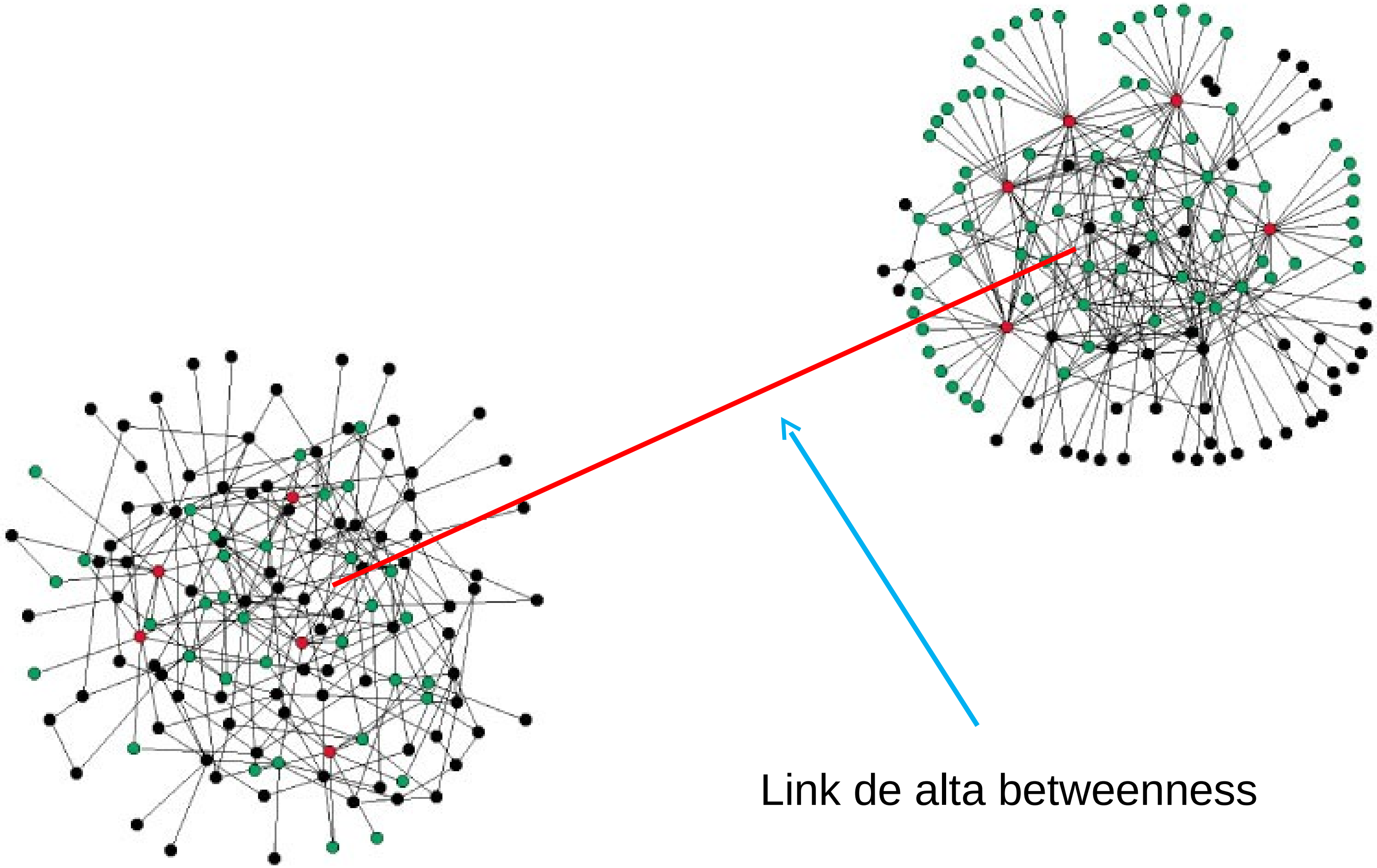
3.2. Simulated annealing analysis

In this section we present a methodology to study the community structure in networks based on the search for that partition that maximizes the value of Q. This is accomplished by resorting to a SA [10] calculation in the space of the partitions of the network under analysis. SA is a generalization of the well known Metropolis Monte Carlo (MMC) procedure. MMC consists in the realization of a Markov Chain in the space of the configurations of the system according to certain transition probabilities chosen in such a way that the asymptotic frequency of each state satisfies the Boltzmann distribution $\exp(-\beta E_i)/Z$ with $\beta = (1/kT)$ where T is the temperature of the system, E_i the energy of state i and Z the canonical partition function. The transition probability q_{ij} reads

$$q_{ij} = \min(1, \exp(-\beta(E_j - E_i)) .$$

In SA (see [10] for details) the same procedure is employed but instead of using the temperature of the system we use a pseudo temperature, τ, which controls the behavior of the transition probability and instead of the energy the observable that we want to maximize. The pseudo temperature τ is monotonously lowered until an extremum of the relevant observable is attained. In our case the Markov Chain is performed in the space of the partitions of the network under consideration. The transition probabilities read $q_{ij} = \min(1, \exp(-\beta'(Q_j - Q_i))$ with $\beta' = 1/\tau$ and E_k has been replaced by Q_k, the modularity of partition k. Moreover, because we are looking for the maximum of the modularity $(Q_j - Q_i)$ stands for $(Q_{initial} - Q_{final})$.

Cual es la idea de esto?

