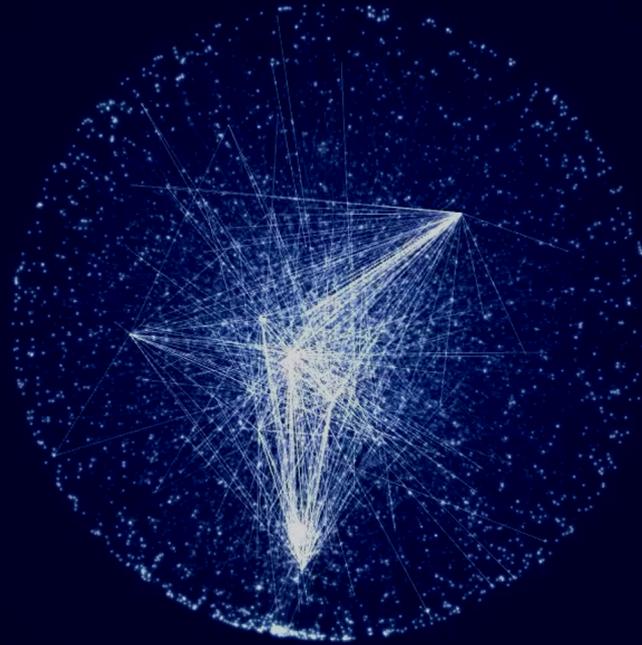


Network Science



Dr. Cristian Candia

*Director del Magister en Data Science y del Computational
Research in Social Science Laboratory,
Instituto de Data Science, Centro de Inv. en Complejidad Social,
Facultades de Ingeniería y Gobierno
Universidad del Desarrollo, Chile*

*Académico Adjunto,
Northwestern University, United States*

Video: Protein-Protein Interactions
Credit: Mauro Martino



Universidad del Desarrollo
Facultad de Ingeniería



Universidad del Desarrollo
Facultad de Gobierno

dataScience UDD

CICS UDD
CENTRO DE INVESTIGACIÓN EN COMPLEJIDAD SOCIAL

CRISSLAB
DECODING BEHAVIOR, ADVANCING SOCIETY

NICO NORTHWESTERN INSTITUTE
ON COMPLEX SYSTEMS

Topología

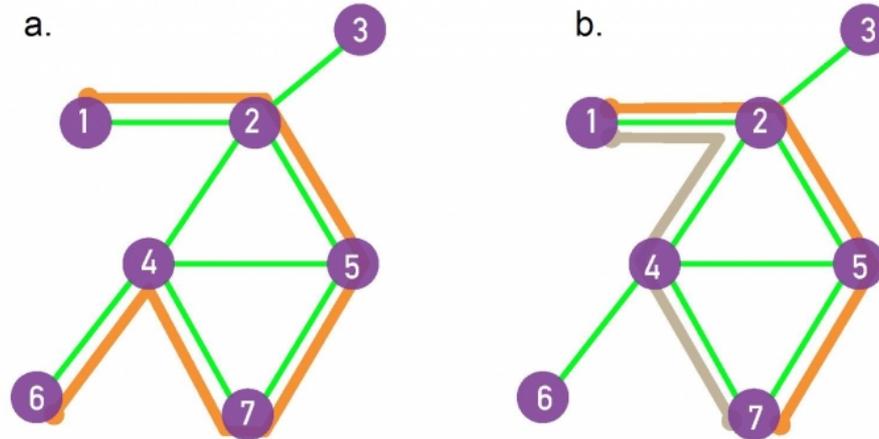
Estas diapositivas se basan en la presentación original del Prof. Albert-László Barabási, de Northeastern University, con autorización. El contenido ha sido traducido para su uso en este curso.

CAMINOS

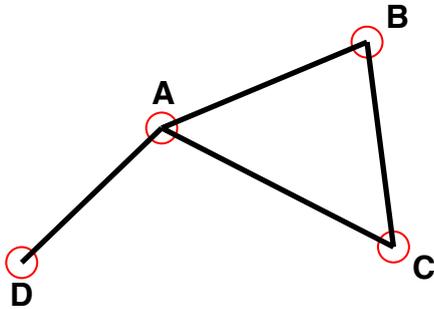
Un camino (path) es una secuencia de nodos en los que cada nodo es adyacente al siguiente

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

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

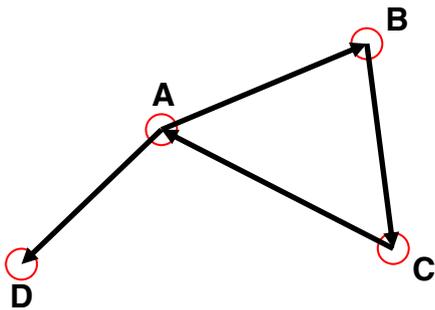


- En una red dirigida, la ruta solo puede seguir la dirección de una flecha.



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

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



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

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

N_{ij} , número de caminos entre dos nodos i y j :

Longitud $n=1$: Si existe un link entre i y j , entonces $A_{ij}=1$ y $A_{ij}=0$ en otro caso.

Longitud $n=2$: Si existe un camino de longitud dos entre i y j , entonces $A_{ik}A_{kj}=1$, y $A_{ik}A_{kj}=0$ en otro caso.

El número de caminos de longitud 2 es:

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

Longitud n : en general, si existe un camino de longitud n entre i y j , entonces $A_{ik}\dots A_{lj}=1$ y $A_{ik}\dots A_{lj}=0$ en otro caso.

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

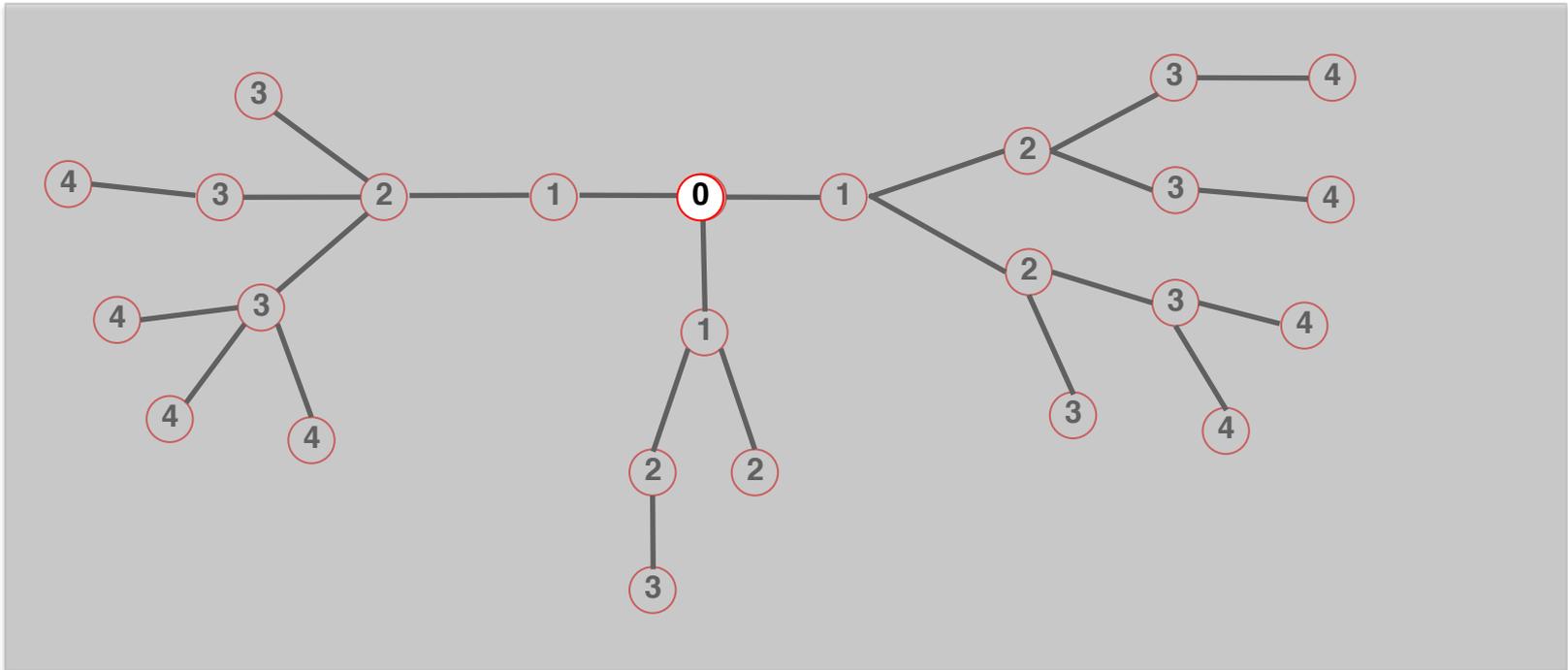
$$N_{ij}^{(n)} = [A^n]_{ij}$$

* Se aplica tanto a redes dirigidas como no dirigidas.

