Inferencia redes biológicas

Una introducción

Roberto Álvarez [1]

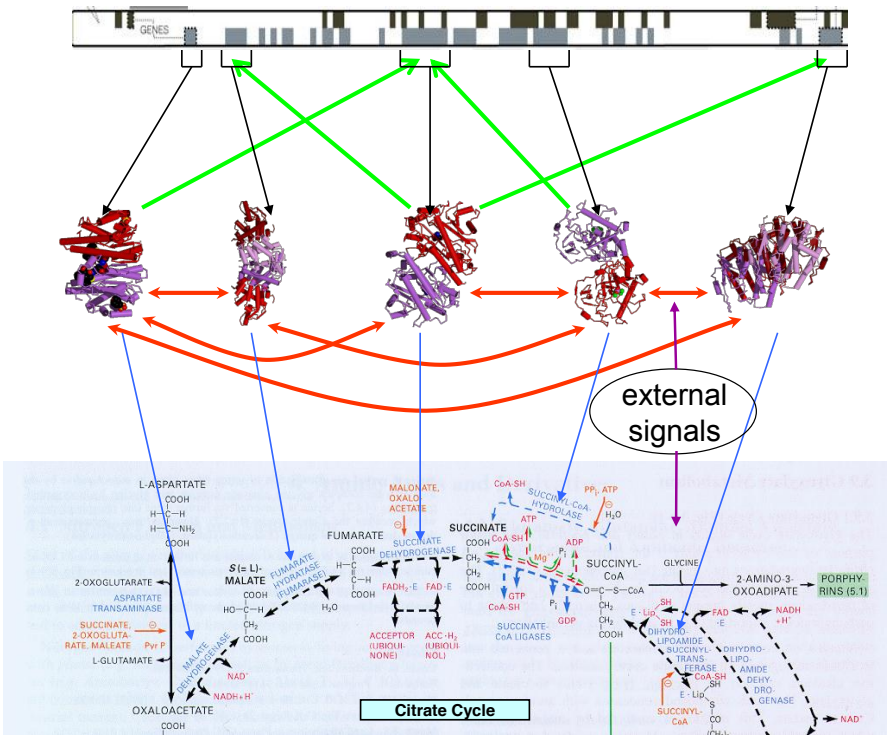
roberto.carlos.alvarez.martinez@gmail.com

[1] Unidad de Microbiología.
Facultad de Ciencias Naturales.
Universidad Autónoma de Querétaro

Outline

1. Introducción
   - La importancia de las redes complejas
2. Redes Biológicas
3. Teoría de grafos
   - Definiciones básicas
4. Visualización
5. Propiedades universales
6. Popiedades estáticas
   - Distribución de conectividades
   - Mundo pequeño
   - Distancias en redes
7. Centralidad de conectividades
8. Betweenness centrality
9. Métodos de clusterización
10. Redes genéticas
    - Redes booleanas
    - Redes de co-expresión

Complejidad



GENOME

gene regulation

PROTEOME

protein-protein interactions

signal transduction

METABOLISM

Bio-chemical reactions

Figura: Figura de R. Albert 2012

El siglo de la complejidad

"I think the next century will be the century of complexity"

Stephen Hawking, Enero 23 de 2000

¿Qué es la complejidad?

Complejidad, una teoría científica que afirma que algunos sistemas observan un comportamiento que no es explicable por análisis convencionales de las partes constitutivas. Fenomenos emergentes que sólo se entienden en la interacción de loc omponentes internos del sistema. A este fenómeno, se le llama comúnmente como propiedad emergente, que parecen ocurrir en muchos sistemas tales como organismos vivos, el mercado de valores o el cerebro humano.

La complejidad

Sistemas complejos: características

- Sin control maestro.
- Auto-organización (espacial, temporal).
- Adaptación.
- Leyes fenomenológicas a distinto niveles.
- Propiedades emergentes.

Propiedades emergentes: ejemplos

Emergent Properties of Networks of Biological Signaling Pathways

Upinder S. Bhalla and Ravi Iyengar*

Many distinct signaling pathways allow the cell to receive, process, and respond to information. Often, components of different pathways interact, resulting in signaling networks. Biochemical signaling networks were constructed with experimentally obtained constants and analyzed by computational methods to understand their role in complex biological processes. These networks exhibit emergent properties such as integration of signals across multiple time scales, generation of distinct outputs depending on input strength and duration, and self-sustaining feedback loops. Feedback can result in bistable behavior with discrete steady-state activities, well-defined input thresholds for transition between states and prolonged signal output, and signal modulation in response to transient stimuli. These properties of signaling networks raise the possibility that information for "learned behavior" of biological systems may be stored within intracellular biochemical reactions that comprise signaling pathways.

Studies on the cyclic adenosine monophosphate (cAMP) signaling pathway led to the identification of several general mechanisms of signal transfer, such as regulation by protein-protein interactions, protein phosphorylation, regulation of enzymatic activity, production of second messengers, and cell surface signal transduction systems (*1*). These mechanisms of signal transfer have subsequently been shown to occur in many pathways, including Ca^{2+} signaling pathways (*2*), tyrosine kinase pathways (*3*), and other protein kinase cascades, and recently in the intracellular protease cascades in apoptosis (*4*). Initially, signaling pathways were studied in a linear fashion, and it was shown that many important biological effects are obtained through linear information transfer. However, it has become increasingly clear that signaling pathways interact with one another and the final biological response is shaped by interaction between pathways. These interactions result in networks that are quite complex and may have properties that are nonintuitive. A systematic analysis of interactions between signaling pathways could be useful in understanding the properties of these networks. We developed models for simple net-

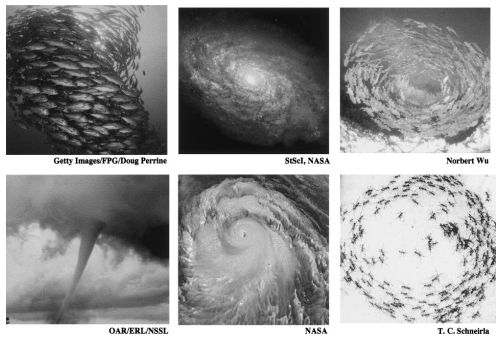
U. S. Bhalla, Department of Pharmacology, Mount Sinai School of Medicine, New York, NY 10029, USA, and National Centre for Biological Sciences, Post Office Box 1234, Bangalore 560012, India. R. Iyengar, Department of Pharmacology, Mount Sinai School of Medicine, New York, NY 10029, USA.

*To whom correspondence should be addressed.

Self-Organized Fish Schools: An Examination of Emergent Properties

JULIA K. PARRISH[1,2,*], STEVEN V. VISCIDO[2], AND DANIEL GRÜNBAUM[3]

[1] *School of Aquatic and Fishery Sciences, Box 355020, University of Washington, Seattle, Washington 98195-5020;* [2] *Zoology Department, University of Washington; and* [3] *School of Oceanography, University of Washington*



THE ROYAL SOCIETY biology letters

The emergent properties of a dolphin social network

David Lusseau†

Department of Zoology, University of Otago, PO Box 56, Dunedin, New Zealand (lusda563@student.otago.ac.nz)

Recd 29.05.03; *Accptd* 02.06.03; *Online* 04.07.03

Many complex networks, including human societies, the Internet, the World Wide Web and power grids, have surprising properties that allow vertices

Herramientas para el estudio de los sistemas complejos

- Autómatas celulares (Ulam, von Neumann 1940, Wolfram 1980)
- Modelos basados en agentes (Ulam, von Neuman 1940, Comway 1970)
- Redes de ecuaciones diferenciales (Leibnitz, Newton 1600)
- Redes complejas (Euler, Erdös, Renyi, Barabasi, Albert, Newman, Strogatz, . . .)

Las redes: (una) herramienta para estudiar los sistemas complejos

Propiedades

- Un número grande de elementos no (necesariamente) idénticos conectados por diversas interacciones. Por ejemplo, redes celulares: interacción de proteínas, reacciones químicas, y regulación de expresión génica.
- Propiedades de la red proporcionan información sobre las interacciones.
- La topología de las redes refleja la robustez de los sistemas y de las dinámicas de flujos entre los elementos.
- Las dinámicas de las interacciones moleculares determina el comportamiento de las células. (Puntos fijos de los sistemas son fenotipos)
- Entender las propiedades emergentes: sincronización, diferenciación y homeostásis.

¿Por qué en este momento?

- Disponibilidad de datos y capacidad de cómputo.
- Interesantes.
- Necesarios.

Datos

Disponibilidad

- Red de conexiones neuronales de *C. elegans*,1990
- Red de Actores, 1998
- Red de citas de artículos científicos, 1998
- World Wide Web, 1999.
- Red metabólica, 2000.
- Red de interacción de proteínas, 2001.

Redes sociales

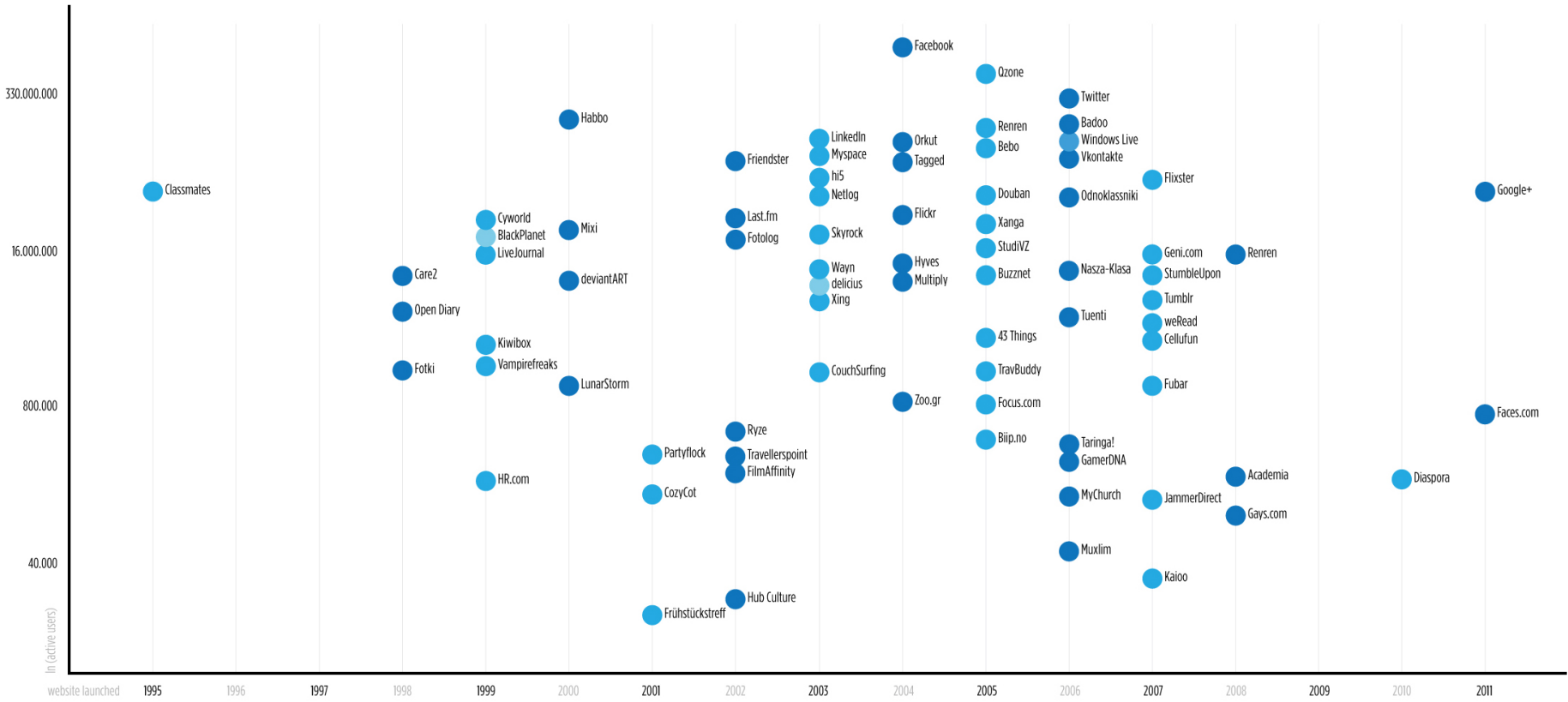


Figura: Barabasi et al. 2011

(Científicamente) interesantes

> Las cosas se parecen, por eso existe la ciencia; las cosas no se parecen, para eso existe la ciencia
>
> Richard Levins

Universalidad

La arquitectura de las redes (realmente) existentes es, al parecer, independiente del sustrato material de sus componentes. Las redes que se presentan en la naturaleza, las sociales y teconológicas tienen propiedades estáticas similares.

Necesidad: comprender los sistemas complejos

Las redes no sólo son esenciales para entender muchos sistemas complejos; también constituyen ya un campo de estudio propio (Ciencia de redes)
En la década pasada los avances más importantes en el área de los sistemas complejos se dieron en el estudio de las redes.

Características de la ciencia de las redes

- Interdisciplinaria.
- Empírica.
- Cuantitativa y matemática.
- Computacional.

Herramientas de la teoría de redes

Toolbox

- Teoría de grafos (gráficas)
- Teoría de redes sociales
- Física Estadística
- Ciencias d ela computación
- Biología
- Estadística
- Bionformática

(Algunos de) los *papers* fundamentales

- 1959: Erdös, P.; Rényi, A. “On Random Graphs. I”. Publicationes Mathematicae 6: 290–297
- 1998: Watts y Strogatz ”Collective dynamics of “small-world” networks. Nature 393 (6684): 440–442
- 1999: Barabasi y Albert “Emergence of scaling in random networks”. Science 286: pp. 509–512
- 2001: Pastor -Satorras y Vespignani “Epidemic spreading in scale-free networks” Physical Review Letters: 86 (14), 3200
- 2002: Girvan y Newman “Community structure in social and biological network”s, Proc. Natl. Acad. Sci. USA 99, 7821–7826.

Reviews

- R. Albert, A-L Barabasi, “Statistichal Mechanics of Complex Networks”, REviews of Modern Physics, 2001
- A-L Barabasi, ZN. Oltvai, “Network biology: understanding the cell’s functional organization”, Nature Reviews Genetics, 2004
- R. Albert, “Scale-free networks in cell biology”, Journal of cell science, 2005

Hitos en la física

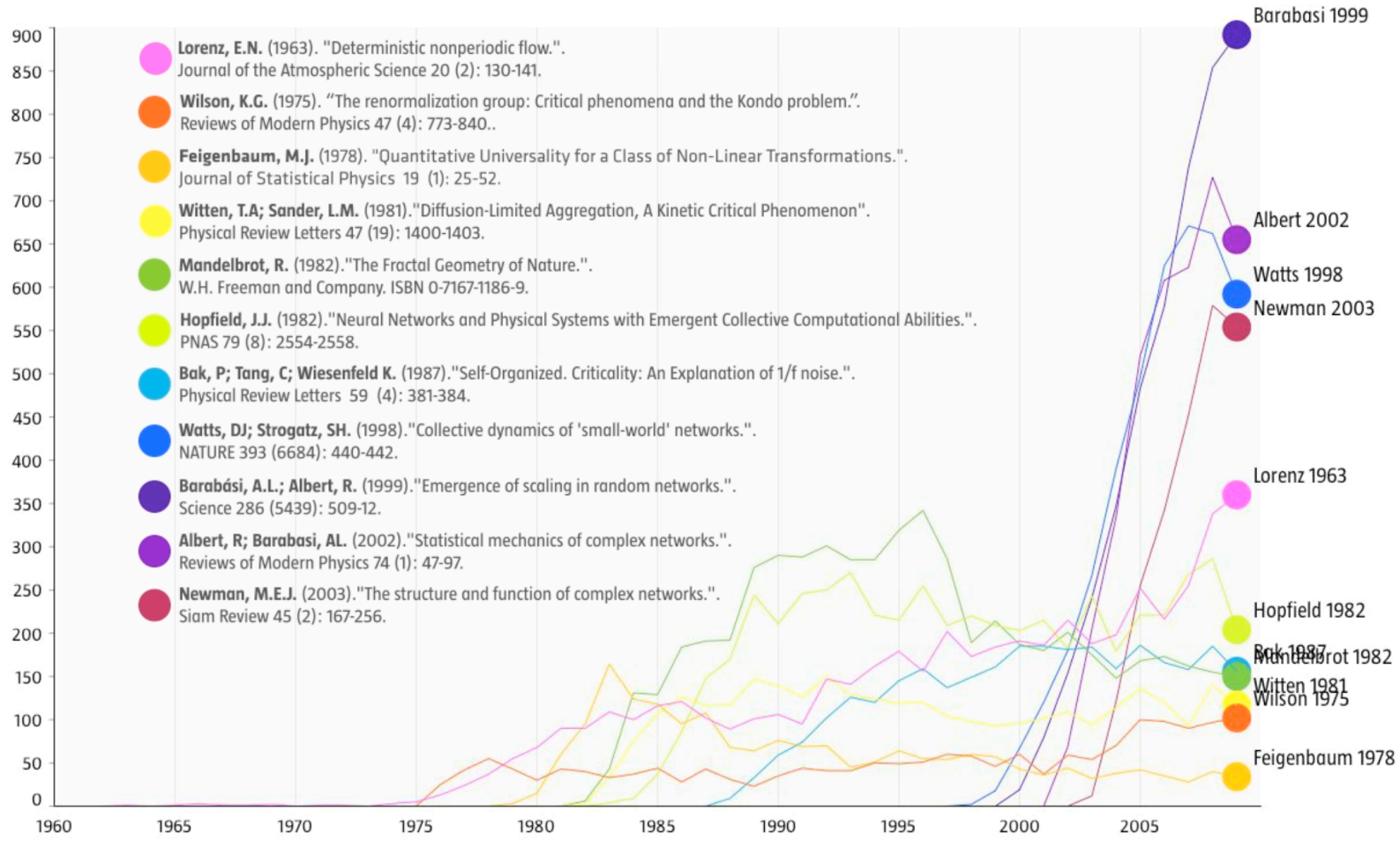
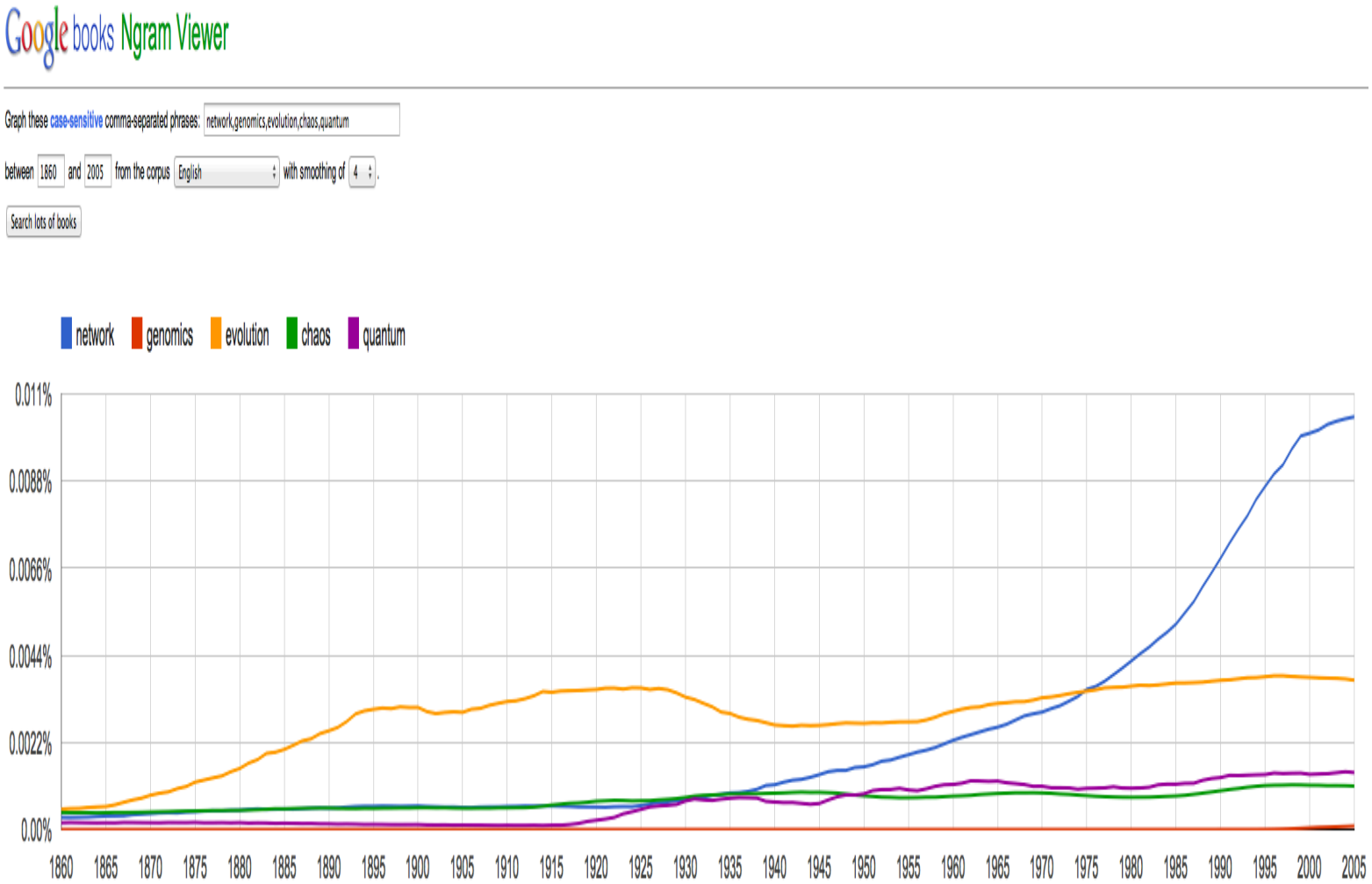
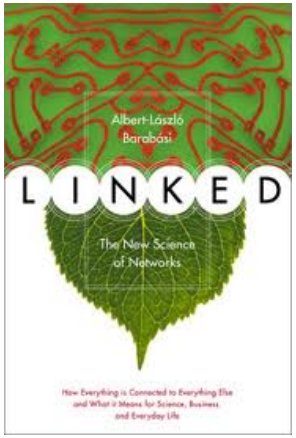


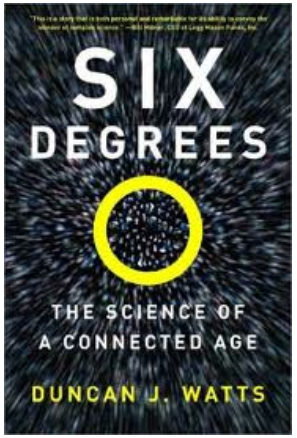
Figura: Barabasi et al. 2011

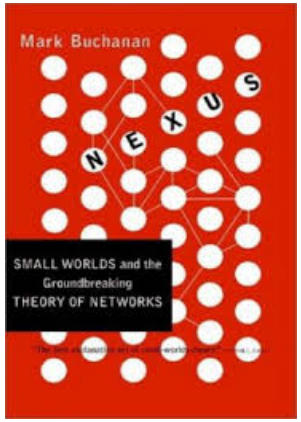
Otros

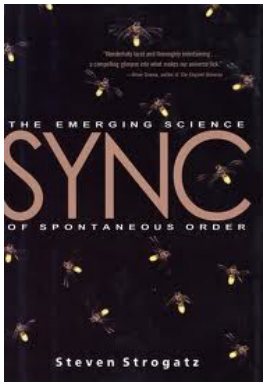

Google books Ngram Viewer
Graph these case-sensitive comma-separated phrases: network,genomics,evolution,chaos,quantum
between 1860 and 2005 from the corpus English with smoothing of 4 .
Search lots of books
network
genomics
evolution
chaos
quantum
0.011%
0.0088%
0.0066%
0.0044%
0.0022%
0.00%
1860 1865 1870 1875 1880 1885 1890 1895 1900 1905 1910 1915 1920 1925 1930 1935 1940 1945 1950 1955 1960 1965 1970 1975 1980 1985 1990 1995 2000 2005

Libros de divulgación

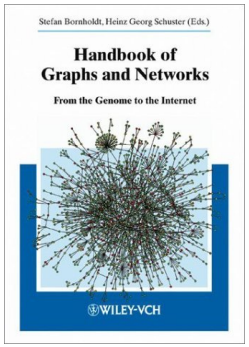




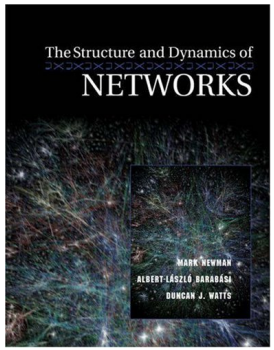




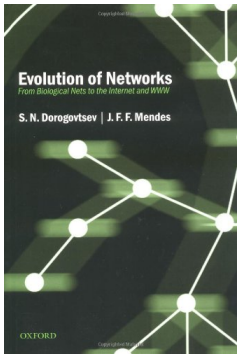
Libros de “texto”



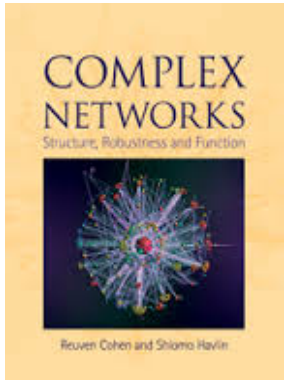
Handbook of Graphs and Networks: From the Genome to the Internet (Wiley-VCH, 2003).



M. Newman, A.-L. Barabasi, D. J. Watts, The Structure and Dynamics of Networks: (Princeton Studies in Complexity) (Princeton University Press, 2006)

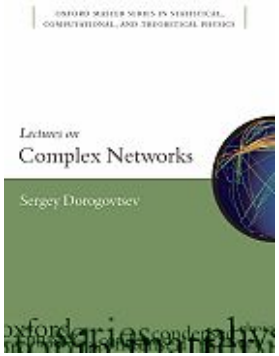


S. N. Dorogovtsev and J. F. F. Mendes, Evolution of Networks: From Biological Nets to the Internet and WWW (Oxford University Press, 2003).

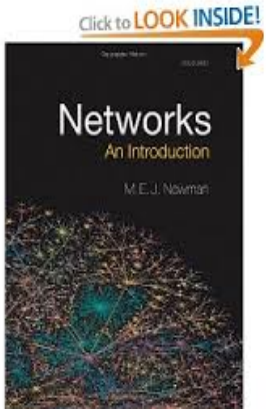


R. Cohen, S. Havlin, Complex Networks. Structure, Robustness and Function, (Cambridge University Press, 2010)

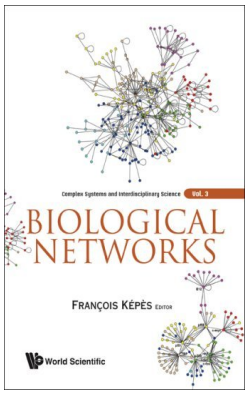
Libros de “texto”



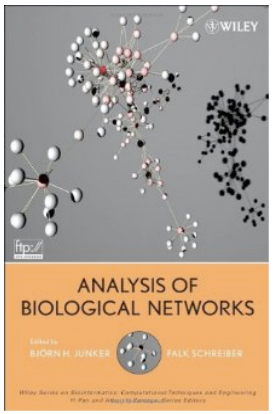
S. Dorogovtsev, Lectures on Complex Networks, Oxford Master Series in Physics, 2010



M.E.J. Newman, Networks. An Introduction, Oxford University Press,2010



F. Kopos, Biological Networks (Complex Systems and Interdisciplinary Science) (World Scientic Publishing Company, 2007)



B. H. Junker, F. Schreiber, Analysis of Biological Networks (Wiley Series in Bioinformatics) (Wiley-Interscience, 2008).

Red (preliminar)

“Definición”

Una red es una conjunto de elementos vinculados entre si.

1. Conjunto de elementos (¿son únicos, iguales, son de la misma clase, etc?)
2. Vínculos (¿ físicos, cómo se definen, son iguales, etc?)

Ejemplos: Internet (Computadoras, conexiones físicas), WWW (páginas web, enlances), Proteínas (interacciones físicas), reacciones químicas (reactantes) , amistades (personas, amistad),

Red (preliminar)

“Definición”

Una red es una conjunto de elementos vinculados entre si.

1. Conjunto de elementos (¿son únicos, iguales, son de la misma clase, etc?)
2. Vínculos (¿ físicos, cómo se definen, son iguales, etc?)

Ejemplos: Internet (Computadoras, conexiones físicas), WWW (páginas web, enlances), Proteínas (interacciones físicas), reacciones químicas (reactantes) , amistades (personas, amistad),

Red (preliminar)

“Definición”

Una red es una conjunto de elementos vinculados entre si.

1. Conjunto de elementos (¿son únicos, iguales, son de la misma clase, etc?)
2. Vínculos (¿ físicos, cómo se definen, son iguales, etc?)

Ejemplos: Internet (Computadoras, conexiones físicas), WWW (páginas web, enlances), Proteínas (interacciones físicas), reacciones químicas (reactantes) , amistades (personas, amistad),

Red (preliminar)

“Definición”

Una red es una conjunto de elementos vinculados entre si.

1. Conjunto de elementos (¿son únicos, iguales, son de la misma clase, etc?)
2. Vínculos (¿ físicos, cómo se definen, son iguales, etc?)

Ejemplos: Internet (Computadoras, conexiones físicas), WWW (páginas web, enlances), Proteínas (interacciones físicas), reacciones químicas (reactantes) , amistades (personas, amistad),

Definición

Un grafo $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ consiste en un conjunto de nodos (*vertex*) $\mathcal{V}$ y un conjunto de enlaces (*edges, arcs, links*), $\mathcal{E}$, en el cual cada enlace está asignado a dos nodos.

Redes biológicas

$$\{E, V\}$$

1. ¿Quiénes son los nodos?
2. ¿Cómo se conectan?

nodos: personas, especies, metabolitos, proteínas, genes, neuronas, etc.
conexiones: contacto (epidemiología),depredador-pesa, reacción química, *binding*, co-expresión/regulación, activación.

Redes biológicas

$$\{E, V\}$$

1. ¿Quiénes son los nodos?
2. ¿Cómo se conectan?

nodos: personas, especies, metabolitos, proteínas, genes, neuronas, etc.
conexiones: contacto (epidemiología),depredador-pesa, reacción química, *binding*, co-expresión/regulación, activación.

Redes biológicas

$$\{E, V\}$$

1. ¿Quiénes son los nodos?
2. ¿Cómo se conectan?

nodos: personas, especies, metabolitos, proteínas, genes, neuronas, etc.
conexiones: contacto (epidemiología),depredador-pesa, reacción química, *binding*, co-expresión/regulación, activación.