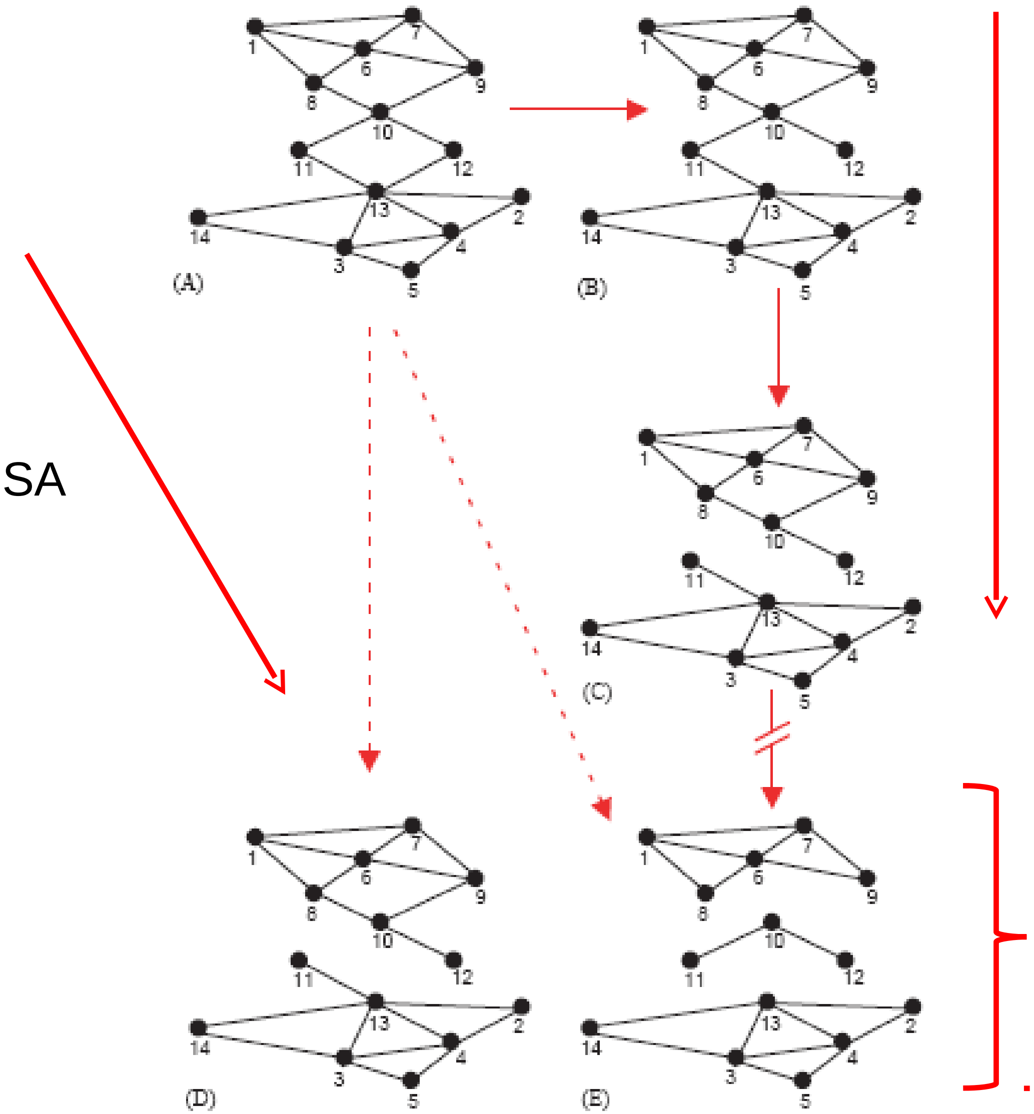


Link de alta betweenness

Gran candidato a cortar!

Camino del betweenness

SA



No son accesibles por este camino

El camino de GN es Irreversible!

Fig. 1. Development of community structures in terms of the ER and SA analysis. Full arrows denote steps in the ER approach. Dotted arrows denote results from SA methodology. Starting from network A by applying ER methodology we first get to network B and, after the second removal of a link, to network C. On the other hand, starting from the same initial network the SA will give network D if we impose the constraint that the final configuration should display two communes. If we do not impose any constraint the result according to SA will be network E. It is important to notice that network E is unreachable from network C. This is the main drawback of the ER approach.

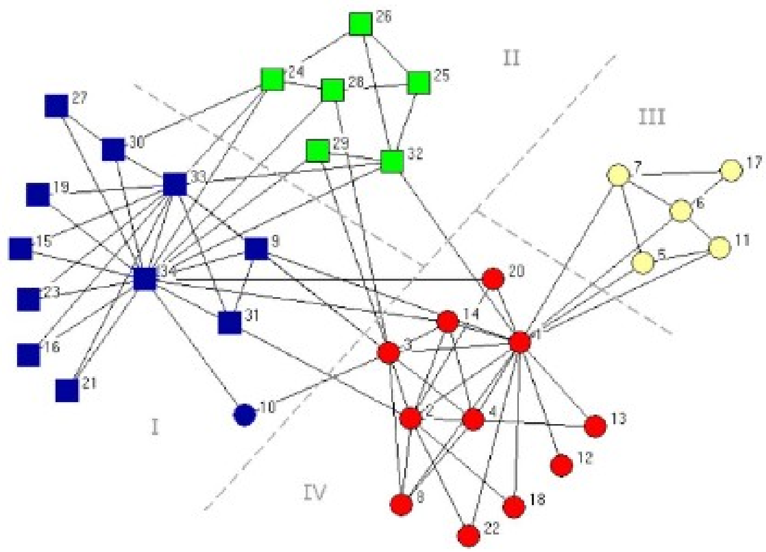


Fig. 2. Community structures for the Zachary network according to SA approach. In this figure, squares and circles denote the members of the two subsets according to observations by Zachary. Broken lines denote the partitions obtained according to SA approach.

SA
Q=0.42

ER
Q=0.36

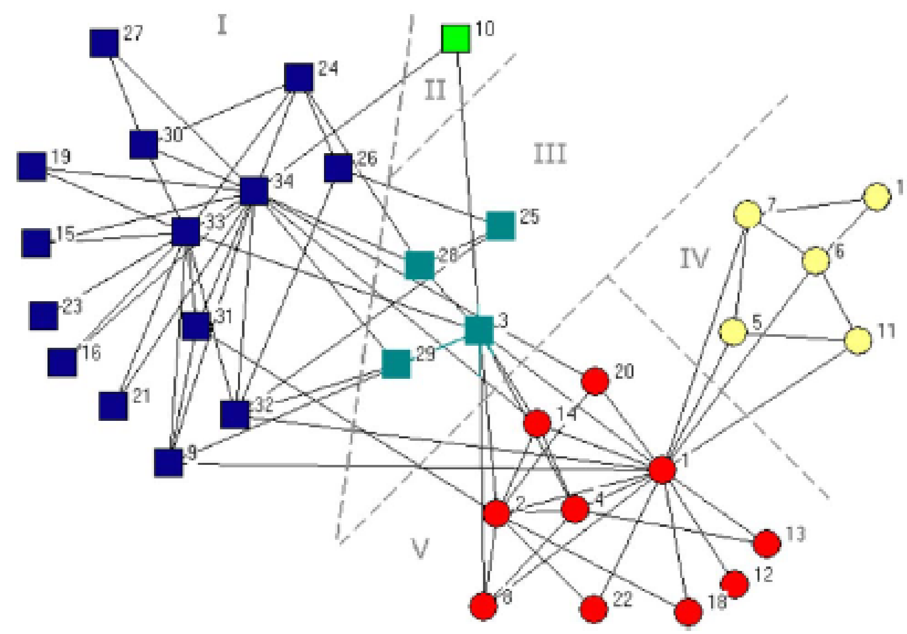


Fig. 3. Community structures for the Zachary network according to ER approach.