ECONTRANDO DISTANCIAS: “BREADTH FIRST SEARCH”

La distancia entre el nodo 0 y nodo 4:

1. Comienza en 0.

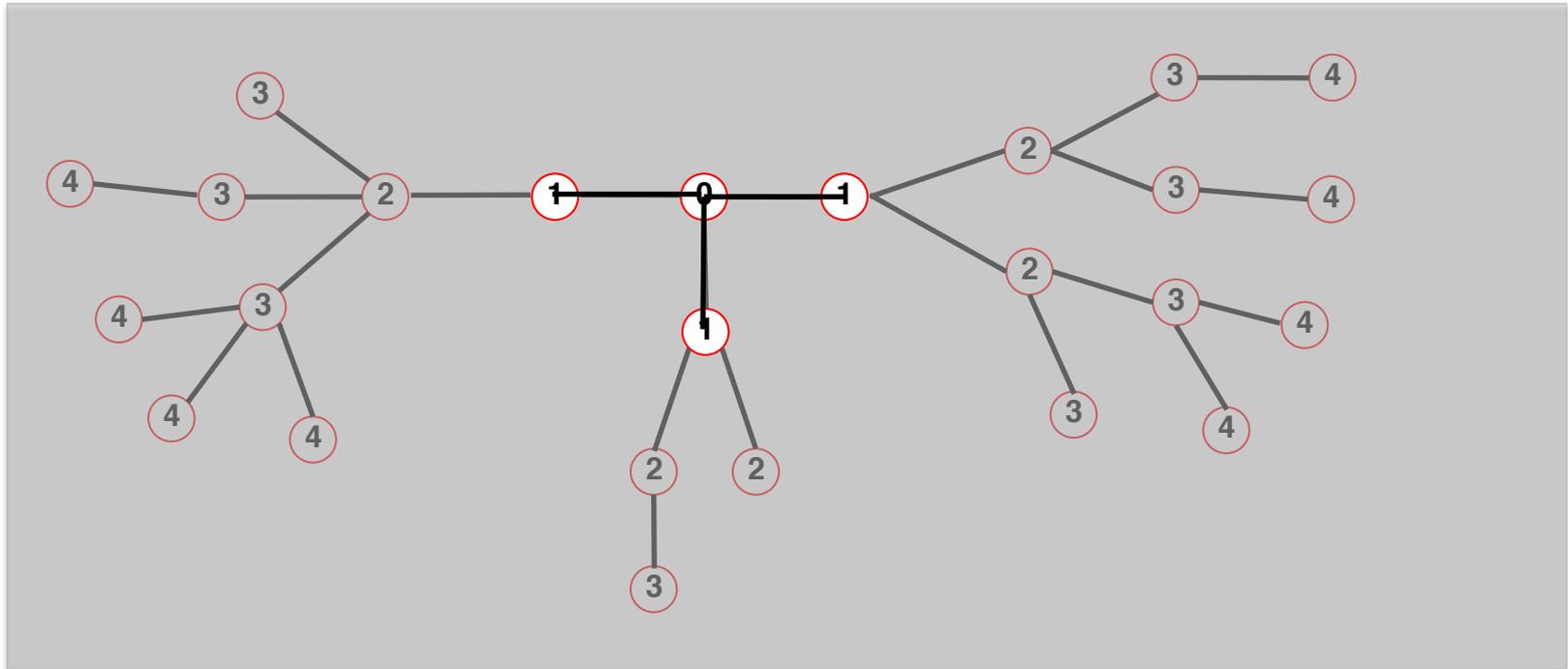


ECONTRANDO DISTANCIAS: “BREADTH FIRST SEARCH”

La distancia entre el nodo 0 y nodo 4:

1. Comienza en 0.

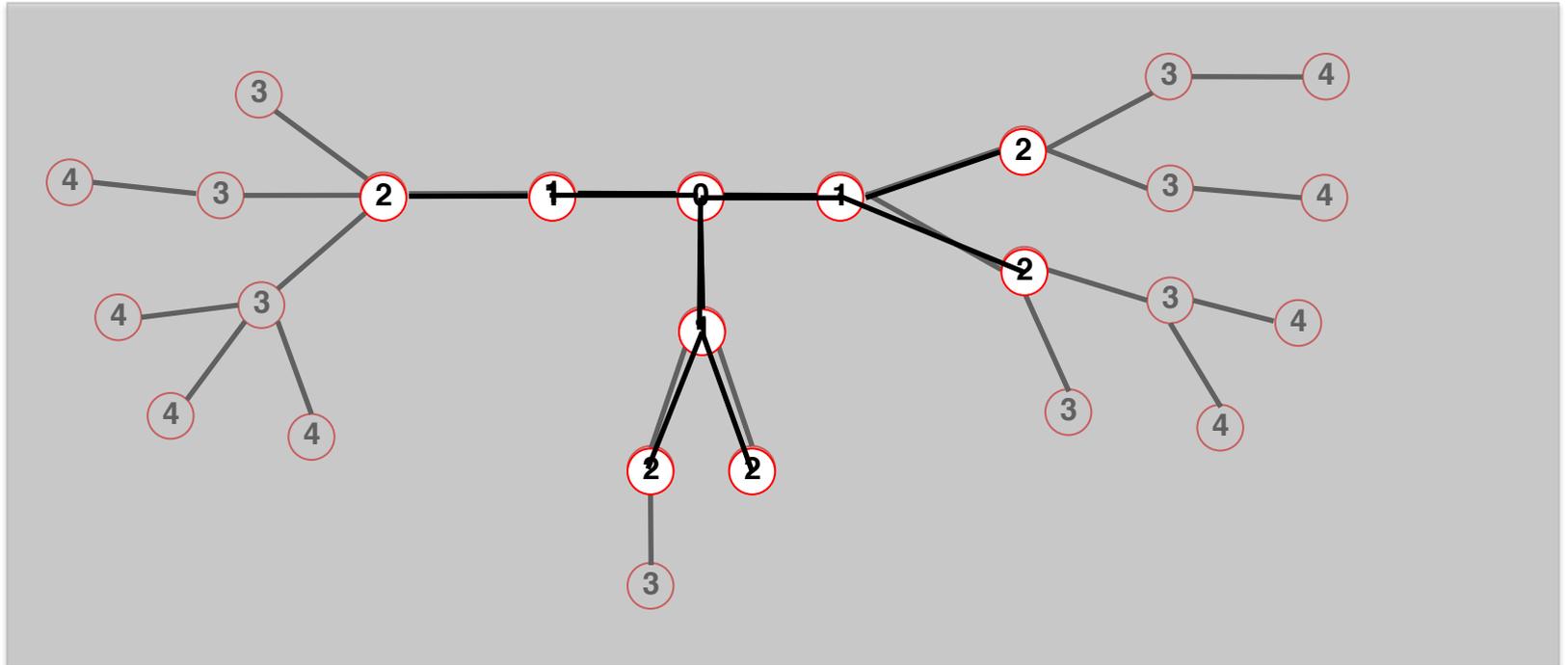
2. Encuentra los nodos adyacentes a 0. Márcalos con la etiqueta 1. Ponlos en una fila.