Redes biológicas

$$\{E, V\}$$

1. ¿Quiénes son los nodos?
2. ¿Cómo se conectan?

nodos: personas, especies, metabolitos, proteínas, genes, neuronas, etc.

conexiones: contacto (epidemiología),depredador-pesa, reacción química, *binding*, co-expresión/regulación, activación.

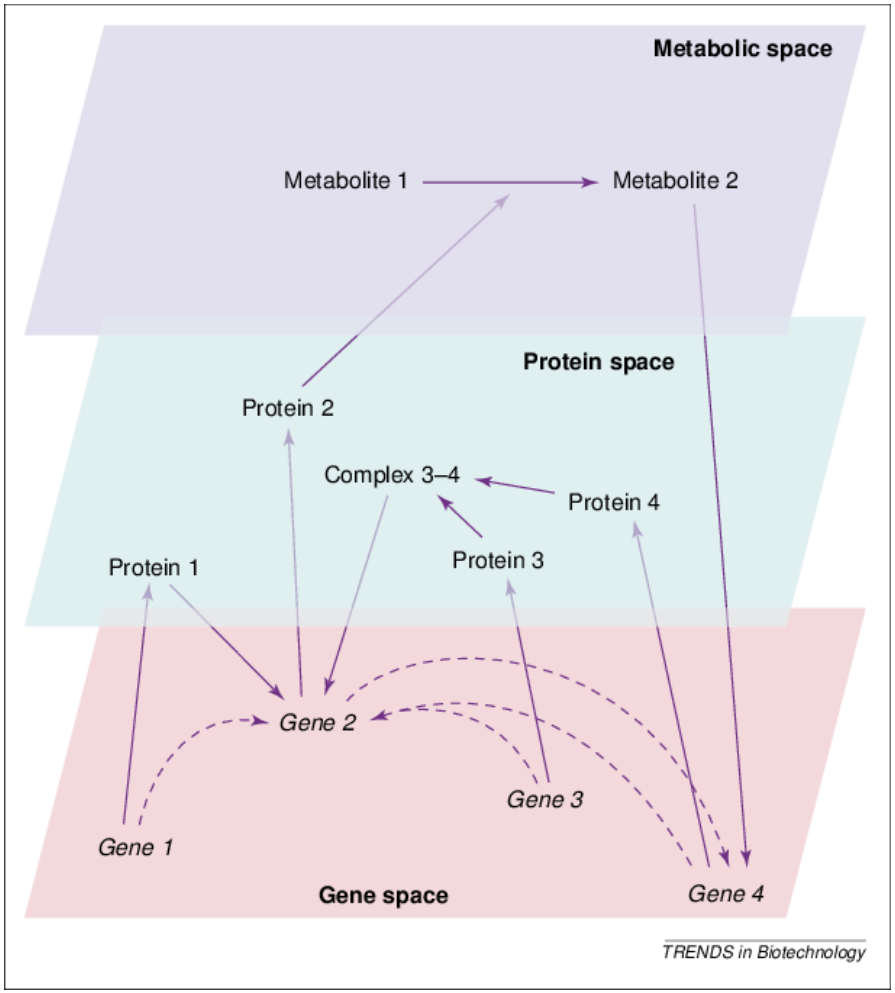
Redes biológicas

$$\{E, V\}$$

1. ¿Quiénes son los nodos?
2. ¿Cómo se conectan?

nodos: personas, especies, metabolitos, proteínas, genes, neuronas, etc.
conexiones: contacto (epidemiología),depredador-pesa, reacción química, *binding*, co-expresión/regulación, activación.

Redes


Metabolic space
Metabolite 1
Metabolite 2
Protein space
Protein 2
Complex 3–4
Protein 4
Protein 3
Protein 1
Gene 2
Gene 3
Gene 1
Gene space
Gene 4
TRENDS in Biotechnology

Red de regulación transcripcional

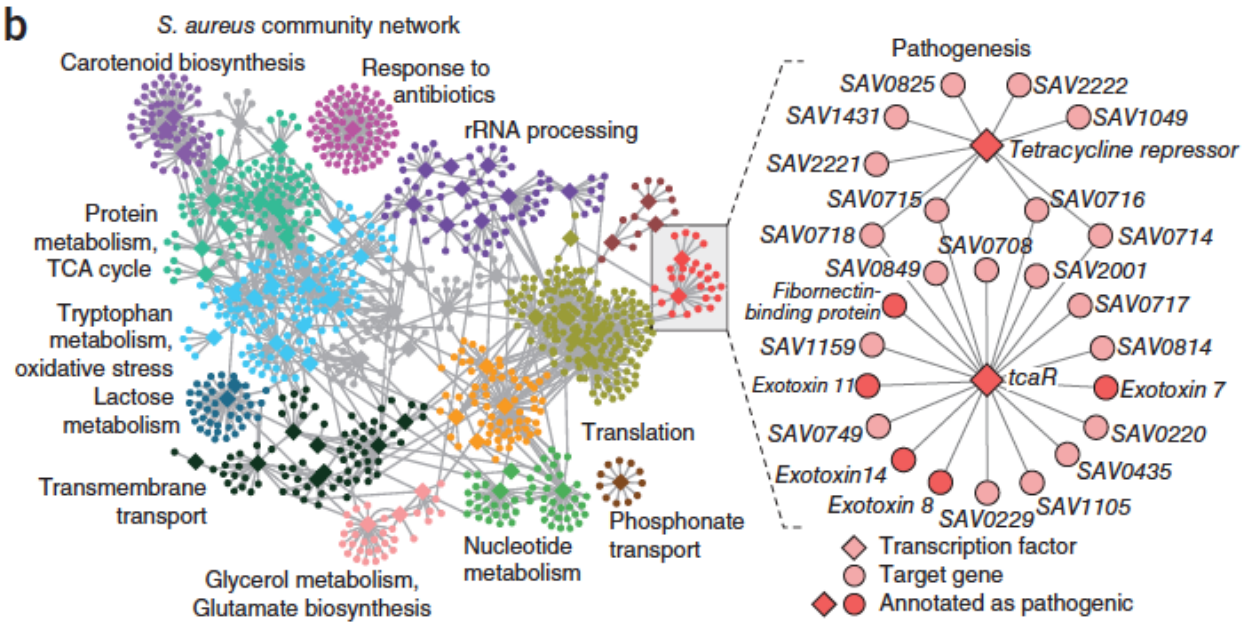
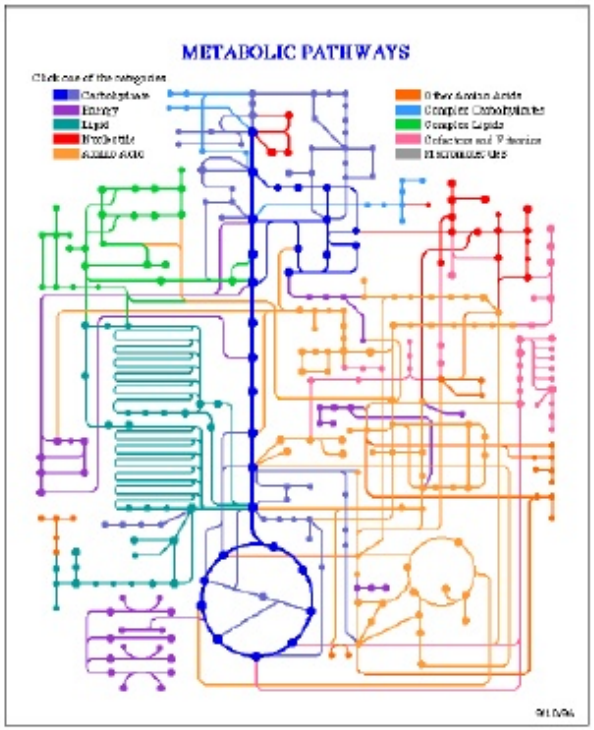


Figura: Redes de *E. coli* y *S. aureus*, $\sim 1,700$ interacciones transcripcionales (Marbach *et al.*, Nature Methods, 2012)

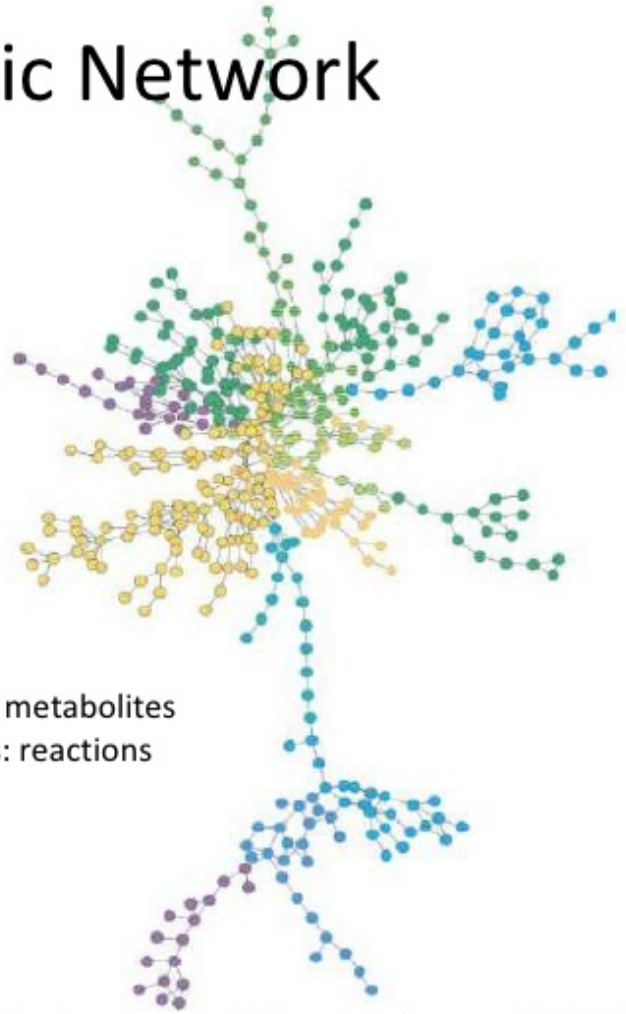
Red de interacción de proteínas

Figure 2 | **Yeast protein interaction network.** A map of protein–protein interactions[18] in *Saccharomyces cerevisiae*, which is based on early yeast two-hybrid measurements[23], illustrates that a few highly connected nodes (which are also known as hubs) hold the network together. The largest cluster, which contains ~78% of all proteins, is shown. The colour of a node indicates the phenotypic effect of removing the corresponding protein (red = lethal, green = non-lethal, orange = slow growth, yellow = unknown). Reproduced with permission from REF. 18 © Macmillan Magazines Ltd.

Red metabólica



Kegg, Wit, Biocyc, Bigg (UCSD)



Guimera and Nunes Amaral 2005

Red cerebro

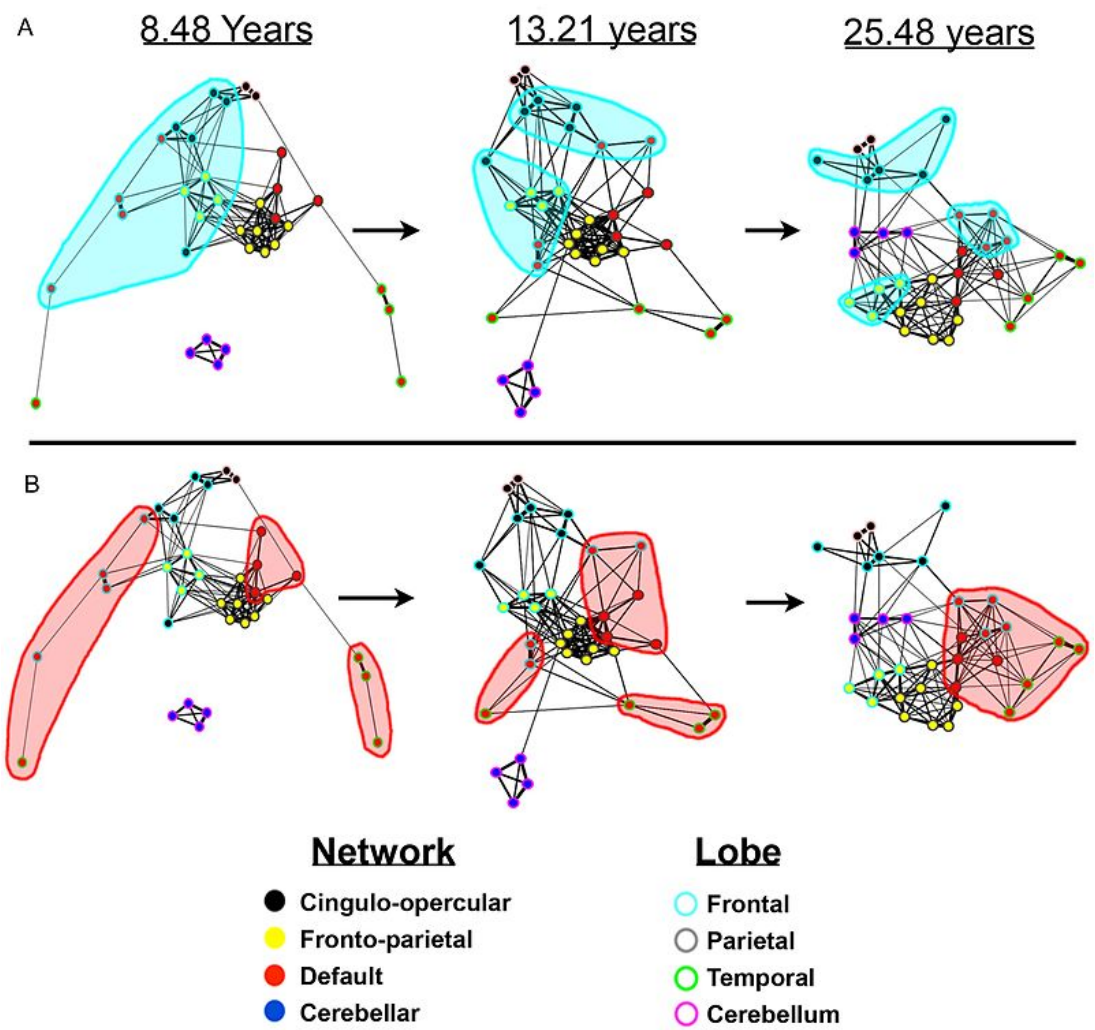
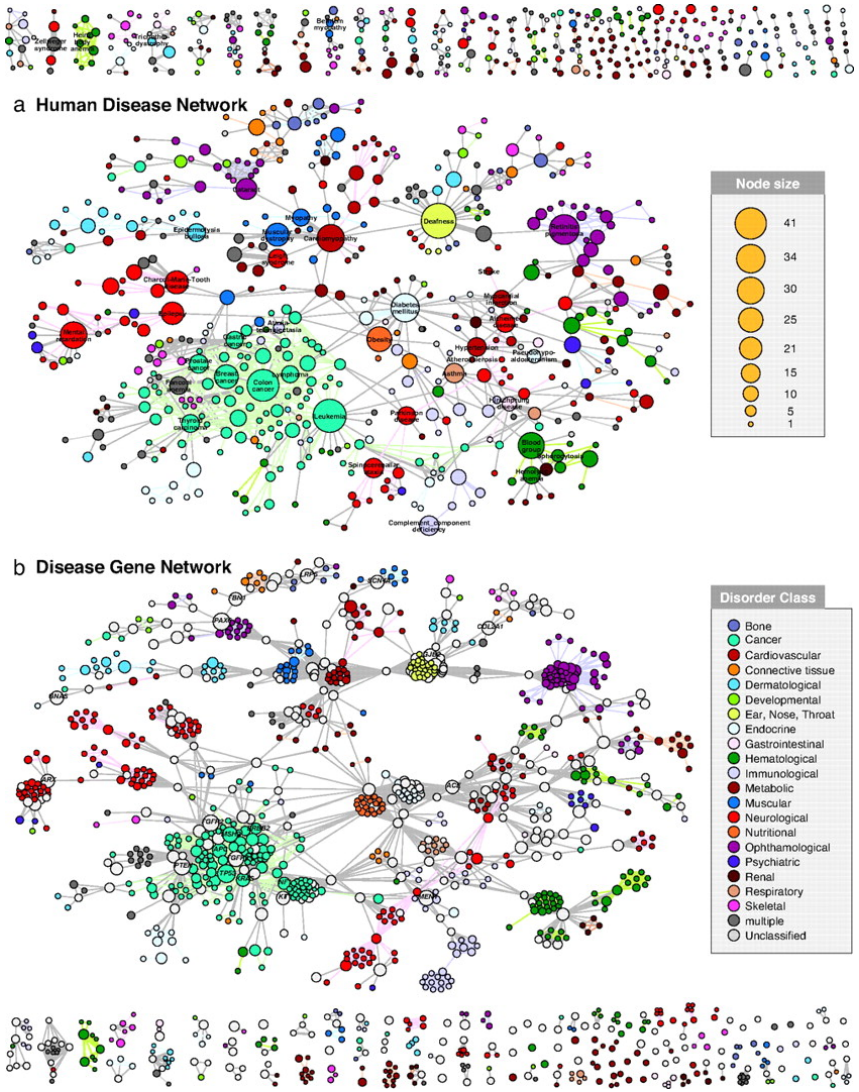


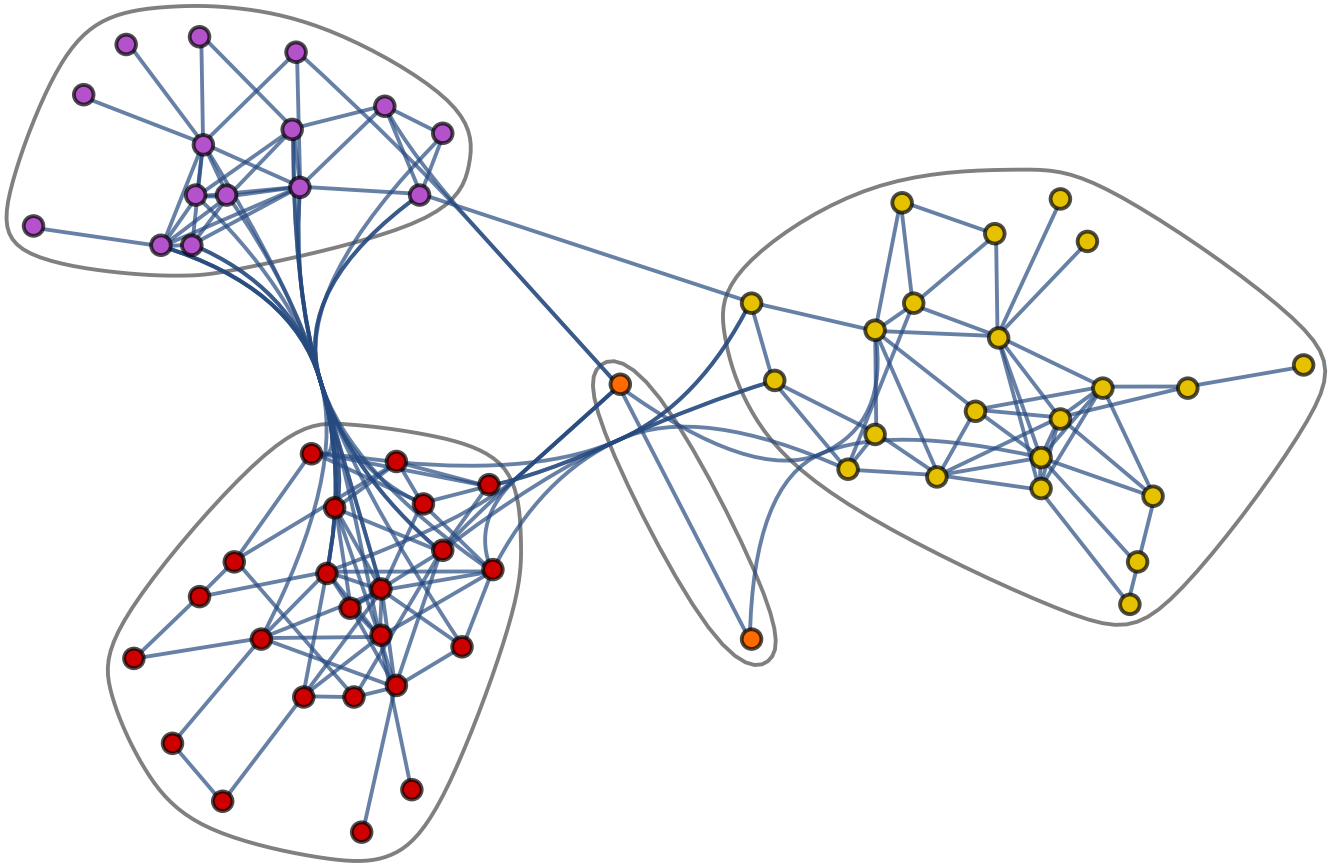
Figura: Fair, Damien A. *et al.*"Functional Brain Networks Develop from a 'Local to Distributed' Organization,PLoS Computational Biology 2009

Red de enfermedades

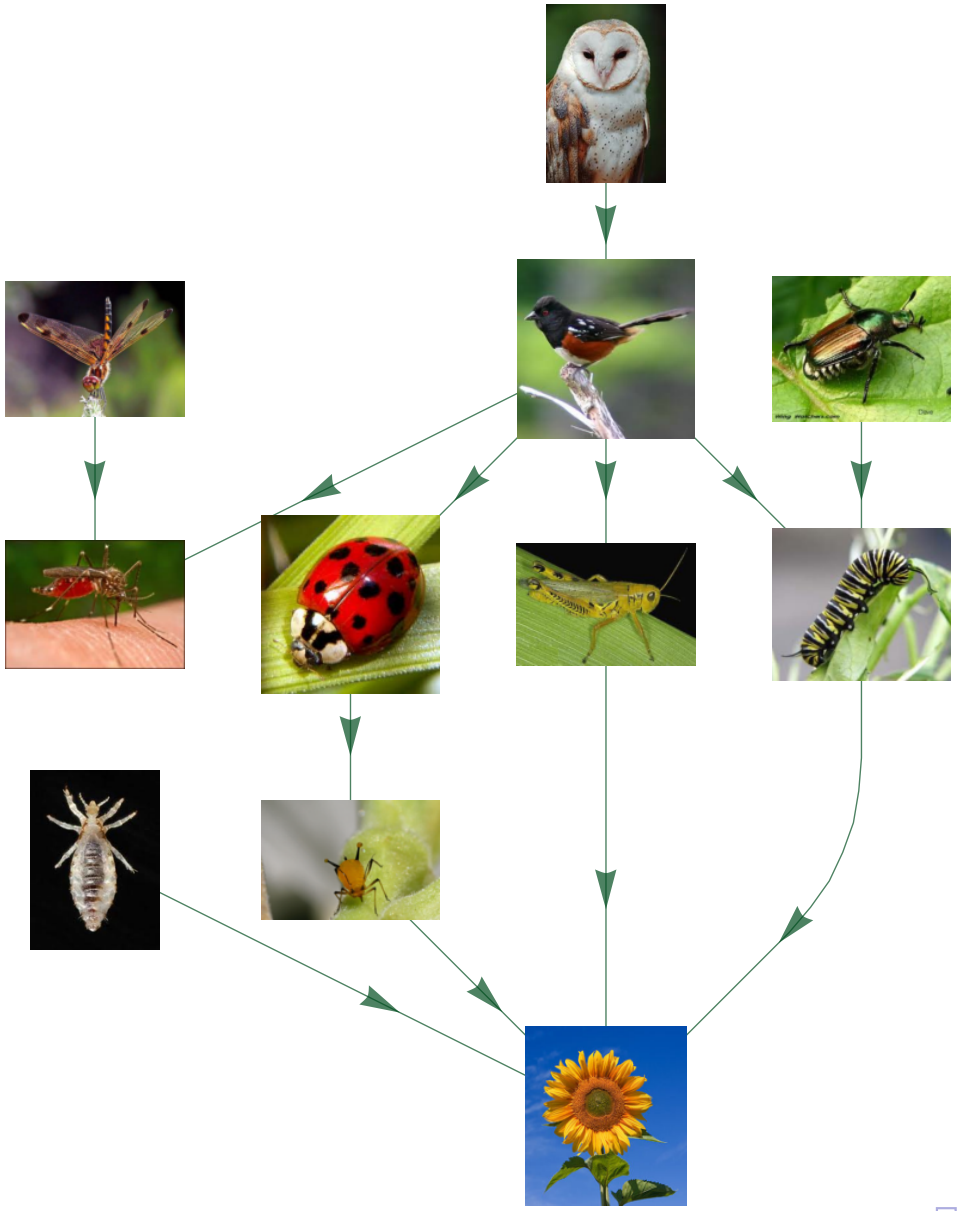


Kwang-Il Goh et al. PNAS 2007;104:8685-8690

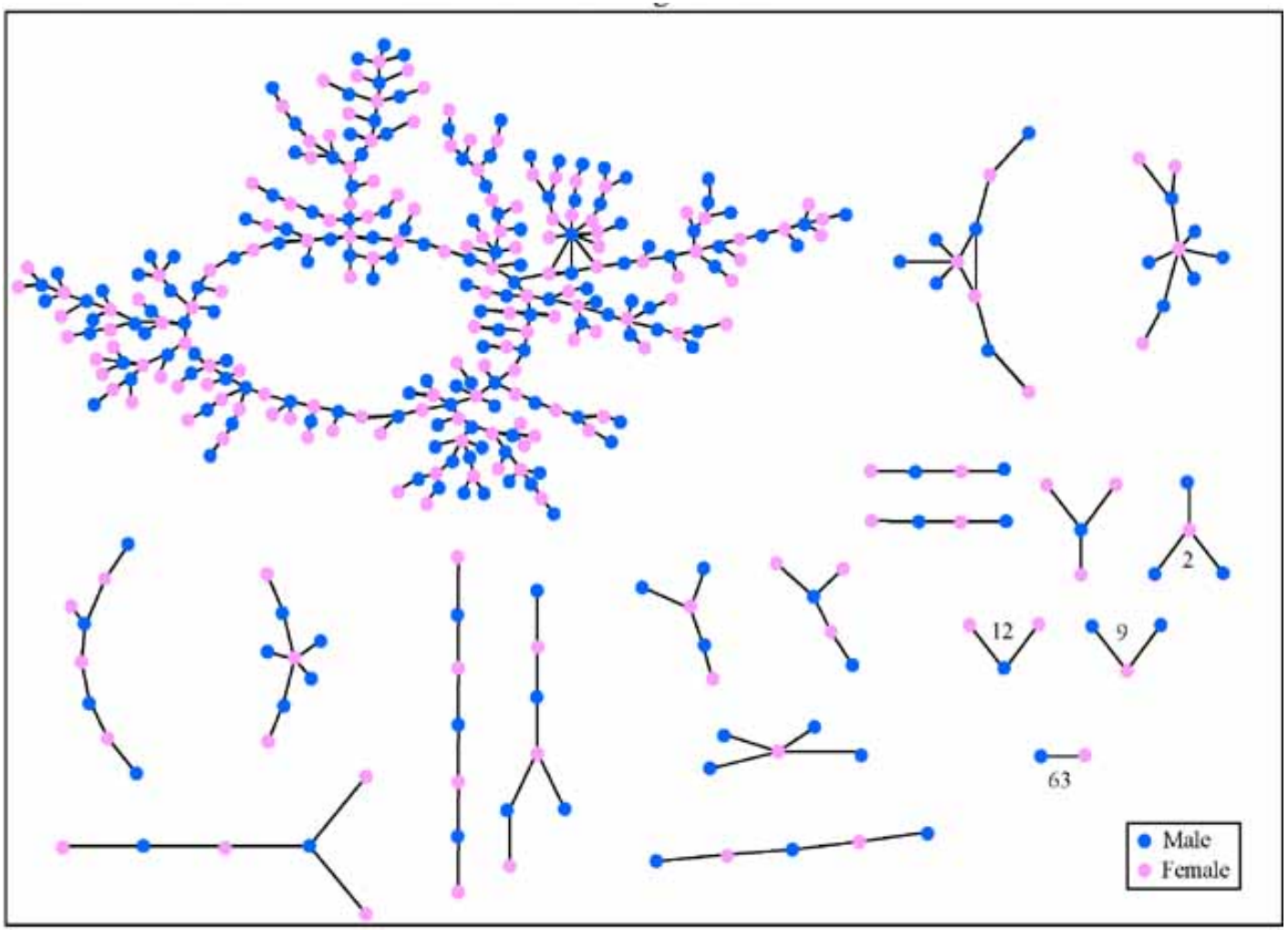
Red de amistades en delfines



Red trófica



Red de contactos sexuales



Each circle represents a student and lines connecting students represent romantic relations occuring 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).

Figura: h

Ejemplos con GeneMANIA y String

Bases de datos

- GeneMANIA
- STRING
- PPI Yeast

- igraph
- Cytoscape
- Gephi

Propiedades comunes

Redes biológicas

- Distribución de conectividades (sin escala).
- Mundo pequeño (*small world*).
- Coeficiente de agrupamiento.

Redes no dirigidas

Definición

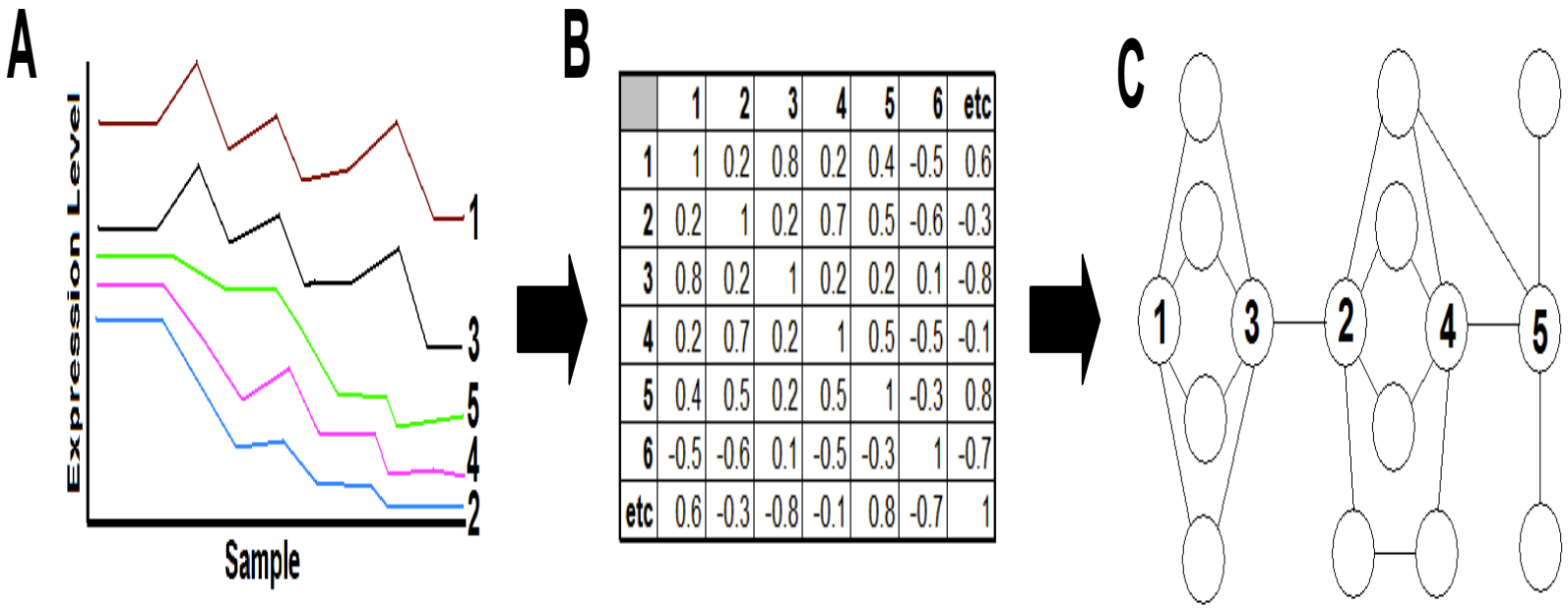
Una red es **no** dirigida si la conexión de i a j es equivalente a la conexión de j a i

$$M_{ij} = M_{ji}$$

Ejemplos:

- Dos actores se conectan si participaron en la misma película. (IMDB).
- Red de coautorías científicas. (Dos autores se conectan si escribieron un paper juntos).
- Red de co-expresión de genes (dos genes se conectan si su correlación es mayor un cierto valor mínimo.)
- Red de interacción de proteínas.