Si se toma en cuenta que los links son pesados

602 A. Medus et al. / Physica A 358 (2005) 593–604

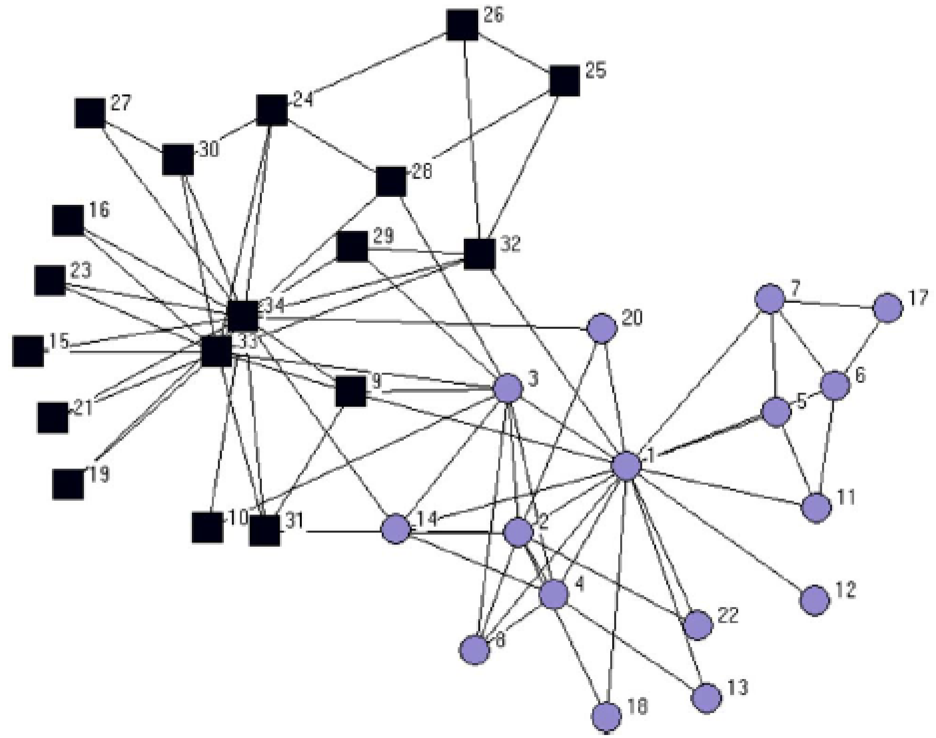


Fig. 4. Actual community structures as recorded by Zachary. Once again squares and circles denote the members of each subset.

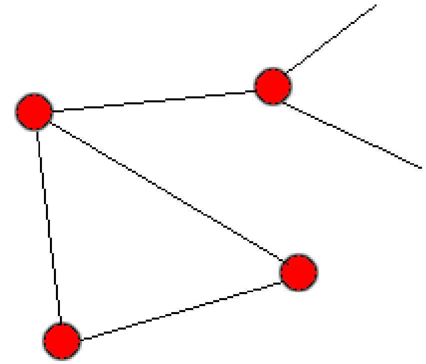
Si se aplica SA con la condición de solo 2 comunas finales, se parece mucho

Definiciones locales de comuna

(F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, y D. Parissi, Proc. Nat. Acad. Sci. 101, 2658-2663 (2004).)

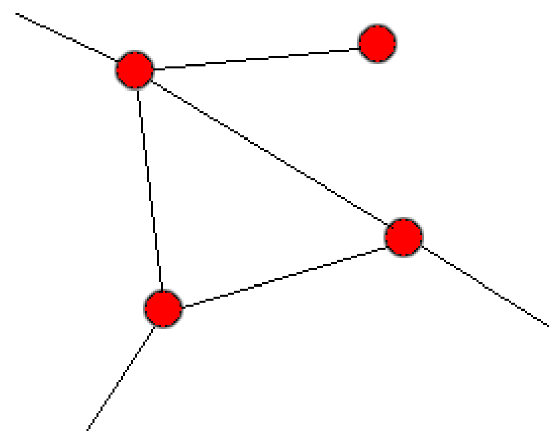
Radicchi débil:

$$\sum_{i \in C} k_i^{in} > \sum_{i \in C} k_i^{out}$$



Radicchi fuerte:

$$k_i^{in} > k_i^{out} \ \forall i \in C$$



Fuerza de una comuna

Buscamos establecer una magnitud que permita comparar el grado de comunalidad para distintos subgrafos.

Dado un subgrafo Cj definimos la denominada “fuerza de la comuna (S)” como:

$$S(C_j) = \sum_{i \in C_j} \frac{k_i^{in} - k_i^{out}}{2L(C_j)}$$

$$\text{siendo } L(C_j) = \frac{1}{2} \sum_{i \in C_j} k_i \text{ y además } -1 \leq S(C_j) \leq 1$$

Esta definición se corresponde con la idea cualitativa que tenemos de comuna y es estrictamente **local**

Nuevo método para la detección de comunas

Factor de mérito para la definición débil de comuna:

$$Q_W = \sum_{j=1}^{M} S(C_j) = \sum_{j=1}^{M} \sum_{i \in C_j} \frac{k_i^{in} - k_i^{out}}{2L(C_j)}$$

con la restricción: $S(C_j) > 0 \forall C_j \subset \{C_j\}_{j=1,\ldots,M}$ para que cada subconjunto de nodos C_j satisfaga la definición débil de comuna.

Factor de mérito para la definición fuerte de comuna:

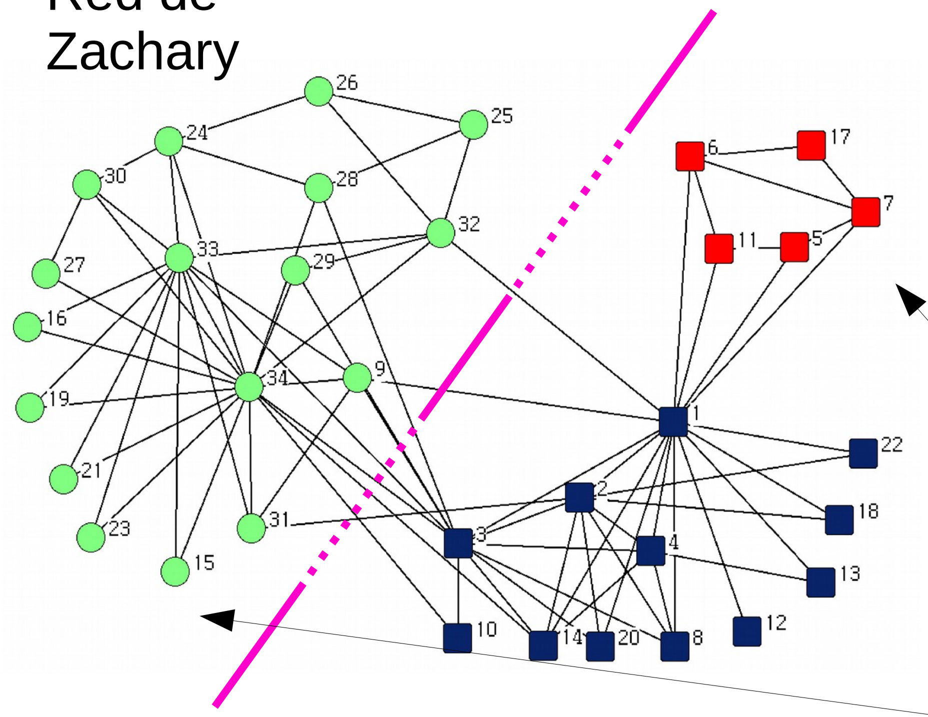
$$Q_S = \sum_{j=1}^{M} S(C_j) = \sum_{j=1}^{M} \sum_{i \in C_j} \frac{k_i^{in} - k_i^{out}}{2L(C_j)}$$

$$k_i^{in} > k_i^{out} \ \forall i \in C_j \text{ para } j = 1, \ldots, M$$

Qs y Qw están idénticamente definidos, solo cambia la condición sobre las comunas. La ausencia de partición equivale a Qs=Qw=1.