ECONTRANDO DISTANCIAS: “BREADTH FIRST SEARCH”

La distancia entre el nodo 0 y nodo 4:

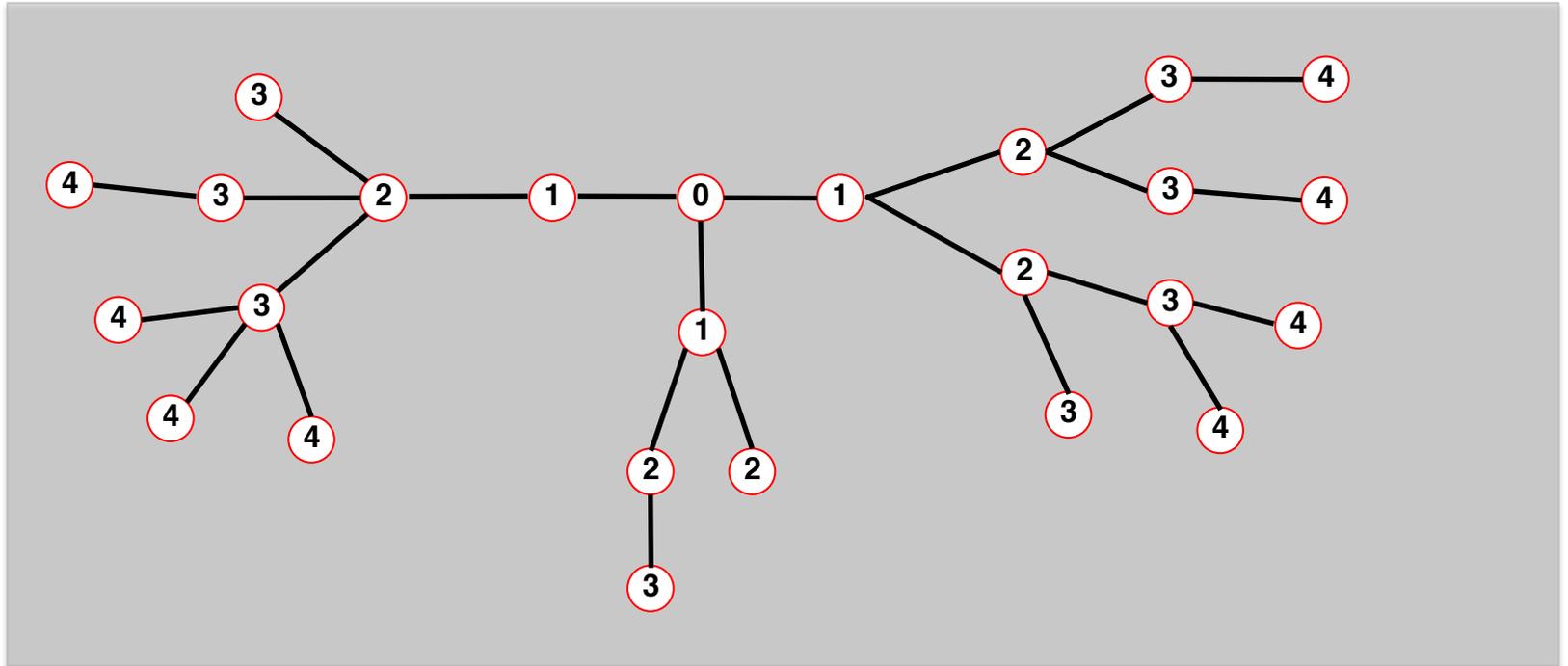
1. Comienza en 0.
2. Encuentra los nodos adyacentes a 1. Márcalos con la etiqueta 1. Pon los en una fila.
3. Toma el primer nodo de la fila. Encuentra los nodos no-marcados adyacentes en el grafo. Márcalos con la etiqueta 2. Ponlos en la fila.



ECONTRANDO DISTANCIAS: “BREADTH FIRST SEARCH”

La distancia entre el nodo 0 y nodo 4:

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



DIÁMETRO DE RED Y DISTANCIA MEDIA

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

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

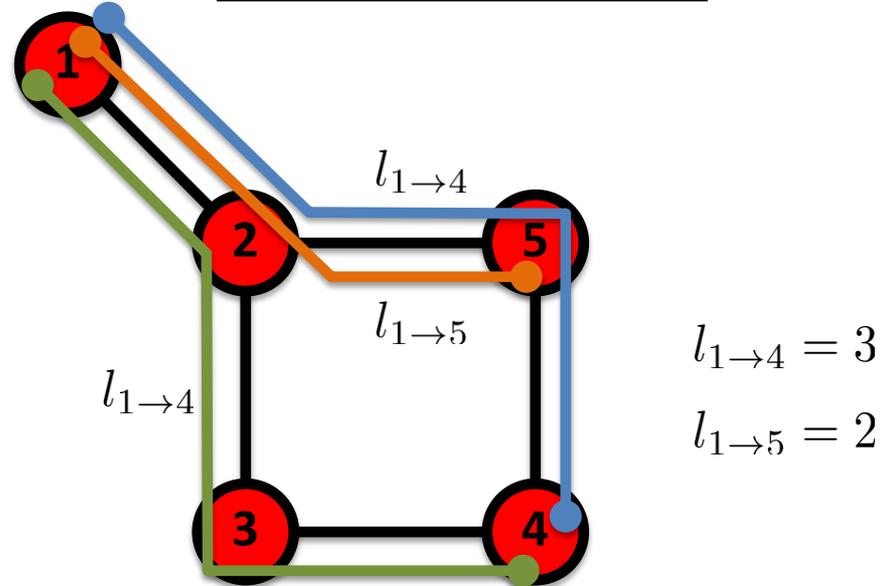
dónde d_{ij} es la distancia desde el nodo i al nodo j

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

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

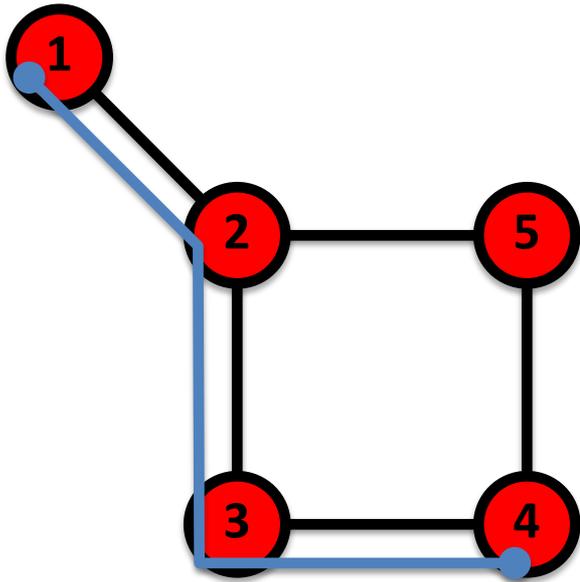
$$\langle d \rangle \equiv \frac{1}{L_{\max}} \sum_{i, j > i} d_{ij}$$

Camino más corto:



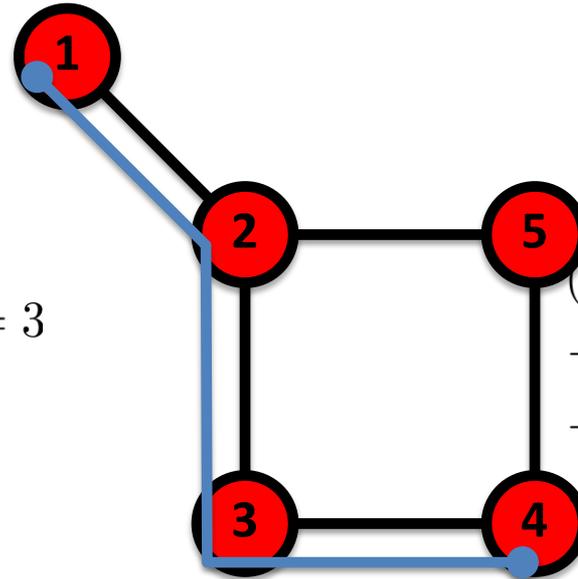
El camino con la longitud más corta entre dos nodos (distancia).