Red de co-expresión



	1	2	3	4	5	6	etc
1	1	0.2	0.8	0.2	0.4	-0.5	0.6
2	0.2	1	0.2	0.7	0.5	-0.6	-0.3
3	0.8	0.2	1	0.2	0.2	0.1	-0.8
4	0.2	0.7	0.2	1	0.5	-0.5	-0.1
5	0.4	0.5	0.2	0.5	1	-0.3	0.8
6	-0.5	-0.6	0.1	-0.5	-0.3	1	-0.7
etc	0.6	-0.3	-0.8	-0.1	0.8	-0.7	1

Redes dirigidas

Definición

Una red es dirigida si la conexión entre i y j **no** es equivalente a la conexión entre j e i. **¡La direccionalidad importa!**

$$M_{ij} \neq M_{ji}$$

Ejemplos:

- Un usuario de twitter conecta al otro si lo “sigue” (following).
- Red de citas en papers científicos. (Dos artículos se conectan si uno cita al otro)
- Red de regulación de genes. Un gen (FT) se conecta a otro si lo regula.

Mixtas

Definición

Una red es mixta si tiene conexiones dirigidas y no dirigidas.

$$M_{ij} \neq M_{ji}$$

,

$$M_{kl} = M_{lk}$$

Ejemplos:

- Proteínas: interacción física, activación, fosforilación

Redes ponderadas

Definición

Una red ponderada o pesada asigna una medida de la certeza o fortaleza de la conectividad en valores continuos.

$$M_{ij} \in \mathcal{R}$$

$$M_{ij} = 1, 0, -1, 0.75, 5, 3.67, \ldots$$

Las redes ponderadas pueden ser mixtas, dirigidas o no dirigidas. Ejemplos:

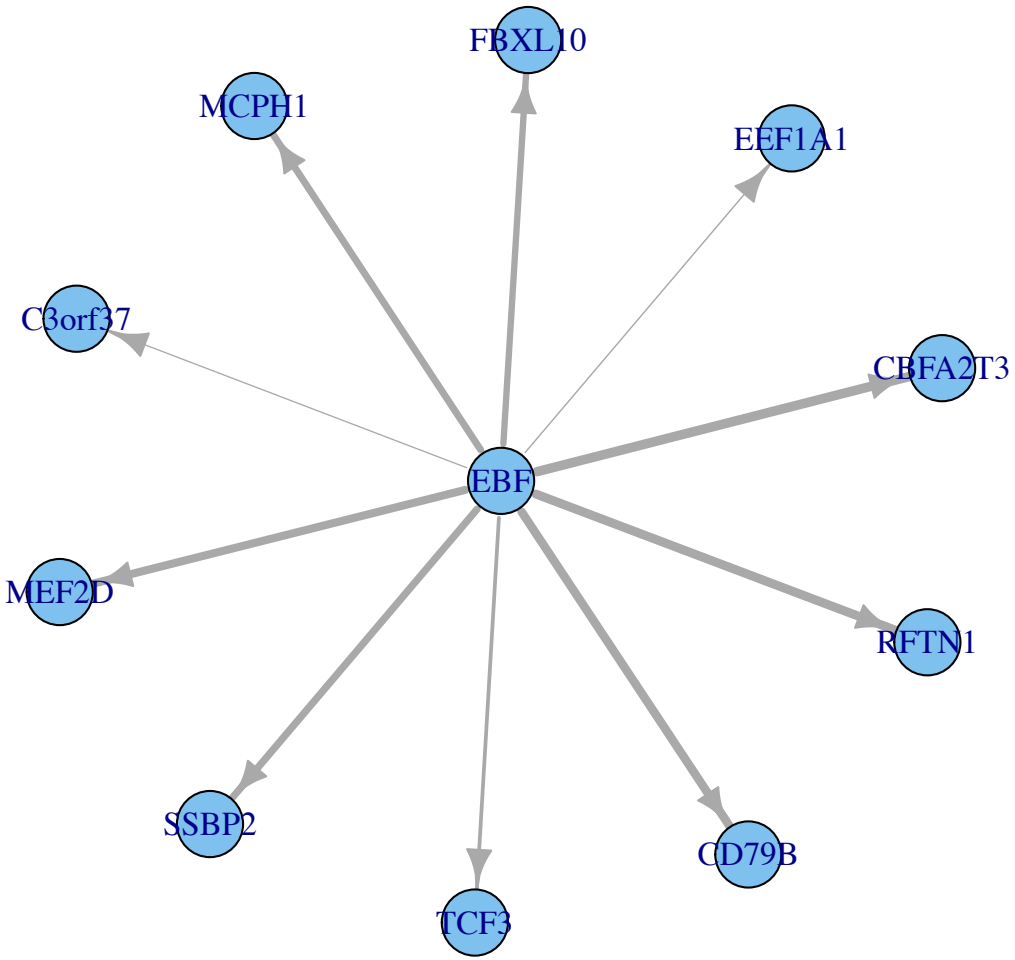
- Red de correlación de genes (no dicotómica)
- Red de regulación transcripcional.

Red ponderada dirigida

Red de regulación del factor de transcripción EBF

Rank	GeneNames	EntrezID	ChIPXpressScore
1	CBFA2T3	863	22.7
2	RFTN1	23180	45.7
3	CD79B	974	56.5
4	MEF2D	4209	71.1
5	SSBP2	23635	87.5
6	MCPH1	79648	92
7	FBXL10	84678	99.5
8	TCF3	6929	114.9
9	EEF1A1	1915	119
10	C3orf37	56941	119.1

Red de regulación transcripcional de EBF

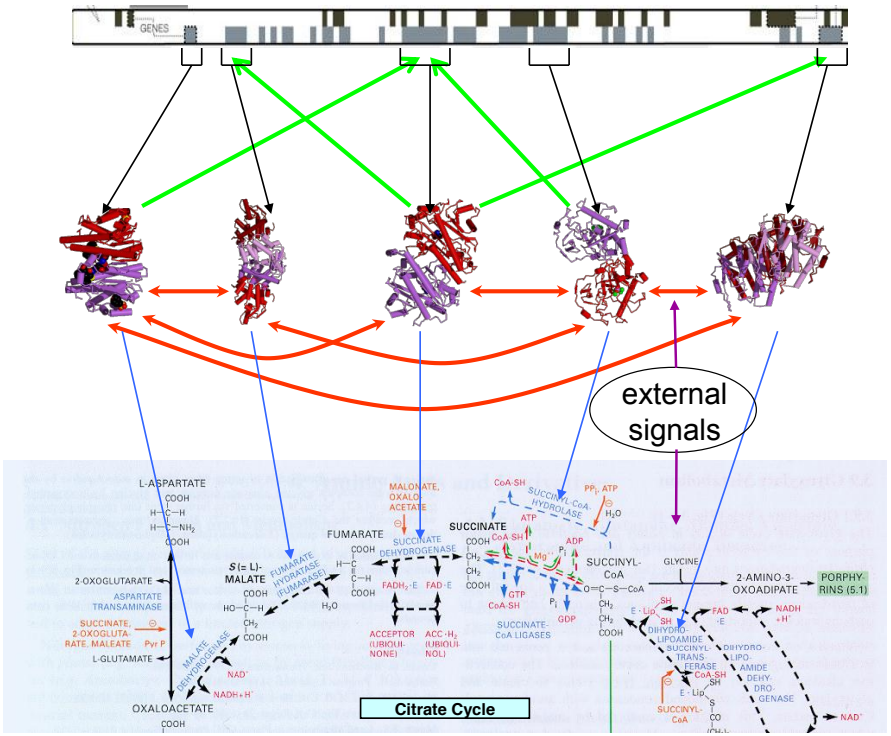


Redes bipartitas

Definición

Una red es bipartita si está compuesta de dos grupos de nodos ajenos entre si. Ejemplos:

- Red de regulación postranscripcional (microRNAs y genes)



GENOME

gene regulation

PROTEOME

protein-protein interactions

signal transduction

METABOLISM

Bio-chemical reactions

Figura: Figura tomada de R. Albert

Ejercicios

1. Se tiene n nodos, ¿cuántas conexiones son necesarias para construir una red completamente conectada ?
2. Construcción de una red desconectada de n nodos. Por cada conexión que agregues tu recibes 1. ¿Cuál es la mayor cantidad de dinero que puedes recibir ?
3. Dado un conjunto de n nodos, ¿cuántas conexiones son necesarias para tener una red completamente conectada?

Distribución de conectividades

Propiedades estáticas

- Distribución de conectividades
- Mundo pequeño (*"small-world"*)
- Coeficiente de agrupamiento

Distribución de conectividades

Definición

$P(k)$, es la probabilidad de que un nodo tenga k conectividades

$$P(k) \approx \frac{k_i}{\sum k_i}$$

La probabilidad de que un nodo tomado al azar tenga k conectividades.

$$P(k) = \frac{Numero\ nodos\ con\ k\ vecinos}{Numero\ nodos}$$

Distribución de conectividades

Definición

$P(k)$, es la probabilidad de que un nodo tenga k conectividades

$$P(k) \approx \frac{k_i}{\sum k_i}$$

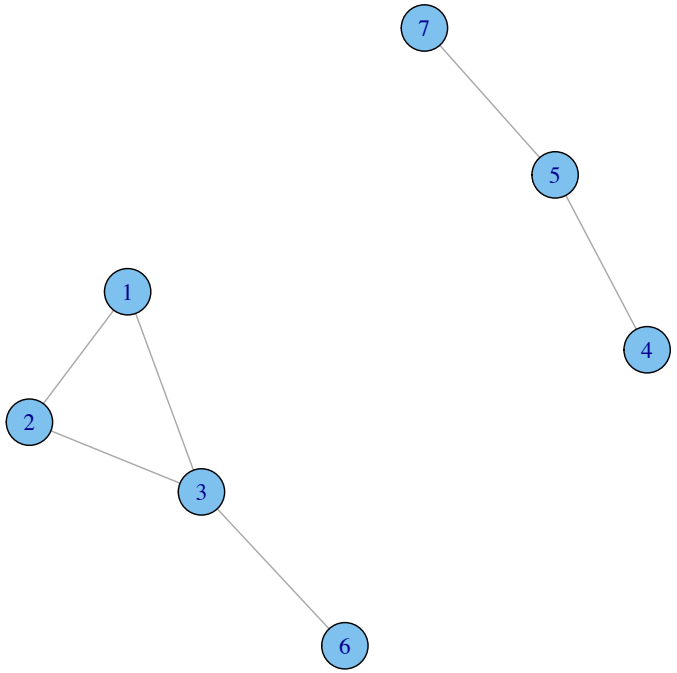
La probabilidad de que un nodo tomado al azar tenga k conectividades.

$$P(k) = \frac{Numero\ nodos\ con\ k\ vecinos}{Numero\ nodos}$$

Distribución de conectividades

Degree distribution

$P(k)$, es la probabilidad de que un nodo al azar tenga k conectividades

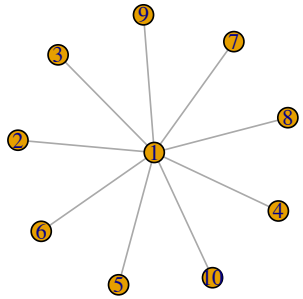


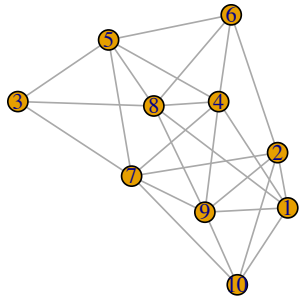
$$(2, 2, 3, 1, 2, 1, 1)$$

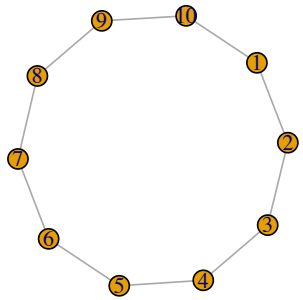
$$P(1) = \frac{3}{7}, P(2) = \frac{3}{7}, \ P(3) = \frac{1}{7}$$

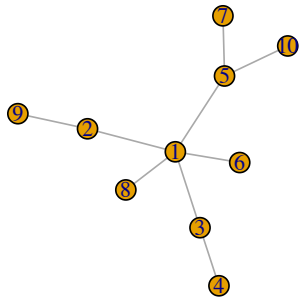
Ejercicio

Calcula la distribución de conectividades de cada una de las siguientes redes







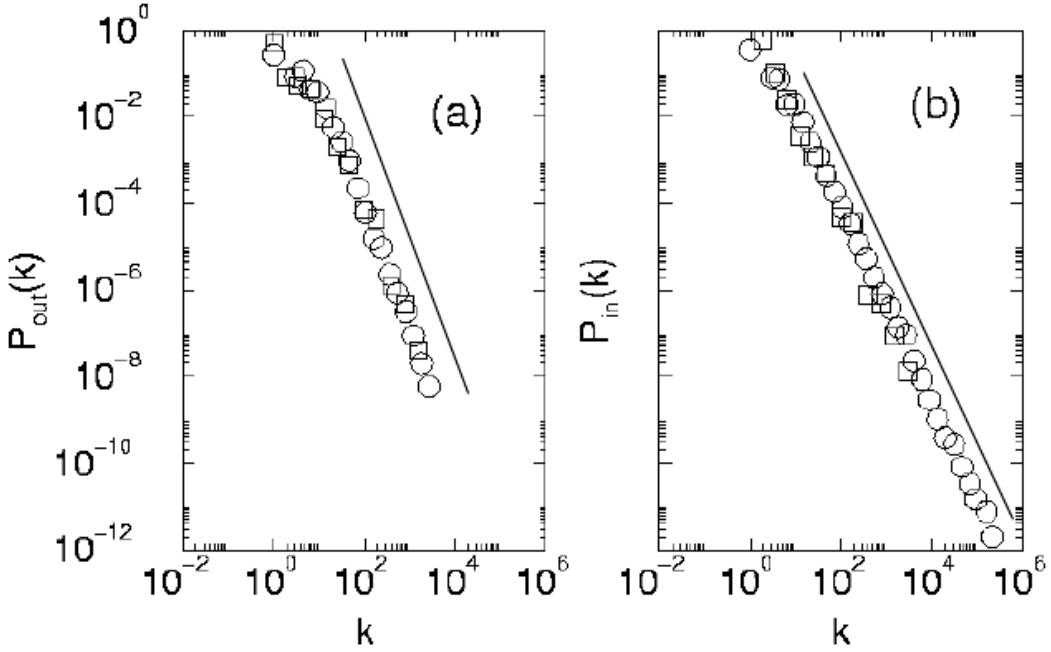


1. Para una red aleatroria, calcula el diámetro para $n = 10, 100, 200, 300, 500, 1000$
2. Calcula el coeficiente de agrupamiento
3. Observa, cómo es la distribución de conectividades
4. Elimina $r = 1, 2, 3 \ldots 10$ nodos al azar y repite los cálculos (Sugerencia: usa la funciń delete.vertices)

Datos reales

- Stanford Large Dataset Collection (http://snap.stanford.edu/data/)
- CCNR (http://www3.nd.edu/ networks/resources.htm)

In- and out-degree distribution of the WWW



nodes: webpages

edges: hyperlinks

$$P_{out}(k) \approx k^{-2.45}$$

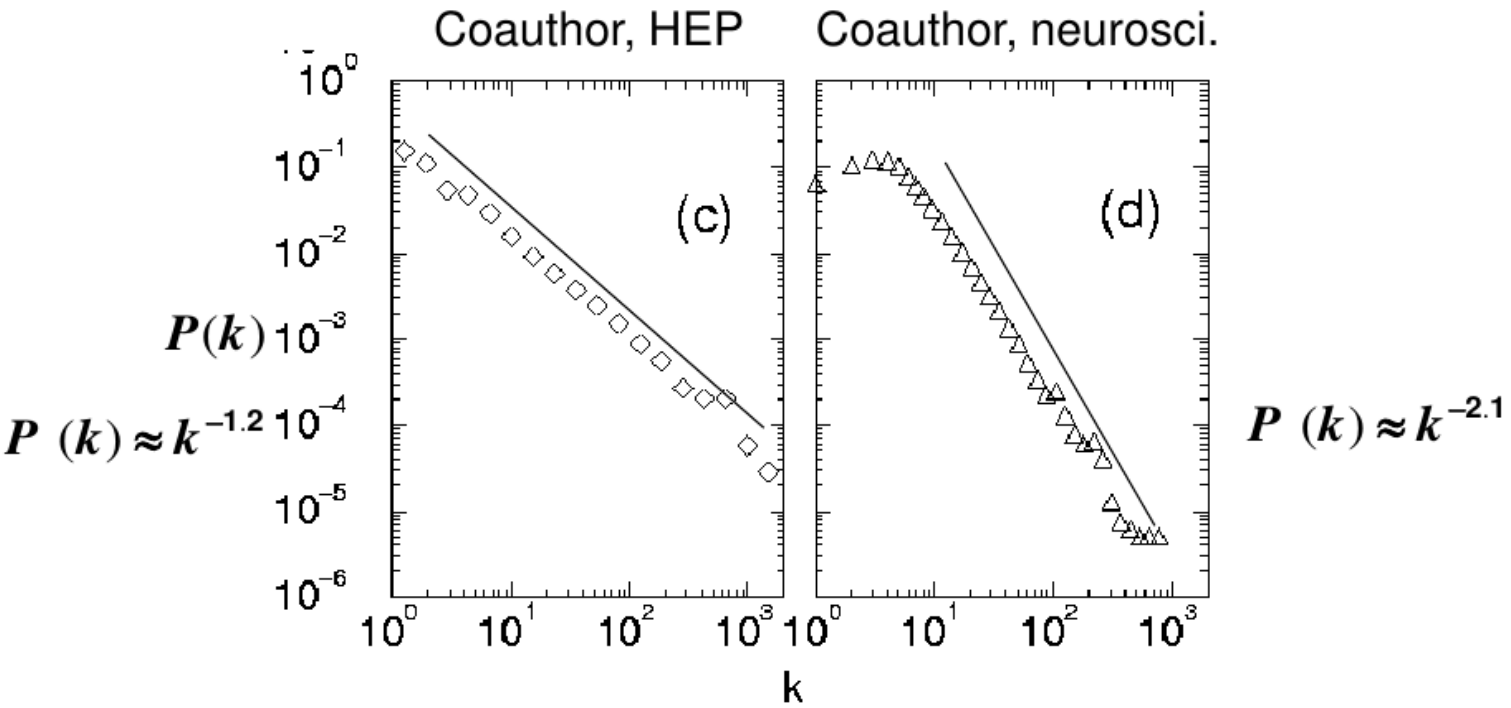
$$P_{in}(k) \approx k^{-2.1}$$

Usage: the degree distribution scales as a power law

R. Albert, H. Jeong, A.-L. Barabási, Nature 401, 130 (1999)

A. Broder *et al.*, Comput. Netw. 33, 309 (1999)

Degree distributions in networks of science collaborations



M. E. J. Newman, Phys. Rev. E 64, 016131 (2001)

A.-L. Barabási et al., cond-mat/0104162 (2001)

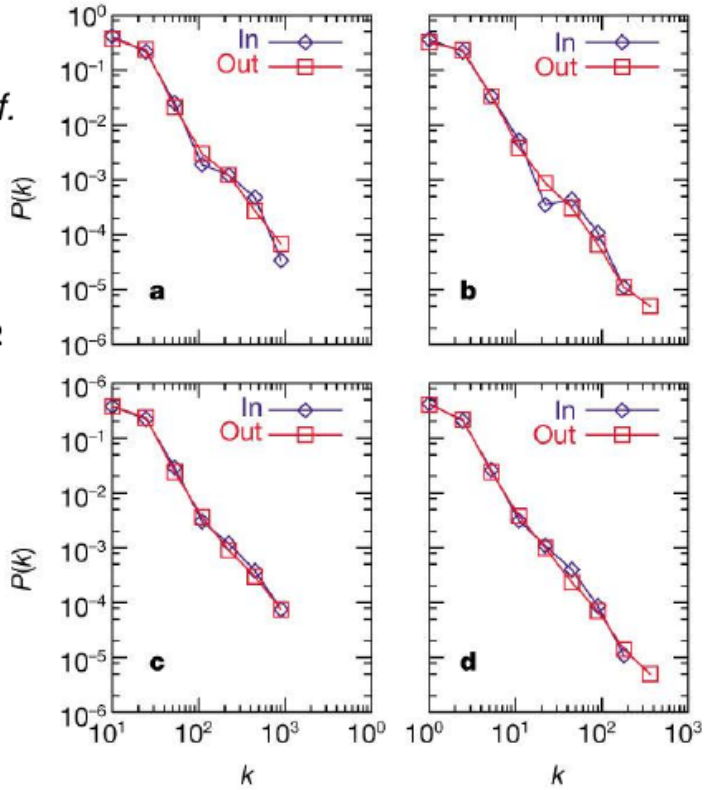
Metabolic networks have a power-law degree distribution

Archaeoglobus f.

$$P_{in}(k) \approx k^{-2.2}$$

$$P_{out}(k) \approx k^{-2.2}$$

C. elegans



E. coli

bipartite

nodes: metabolites, reactions

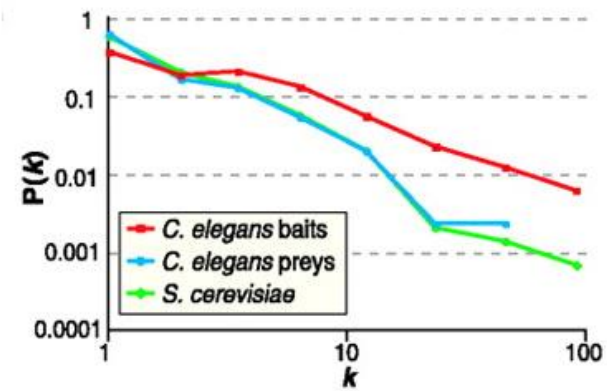
directed edges,

out: reactant (substrate)

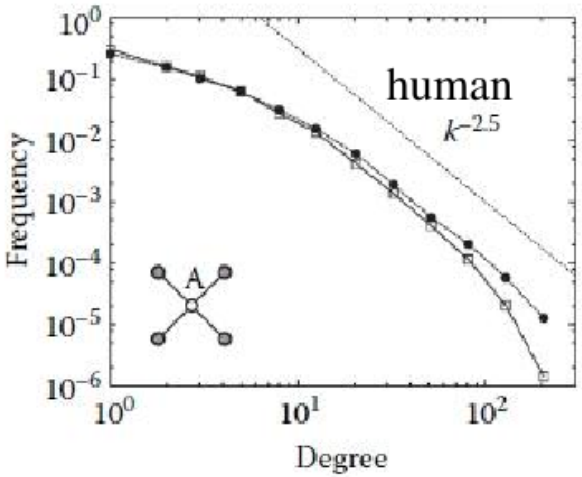
in: product of reaction

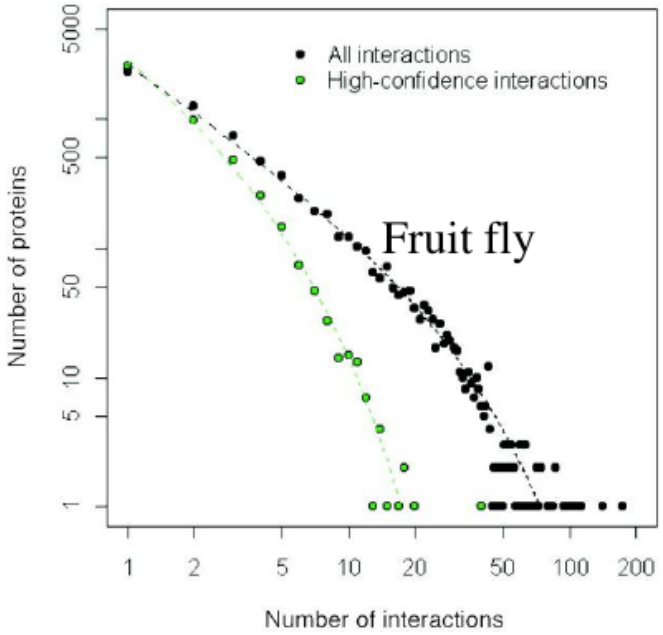
H. Jeong et al., Nature 407, 651 (2000)

Degree distribution of protein networks



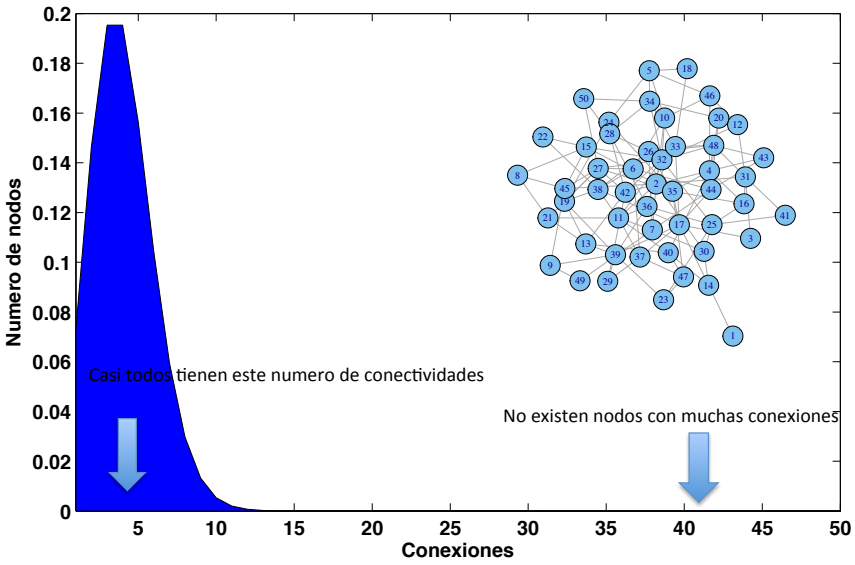
$$P(k) \approx Ak^{-\gamma}$$





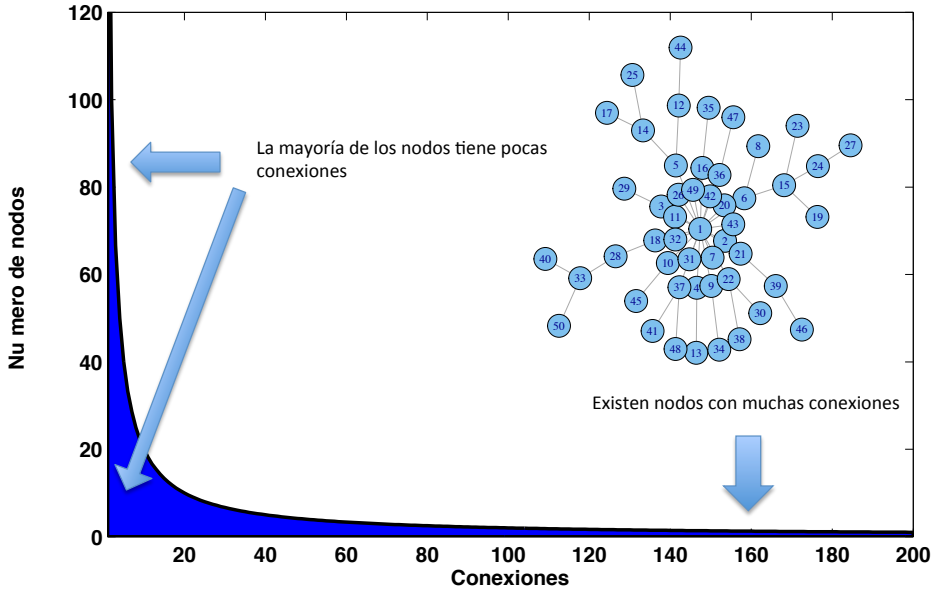
$$P(k) \approx Ak^{-\gamma} \exp(-\beta k)$$

Giot et al. Science 2003 – *Drosophila m.*
Li et al. Science 2004 – *C. elegans*
Rual et al. Nature 2005 – human
Stelzl et al. Cell 2005 - human



Aleatorias.