Ejemplos

Red de Zachary



Definición fuerte: 2 comunas de 29 (S=0.943) y 5 (S=0.5) nodos.

$$S(C_1 \rightarrow 17) = 0.744$$
$$S(C_2 \rightarrow 12) = 0.548$$
$$S(C_3 \rightarrow 5) = 0.5$$

$$Q_W = 1.792$$

Red de delfines “nariz de botella”

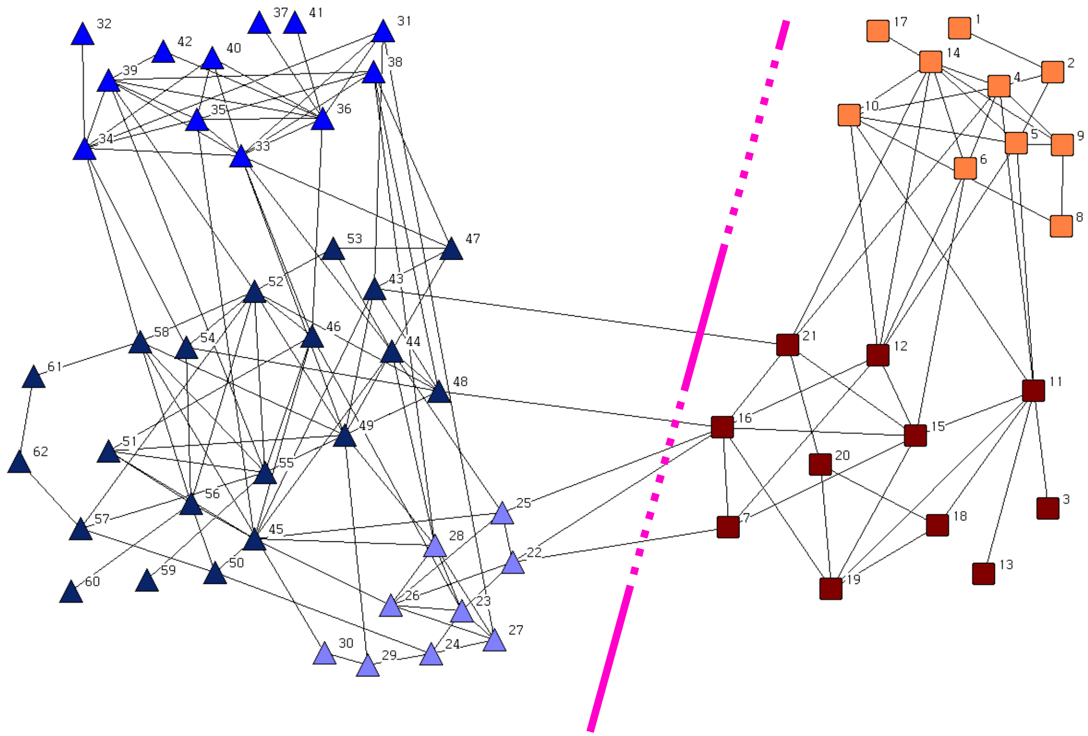
COMUNA= 1
Poblacion total = 11.
Fuerza de la comuna=
0.632653061
Total de nodos "strong connected"=
11

COMUNA= 2
Poblacion total = 9.
Fuerza de la comuna=
0.302325581
Total de nodos "strong connected"=
6

COMUNA= 3
Poblacion total = 12.
Fuerza de la comuna=
0.523809524
Total de nodos "strong connected"=
11

COMUNA= 4
Poblacion total = 10.
Fuerza de la comuna=
0.391304348
Total de nodos "strong connected"=
7

COMUNA= 5



$$Q_W(5 \text{ comunas}) = 2.407$$

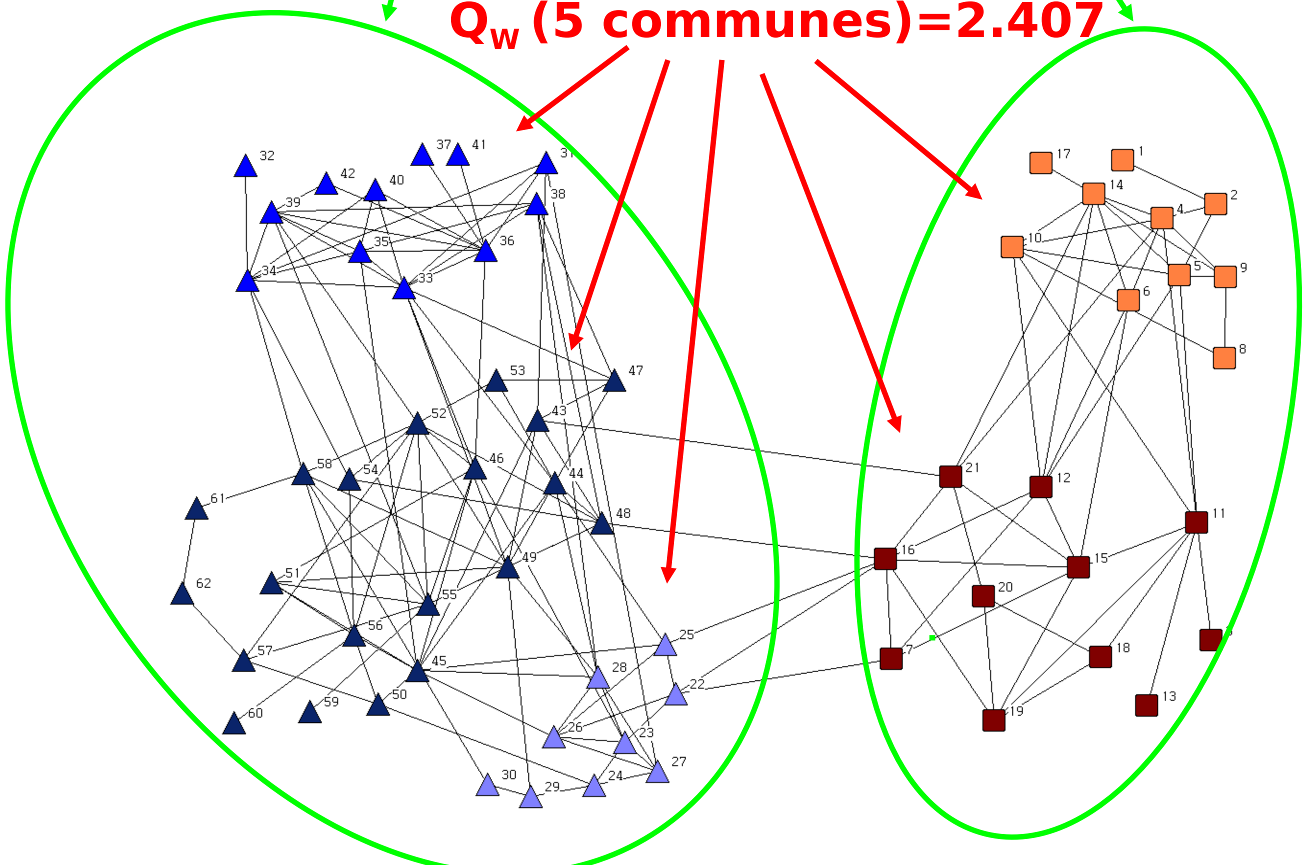
$$Q_S(2 \text{ comunas}) = 1.849$$

La partición en 2 comunas es la óptima para la definición fuerte de comuna (Qs) y se corresponde con la observada por Lusseau.