Diámetro



La distancia más larga del grafo

Longitud de camino promedio

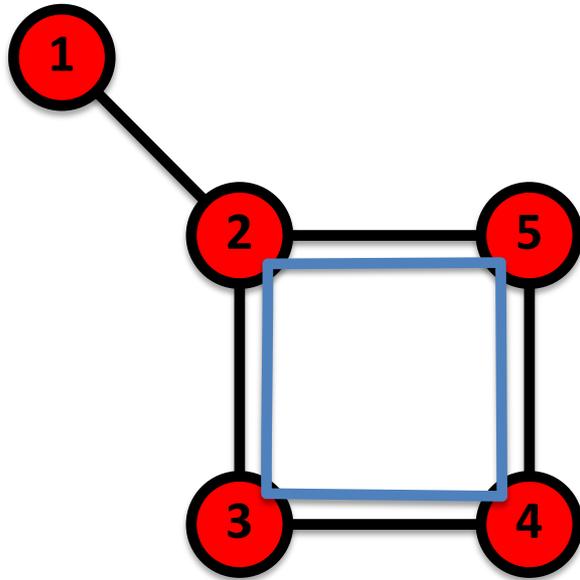


$$l_{1 \rightarrow 4} = 3$$

$$(l_{1 \rightarrow 2} + l_{1 \rightarrow 3} + l_{1 \rightarrow 4} + l_{1 \rightarrow 5} + l_{2 \rightarrow 3} + l_{2 \rightarrow 4} + l_{2 \rightarrow 5} + l_{3 \rightarrow 4} + l_{3 \rightarrow 5} + l_{4 \rightarrow 5}) / 10 = 1.6$$

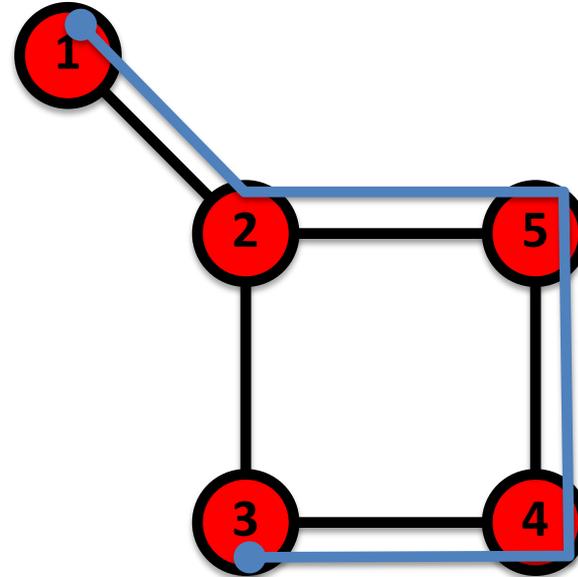
El promedio de los caminos más cortos para todos los pares de nodos

Ciclo



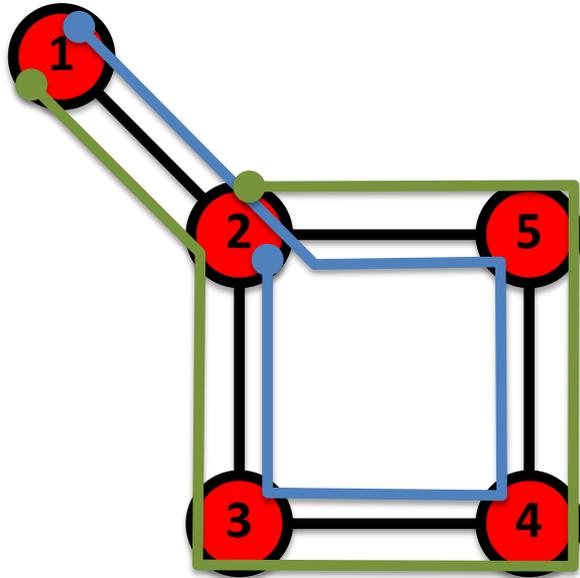
Una ruta con el mismo nodo de inicio y final.

Camino de auto-evasión



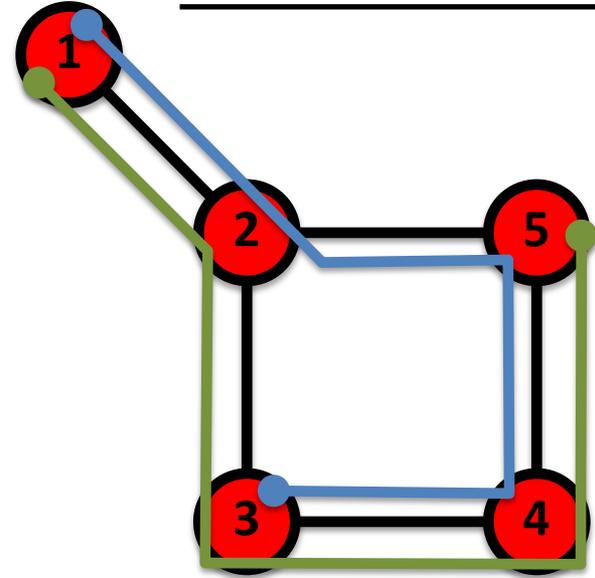
Un camino que no se intersecta.

Camino Euleriano

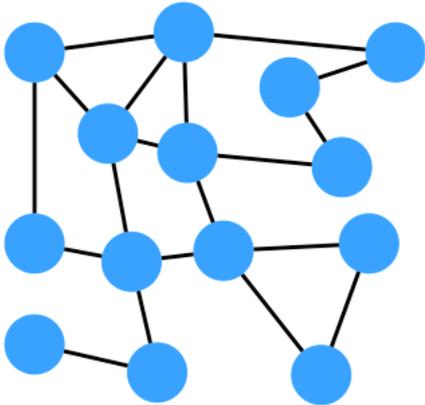


Un camino que atraviesa cada enlace exactamente una vez.

Camino Hamiltoniano



Una ruta que visita cada nodo exactamente una vez.

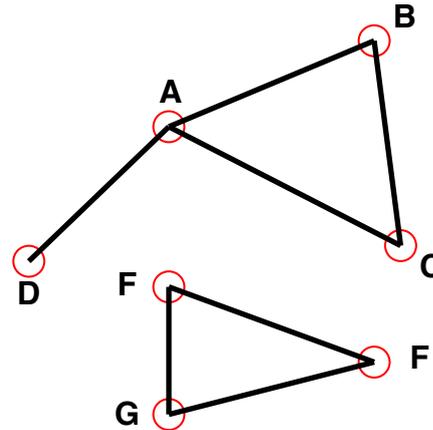
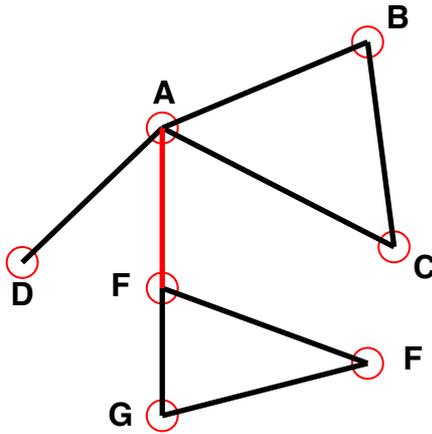


[**Importante**] La red proporciona una métrica natural para el sistema. Es decir, nos permite cuantificar distancias entre elementos.

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

CONECTIVIDAD DE GRAFOS NO-DIRIGIDOS

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

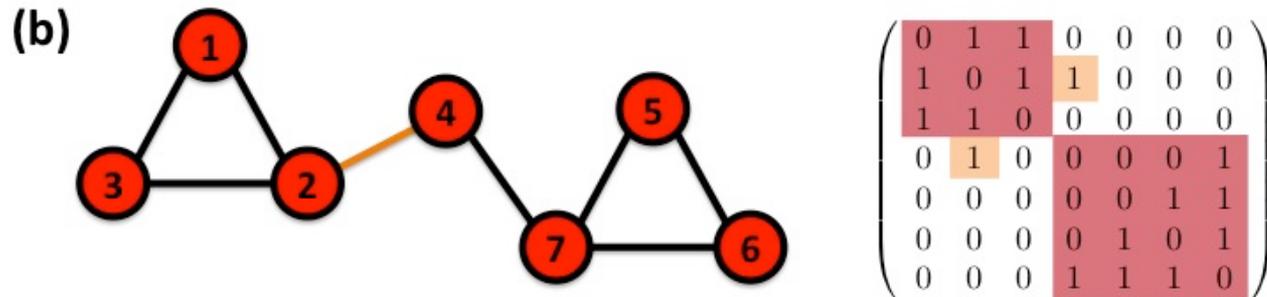
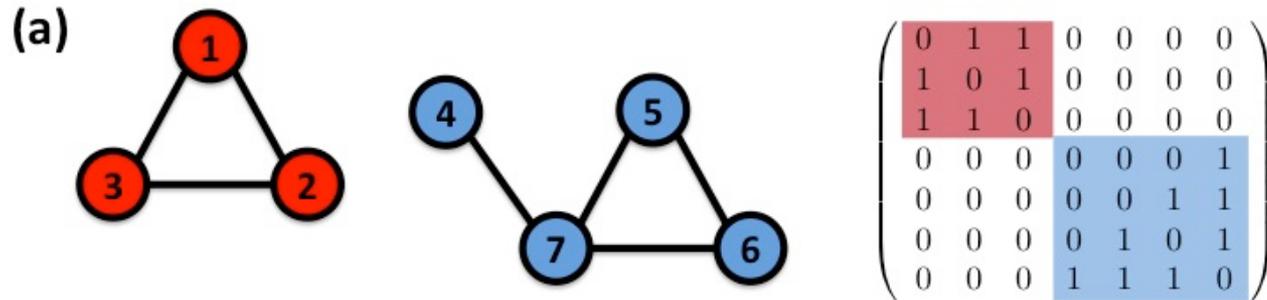


Componente más grande:
Componente gigante

El resto: **Aislados**

Puente: si lo borramos, la gráfica se desconecta.