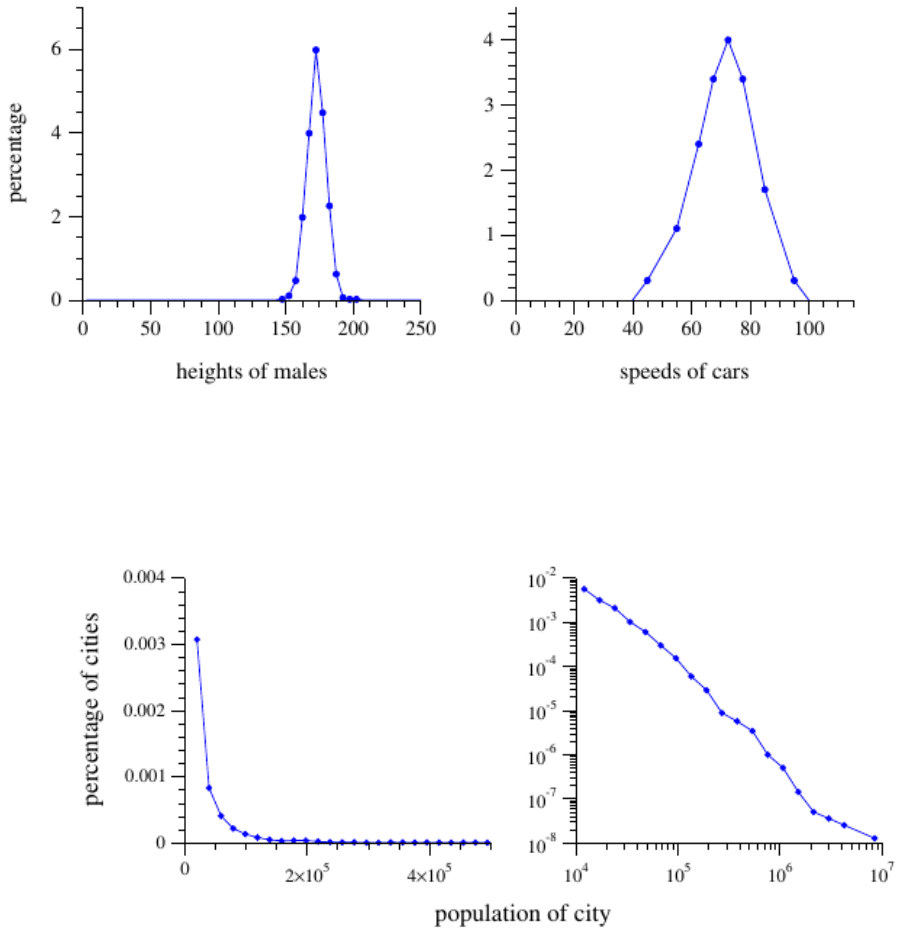
Re



Complejas

Leyes de potencia

Sin escala característica (*free scale*)

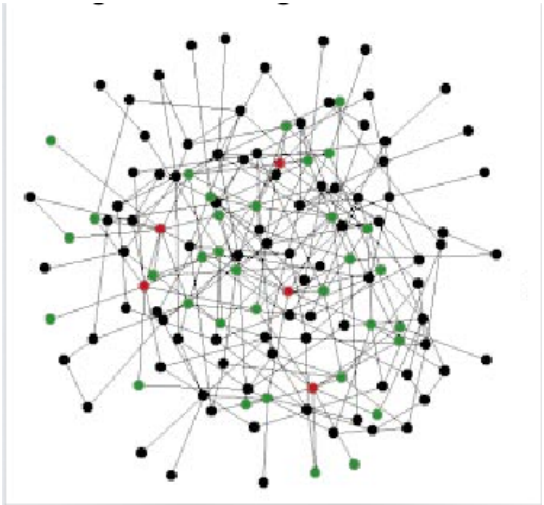


Power laws, Pareto distributions and Zipf's law, M.E.J. Newman, 2008

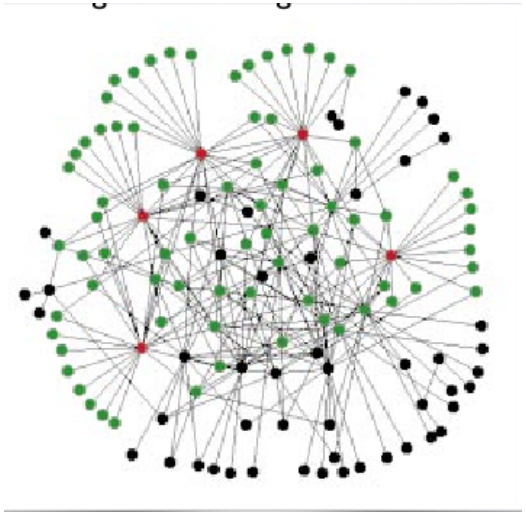
Comparativo con redes aleatorias

Error and attack tolerance of complex networks 2000

Réka Albert, Hawoong Jeong & Albert-László Barabási

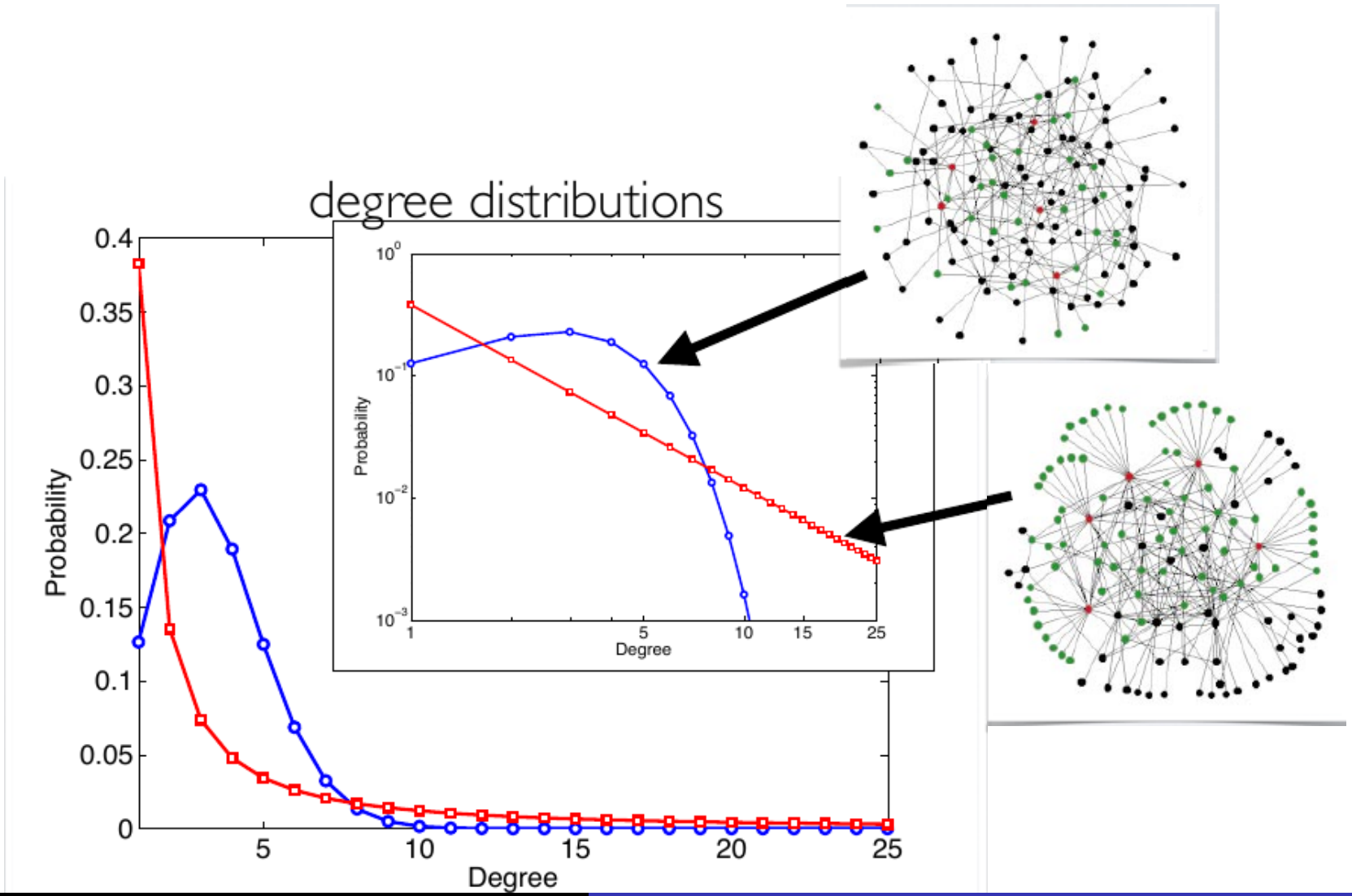


Red aleatoria (homogénea)



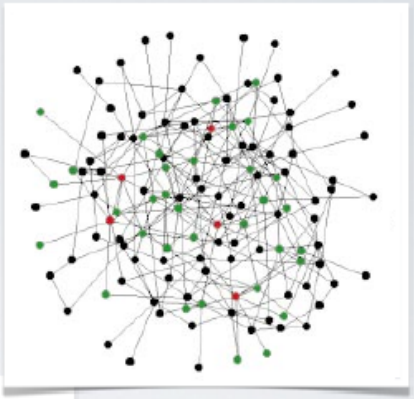
Red heterogenea

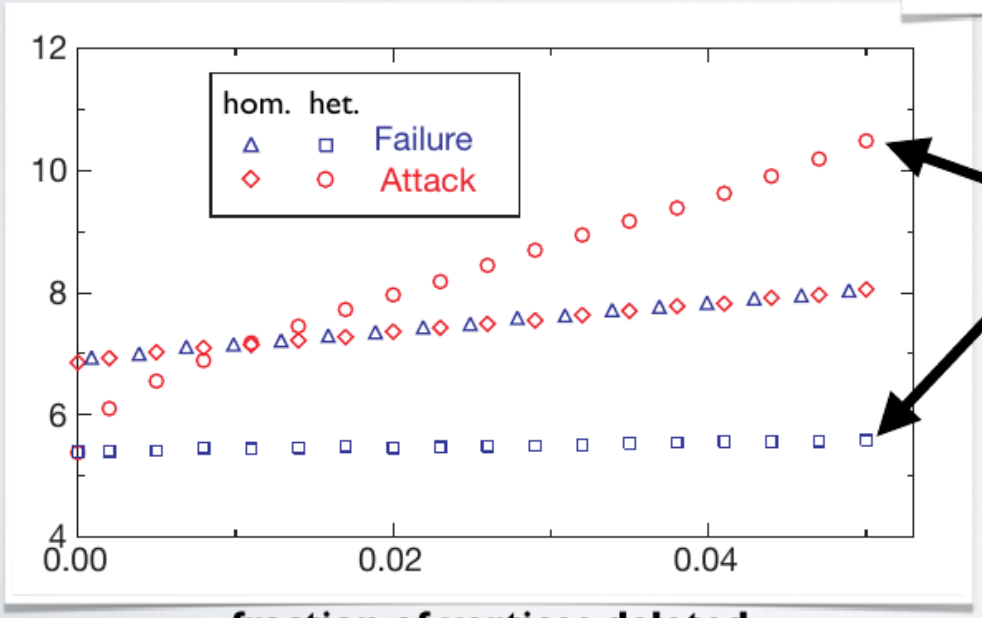
Comparativo con redes aleatorias

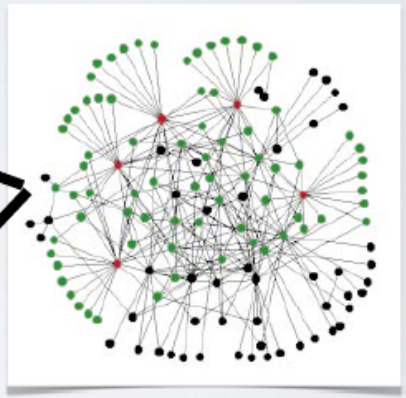


strategy: delete vertices

1. uniformly at random ("failure")
2. in order of degree ("attack")







Aaron Clauset 2013

1. Para una red aleatoria (np, fijo), calcula el diámetro para $n = 10, 100, 200, 300, 500, 1000, 10000$
2. Calcula el coeficiente de agrupamiento.
3. Observa cómo es la distribución de conectividades.
4. A partir de la red más grande: elige $r = 0.01, 0.02, 0.03 \ldots 0.1$ fracciones de nodos al azar y repite los cálculos (Sugerencia: usa la función delete.vertices)

brief communications

Lethality and centrality in protein networks

The most highly connected proteins in the cell are the most important for its survival.

Proteins are traditionally identified on the basis of their individual actions as catalysts, signalling molecules, or building blocks in cells and microorganisms. But our post-genomic view is expanding the protein's role into an element in a network of protein–protein interactions as well, in which it has a contextual or cellular function within functional modules[1,2]. Here we provide quantitative support for this idea by demonstrating that the phenotypic consequence of a single gene deletion in the yeast Saccharomyces cerevisiae is affected to a large extent by the topological position of its protein product in the complex hierarchical web of molecular interactions.

The S. cerevisiae protein–protein interaction network we investigate has 1,870 proteins as nodes, connected by 2,240 identified direct physical interactions, and is derived from combined, non-overlapping data[3,4], obtained mostly by systematic two-hybrid analyses[3]. Owing to its size, a complete map of the network (Fig. 1a), although informative, in itself offers little insight into its large-scale characteristics. Our first goal was therefore to identify the architecture of this network, determining whether it is best described by an inherently uniform exponential topology, with proteins on average possessing the same number of links, or by a highly heterogeneous scale-free topology, in which proteins have widely different connectivities[5].

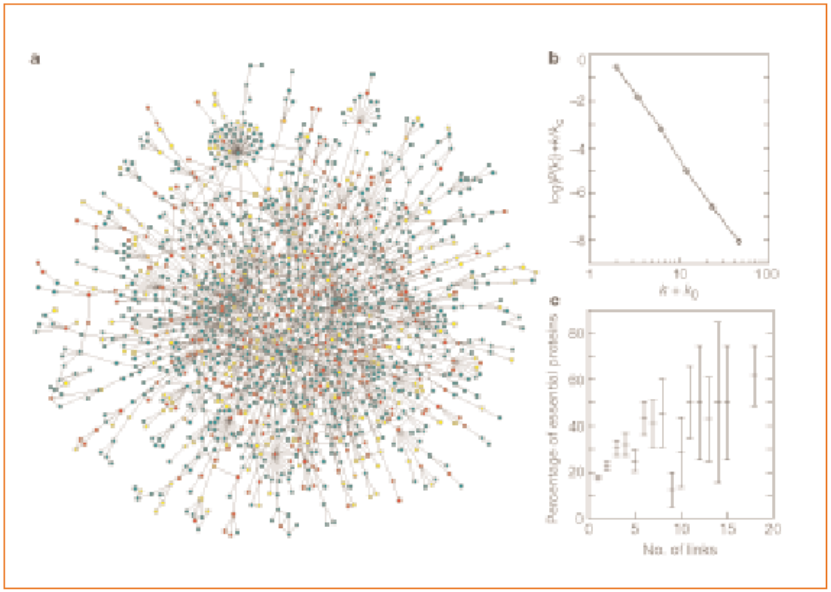
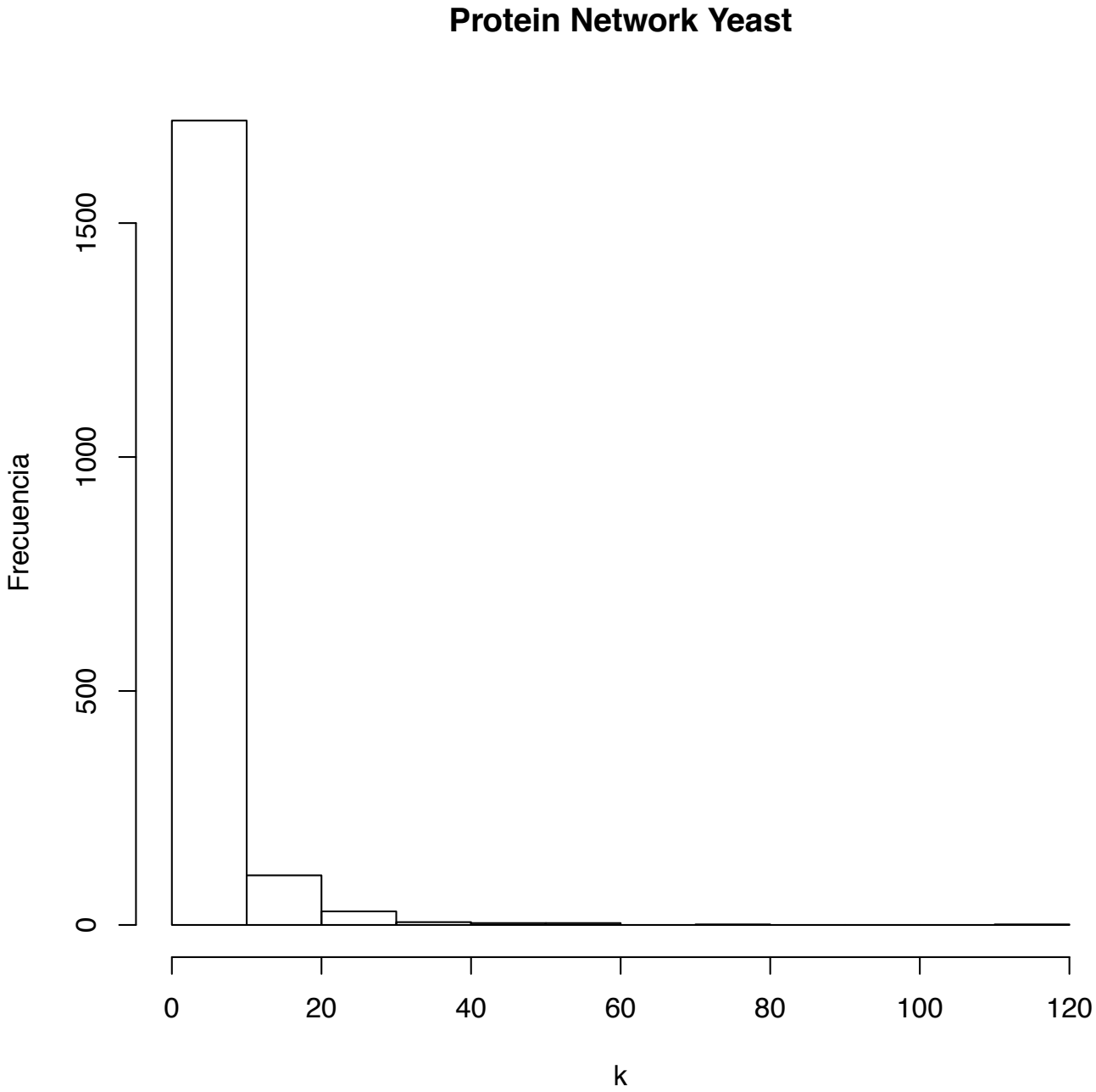
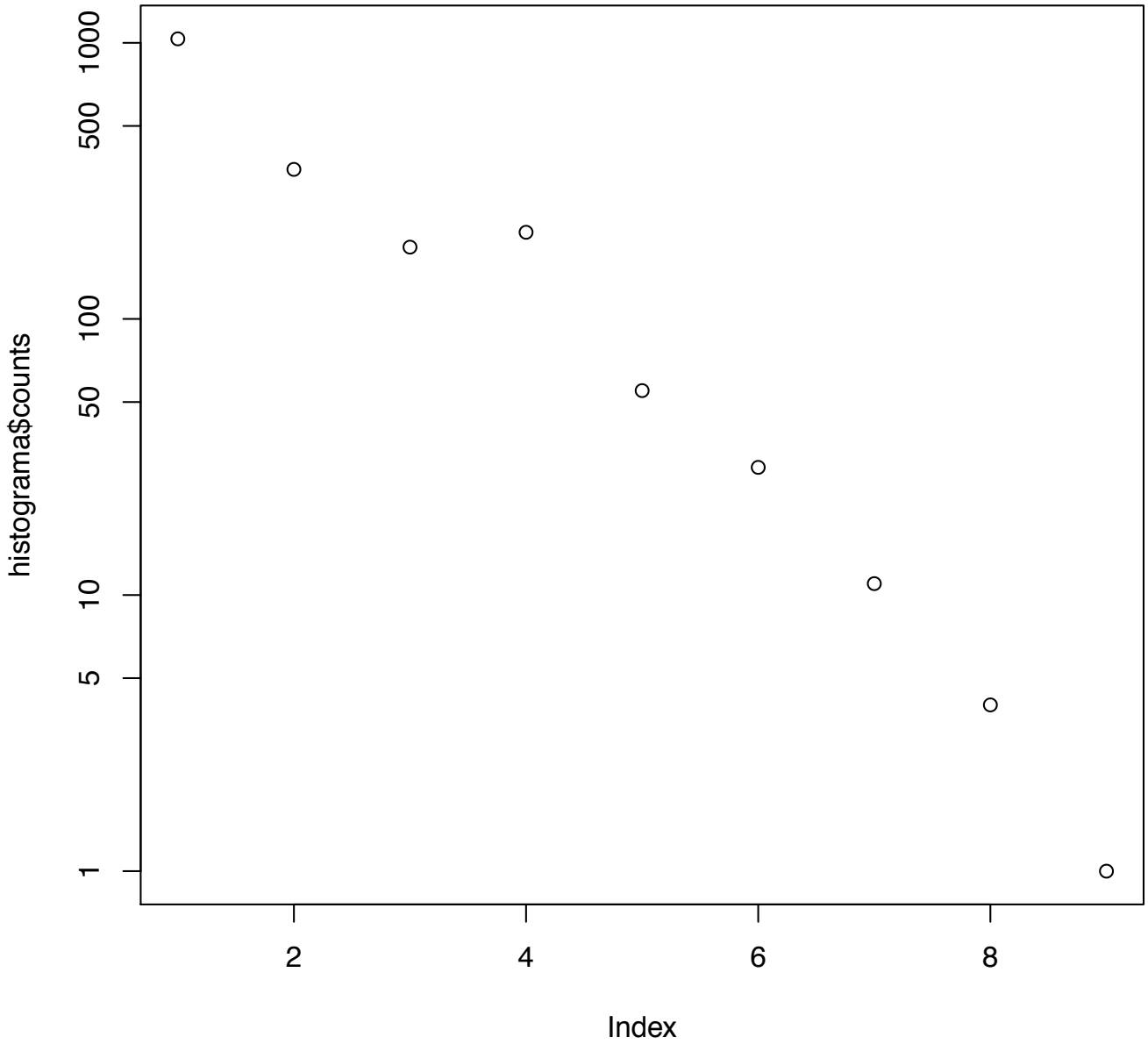


Figure 1 Characteristics of the yeast proteome. **a,** Map of protein–protein interactions. The largest cluster, which contains ~78% of all proteins, is shown. The colour of a node signifies the phenotypic effect of removing the corresponding protein (red, lethal; green, non-lethal; orange, slow growth; yellow, unknown). **b,** Connectivity distribution P(k) of interacting yeast proteins, giving the probability that a given protein interacts with k other proteins. The exponential cut-off[6] indicates that the number of proteins with more than 20 interactions is slightly less than expected for pure scale-free networks. In the absence of data on the link directions, all interactions have been considered as bidirectional. The parameter controlling the short-length scale correction has value $k_0 \approx 1$. **c,** The fraction of essential proteins with exactly k links versus their connectivity, k, in the yeast proteome. The list of 1,572 mutants with known phenotypic profile was obtained from the Proteome database[13]. Detailed statistical analysis, including $r = 0.75$ for Pearson's linear correlation coefficient, demonstrates a positive correlation between lethality and connectivity. For additional details, see http://www.nd.edu/~networks/cell.

experiments which identified a striking ... The simultaneous emergence of an


Protein Network Yeast
Frecuencia
0
500
1000
1500
0
20
40
60
80
100
120
k

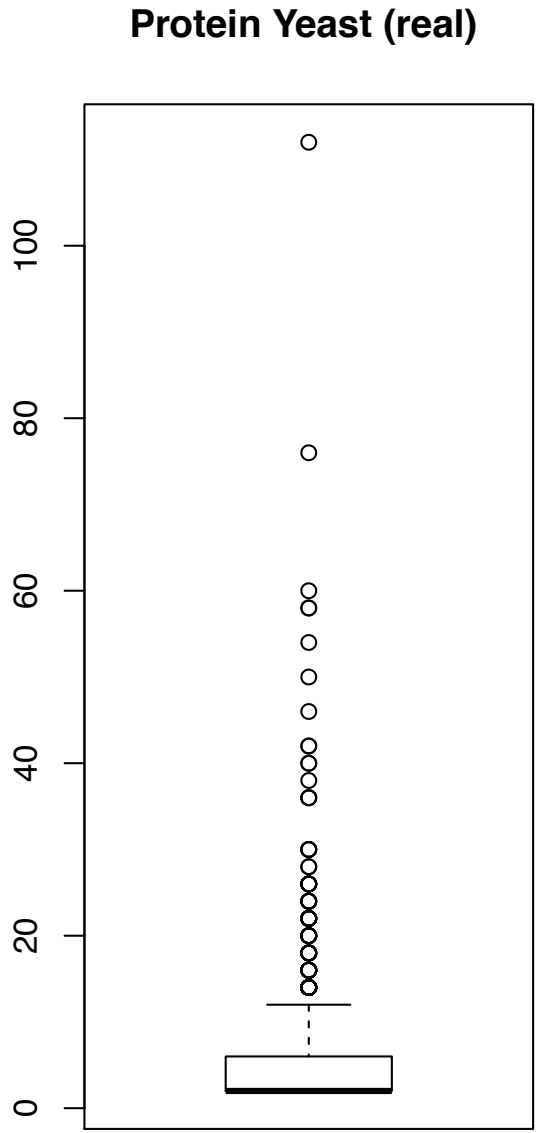

histograma$counts
1
5
10
50
100
500
1000
2
4
6
8
Index

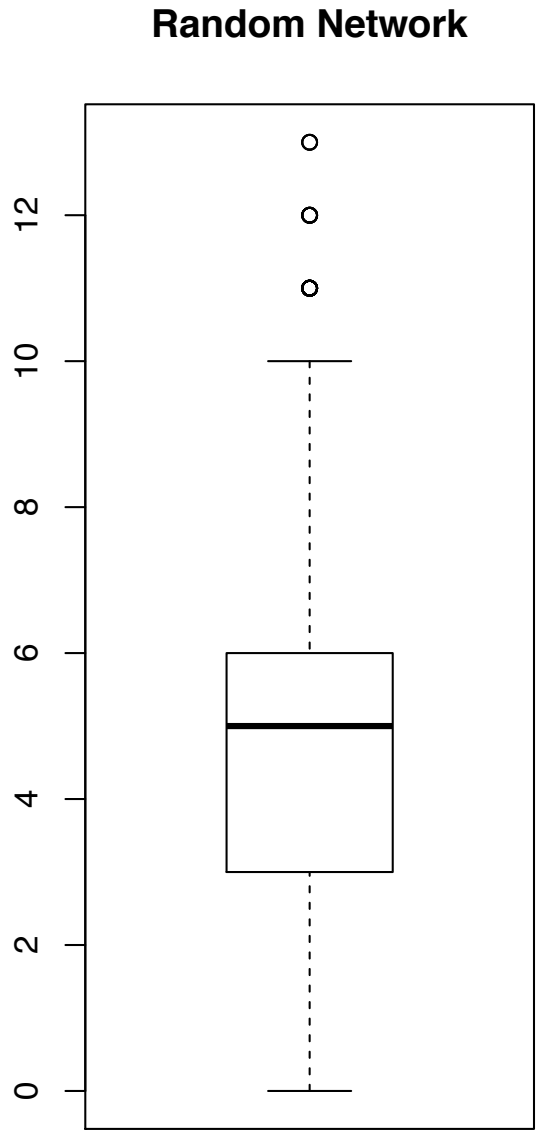
Ley de potencia

$$P(k) = Bk^{-\alpha} = \frac{B}{k^{\alpha}}$$

$$\log P(k) = -\alpha \log(k) + \log B$$

$$Y = -\alpha X + \mathcal{B}$$


Protein Yeast (real)
0
20
40
60
80
100

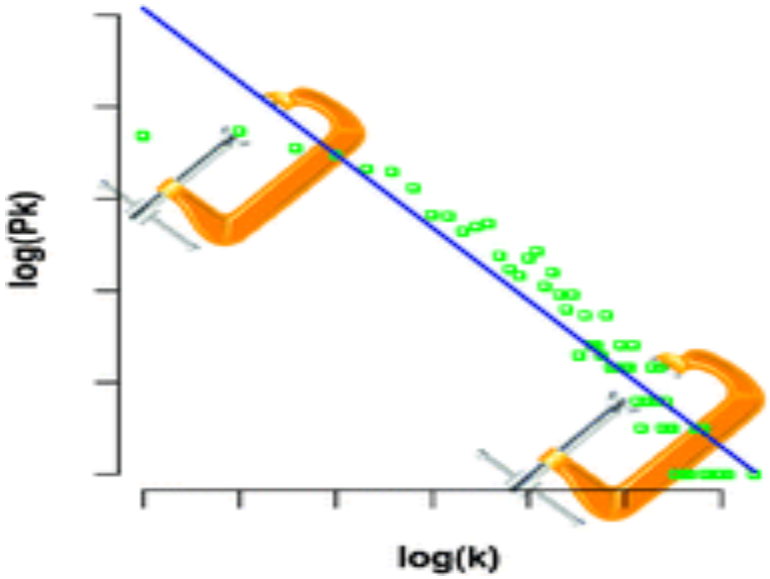

Random Network
0
2
4
6
8
10
12

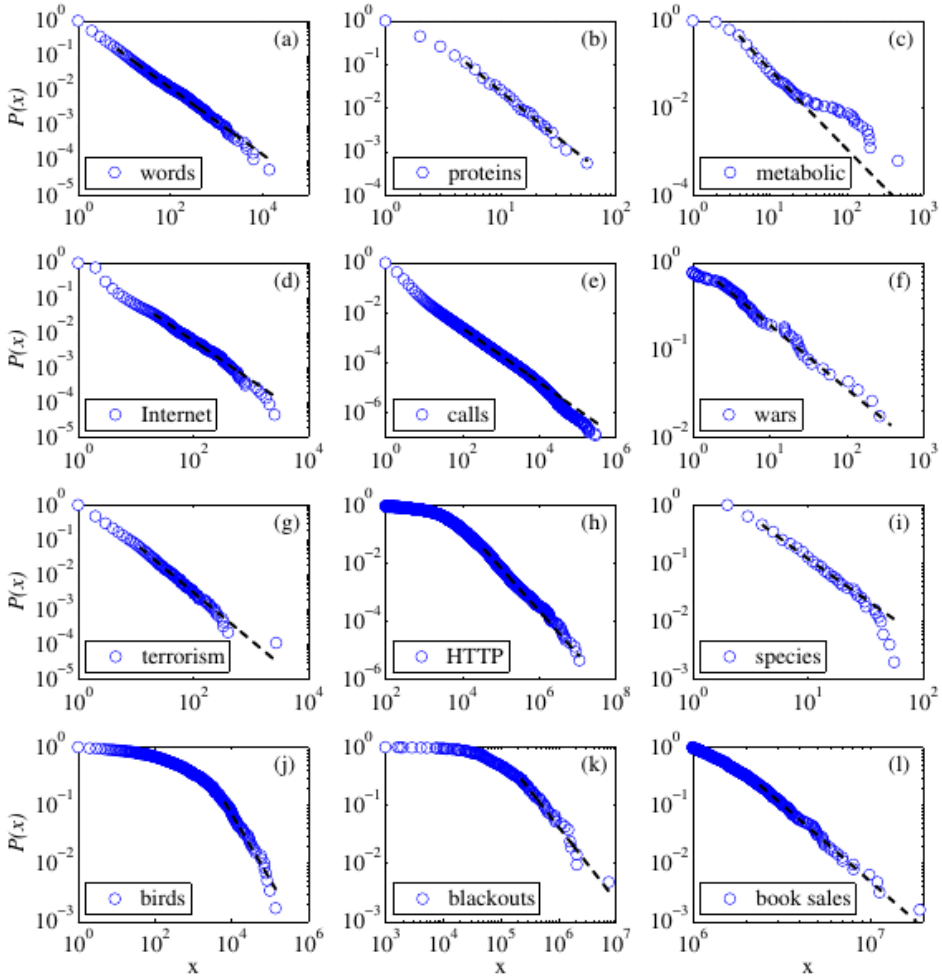
REVIEW www.rsc.org/molecularbiosystems | Molecular BioSystems