Bottle Nose Dolphins Network

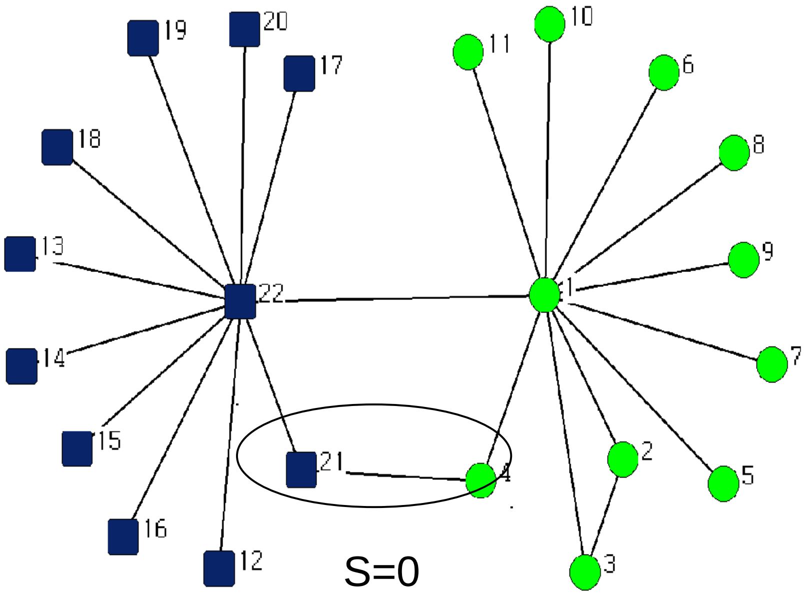
Q_S (2 communes) = 1.849

Q_W (5 communes)=2.407



The Q_S solution corresponds exactly to the observed community structure by Lusseau.

Red estrella



Las comunas obtenidas mediante la modularidad no siempre satisfacen la definición cualitativa de comuna

$$S(C_2) = 0.818 \qquad S(C_1) = 0.833 \longrightarrow Q_W = 1.652$$

Mediante la optimización de Qs no se obtiene ninguna partición de la red.

Red anillo

Definición débil:

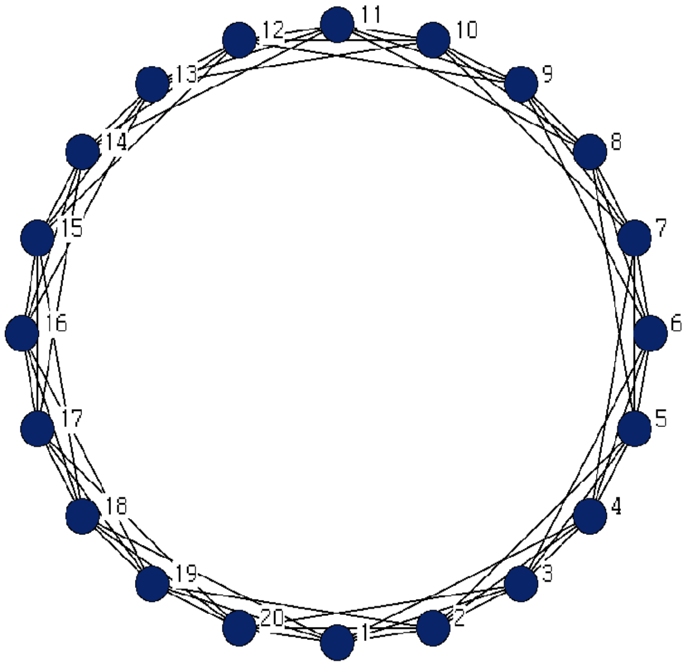
Dos comunas de 10 nodos cada una.

$$Q_W(2 \text{ comunas}) = 1.2$$

Modularidad Q de Newman:

Tres comunas: dos de 7 y una de 6 nodos.

$$Q = 0.365$$



20 nodos, k=6

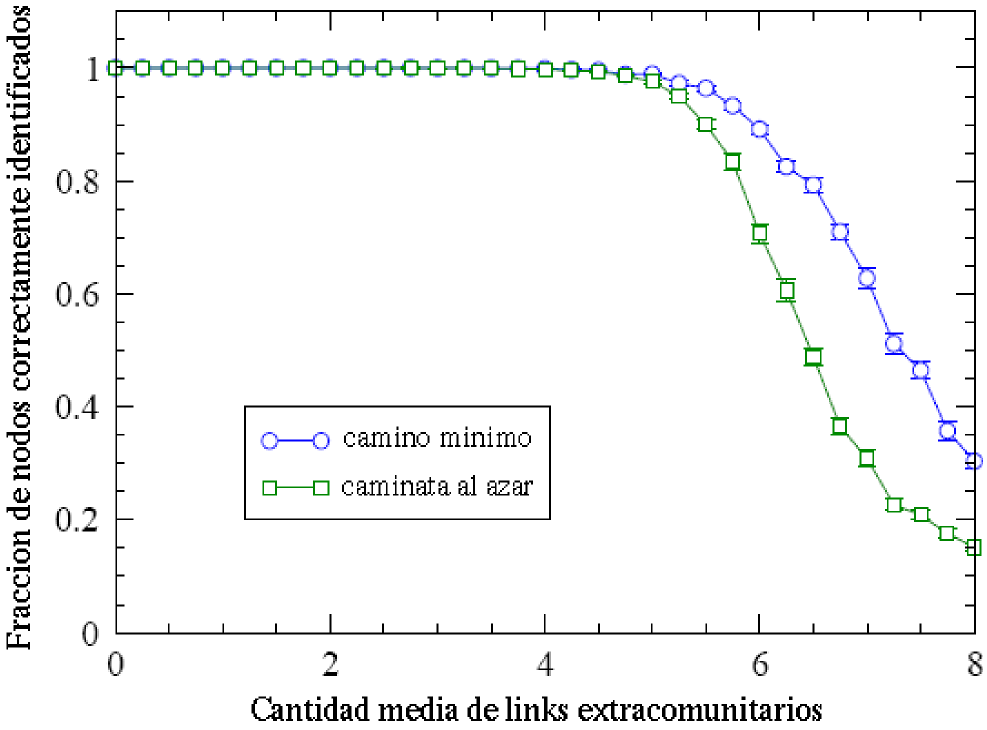
Definición fuerte:

No hay partición.

Test computacional de Newman

128 nodos con grado medio 16, divididos en 4 comunas de 32 nodos cada una. Cada nodo posee <Kin> links hacia otros nodos de su propia comuna el resto (<Kout>) hacia nodos que no pertenecen a su comuna.

Mediante la optimización de Q se detectan 4 comunas aun para Kout>7. Esto no ocurre cuando se utiliza Qw o Qs.



Esta clase de test solo sirven para evaluar algoritmos para la detección de comunas.