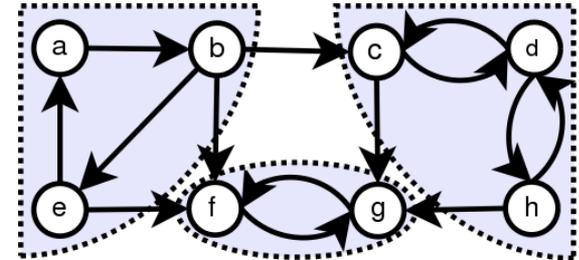
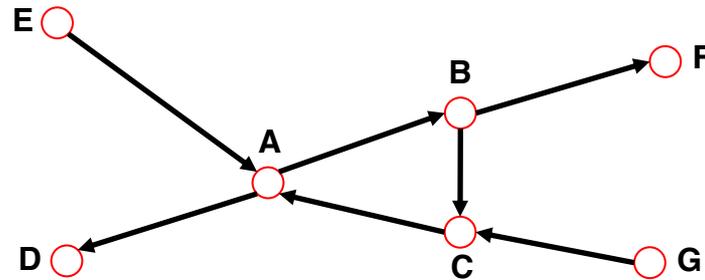
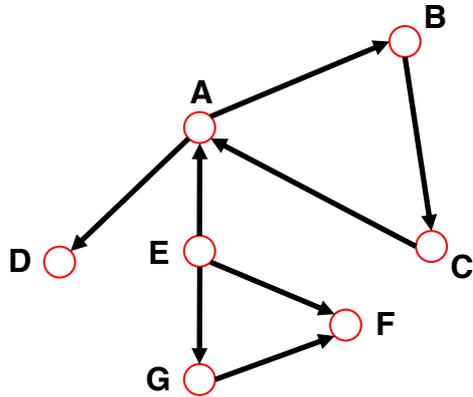
La matriz de adyacencia de una red con varios componentes se puede escribir en forma de diagonal de bloques, de modo que los elementos distintos de cero se limiten a los cuadrados, y todos los demás elementos sean cero:



CONECTIVIDAD DE GRAFOS DIRIGIDOS

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

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



In-component: nodos que pueden alcanzar el scc (strongly connected component),

Out-component: Nodos a los que se puede acceder desde el scc.

Coeficiente de Agrupamiento

* Coeficiente de clustering:

¿Qué fracción de tus vecinos están conectados?

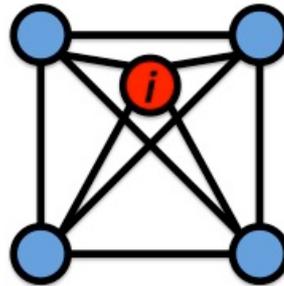
- * Nodo i con grado k_i
- * C_i entre $[0,1]$
- * e_i (links entre los vecinos del nodo i)

Dirigido

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

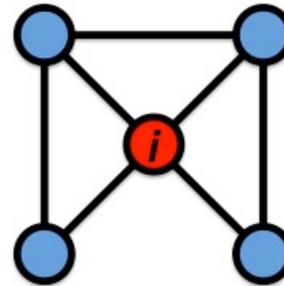
No-Dirigido

$$C_i = \frac{2|\{e_{jk}\}|}{k_i(k_i - 1)} : v_j, v_k \in N_i, e_{jk} \in E.$$



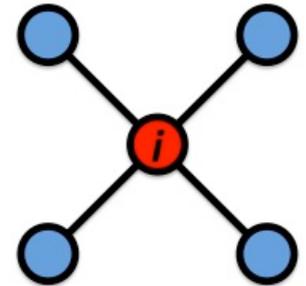
$$C_i = 1$$

$$2 \cdot 6 / (4 \cdot 3) = 1$$



$$C_i = 1/2$$

$$2 \cdot 3 / (4 \cdot 3) = 1/2$$



$$C_i = 0$$

Watts & Strogatz, Nature 1998.

* Coeficiente de clustering:

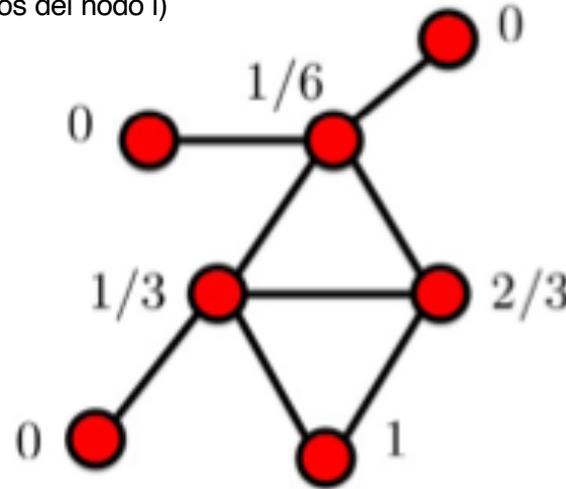
¿Qué fracción de tus vecinos están conectados?

- * Nodo i con grado k_i
- * C_i entre $[0,1]$
- * e_i (links entre los vecinos del nodo i)

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

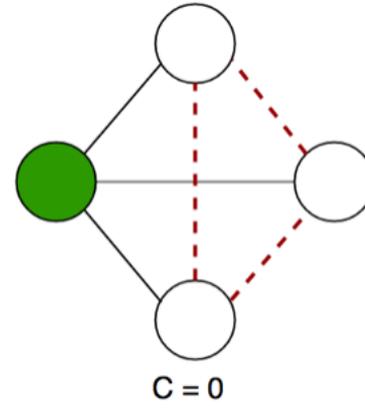
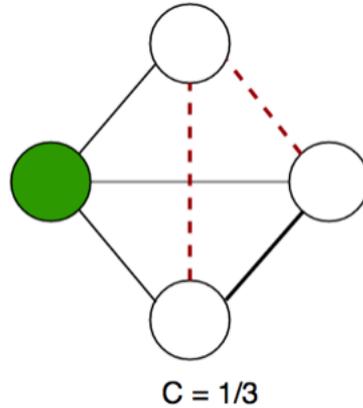
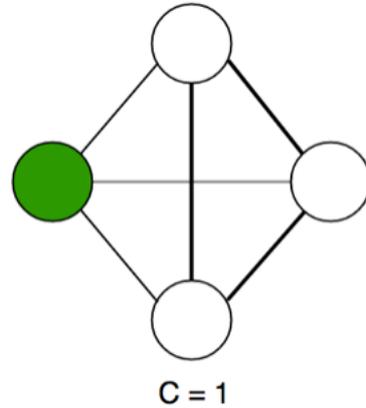
$$k_i = \sum_j A_{ij}$$

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



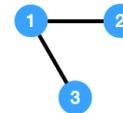
$$\langle C \rangle = \frac{13}{42} \approx 0.310$$

COEFICIENTE DE CLUSTERING

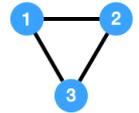


$$C = \frac{\text{number of closed triplets}}{\text{number of all triplets (open and closed)}}.$$

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

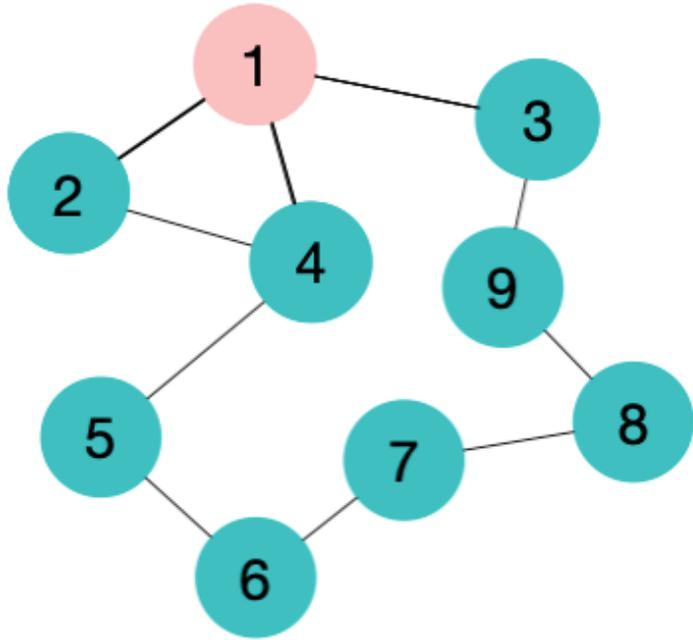


Open Triangle
(triplet)