The powerful law of the power law and other myths in network biology†

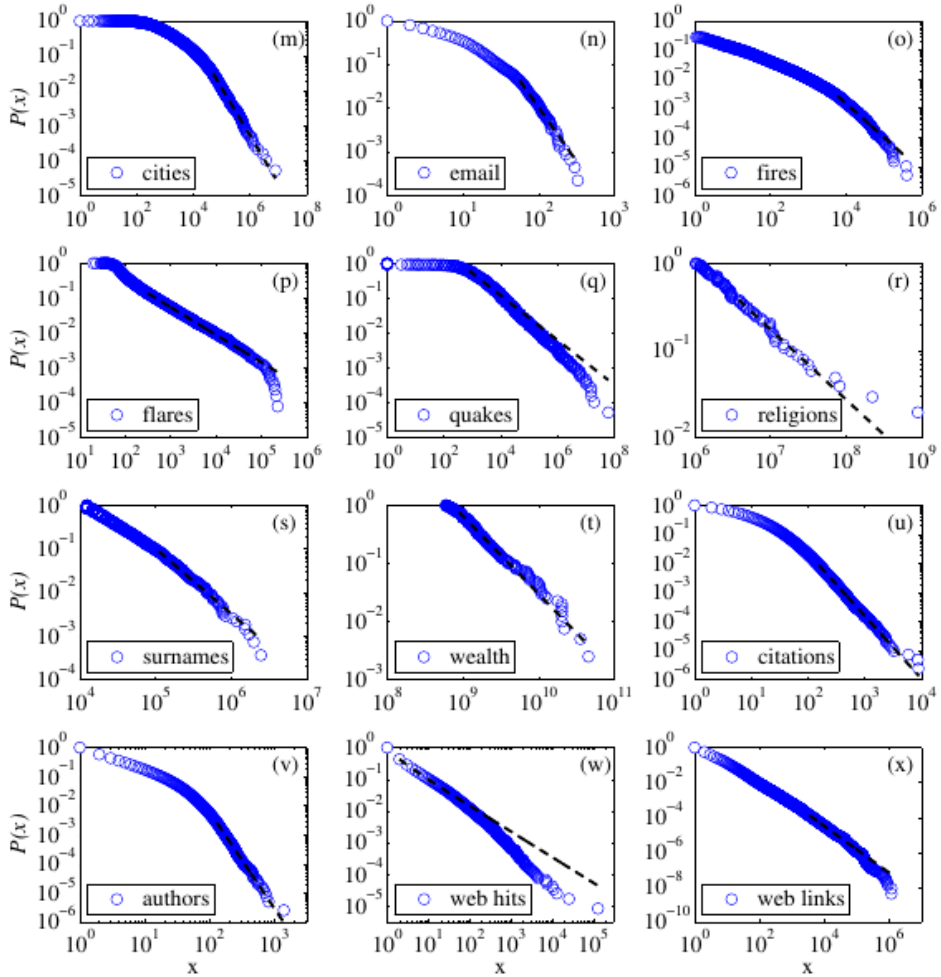
Gipsi Lima-Mendez* and Jacques van Helden*

Received 5th May 2009, Accepted 12th August 2009
First published as an Advance Article on the web 2nd October 2009
DOI: 10.1039/b908681a





Power laws distribution in empirical data Clauset , Shalizi ,M.E.J. Newman, 2008



Power laws distribution in empirical data A. Clauset , C. Shalizi ,M.E.J. Newman, 2008

Colas largas

- Pareto $x^{-\alpha-1}$
- Beta Prima $\frac{x^{p-1}(1+x)^{-(p+q)}}{B(p,q)}$
- Dagum $ab^{-ap}px^{ap-1}(1+(\frac{x}{b})^a)^{-(1+p)}$
- Davis
- (Correción tamaño finito) Exponencial con corte $x^{-\alpha}e^{-\beta x}$
- FBC $\frac{(1-x)^b}{x^a}$

Comparativo robustez

Ejercicio

Generar una red sin escala con $1 \leq \alpha \leq 3$ (barabasi.game(n, alpha)) con $n \geq 1000$. Remover $0.01, 0.02, \ldots 0.1$ fracción de los nodos y calcular las distancias promedio (average.path.length)

Small-world

6 grados

Red social
Red de conocidos y amigos
Stanley miligram

Kevin Bacon

Small-world

Google bacon number el santo

Web Images More Search tools

About 54,000,000 results (0.75 seconds)

Santo's Bacon number is 3

Santo and Armando Silvestre appeared in Santo Contra los Zombis.

Armando Silvestre and Glenn Ford appeared in Rage.

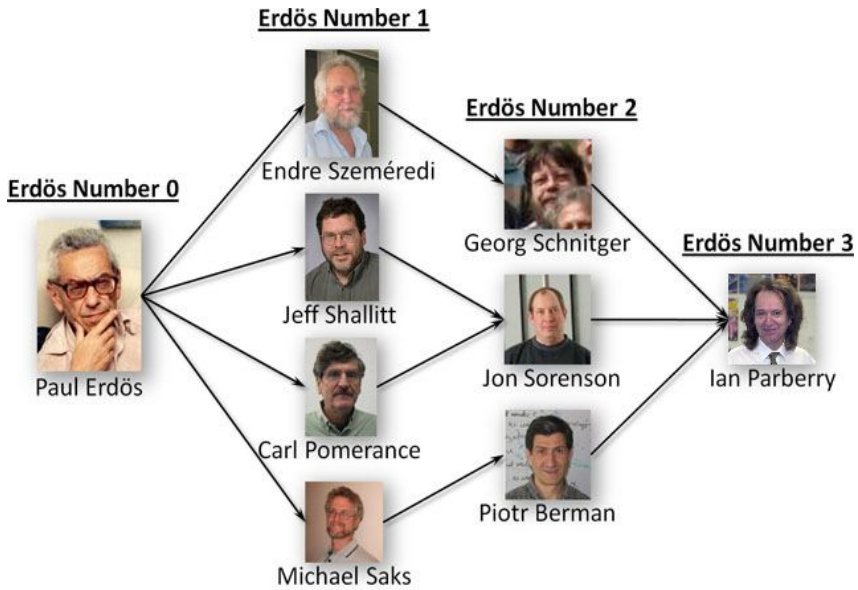
Glenn Ford and Kevin Bacon appeared in The Gift.

Kevin Bacon

http://www.youtube.com/watch?v=n9u-TITxwoM

Red de colaboraciones científicas

Número de Erdös



Número de Erdos + Bacon

4 + 2

5 + 2

Ultra small-world

(Ultra-) Small-world

- Cohen and Havlin ,PRL , 2003 (Scale-free networks)

$$L \approx \log(\log N)$$

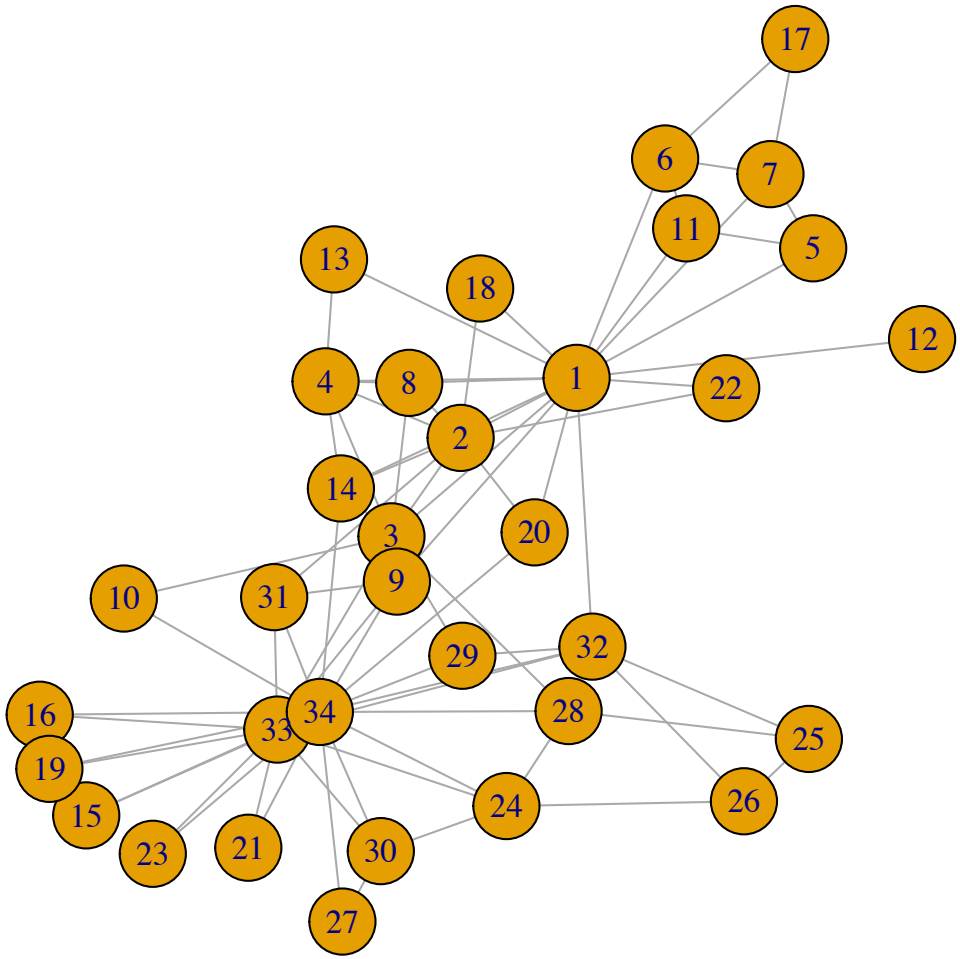
Examples of Scale-Free Networks

NETWORK	NODES	LINKS
Cellular metabolism	Molecules involved in burning food for energy	Participation in the same biochemical reaction
Hollywood	Actors	Appearance in the same movie
Internet	Routers	Optical and other physical connections
Protein regulatory network	Proteins that help to regulate a cell's activities	Interactions among proteins
Research collaborations	Scientists	Co-authorship of papers
Sexual relationships	People	Sexual contact
World Wide Web	Web pages	URLs

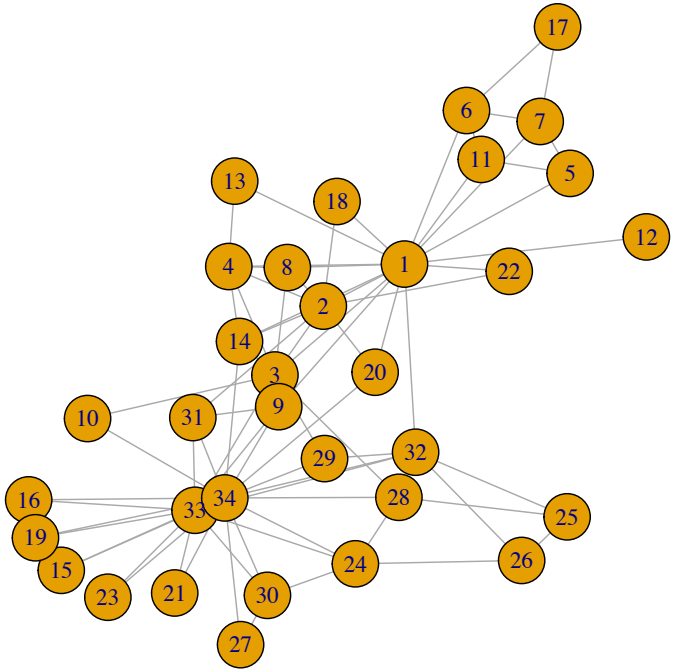
Barabasi and Boneabu, Sci. Am. 2003

Red	Tipo	n	m	C	l	α
Metabolica	u	765	3686	0.67	2.56	2.2
Proteinas	u	2115	2240	0.071	6.80	2.4
Neuronal	d	307	2359	0.28	3.97	–
Co-aut(bio)	u	1 520 251	11803064	0.60	4.92	–
Co-aut(fis)	u	52909	245300	0.56	6.19	–
Rel .Est.	u	573	477	0.001	16.01	–
Actores	u	449913	25516482	0.88	3.48	2.3
WWW	d	$2\ X10^8$	$2X10^{10}$	–	16.1	2.1
Internet	u	10697	31992	0.39	3.31	2.5
Eléctrica	u	4941	6594	0.080	18.99	–

Distancias en redes



Distancias en redes



$$d(1,2) = d(2,1) = 1$$

$$d(1,24) = d(24,1) = 3$$

Matriz de distancias

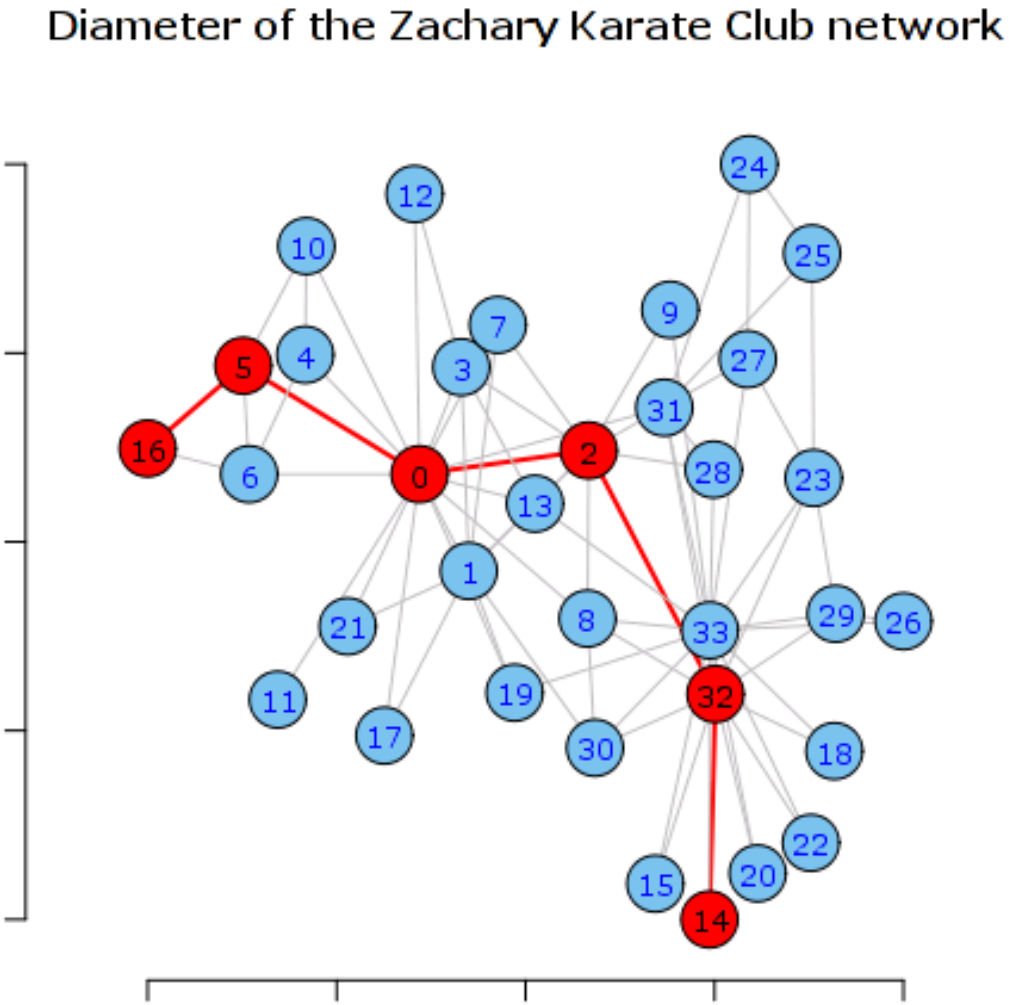
$$\mathrm{D}_{m,n} = \begin{pmatrix} d(1,1) & d(1,2) & \cdots & d(1,34) \\ d(2,1) & d(2,2) & \cdots & d(2,34) \\ \vdots & \vdots & \ddots & \vdots \\ d(34,1) & d(34,2) & \cdots & d(34,34) \end{pmatrix}$$

Matriz de distancias Zachary

$$D_{m,n} = \begin{pmatrix} 0 & 1 & \cdots & 2 \\ 1 & 0 & \cdots & 2 \\ \vdots & \vdots & \ddots & \vdots \\ 2 & 2 & \cdots & 0 \end{pmatrix}$$

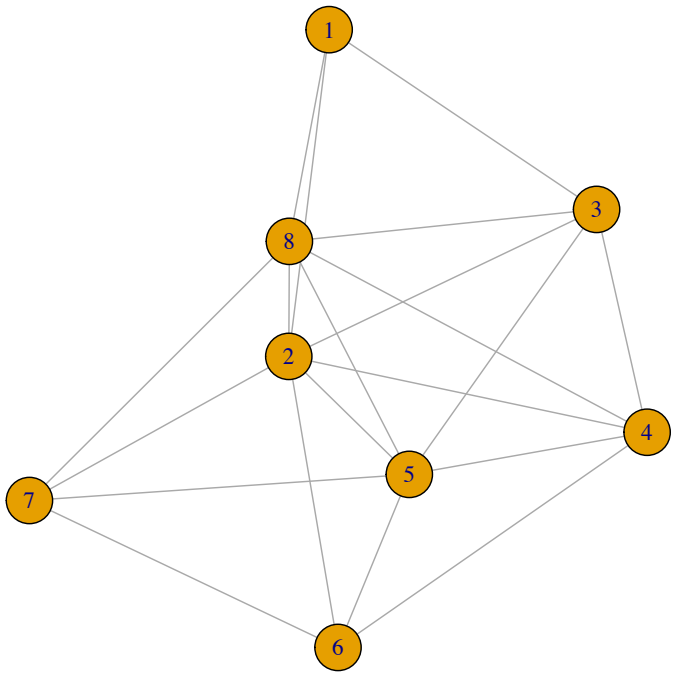
Diámetro

El diámetro de una red es la mayor distancia entre sus nodos

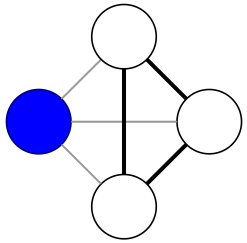


Ejercicio

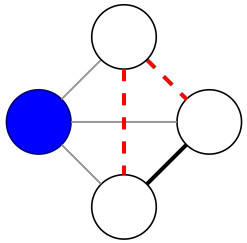
1. Calcular la distancia de cada nodo
2. Calcular la matriz de distancias
3. Encontrar el diámetro
4. ¿Cuál es el camino más largo?



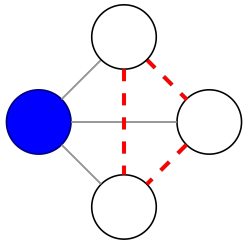
Coeficiente de clusterización (transitivity)



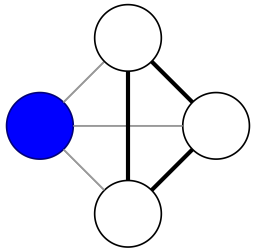
c = 1



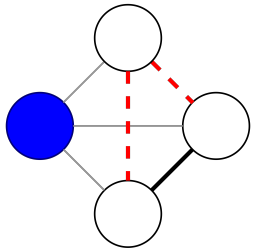
c = 1/3



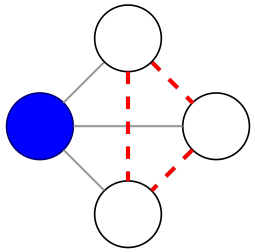
c = 0



c = 1



c = 1/3



c = 0

$$C_i = \frac{2(\#\textit{conexiones entre vecinos})}{k_i(k_i - 1)} \qquad (1)$$

redes no dirigidas

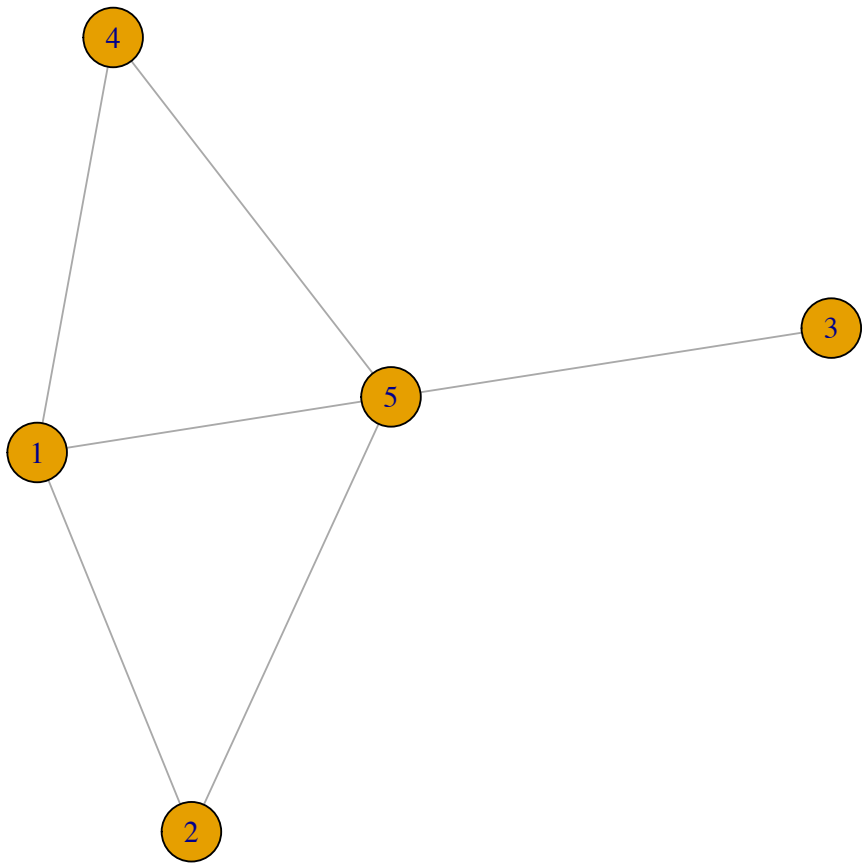
$$C_i = \frac{\#\textit{conexiones entre vecinos}}{k_i(k_i - 1)} \qquad (2)$$

redes dirigidas

$$C = \bar{C}_i = \sum_{i=1}^{n} \frac{C_i}{n} \qquad (3)$$

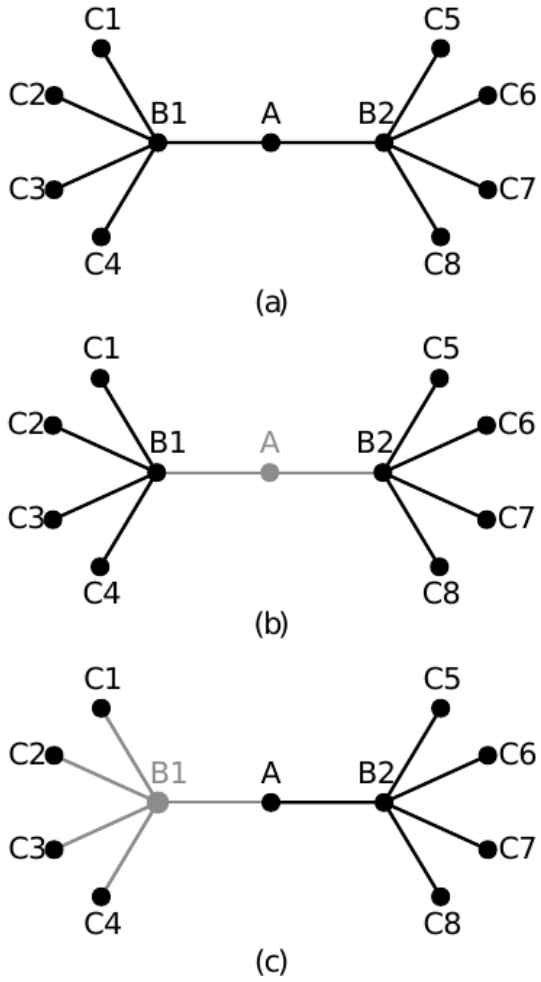
Ejemplo

Calcula el coeficiente de clusterización para cada nodo de la siguiente red

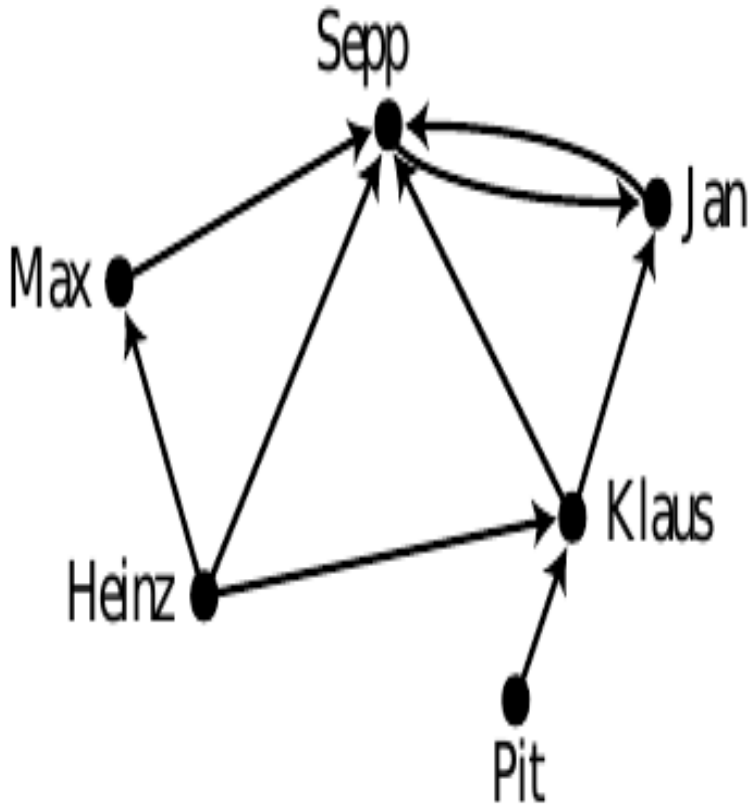


Importancia de los nodos

(Algunas) medidas de centralidad

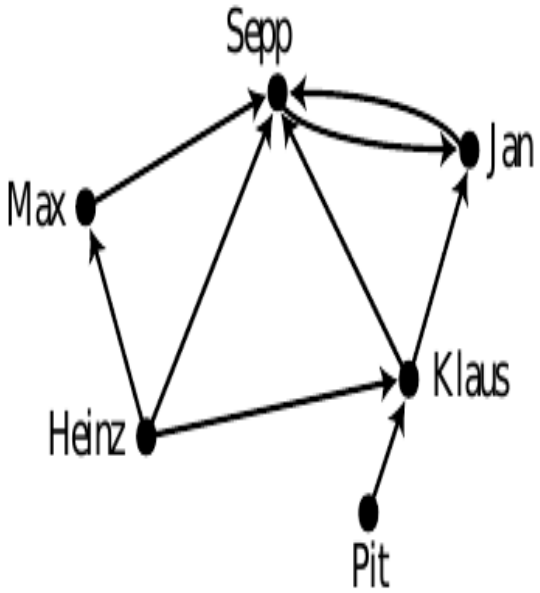


Degree centrality



Degree centrality

Una forma de destacar la importancia de un nodo es mediante el número de conexiones (*degree*).
A esta forma de medidas de centralidad se le conoce como *degree centrality*



Person	Number of votes received
Sepp	4
Jan	2
Klaus	2
Max	1
Heinz	0
Pit	0

Degree centrality

Redes biológicas

En redes de interacción de proteínas, los valores grandes de centralidad incican que se tratan de genes esenciales para para el oganismo (*knock out*). (*Saccharomyces cerevisiae*).

- Con diferentes medidas de centralidad comparadas (degree, closeness, and betweenness) se pueden identificar las proteínas esenciales, para 3 diferentes organismos:*Sacharomyces cerevisiae, Caenorhabditis elegans y Drosophila melanogaster*.
- Se han encontrado resultados similares para redes de co-expresión de *S. cerevisiae, E. coli, y C. elegans*. Los genes con un alto grado de conexión en la red es más probable que sean esenciales.

brief communications

Lethality and centrality in protein networks

The most highly connected proteins in the cell are the most important for its survival.

Proteins are traditionally identified on the basis of their individual actions as catalysts, signalling molecules, or building blocks in cells and microorganisms. But our post-genomic view is expanding the protein's role into an element in a network of protein–protein interactions as well, in which it has a contextual or cellular function within functional modules[1,2]. Here we provide quantitative support for this idea by demonstrating that the phenotypic consequence of a single gene deletion in the yeast Saccharomyces cerevisiae is affected to a large extent by the topological position of its protein product in the complex hierarchical web of molecular interactions.