Comunidades alternativas

Definimos comunidad

(i) Community in strong sense. C is a community in the strong sense if

$$k_i^{in} > k_i^{out} \quad \forall\, i \in C. \qquad (1)$$

(ii) Community in weak sense. C is a community in weak sense if

$$\sum_{i \in C} k_i^{in} > \sum_{i \in C} k_i^{out}. \qquad (2)$$

S. Fortunato and M. Barthélemy, Proc. Natl. Acad. Sci. U.S.A. **104**, 36 (2007).

In words: a subgraph $C \subset G$ will be a community in the strong sense if each of its nodes has more links connecting it with nodes in C than those that connect it with other nodes not belonging to C. In a similar way, $C \subset G$ will be a community in the weak sense if the sum of the number of links that interconnect nodes inside C is larger than the sum of all links that connect nodes in C with nodes not belonging to C. These community definitions are simple, intuitive, and *local*: given a subgraph $C \subset G$ we can decide if it constitutes a community, in either strong or weak sense, without knowledge of the entire structure of G.

PHYSICAL REVIEW E **79**, 066111 (2009)

Alternative approach to community detection in networks

A. D. Medus* and C. O. Dorso†

Departamento de Física, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Pabellón 1, Ciudad Universitaria,

Following the weak and strong definitions of community, the more internal links a community has, with respect to the external ones, the "stronger" it will be. If $k_i = k_i^{in} + k_i^{out}$ is the degree of node $i \in C_j$, where k_i^{in} and k_i^{out} are the number of internal and external links for node i, we define the "community strength" (S) that measures the normalized difference between internal and external links for nodes in C_j:

$$S(C_j) = \sum_{i \in C_j} \frac{k_i^{in} - k_i^{out}}{2L(C_j)},$$

where $L(C_j) = \frac{1}{2}\sum_{i \in C_j} k_i$. Then, $-1 \leq S(C_j) \leq 1$, and it achieves its maximum value 1 when $k_i^{out} = 0 \ \forall i \in C_j$.

Now we introduce a merit factor Q_W for the weak community definition as the sum of $S(C_j)$ over all subgraphs $C_j \subset G$:

$$Q_W = \sum_{j=1}^{m} S(C_j) = \sum_{j=1}^{m} \sum_{i \in C_j} \frac{k_i^{in} - k_i^{out}}{2L(C_j)} \qquad (5)$$

with the constraint that each subgraph $C_j \subset \{C_j\}_{1 \le j \le m}$ must satisfy the weak community definition, i.e.,

$$S(C_j) > 0 \quad \forall\, C_j \subset \{C_j\}_{1 \le j \le m}. \qquad (6)$$

As in the case of Q_N: the bigger Q_W is, the better the m-subgraphs partition $\{C_j\}_{1 \le j \le m}$ of G will be, in the sense of weak community definition. Then, it is possible to implement the optimization algorithms developed for Q_N for this merit factor Q_W.

In the same spirit we now define a merit factor Q_S according to the strong community definition:

$$Q_S = \sum_{j=1}^{m} S(C_j) = \sum_{j=1}^{m} \sum_{i \in C_j} \frac{k_i^{in} - k_i^{out}}{2L(C_j)} \tag{7}$$

with the constraint

$$(k_i^{in} - k_i^{out}) > 0 \quad \forall\, i \in C_j. \tag{8}$$

Now, our definition of optimal partition can be stated in the following way:

Definition. The optimal m-subgraphs partition $\{C_j\}_{1 \le j \le m}$ of a graph G in the strong (weak) sense is that one with maximal merit factor $Q_S(Q_W)$.

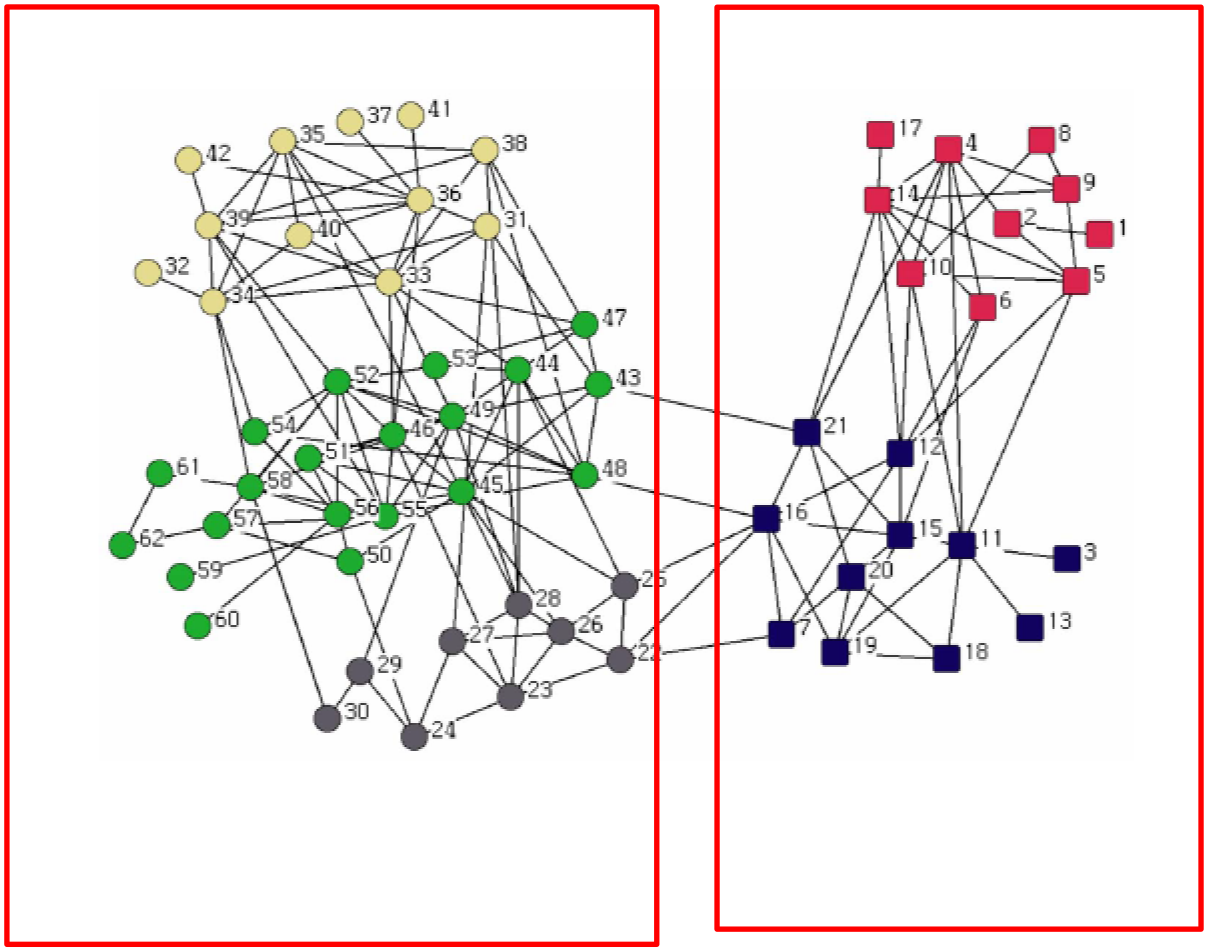


FIG. 4. (Color online) Bottlenose dolphin network. This network has a size of 62 nodes and it is known from direct observation that it has two communities. In this figure squares and circles denote the communities detected by our strong community approach and the colors (shades of gray or colors online) show the results of the weak community approach. Notice that the optimization according to Q_W merely subdivides the communities obtained through Q_S optimization [17].