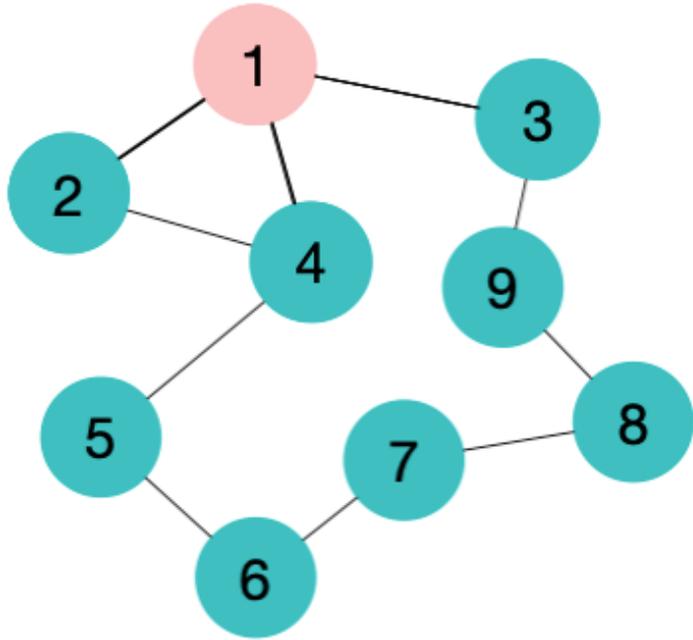
Closed Triangle
(closed triplet)

COEFICIENTE DE CLUSTERING

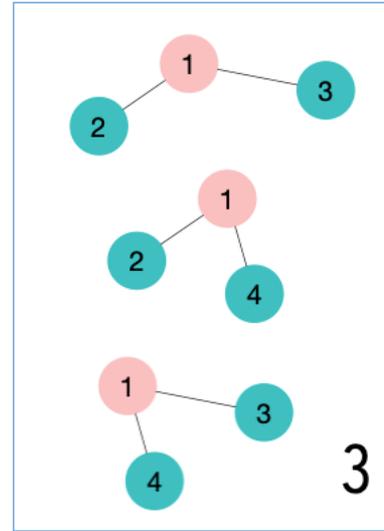


¿Cuál es el CC del nodo 1?

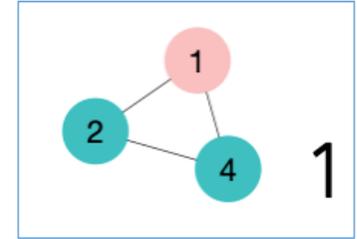
COEFICIENTE DE CLUSTERING



Tripletes



Triángulos Cerrados



Coeficiente de Clustering:

$$1/3=0.33$$

Clustering Coefficients for Correlation Networks



Naoki Masuda^{1*}



Michiko Sakaki^{2,3}



Takahiro Ezaki⁴ and



Takamitsu Watanabe⁵

¹ Department of Engineering Mathematics, University of Bristol, Bristol, United Kingdom

² School of Psychology and Clinical Language Sciences, University of Reading, Reading, United Kingdom

³ Research Institute, Kochi University of Technology, Kochi, Japan

⁴ PRESTO, Japan Science and Technology Agency, Kawaguchi, Japan

⁵ Institute of Cognitive Neuroscience, University College London, London, United Kingdom

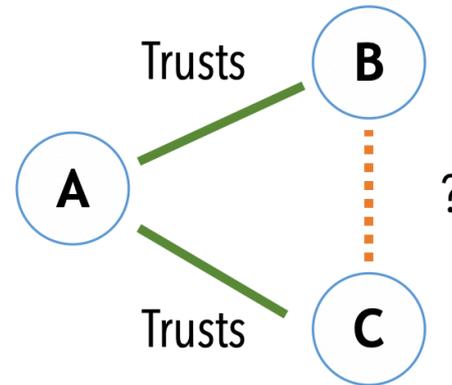
Graph theory is a useful tool for deciphering structural and functional networks of the brain on various spatial and temporal scales. The clustering coefficient quantifies the abundance of connected triangles in a network and is a major descriptive statistics of networks. For example, it finds an application in the assessment of small-worldness of brain networks, which is affected by attentional and cognitive conditions, age, psychiatric disorders and so forth. However, it remains unclear how the clustering coefficient should be measured in a correlation-based network, which is among major representations of brain networks. In the present article, we propose clustering coefficients tailored to correlation matrices. The key idea is to use three-way partial correlation or partial mutual information to measure the strength of the association between the two neighboring nodes of a focal node relative to the amount of pseudo-correlation expected from indirect paths between the nodes. Our method avoids the difficulties of previous applications of clustering coefficient (and other) measures in defining correlational networks, i.e., thresholding on the correlation value, discarding of negative correlation values, the pseudo-correlation problem and full partial correlation matrices whose estimation is computationally difficult. For proof of concept, we apply the proposed clustering coefficient measures to functional magnetic resonance imaging data obtained from healthy participants of various ages and compare them with conventional clustering coefficients. We show that the clustering coefficients decline with the age. The proposed clustering coefficients are more strongly correlated with age than the conventional ones are. We also show that the local variants of the proposed clustering coefficients (i.e., abundance of triangles around a focal node) are useful in characterizing individual nodes. In contrast, the conventional local clustering coefficients were strongly correlated with and therefore may be confounded by the node's connectivity. The proposed methods are expected to help us to understand clustering and lack thereof in correlational brain networks, such as those derived from functional time series and across-participant correlation in neuroanatomical properties.

Strength of Weak Ties



Mark Granovetter

Triadic closure in Social Relationships



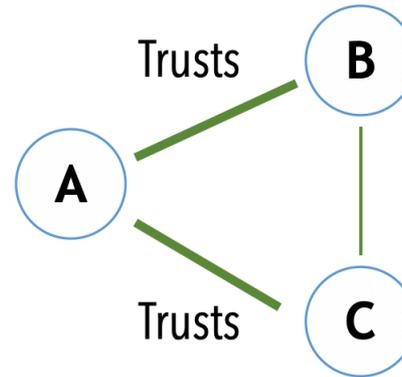
¿Podemos predecir vínculos entre individuos?
La idea de cierre triádico se vincula con la idea de Coeficiente de Clúster.

Strength of Weak Ties



Mark Granovetter

Cierre Triádico en las Relaciones Sociales



B y C deberían tener cierta confianza latente sobre la base de la relación con A

La primera premisa de la teoría es que cuanto más fuerte es el vínculo entre dos personas, más probable es que sus mundos sociales se superpongan.

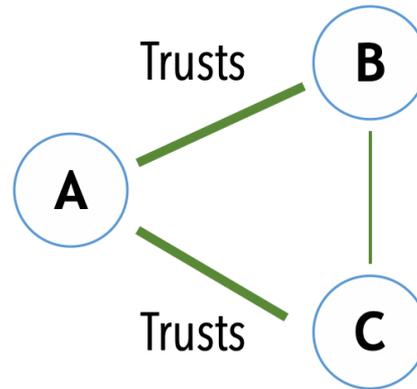
Las causas subyacentes de la formación de vínculos tienen este tipo de transitividad incorporada.

Strength of Weak Ties

Cierre Triádico en las Relaciones Sociales



Mark Granovetter



B y C deberían tener cierta confianza latente sobre la base de la relación con A

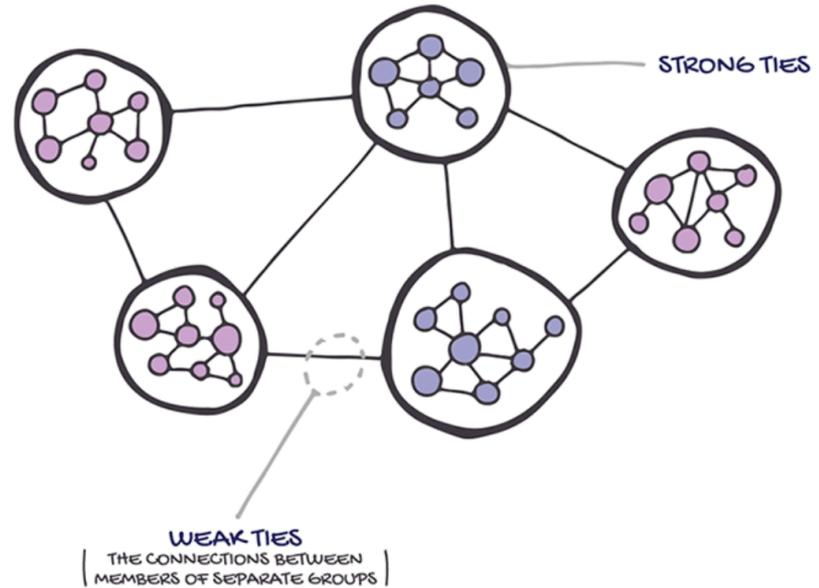
La segunda premisa de SWT es que los lazos de unión son una fuente potencial de ideas novedosas.