The S. cerevisiae protein–protein interaction network we investigate has 1,870 proteins as nodes, connected by 2,240 identified direct physical interactions, and is derived from combined, non-overlapping data[3,4], obtained mostly by systematic two-hybrid analyses[3]. Owing to its size, a complete map of the network (Fig. 1a), although informative in itself offers little

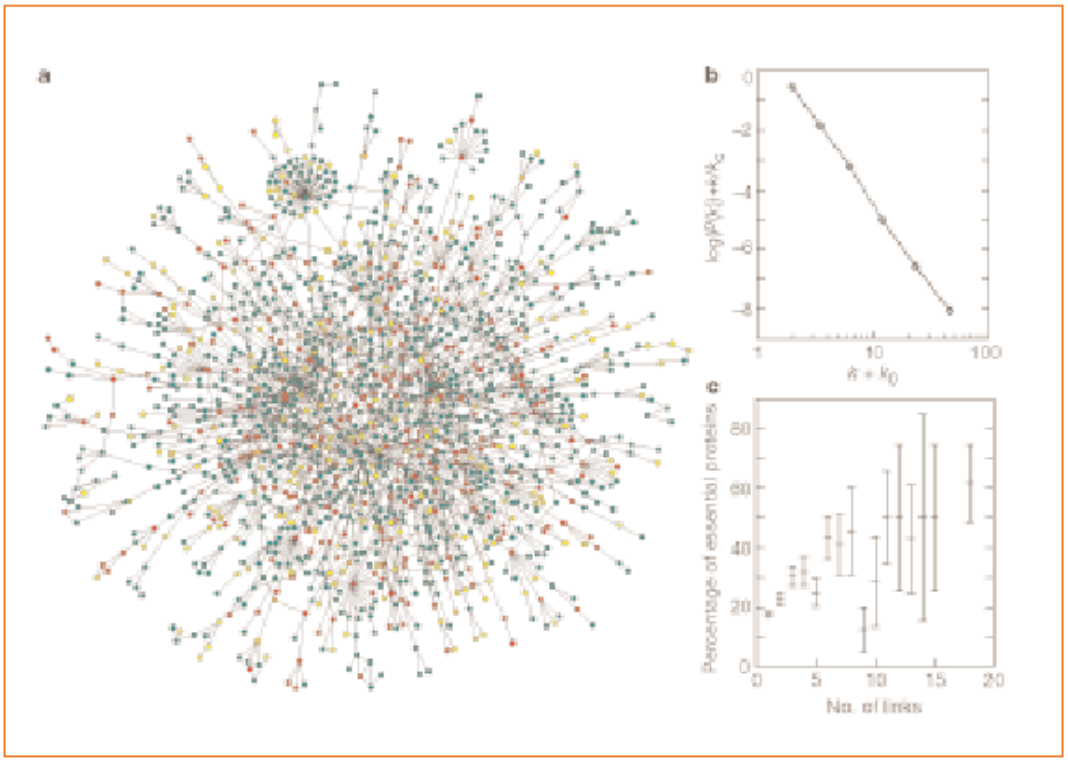
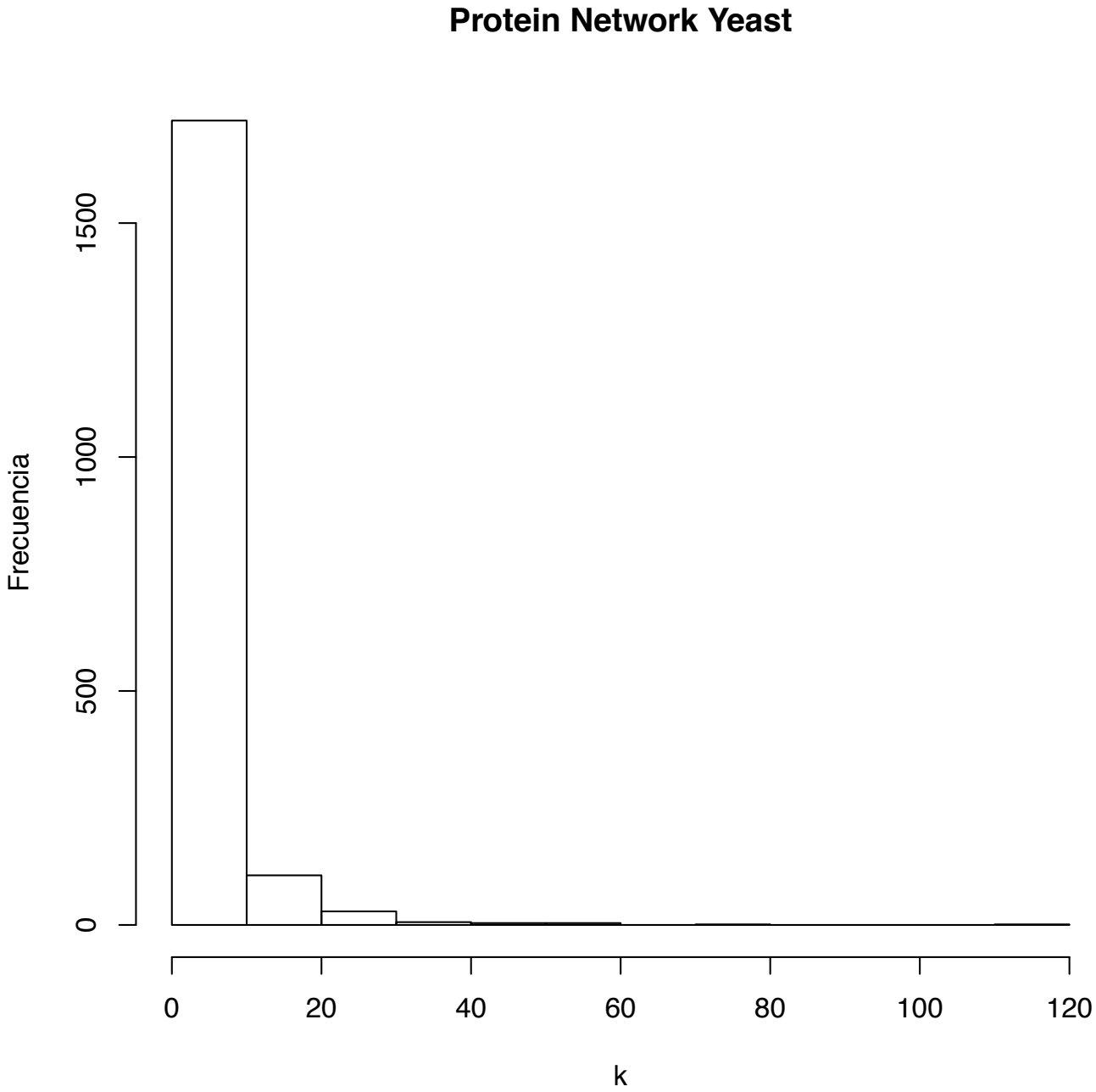
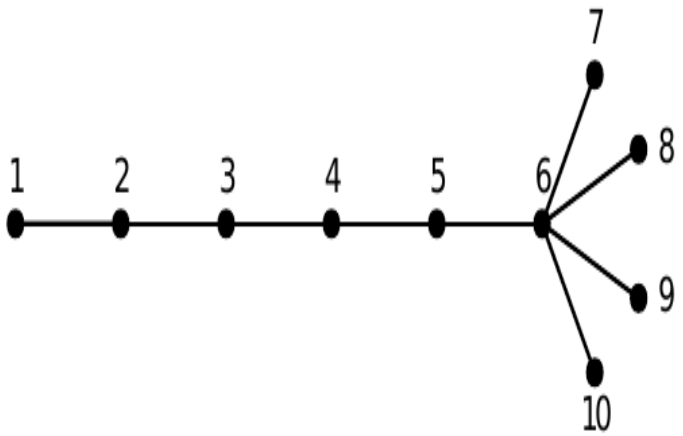


Figure 1 Characteristics of the yeast proteome. **a,** Map of protein–protein interactions. The largest cluster, which contains ~78% of all


Protein Network Yeast
Frecuencia
0
500
1000
1500
0
20
40
60
80
100
120
k

Excentricidad

Es la máxima distancia de un nodo a alguno de la red.



$e(1) = 6$, $e(2) = 5 \ldots$

Excentricidad

Definición

Sea un grafo $G = (V, E)$ conectado y no dirigido, la excentricidad del nodo s es

$$C_{exc}(s) = \max\{dist(s, t) : t \in V\}$$

Closeness Centrality

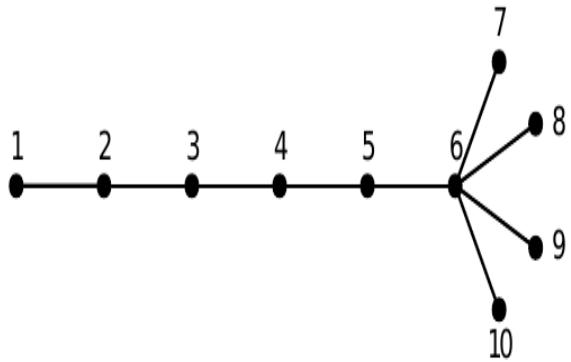
Definición

Sea un grafo $G = (V, E)$ conectado y no dirigido, la closeness centrality del nodo s es

$$C_{close}(s) = \frac{1}{\sum_t dist(s, t)}$$

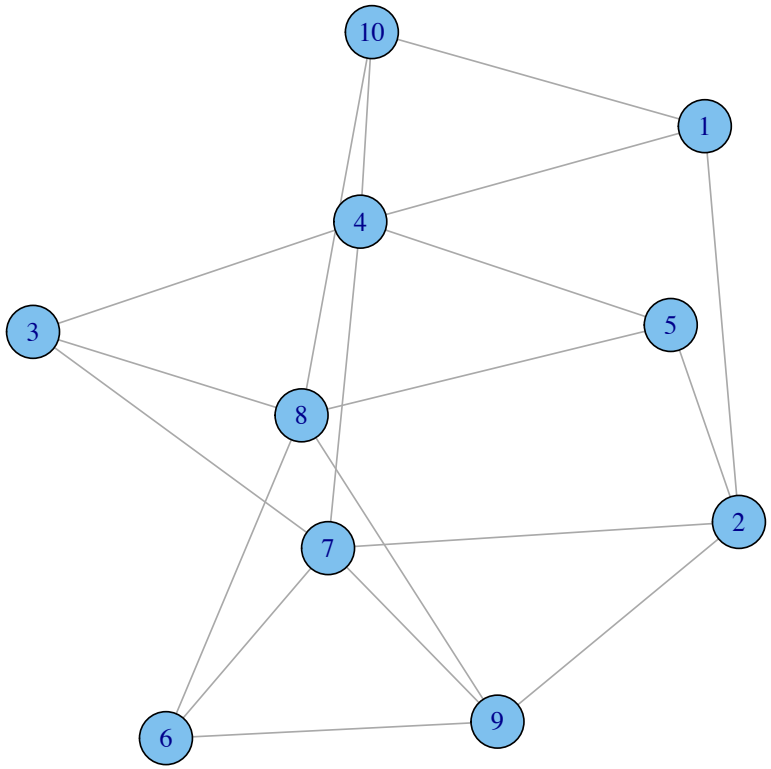
Closeness Centrality

Es el inverso de la suma de las distancias de un nodo a todos de la red.



$d(1,2) = 1$, $d(1,3) = 2$, $d(1,4) = 3$, $d(1,5) = 4$, $d(1,6) = 5$,$d(1,7) = d(1,8) = d(1,9) = d(1,10) = 6$

$1 + 2 + 3 + 4 + 5 + 6 * 4 = 39$, $\frac{1}{39} = 0.025$



La fración del número de caminos más cortos que pasan por un nodo.

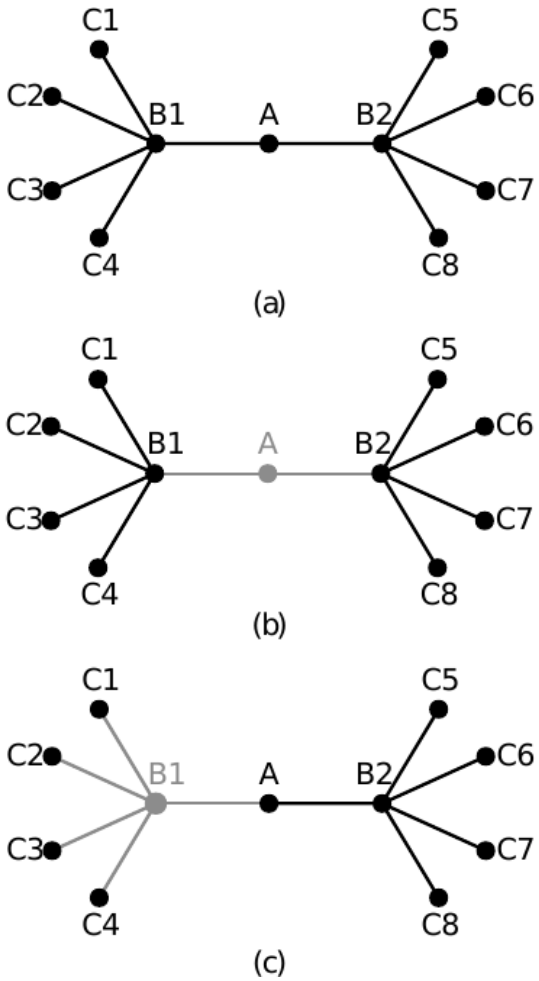
Betweenness centrality

Definición

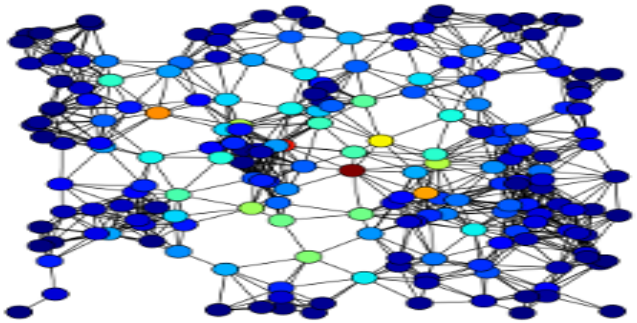
$$g(v) = \sum_{s,t} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

con σ_{st} el número de caminos más cortos del nodo s al nodo t, $\sigma_{st}((v)$, el número de estos que pasa por v

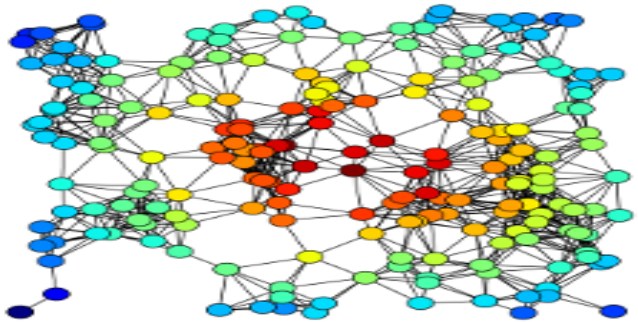
Betweenness centrality



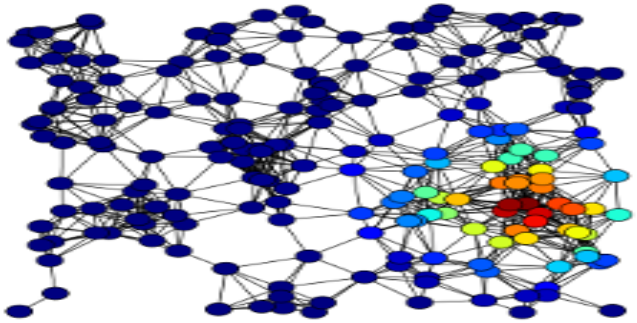
Las medidas de centralidad miden cosas distintas



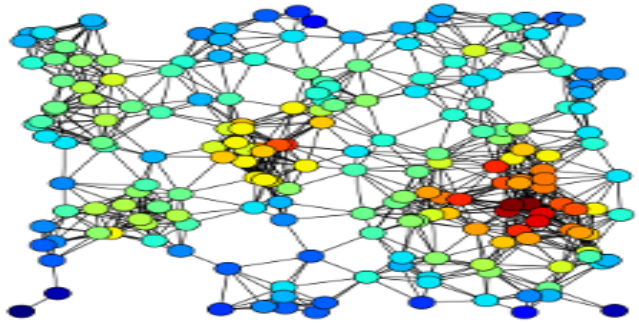
A



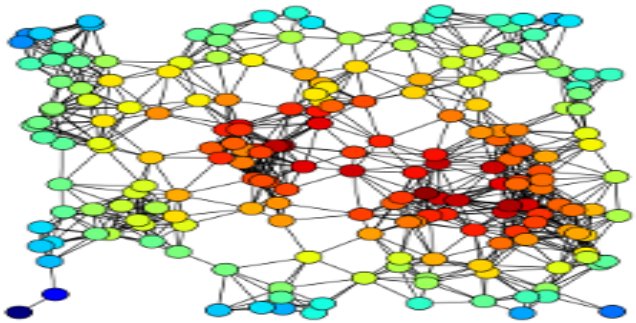
B



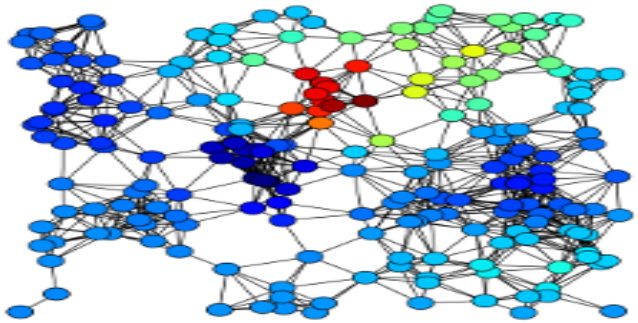
C



D

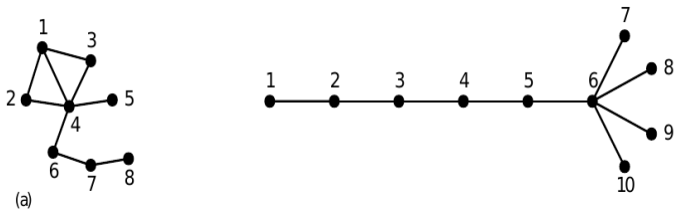


E



F

Medidas de centralidad



Ejercicio

- Calcula las medidas de centralidad para los nodos de las redes mostradas en la figura
- Usa igraph para calcular las medidas de centralidad de la red de interacción proteína-proteína de la red real.

Métodos de clusterización

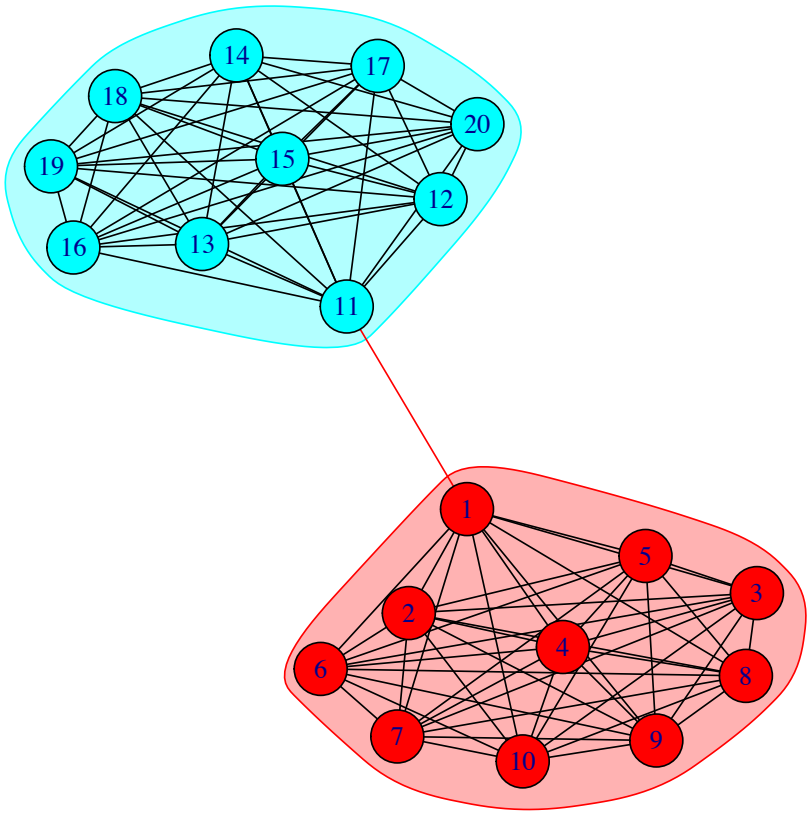
Definición

El proceso de agrupar objetos en conjuntos (clústers), de tal forma, que cada conjunto consista de elementos similares de acuerdo a una característica específica.

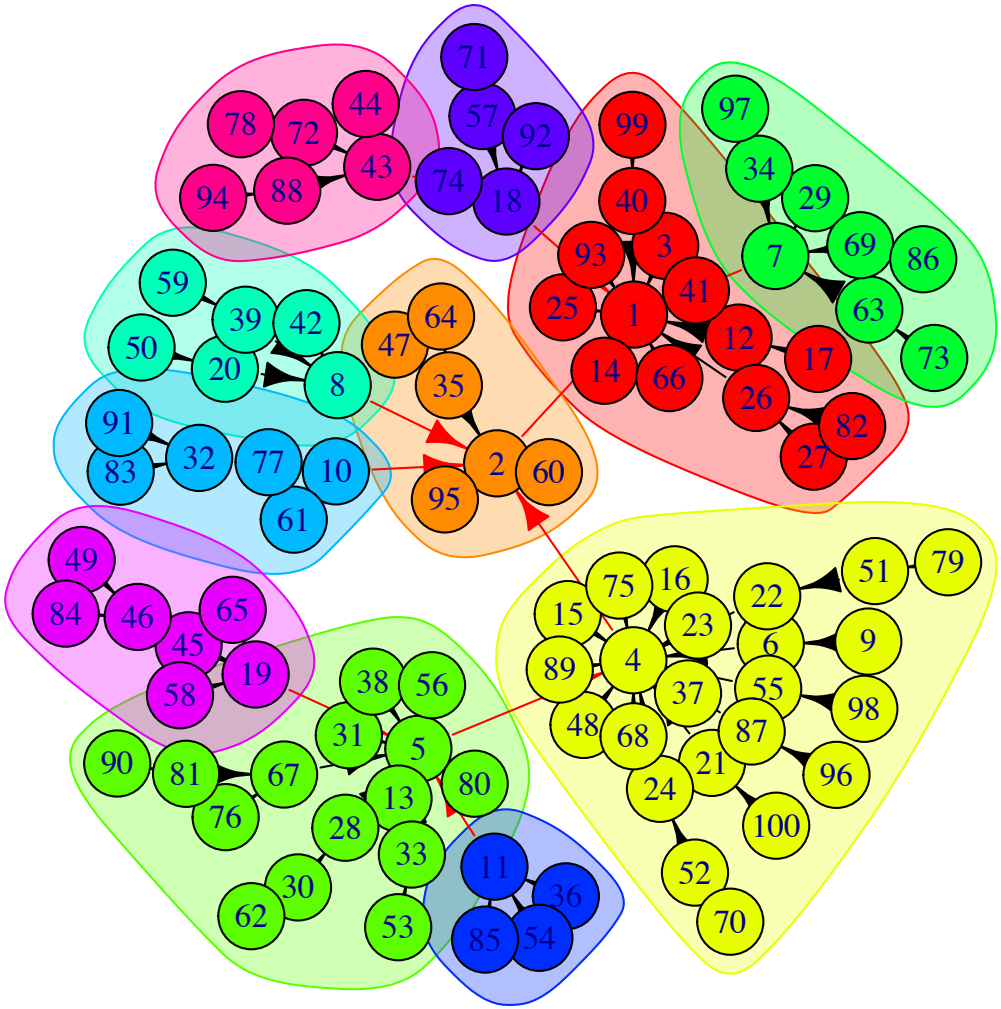
- Distancia: elementos están agrupados si son cercanos.

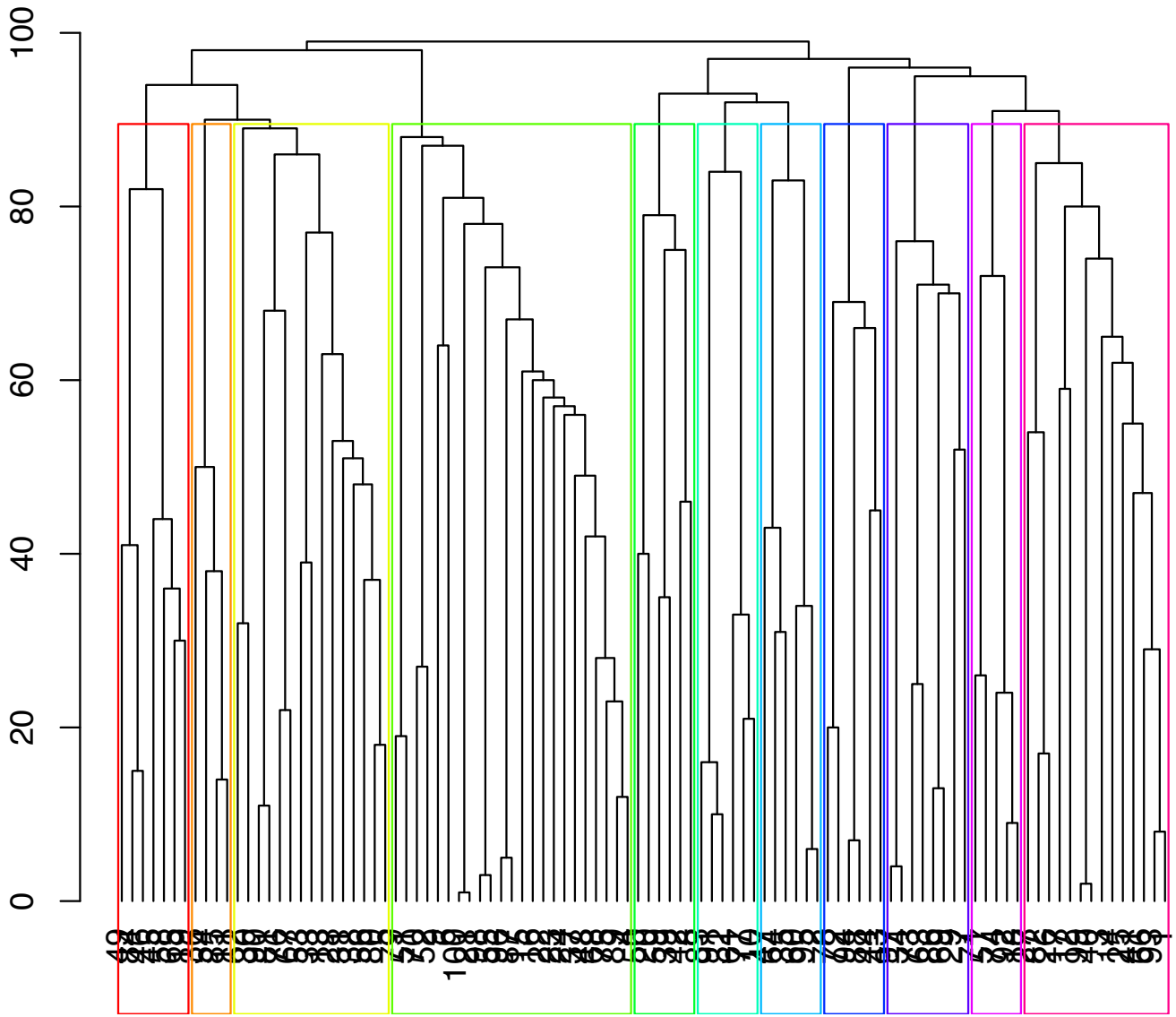
Edge betweenness

M Newman and M Girvan: Finding and evaluating community structure in networks, Physical Review E 69, 026113 (2004)