Random Graphs

Erdös & Renyi

Tomemos un numero de n vertices y conectemos cada par con una probabilidad p , esto define el ensemble $G_{n,p}$ en el cual el grafo con m lados aparece con una probabilidad

$$G_{n,p} \rightarrow p^m (i - p)^{M-n}$$

donde $M = \frac{1}{2} n(n-1)$

1

Tambien definieron el modelo $G_{n,m}$ que es el enesemble con n vertices y exactamente m lados todos ellos con la misma

probabilidad (o sea que este es del tipo microcanonico y el anterior del tipo canonico)

Supongamos que calculamos las cosas en el limite de n muy grande, manteniendo el grado medio constante $z = p(n-1)$, luego estamos en el caso de Poisson

$$p_k = \binom{n}{k} p^k (1-p)^{n-k} \simeq \frac{z^k e^{-z}}{k!}$$

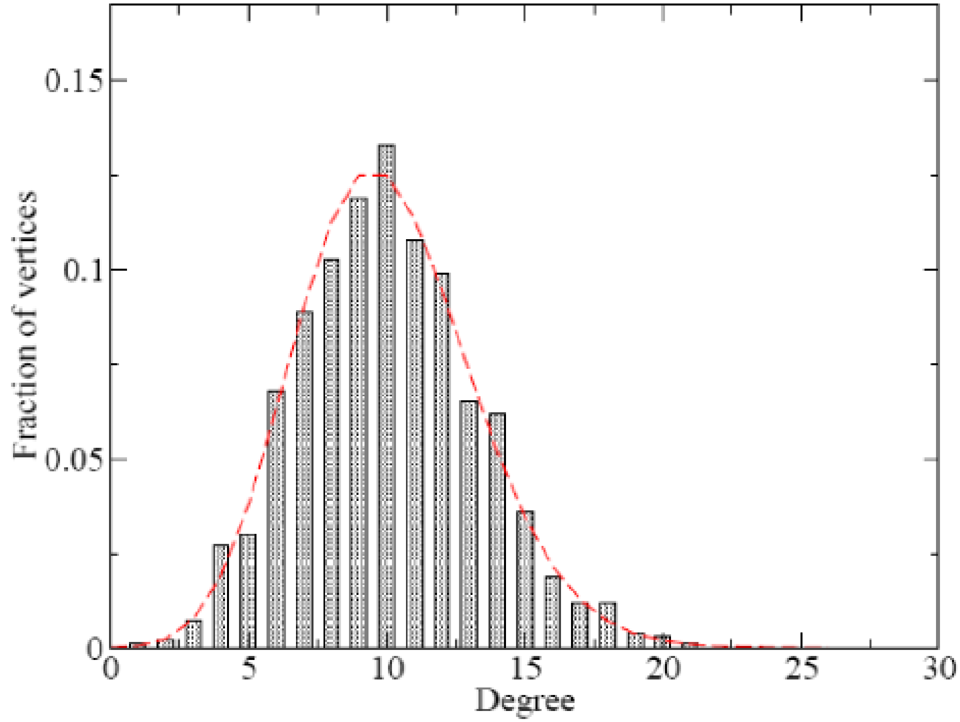


Figure 1.9. The degree distribution for an ER random network where $n = 1000$ and $p = 0.1$, with the real distribution plotted as a bar graph and the Poisson approximation plotted as the dashed line.

En este caso a medida que vamos incrementando p a partir de un valor bajo empiezan a aparecer "estructuras" conexas.

El resultado fundamental es que (no resulta insperado)

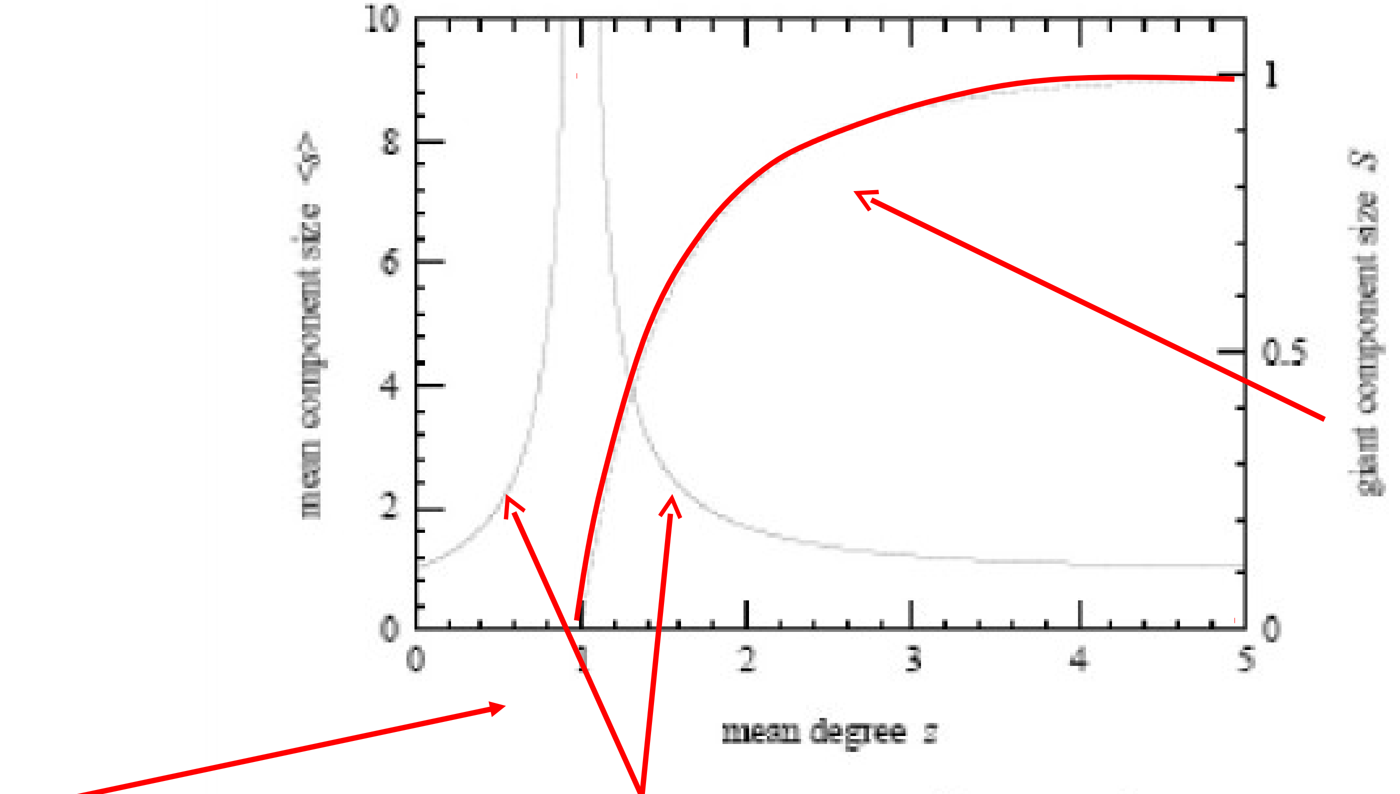


FIG. 10 The mean component size (solid line), excluding the giant component if there is one, and the giant component size (dotted line), for the Poisson random graph, Eqs. (20) and (21).