A través de un lazo puente, una persona puede escuchar cosas que aún no están circulando entre sus amigos cercanos.

Strength of Weak Ties



Mark Granovetter



Los lazos débiles son los responsables de la difusión de las innovaciones.

The Strength of Weak Ties¹

Mark S. Granovetter

Johns Hopkins University



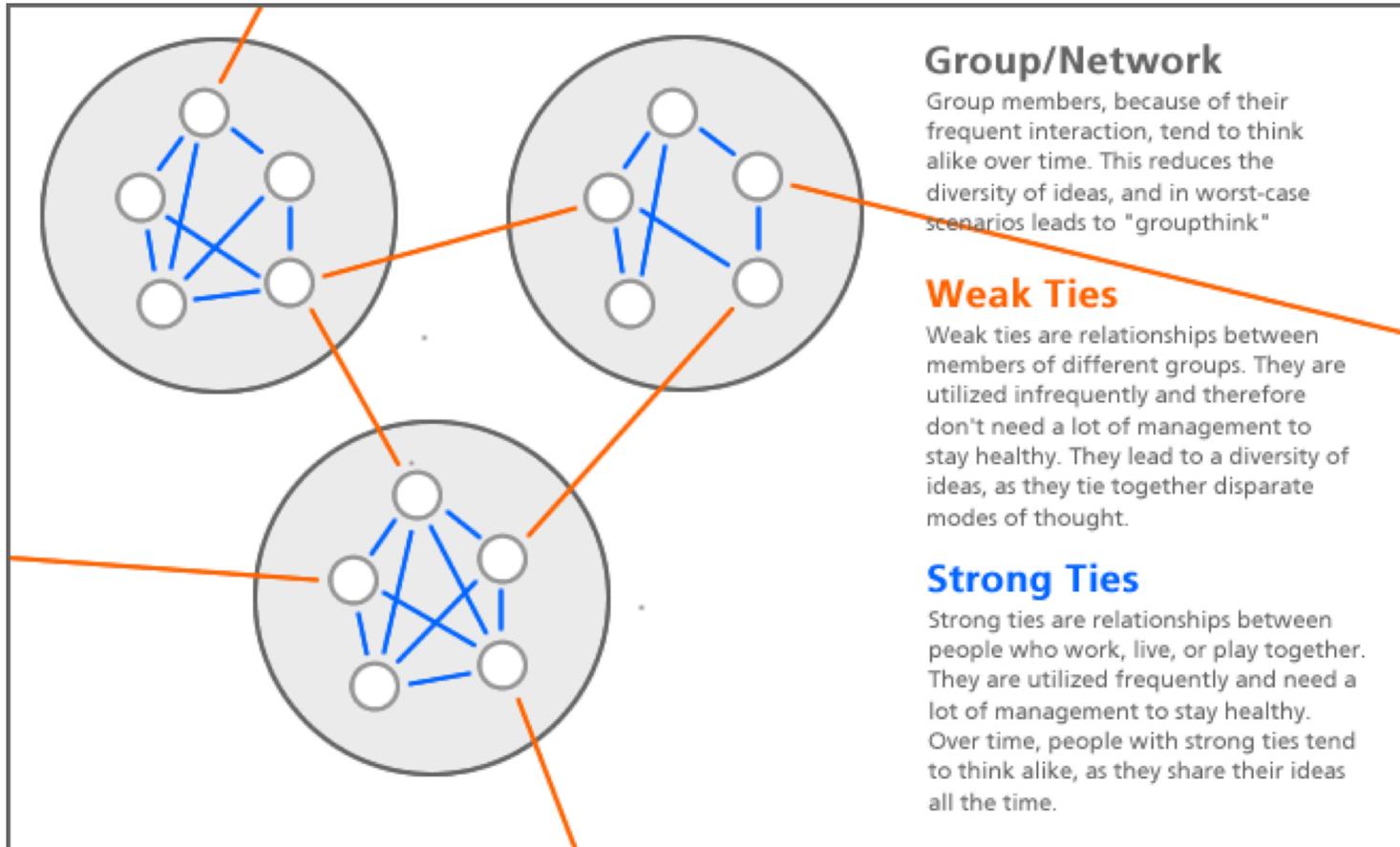
Mark Granovetter

Analysis of social networks is suggested as a tool for linking micro and macro levels of sociological theory. The procedure is illustrated by elaboration of the macro implications of one aspect of small-scale interaction: the strength of dyadic ties. It is argued that the degree of overlap of two individuals' friendship networks varies directly with the strength of their tie to one another. The impact of this principle on diffusion of influence and information, mobility opportunity, and community organization is explored. Stress is laid on the cohesive power of weak ties. Most network models deal, implicitly, with strong ties, thus confining their applicability to small, well-defined groups. Emphasis on weak ties lends itself to discussion of relations *between* groups and to analysis of segments of social structure not easily defined in terms of primary groups.

Uno de los artículos más citados en Sociología.

Los conceptos expuestos en su obra seminal son fundamentales...

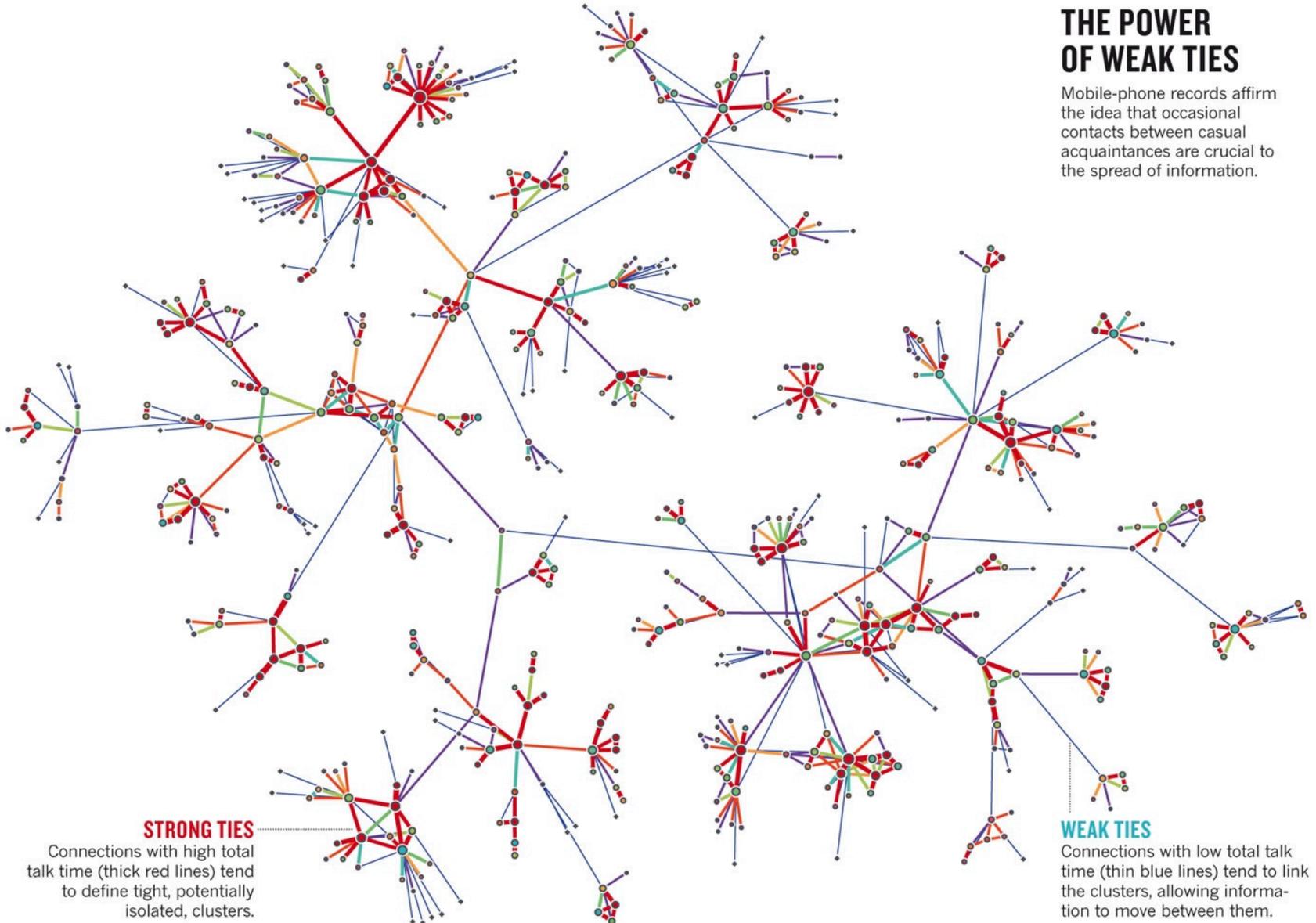
Altamente influyente en marketing, sistemas de información y política, ayudando a los investigadores a ofrecer productos que pueden llegar a audiencias poco probables.