Edge betweenness

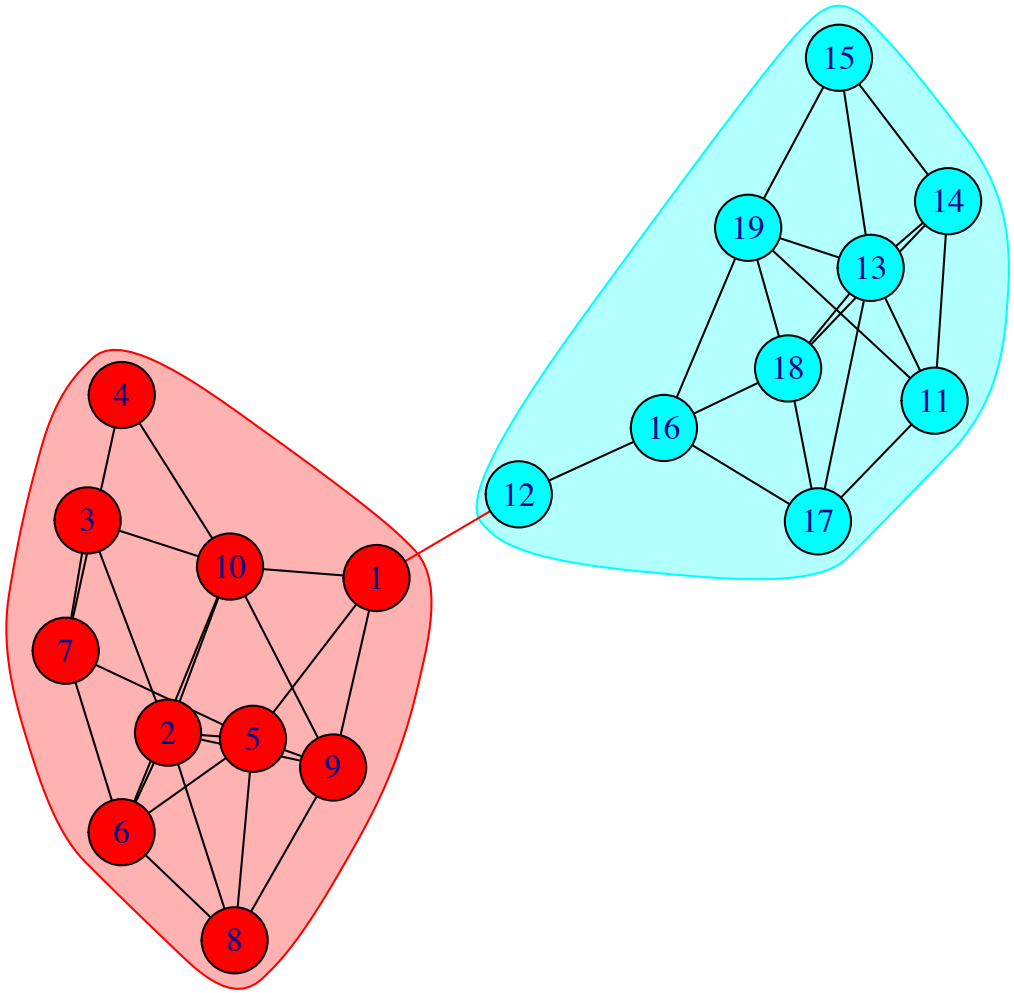


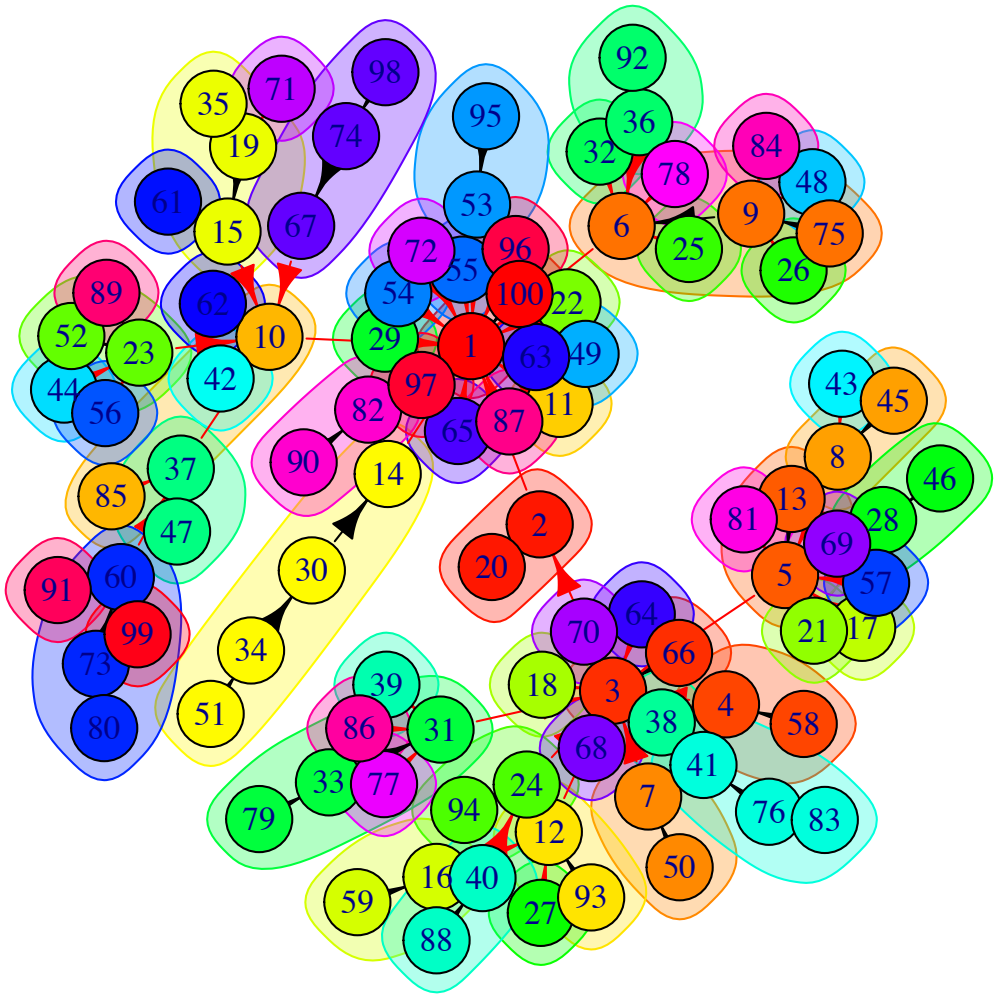

100
80
60
40
20
0

Label propagation

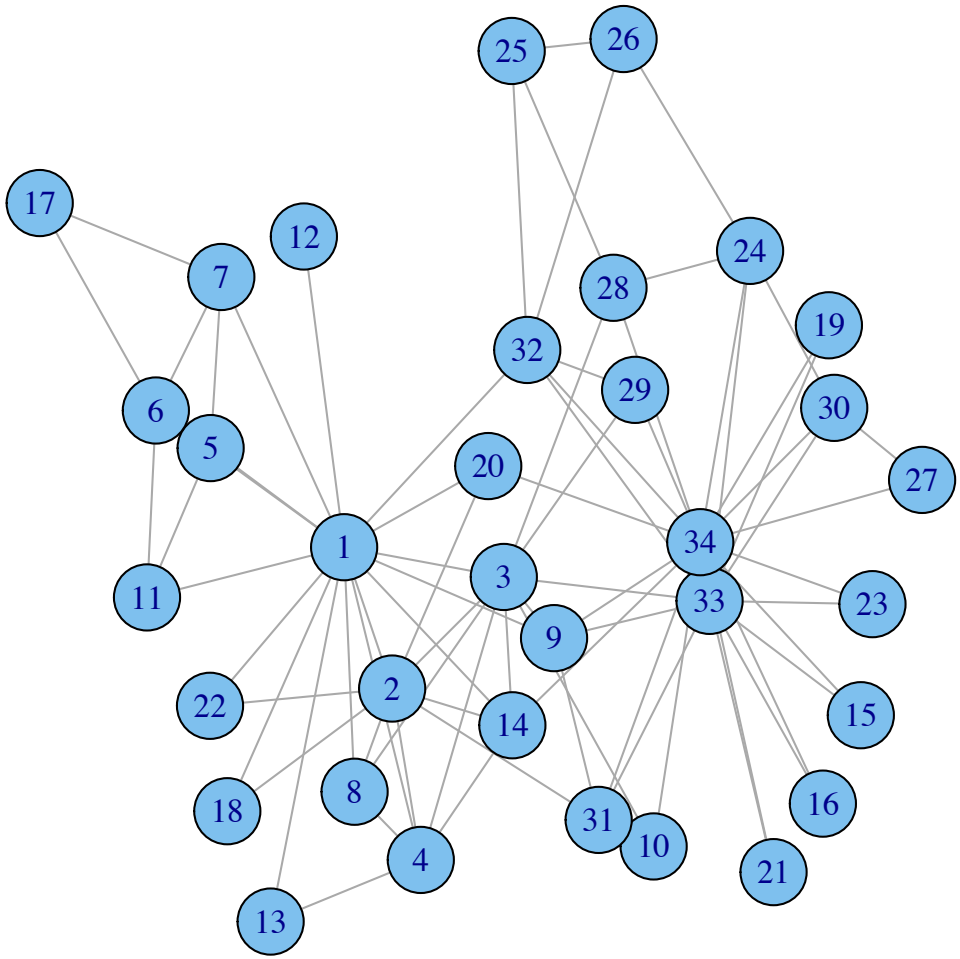
Detección de comunidades: Raghavan, U.N. and Al- bert, R. and Kumara, S.: Near linear time algorithm to detect community structures in large-scale networks. Phys Rev E 76, 036106. (2007).

"*In our algorithm every node is initialized with a unique label and at every step each node adopts the label that most of its neighbors currently have. In this iterative process densely connected groups of nodes form a consensus on a unique label to form communities.*"

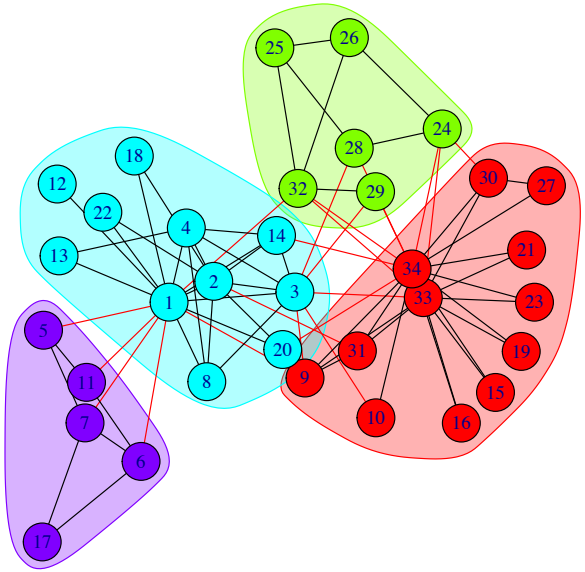

15
14
19
13
18
11
16
12
17
4
3
10
1
7
2
5
9
6
8

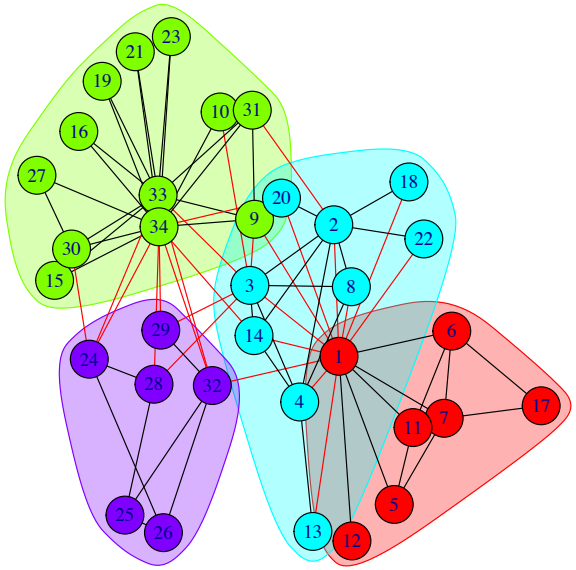

92
36
32
78
84
48
6
25
9
75
26
95
53
72
96
55
54
100
22
29
1
63
49
97
11
82
65
87
90
14
2
20
43
45
8
46
81
13
28
69
5
57
21
17
70
64
66
18
3
4
58
38
68
41
7
76
83
50
39
86
31
33
77
79
24
94
12
16
40
59
27
93
88
35
71
98
74
19
61
15
67
89
62
10
52
23
42
44
56
37
85
47
91
60
99
73
80
30
34
51

Comparativo



Comparativo





Muchos métodos de clusterización

fast greedy, infomap,leading eigenvector, optimal, spinglass, walktrap,e ntre muchas más.

¿Es modular una red (real)?

- Clustering implica modularidad
- Funcionalidad implica modularidad
- Redes de mundo pequeño tienden a eliminar la modularidad
- Hubs (concentradores) tienden a eliminar la modularidad

¿Es modular una red (real)?

- Clustering implica modularidad
- Funcionalidad implica modularidad
- Redes de mundo pequeño tienden a eliminar la modularidad
- Hubs (concentradores) tienden a eliminar la modularidad

¿Es modular una red (real)?

- Clustering implica modularidad
- Funcionalidad implica modularidad
- Redes de mundo pequeño tienden a eliminar la modularidad
- Hubs (concentradores) tienden a eliminar la modularidad

¿Es modular una red (real)?

- Clustering implica modularidad
- Funcionalidad implica modularidad
- Redes de mundo pequeño tienden a eliminar la modularidad
- Hubs (concentradores) tienden a eliminar la modularidad

Efecto San Mateo

"Porque al que tiene se le dará más y tendrá en abundancia, pero al que no tiene, se le quitará aun lo que tiene.

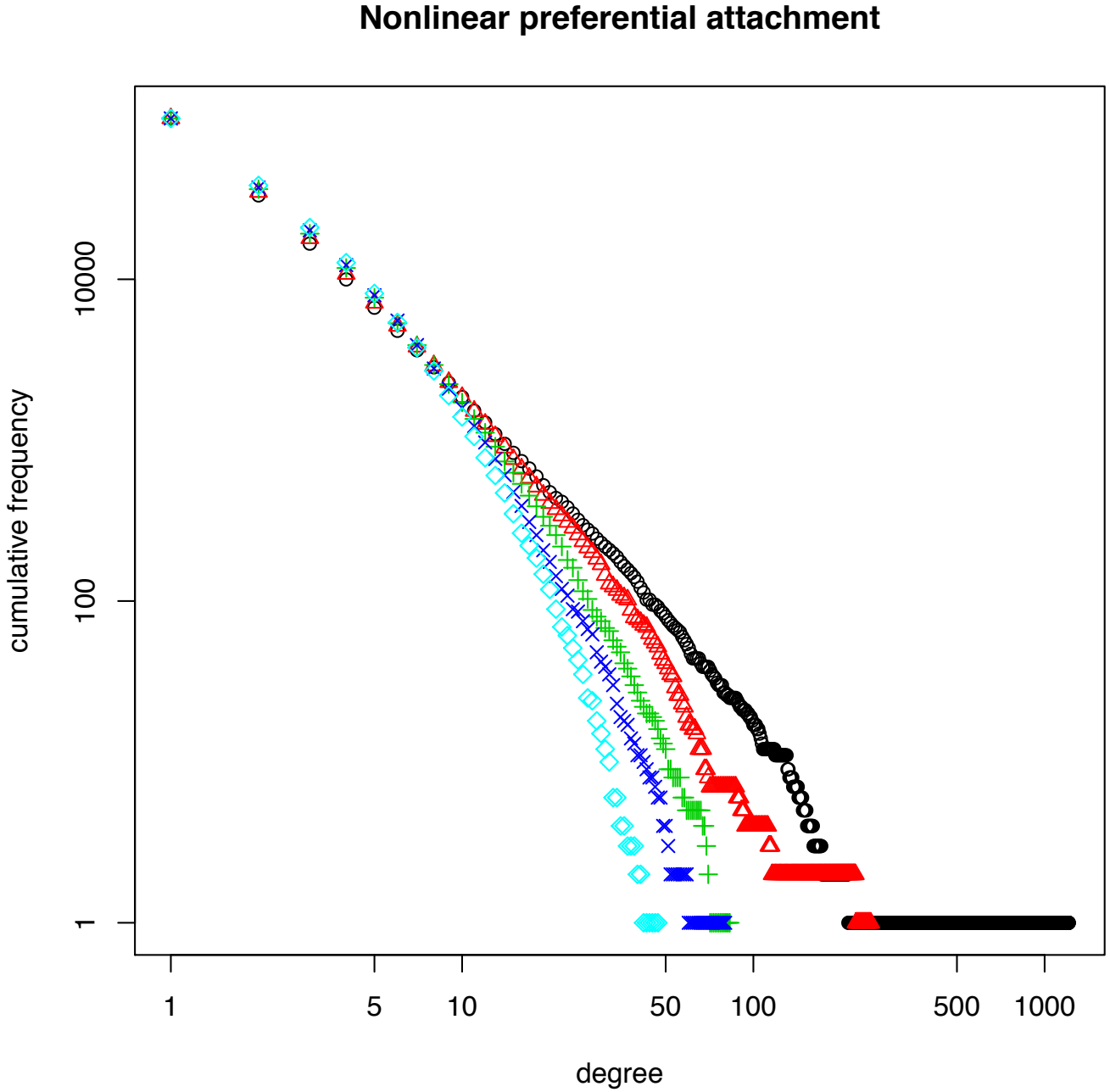
Variantes

Linear growth, linear pref. attachment	$\gamma=3$	Barabási and Albert, 1999
Nonlinear preferential attachment $\Pi(k_i)\sim k_i^{\alpha}$	no scaling for $\alpha\neq 1$	Krapivsky, Redner, and Leyvraz, 2000
Asymptotically linear pref. attachment $\Pi(k_i)\sim a_\infty k_i$ as $k_i\to\infty$	$\gamma\to 2$ if $a_\infty\to\infty$ $\gamma\to\infty$ if $a_\infty\to 0$	Krapivsky, Redner, and Leyvraz, 2000
Initial attractiveness $\Pi(k_i)\sim A+k_i$	$\gamma=2$ if $A=0$ $\gamma\to\infty$ if $A\to\infty$	Dorogovtsev, Mendes, and Samukhin, 2000a, 2000b
Accelerating growth $\langle k\rangle\sim t^{\theta}$ constant initial attractiveness	$\gamma=1.5$ if $\theta\to 1$ $\gamma\to 2$ if $\theta\to 0$	Dorogovtsev and Mendes, 2001a
Internal edges with probab. p	$\gamma=2$ if $q=\frac{1-p+m}{1+2m}$	
Rewiring of edges with probab. q	$\gamma\to\infty$ if $p,q,m\to 0$	Albert and Barabási, 2000
c internal edges or removal of c edges	$\gamma\to 2$ if $c\to\infty$ $\gamma\to\infty$ if $c\to -1$	Dorogovtsev and Mendes, 2000c
Gradual aging $\Pi(k_i)\sim k_i(t-t_i)^{-\nu}$	$\gamma\to 2$ if $\nu\to-\infty$ $\gamma\to\infty$ if $\nu\to 1$	Dorogovtsev and Mendes, 2000b
Multiplicative node fitness $\Pi_i\sim\eta_i k_i$	$P(k)\sim\frac{k^{-1-C}}{\ln(k)}$	Bianconi and Barabási, 2001a
Edge inheritance	$P(k_{in})=\frac{d}{k_{in}^{\sqrt{2}}}\ln(ak_{in})$	Dorogovtsev, Mendes, and Samukhin, 2000c
Copying with probab. p	$\gamma=(2-p)/(1-p)$	Kumar *et al.*, 2000a, 2000b
Redirection with probab. r	$\gamma=1+1/r$	Krapivsky and Redner, 2001
Walking with probab. p	$\gamma\simeq 2$ for $p>p_c$	Vázquez, 2000
Attaching to edges	$\gamma=3$	Dorogovtsev, Mendes, and Samukhin, 2001a
p directed internal edges $\Pi(k_i,k_j)\propto(k_i^{in}+\lambda)(k_j^{out}+\mu)$	$\gamma_{in}=2+p\lambda$ $\gamma_{out}=1+(1-p)^{-1}+\mu p/(1-p)$	Krapivsky, Rodgers, and Redner, 2001

Ejercicio

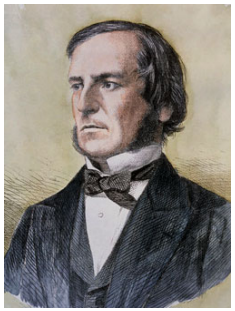
Calcula distintas potencias de conexión prefencial y grafica la distribución de conectividades.

```
powers <- c(0.9, 0.8, 0.7, 0.6) for (p in seq(power
g <- barabasi.game(100000, power=powers[p]) dd <-
degree.distribution(g, mode=ïn", cumulative=TRUE)
points(dd, col=p+1, pch=p+1)
legend(1, 1e-5, c(1,powers), col=1:5, pch=1:5, ncol
yjust=0, lty=0
```


Nonlinear preferential attachment
cumulative frequency
10000
100
1
1
5
10
50
100
500
1000
degree

Algebra booleana

George Boole (1815-1864), *The laws of thought (1854)*, algebra booleana.



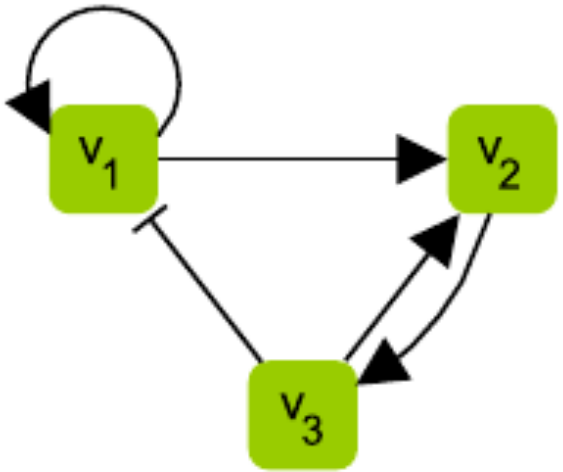
p	q	p AND q
T	T	T
T	F	F
F	T	F
F	F	F

p	q	p OR q
T	T	T
T	F	T
F	T	T
F	F	F

p	NOT p
T	F
F	T

Redes genéticas regulatorias (GRN)

(*a*) Network structure



(*b*) Boolean functions

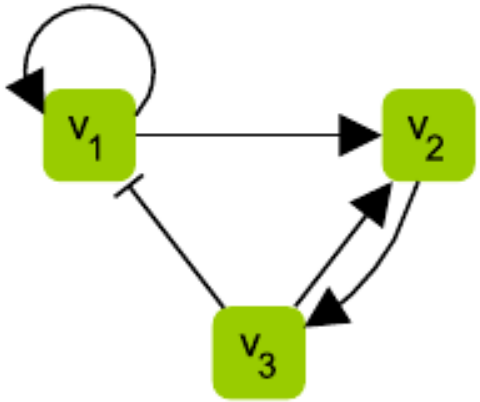
$B_1(\sigma_1, \sigma_3) = \sigma_1$ OR NOT σ_3

$B_2(\sigma_1, \sigma_3) = \sigma_1$ AND σ_3

$B_3(\sigma_2) = \sigma_2$

Redes genéticas regulatorias (GRN)

(a) Network structure



(b) Boolean functions

$B_1(\sigma_1, \sigma_3) = \sigma_1$ OR NOT σ_3

$B_2(\sigma_1, \sigma_3) = \sigma_1$ AND σ_3

$B_3(\sigma_2) = \sigma_2$

(c) Truth tables

$B_1(\sigma_1, \sigma_3)$

σ_1	σ_3	σ_1
0	0	1
0	1	0
1	0	1
1	1	1

$B_2(\sigma_1, \sigma_3)$

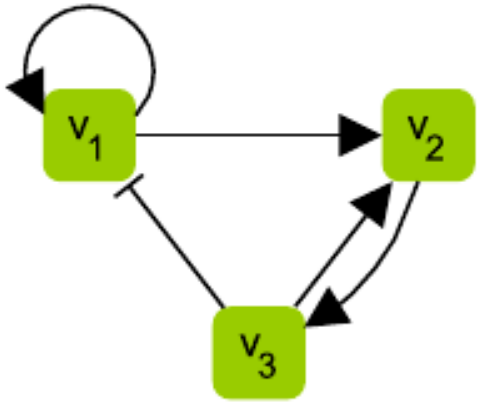
σ_1	σ_3	σ_2
0	0	0
0	1	0
1	0	0
1	1	1

$B_3(\sigma_2)$

σ_2	σ_3
0	0
1	1

Redes genéticas regulatorias (GRN)

(a) Network structure



(b) Boolean functions

$B_1(\sigma_1, \sigma_3) = \sigma_1$ OR NOT σ_3

$B_2(\sigma_1, \sigma_3) = \sigma_1$ AND σ_3

$B_3(\sigma_2) = \sigma_2$

(c) Truth tables

$B_1(\sigma_1, \sigma_3)$

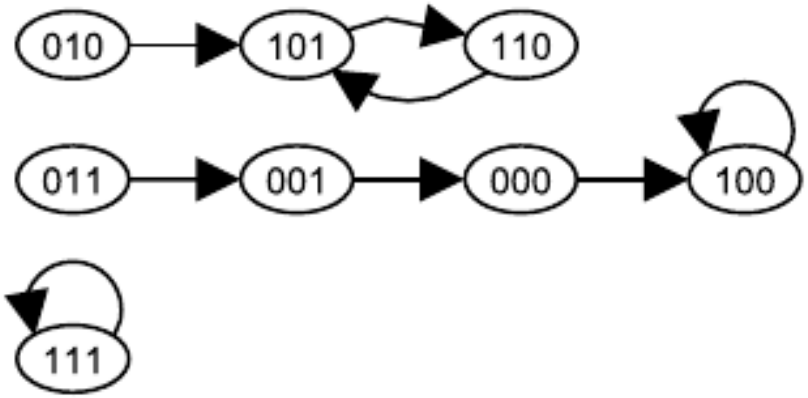
σ_1	σ_3	σ_1
0	0	1
0	1	0
1	0	1
1	1	1

$B_2(\sigma_1, \sigma_3)$

σ_1	σ_3	σ_2
0	0	0
0	1	0
1	0	0
1	1	1

$B_3(\sigma_2)$

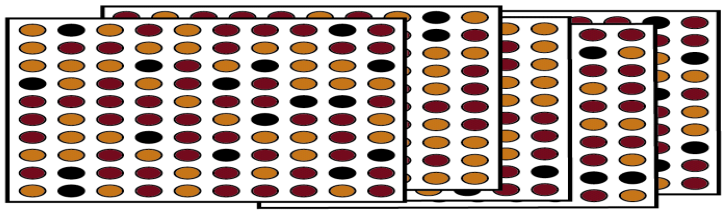
σ_2	σ_3
0	0
1	1



Co expresión

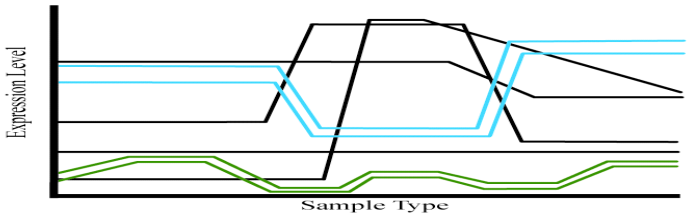
Figure 1

A Array Data



Data contains correlations

B Correlation Analysis



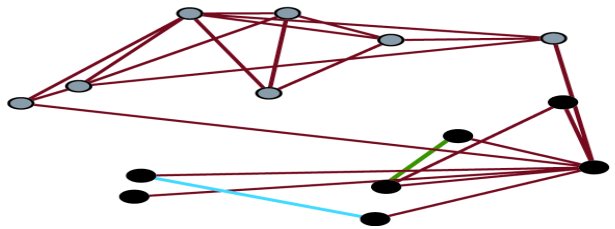
Correlation coefficients for all genes

C Correlation Matrix

	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12	G13	G14
G1	1	0.9	0.9	0.9	0.9	0.8	0.9	0.1	0.9	0.1	0.1	0.8	0.2	0.2
G2	0.9	1	0.9	0.3	0.3	0.7	0.0	0.5	0.3	0.1	0.1	0.2	0.4	0.3
G3	0.9	0.9	1	0.9	0.0	0.2	0.5	0.7	0.6	0.5	0.2	0.6	0.1	0.0
G4	0.9	0.3	0.9	1	0.5	0.3	0.6	0.3	0.0	0.5	0.1	0.2	0.2	0.6
G5	0.9	0.3	0.0	0.5	1	0.1	0.6	0.1	0.3	0.3	0.3	0.5	0.2	0.5
G6	0.8	0.7	0.2	0.3	0.1	1	0.9	0.2	0.1	0.1	0.5	0.3	0.1	0.1
G7	0.9	0.0	0.5	0.6	0.6	0.9	1	0.3	0.1	0.5	0.1	0.3	0.5	0.2
G8	0.1	0.5	0.7	0.3	0.1	0.2	0.3	1	0.9	0.9	0.9	0.8	0.8	0.9
G9	0.9	0.3	0.6	0.0	0.3	0.1	0.1	0.9	1	0.8	0.1	0.3	0.5	0.3
G10	0.1	0.1	0.5	0.5	0.3	0.1	0.5	0.9	0.8	1	0.8	1.0	0.2	0.3
G11	0.1	0.1	0.2	0.1	0.3	0.5	0.1	0.9	0.1	0.8	1	0.5	0.8	0.9
G12	0.8	0.2	0.6	0.2	0.5	0.3	0.3	0.8	0.3	1.0	0.5	1	0.8	0.1
G13	0.2	0.4	0.1	0.2	0.2	0.1	0.5	0.8	0.5	0.2	0.8	0.8	1	0.9
G14	0.2	0.3	0.0	0.6	0.5	0.1	0.2	0.9	0.3	0.3	0.9	0.1	0.9	1

Convert into Adjacency Matrix and Network

D Coexpression Network



Correlación

Definición

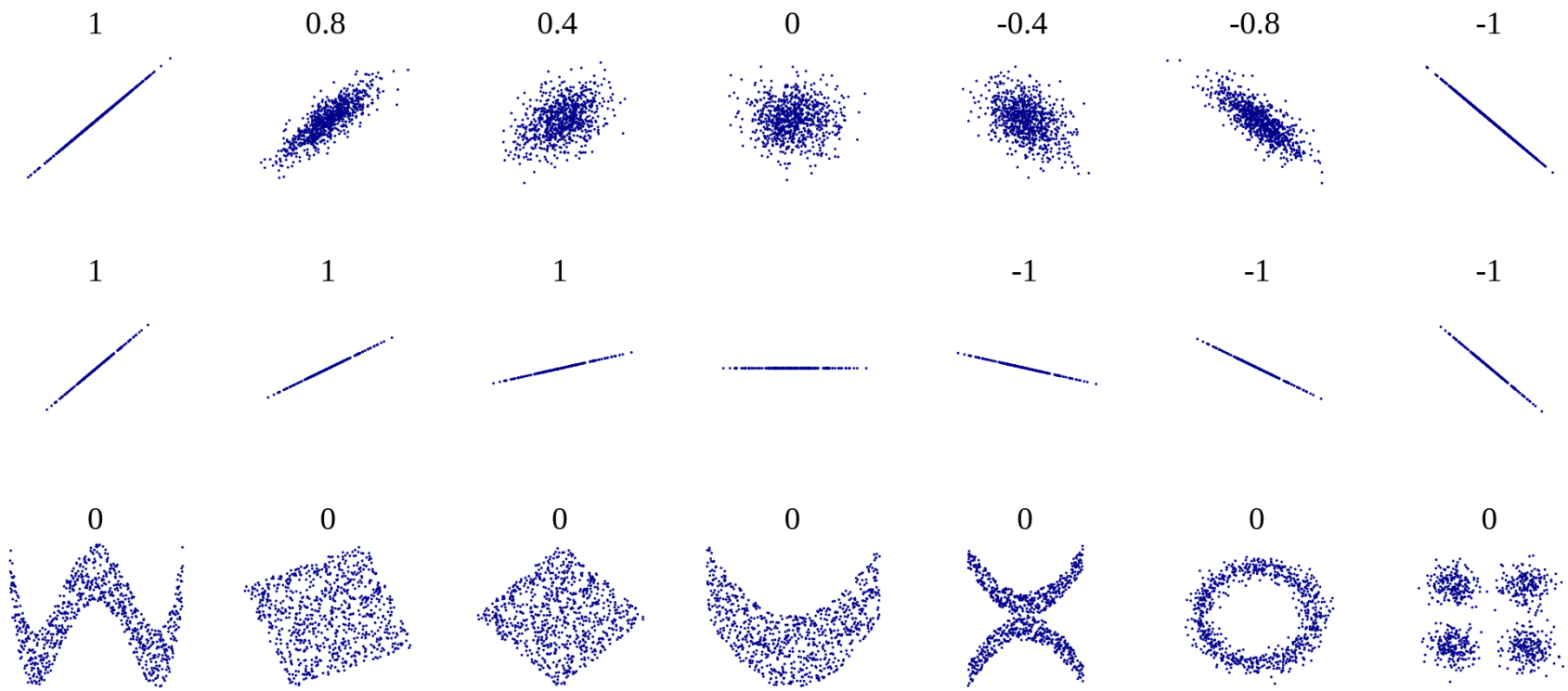
$$\rho_{X,Y} = \operatorname{corr}(X, Y) = \frac{\operatorname{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y},$$

Coeficiente de correlación de Pearson
Para una muestra de valores de dos variables
$X = X_1, X_2, X_3 \ldots X_n$ y $Y = Y_1, Y_2, Y_3 \ldots, Y_n$

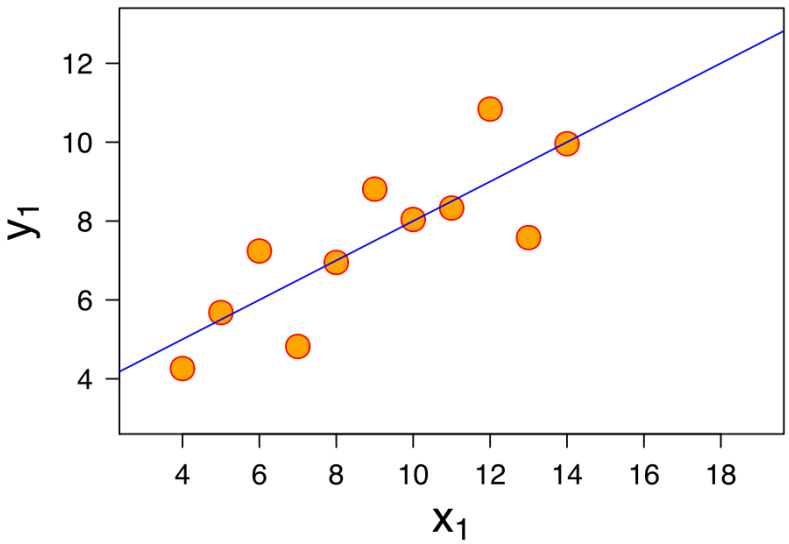
$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