Es decir que existe una probabilidad critica a partir de la cual una fraccion macroscopica de nodos se agrupan en lo que se llama "el componente gigante"

Sea u la fraccion de nodos que no pertenecen al componente gigante.

la probabilidad de que un nodo no pertenezca a la componente gigante es igual a al proba que ninguno de sus vecinos pertenezca a dicha componente

Si el nodo tiene grado k esta probabilidad es u^k

Entonces

$$u = \sum_{k=0}^{\infty} p_k u^k = e^{-z} \sum_{k=0}^{\infty} (zu)^k \frac{1}{k!} = e^{-z} e^{zu} = \exp(z(u-1))$$

pues $\sum_{k=0}^{\infty} \frac{x^k}{k!} = e^x$

Entonces la fraccion S de nodos que pertenecen a la componente gigante es

$$S = (1-u) = (1-e^{-zS})$$

Se demuestra que

$$\langle s \rangle = \frac{1}{1-z-zS}$$

Entonces estamos en algo que se parece a una transicion de fase donde S juega el rol de un parametro de orden y $\langle s \rangle$ de las fluctuaciones.

Aparecen entonces exponentes criticos

$$S \sim (z-1)^{\beta}$$

$$\langle s \rangle \sim |z-1|^{-\gamma}$$

con β y γ =1

La transicion ocurre en $z = 1$ y ademas esto es fijar p pues
$z = p(n - 1)$

Si se estudia la distribucion de fragmentos en $z = 1$ se encuentra que estan distribuidos con $\tau = 3/2$ cuando uno elige nodos del modo que hicimos en percolacion.

El random graph describe solo el camino minimo medio de los grafos reales.

Random graph generalizado

El modelo configuracional!!!!!

i) especificamos una distribucion de grados p_k

ii) elegimos una secuancia de grados k_i con $i = 1, 2, 3, 4\ldots$

Cada vertice i tendra entonces k_i "palitos" saliendo

iii) elegimos pares de "palitos" al azar y los unimos entre si.

Se demuestra que este proceso genera todo posible grafo

El modelo configuracional queda entonces definido por el ensemble de grafos con igual probabilidad.

Modelos de crecimiento de networks

Hasta ahora vimos como eran las propiedades de networks con caracteristicas dadas

Ademas vimos como eran los networks reales

Como es que surgen los networks reales?

Focalizamos en el power law distribution of grades

a) Herbert Simon "rich get richer" (1955)

[69] Bornholdt, S. and Ebel, H., World Wide Web scaling exponent from Simon's 1955 model, *Phys. Rev. E* **64**, 035104 (2001).

b) Price (1965) aplico las ideas de Simon al crecimiento de networks

[344] Price, D. J. de S., A general theory of bibliometric and other cumulative advantage processes, *J. Amer. Soc. Inform. Sci.* **27**, 292–306 (1976).

Price appears to have been the first to discuss cumulative advantage specifically in the context of networks, and in particular in the context of the network of citations between papers and its in-degree distribution. His idea was that the rate at which a paper gets new citations should be proportional to the number that it already has. This is easy to justify in a qualitative way. The probability that one comes across a particular paper whilst reading the literature will presumably increase with the number of other papers that cite it, and hence the probability that you cite it yourself in a paper that you write will increase similarly. The same argument can be applied to other networks also, such as the Web. It is not clear that the dependence of citation probability on previous citations need be strictly linear, but certainly this is the simplest assumption one could make and it is the one that Price, following Simon, adopts. We now describe in detail Price's model and his exact solution of it, which uses what we would now call a *master-equation* or *rate-equation* method.

Sea un grafo dirigido

a) con n vertices,

b) con p_k la fraccion de vertices con de k 'in-degree'

$$\sum_k p_k = 1$$

c) vertices nuevos se agregan continuamente

Los vertices nuevos tendran un 'out-degree' con un valor medio m

Como deben unirse a algo, se satisface

$$\sum_k k \cdot p_k = m$$

Dado un nuevo vertice con un cierto 'out-degree' lo uniremos a los viejos vertices con una probabilidad proporcional al 'in-degree' de los mismos

Como cuando se empieza el 'in-degree' es cero se supone que el proceso anterior sera porpocional a $k + k_0$, (tomo $k_0 = 1$)

Entonces la proba de que se una a un vertice de 'in-degree' k es

$$
\begin{aligned}
\frac{(k+1)n_k}{\sum_k (k+1)n_k}\frac{n}{n} &= \frac{(k+1)p_k}{\sum_k (k+1)p_k} \\
&= \frac{(k+1)p_k}{\sum_k kp_k + \sum_k p_k} \\
&= \frac{(k+1)p_k}{m+1}
\end{aligned}
$$

El numero medio de nuevas citas a vertices con 'in-degree' k es

$$\frac{(k+1)p_k}{(m+1)} \cdot m = \frac{m}{(m+1)} \cdot (k+1)p_k$$

y esta es la perdida de nodos del tipo 'in-degree' k.

Sea $p_{k,n}$ la proba al paso con n vertices

Con esto podemos plantear para la variacion del numero de nodos con 'in-degree' k

al ir de (n) a $(n+1)$ vertices y tomando en cuanta los aportes desde $(k-1)$ y los aportes a $(k+1)$

para $k \geq 1$

$$(n+1)p_{k,n+1} - np_{k,n} = [kp_{k-1,n} - (k+1)p_{k,n}]\frac{m}{m+1}$$

para $k = 0$

$$(n+1)p_{0,n+1} - np_{0,n} = [1 - p_{0,n}]\frac{m}{m+1}$$

La solucion estacionaria corresponde a $p_{k,n+1} = p_{k,n}$

De donde

$$(n+1)p_k - np_k = [kp_{k-1} - (k+1)p_k]\frac{m}{m+1}$$

$$(n+1)p_0 - np_0 = [1 - p_0]\frac{m}{m+1}$$

$$p_k = [kp_{k-1} - (k+1)p_k]\frac{m}{m+1}$$
$$p_0 = 1 - p_0\frac{m}{m+1}$$

De donde para el primer caso

$$p_k = [kp_{k-1} - (k+1)p_k]\tfrac{m}{m+1} \Rightarrow p_k(\tfrac{m+1}{m}(k+1)) = kp_{k-1} \Rightarrow$$

$$p_k = p_{k-1}\frac{k}{(k+2+1/m)}$$

Para el otro

$$p_0 = \frac{m+1}{(2m+1)}$$

De este modo a partir de p_0 se construyen los otros p_k y entonces

$$p_1 = p_0\frac{1}{(3+1/m)} = \frac{m+1}{(2m+1)}\frac{m}{(3m+1)}$$

en general

$$p_k = p_0 \frac{k(k-1)..1}{(k+2+1/m)\dots(3+1/m)}$$
$$= \left(1+\frac{1}{m}\right)B(k+1, 2+1/m)$$

con $B(k+1, 2+1/m) = B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$

Si a crece y b se mantiene fijo, asintoticamente va a a^{-b} o sea

$$p_k \sim k^{-(2+1/m)}$$

qed

Observar que k_0 no aparece en la solucion con lo que se confirma

que haberlo elegido =1...

Albert & **Barabasi**

Siguen la linea de Price, pero el grafo es no dirigido

Comienzan con nodos con grado m