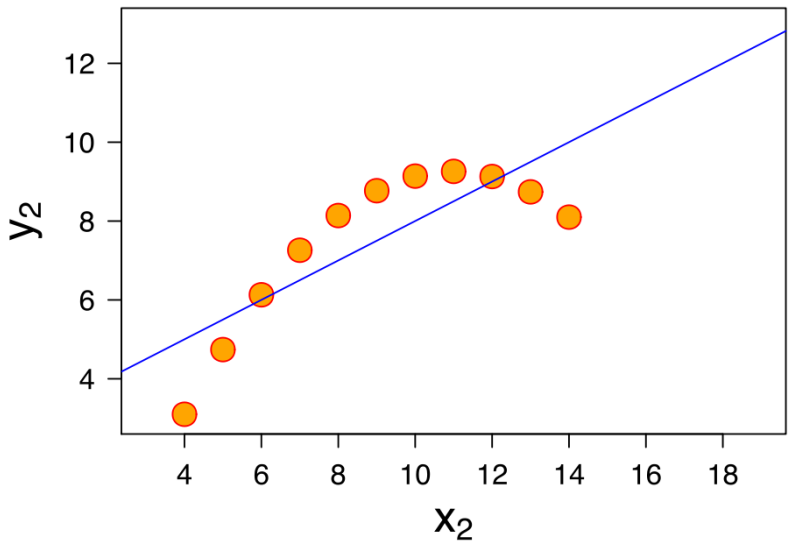
$$-1 \leq r \leq 1$$

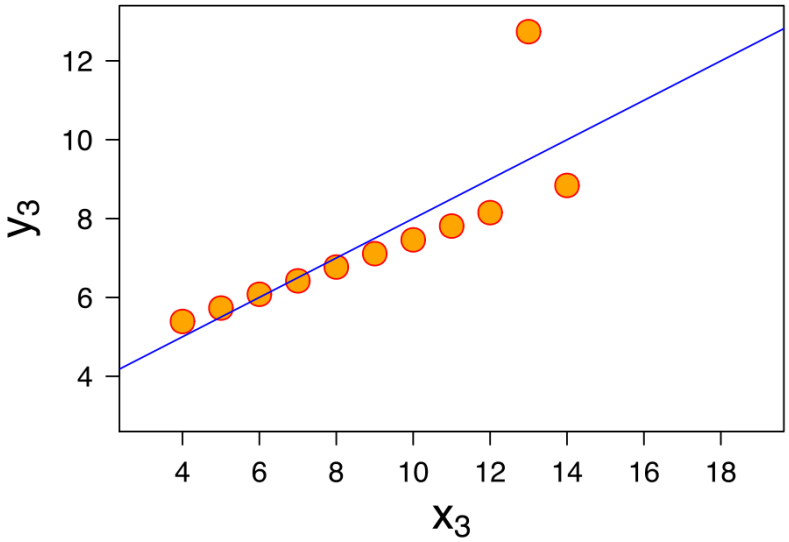
Correlación (de Pearson) y linealidad

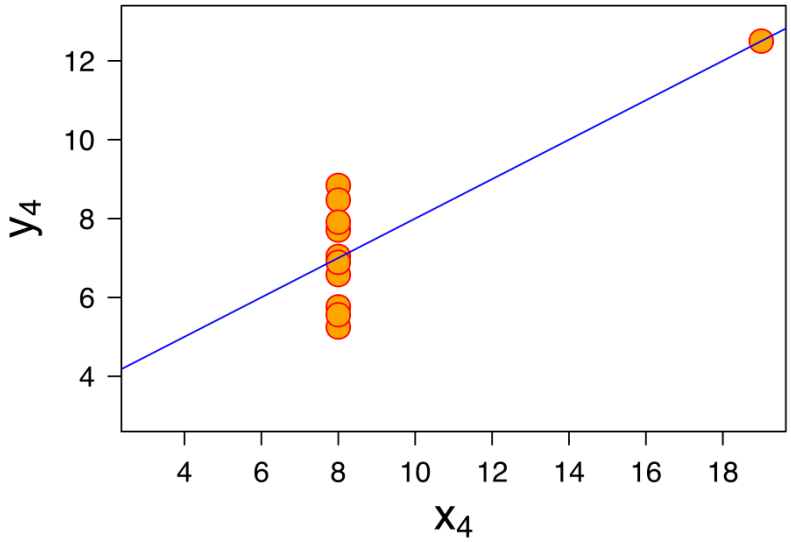


Correlación (de Pearson) y linealidad 2









Correlación (de Pearson) y linealidad 3

$$Y = aX^{\alpha}$$

$$\log Y = \log a + \alpha logX$$

$$\mathcal{Y} = \mathcal{A} + \alpha\mathcal{X}$$

$$Y = e^{-ax}\frac{(1-x)^{\beta}}{x^{\alpha}}$$

$$\log Y = -ax + \beta\log(1-x) - \alpha\log x$$

$$\mathcal{Y} = -a\mathcal{X}_1 + \beta\mathcal{X}_2 - \alpha\mathcal{X}_3$$

Correlación (de Pearson) y linealidad 3

$$Y = aX^{\alpha}$$

$$\log Y = \log a + \alpha logX$$

$$\mathcal{Y} = \mathcal{A} + \alpha\mathcal{X}$$

$$Y = e^{-ax}\frac{(1-x)^{\beta}}{x^{\alpha}}$$

$$\log Y = -ax + \beta \log(1-x) - \alpha \log x$$

$$\mathcal{Y} = -a\mathcal{X}_1 + \beta\mathcal{X}_2 - \alpha\mathcal{X}_3$$

Ejercicio

- Dado un conjunto $X = x_1, \ldots x_n$ generar $Y = aX + b$ con a, b arbitrarios. Calcular la correlación.
- $Y = X^2$, $Y = X^{10}, Y = exp^X, Y = \sin(X)$

Teoría de la información

Entropía Shanon

Sea X una variable aleatoria discreta, con valores $X = X_1, X_2, X_3 \ldots, X_r$, con probabilidad $p_1, p_2, p_3 \ldots, p_r$

$$H(X) = \sum_{i}^{r} p_i \log p_i$$

$$0 \leq H(X) \leq \log r$$

Información mutua

Definición

Sean X , Y dos variables aleatorias discretas

$$I(X;Y) = \sum_{x,y} P_{XY}(x,y) \log \frac{P_{XY}(x,y)}{P_X(x)P_Y(y)}$$

$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

$$0 \leq I(X;Y) < \infty$$

Información mutua

Usos

- Predicción de estructura secundaria de RNA.
- Pérfiles filogenéticos
- Determinación de la similaridad entre dos diferentes métodos de agrupamiento.
- En micro arreglos de expresión se usa la mi entre genes para construir redes (ARACNE)

Ejercicio

Instalar y utilizar la librería de R parmigene

- Dado un conjunto $X = x_1, \ldots x_n$ generar y $Y = aX + b$ con a, b arbitrarios. Calcular la MI
- $Y = X^2$, $Y = X^{10}, Y = exp^X, Y = \sin(X)$

ARACNe

ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks

Basso *et al* Nature Genetics (2005)

1. Calcular la MI entre genes.
2. Remover interacciones a terceros vecinos.

Cálculo de la información mutua

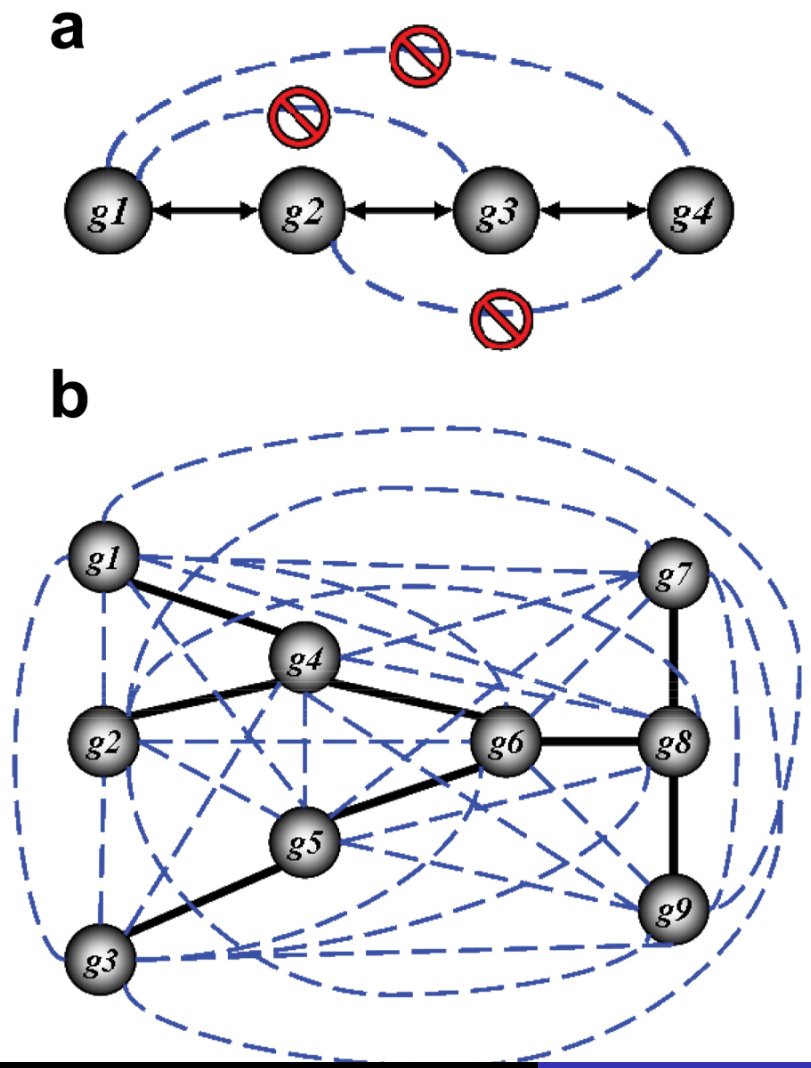
$$\begin{pmatrix} Genes & Expresion1 & \dots & ExpresionM \\ G1 & E_{11} & \dots & E_{1M} \\ G2 & E_{21} & \dots & E_{2M} \\ G3 & E_{31} & \dots & E_{3M} \\ \vdots & \vdots & \vdots & \vdots \\ Gn & E_{N1} & \dots & E_{NM} \end{pmatrix}$$

Cálculo de la información mutua

$$\begin{pmatrix} MI(G1,G1) & \dots & MI(G1,GN) \\ MI(G2,G1) & \dots & MI(G2,GN) \\ \vdots & \vdots & \vdots \\ MI(GN,G1) & \dots & MI(GN,GN) \end{pmatrix}$$

DPI (Data Processing Inequality)

$$MI(G1, G3) \leq min[MI(G1, G2); MI(G2, G3)]$$



Ejercicio

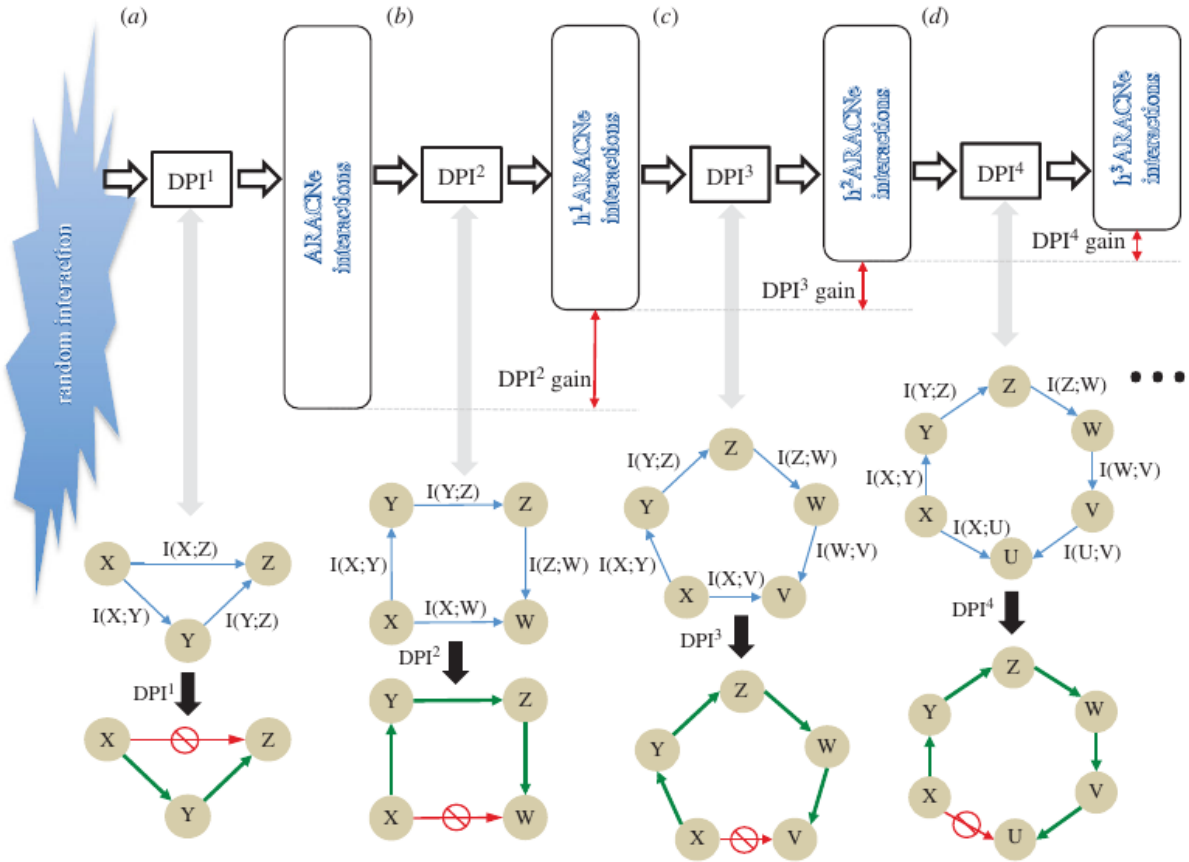
Generar una matriz de “expresión” aleatoria de 20×100 , aplicar el algoritmo ARACNE, graficar la red y calcular clustering, diámetro y distribución de coenctividades. Usar el modelo aditivo con distintos valores de ϵ

hARACNe

Interface Focus
rsfs.royalsocietypublishing.org

hARACNe: improving the accuracy of regulatory model reverse engineering via higher-order data processing inequality tests

In Sock Jang[1], Adam Margolin[1] and Andrea Califano[2]



Maximal information coefficient

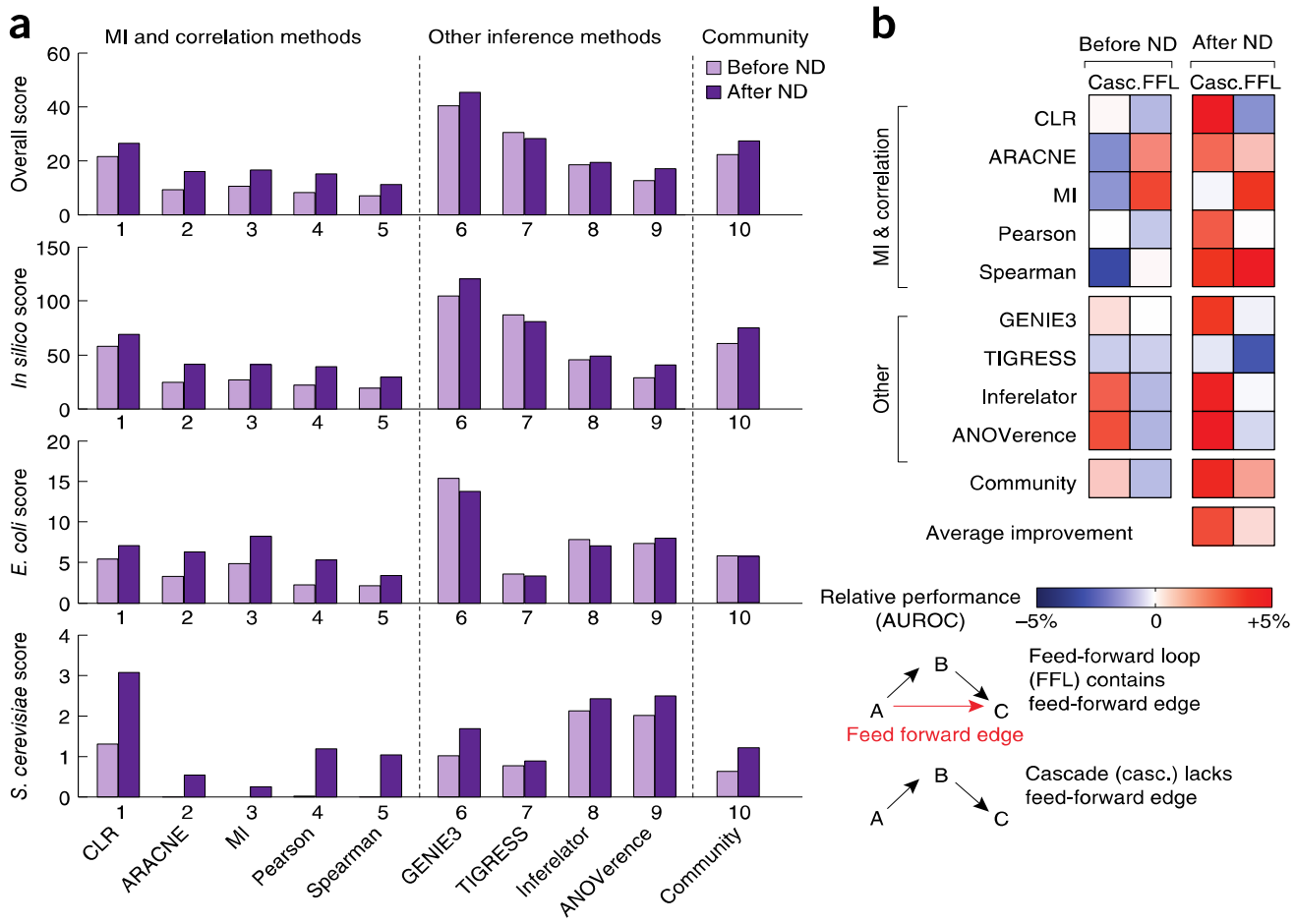
Detecting Novel Associations in Large Data Sets

David N. Reshef,[1,2,3*†] Yakir A. Reshef,[2,4*†] Hilary K. Finucane,[5] Sharon R. Grossman,[2,6] Gilean McVean,[3,7] Peter J. Turnbaugh,[6] Eric S. Lander,[2,8,9] Michael Mitzenmacher,[10‡] Pardis C. Sabeti[2,6‡]

Relationship Type	MIC	Pearson	Spearman	Mutual Information (KDE)	Mutual Information (Kraskov)	CorGC (Principal Curve-Based)	Maximal Correlation
Random	0.18	-0.02	-0.02	0.01	0.03	0.19	0.01
Linear	1.00	1.00	1.00	5.03	3.89	1.00	1.00
Cubic	1.00	0.61	0.69	3.09	3.12	0.98	1.00
Exponential	1.00	0.70	1.00	2.09	3.62	0.94	1.00
Sinusoidal (Fourier frequency)	1.00	-0.09	-0.09	0.01	-0.11	0.36	0.64
Categorical	1.00	0.53	0.49	2.22	1.65	1.00	1.00
Periodic/Linear	1.00	0.33	0.31	0.69	0.45	0.49	0.91
Parabolic	1.00	-0.01	-0.01	3.33	3.15	1.00	1.00
Sinusoidal (non-Fourier frequency)	1.00	0.00	0.00	0.01	0.20	0.40	0.80
Sinusoidal (varying frequency)	1.00	-0.11	-0.11	0.02	0.06	0.38	0.76

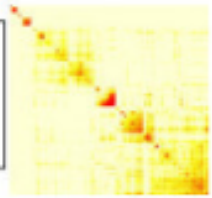
Network deconvolution as a general method to distinguish direct dependencies in networks

Soheil Feizi[1–3], Daniel Marbach[1,2], Muriel Médard[3] & Manolis Kellis[1,2]

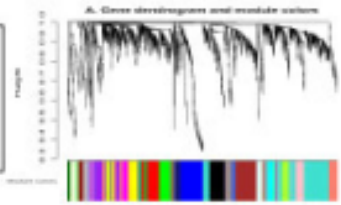


Redes genéticas ponderadas

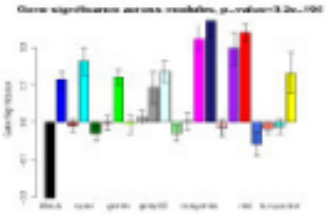
Construct a gene co-expression network
Rationale: make use of interaction patterns among genes
Tools: correlation as a measure of co-expression



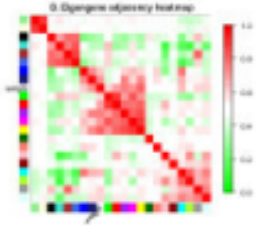
Identify modules
Rationale: module (pathway) based analysis
Tools: hierarchical clustering, Dynamic Tree Cut



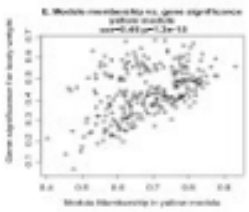
Relate modules to external information
Array Information: clinical data, SNPs, proteomics
Gene Information: ontology, functional enrichement
Rationale: find biologically interesting modules



Study module relationships
Rationale: biological data reduction, systems-level view
Tools: Eigengene Networks



Find the key drivers in *interesting* modules
Rationale: experimental validation, biomarkers
Tools: intramodular connectivity, causality testing



Redes genéticas ponderadas

Weighted Gene Coexpression Network Analysis (WGCNA)

Steve Horvath, PhD

Conservation and evolution of gene coexpression networks in human and chimpanzee brains

Michael C. Oldham*[†1], Steve Horvath[3†], and Daniel H. Geschwind[1†‡]

ARTICLES

nature neuroscience

Functional organization of the transcriptome in human brain

Michael C Oldham[1,2], Genevie
Steve Horvath[4–6] & Daniel H

molecular systems biology

The organization of the transcriptional network in specific neuronal classes

Kellen D Winden[1,2,3], Michael C Old
Steve Horvath[4,10] and Daniel H Ges

nature

LETTERS

Human-specific transcriptional regulation of CNS development genes by FOXP2

Genevieve Konopka[1,2], Jamee M. Bomar[1,2], Kellen Winden[1,2], Giovanni Coppola[3], Zophonias O. Jonsson[3], Fuying Gao[3], Sophia Peng[3], Todd M. Preuss[4], James A. Wohlschlegel[5] & Daniel H. Geschwind[1,2,3,4]

Algoritmo MCL

Markov Clustering

- Stjn van Dongen (2000)
- Disponible gratuitamente en www.micans.org/mcl

MCL

Importancia

- STRING
- Cluster Maker en Cytoscape
- GPLEXUS

Ejercicio

- Clusterizar la red de interacción de proteínas, (igraph (varios métodos) o MCL)
- Generar una red “sin escala” (n $1X10^4$), clusterizarla con algún método en igraph y exportarla en formato ABC.
- Clusterizarla con MCL

Comparativo entre clústers

compare(comm1, comm2, method = c("vi", "nmi","split.join", rand",“adjusted.rand"))

- “vi” variation of information (VI) metric, Meila (2003),
- ‘nmi’ normalized mutual information measure, Danon et al. (2005)
- ‘split.join’ is the split-join distance, Dongen (2000)
- ‘rand’ Rand index, Rand (1971)
- ‘adjusted.rand’ adjusted Rand index, Hubert y Arabie (1985)

Ejemplo

```
g <- graph.famous("Zachary")
sg <- spinglass.community(g)
le <- leading.eigenvector.community(g)
compare(sg, le, method=rand")
compare(membership(sg), membership(le))
```

Hacer una heatmap comparando los clústers de cuando menos 5 métodos en igraph.

Joint-split

"For each pair of clusterings $C1$, $C2$, two numbers are given, say $d1$ and $d2$. Then $d1 + d2$ equals the number of nodes that have to be exchanged in order to transform any of the two clusterings into the other, and you can think of $(d1 + d2)/2N$as the percentage that the two clusterings differ."

Rand index

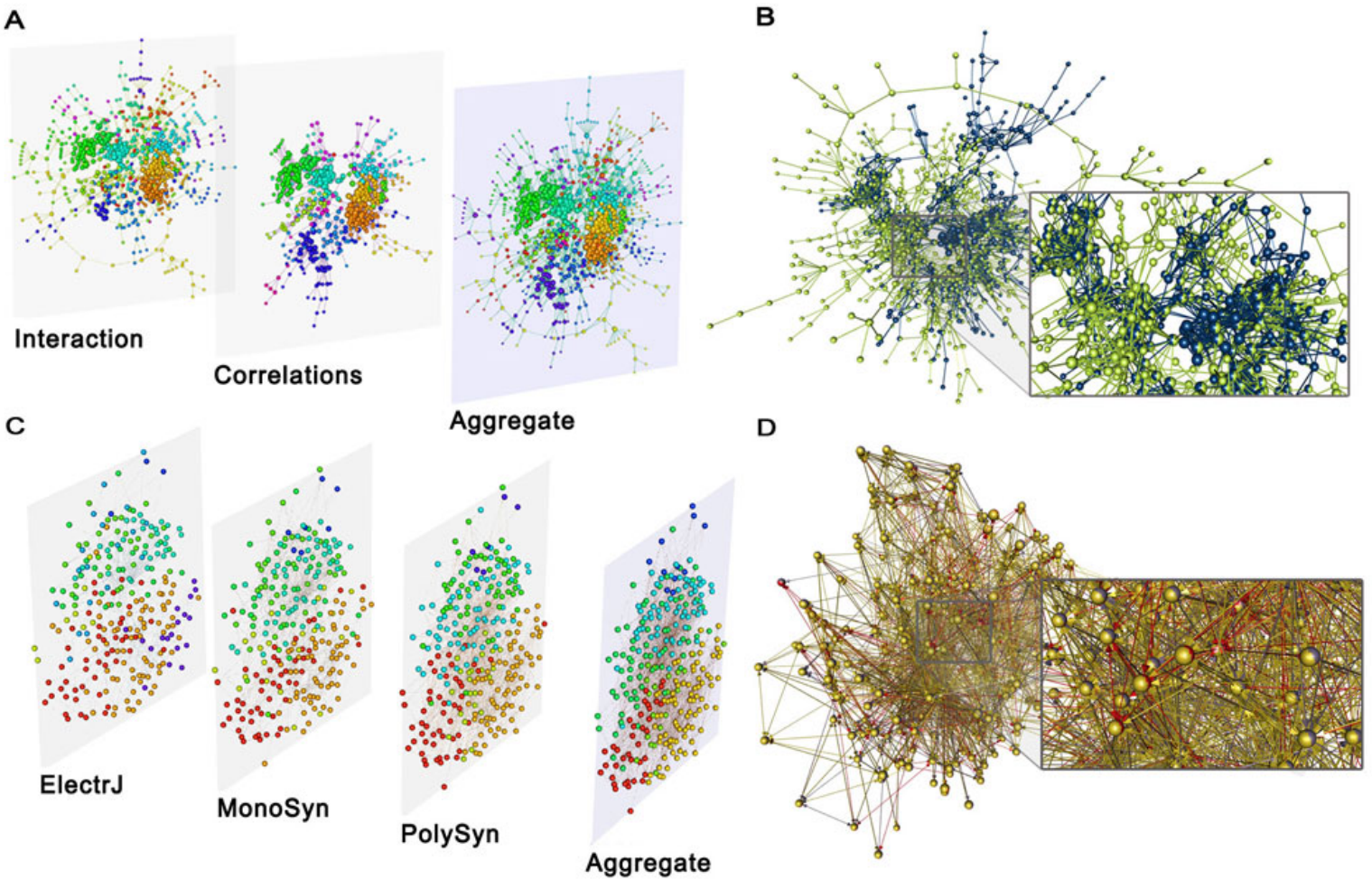
Given a set of n elements $S = \{o_1, \ldots, o_n\}$ and two partitions of S to compare, $X = \{X_1, \ldots, X_r\}$, a partition of S into r subsets, and $Y = \{Y_1, \ldots, Y_s\}$, a partition of S into s subsets, define the following:

- a, the number of pairs of elements in S that are in the same set in X and in the same set in Y
- b, the number of pairs of elements in S that are in different sets in X and in different sets in Y
- c, the number of pairs of elements in S that are in the same set in X and in different sets in Y
- d, the number of pairs of elements in S that are in different sets in X and in the same set in Y

The Rand index, R, is:

$$R = \frac{a+b}{a+b+c+d} = \frac{a+b}{\binom{n}{2}}$$

Redes múltiples



Redes múltiples

Name	Aligned	Disj.	Eq. Size	Diag.	Lcoup.	Cat.	$\|L\|$	d	Example refs.
Multilayer network				✓	✓	✓	Any	1	76
	✓†		✓†				Any	1	103
Multiplex network	✓†		✓†	✓			Any	1	102 103
	✓		✓	✓	✓	✓	Any	1	34 45 66 80 162 250 364
	✓		✓	✓	✓	✓	2	1	199 229 231
				✓	✓	✓	Any	1	91
	✓		✓	✓	✓		Any	1	92 309 310
Multivariate network	✓		✓	✓	✓	✓	Any	1	262
Multinetwork	✓		✓	✓	✓	✓	Any	1	19
	✓		✓	✓	✓	✓	Any	2	20
Multirelational network	✓		✓	✓	✓	✓	Any	1	72 152 319 350
Multirelational data	✓		✓	✓	✓	✓	Any	1	205 249
Multilayered network	✓		✓	✓	✓	✓	Any	1	60 61 63 309
Multidimensional network	✓		✓	✓	✓	✓	Any	1	24 42 44 89 180 333
	✓		✓	✓	✓	✓	Any	3	181
Multislice network	✓†		✓†	✓			Any	1	29 74 237 238
Multiplex of interdep. networks	✓		✓	✓	✓	✓	Any	1	143
Hypernetwork	✓		✓	✓	✓	✓	Any	1	169 314
Overlay network	✓		✓	✓	✓	✓	2	1	125 217
Composite network	✓		✓	✓	✓	✓	2	1	355
Multilevel network**		✓					Any	1	196 344
				✓	✓	✓	Any	1	91 95
Multiweighted graph	✓		✓	✓	✓	✓	Any	1	275
Heterogeneous network		✓					2	1	72 373
Multitype network		✓					Any	1	10 156 339
Interconnected networks		✓	✓				2	1	107 211
		✓					2	1	282 289
Interdependent networks*		✓	✓				2	1	68
*		✓					2	1	257
			✓				2	1	221
		✓					2	1	26 65
	✓		✓	✓	✓	✓	Any	1	35
Partially interdep. networks*		✓					2	1	372
Network of networks*			✓				Any	1	127
Coupled networks				✓	✓	✓	Any	1	365
Interconnecting networks				✓	✓	✓	2	1	363
Interacting networks		✓					Any	1	111 200
		✓					2	1	65
Heterogenous information net**							Any	2	99 324 326
		✓					Any	1	327
Meta-matrix, meta-network							Any	2	78 79 336