<http://bokardo.com/archives/weak-ties-and-diversity-in-social-networks/>

THE POWER OF WEAK TIES

Mobile-phone records affirm the idea that occasional contacts between casual acquaintances are crucial to the spread of information.



STRONG TIES

Connections with high total talk time (thick red lines) tend to define tight, potentially isolated, clusters.

WEAK TIES

Connections with low total talk time (thin blue lines) tend to link the clusters, allowing information to move between them.

A causal test of the strength of weak ties

KARTHIK RAJKUMAR [ID](#), GUILLAUME SAINT-JACQUES [ID](#), IAVOR BOJINOV [ID](#), ERIK BRYNJOLFSSON [ID](#), AND SINAN ARAL [ID](#) [Authors Info & Affiliations](#)

SCIENCE · 15 Sep 2022 · Vol 377, Issue 6612 · pp. 1304-1310 · DOI:10.1126/science.abl4476

↓ 7,997 [”](#) 1

CHECK ACCESS

RELATED PERSPECTIVE

Weak ties, failed tries, and success

BY DASHUN WANG, BRIAN UZZI

The influence of weak associations

Abstract

Supplementary Materials

References and Notes

e Letters (1)

The influence of weak associations

The strength of weak ties is an influential social-scientific theory that stresses the importance of weak associations (e.g., acquaintance versus close friendship) in influencing the transmission of information through social networks. However, causal tests of this paradoxical theory have proved difficult. Rajkumar *et al.* address the question using multiple large-scale, randomized experiments conducted on LinkedIn’s “People You May Know” algorithm, which recommends connections to users (see the Perspective by Wang and Uzzi). The experiments showed that weak ties increase job transmissions, but only to a point, after which there are diminishing marginal returns to tie weakness. The authors show that the weakest ties had the greatest impact on job mobility, whereas the strongest ties had the least. Together, these results help to resolve the apparent “paradox of weak ties” and provide evidence of the strength of weak ties theory. —AMS



Tres métricas centrales en Ciencia de Redes

Distribución de grado:

P(k)

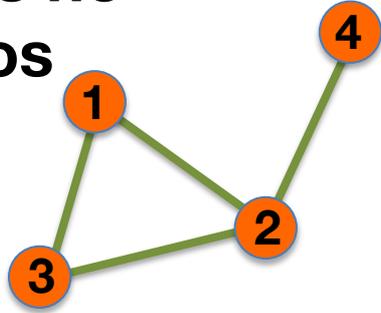
Longitud de camino:

$\langle d \rangle$

Coeficiente de Agrupamiento:

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

Enlaces no dirigidos



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

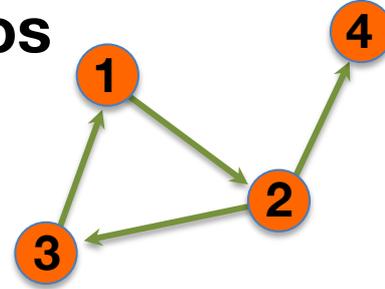
$$A_{ii} = 0$$

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

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

Actor network, protein-protein interactions

Enlaces dirigidos



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

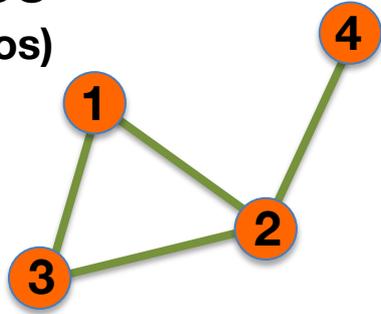
$$A_{ii} = 0$$

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

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

WWW, citation networks

Sin peso (no dirigidos)



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

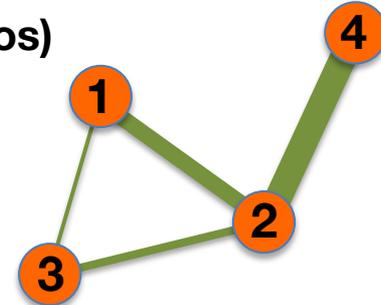
$$A_{ii} = 0$$

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

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

$$\langle k \rangle = \frac{2L}{N}$$

Con peso (no dirigidos)



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

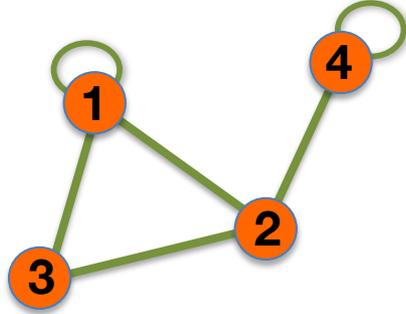
$$A_{ii} = 0$$

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

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

$$\langle k \rangle = \frac{2L}{N}$$

Auto-interacciones



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

$$A_{ii} \neq 0$$

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

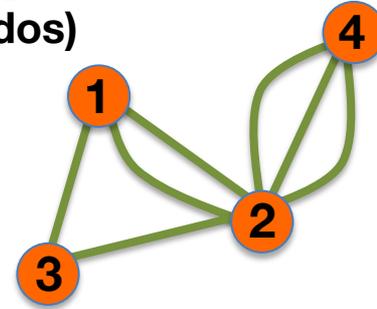
$$L = \frac{1}{2} \sum_{i,j=1, i \neq j}^N A_{ij} + \sum_{i=1}^N A_{ii}$$

?

Protein interaction network, www

Multi-grafos

(no dirigidos)



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

$$A_{ii} = 0$$

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

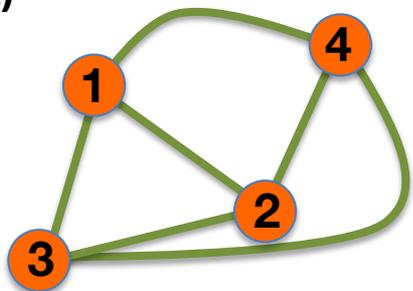
$$L = \frac{1}{2} \sum_{i,j=1}^N \text{nonzero}(A_{ij})$$

$$\langle k \rangle = \frac{2L}{N}$$

Social networks, collaboration networks

Red Completa

(no dirigida)

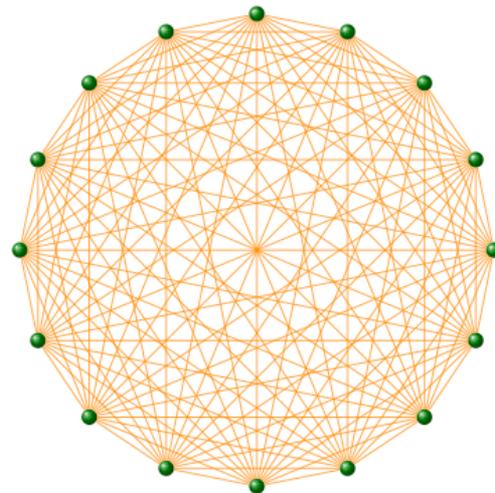


$$A_{ij} = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$

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

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

Actor network, protein-protein interactions



Grafología: Las redes reales pueden tener multiples características

WWW > dirigida, multi-grafo, con auto-interacciones

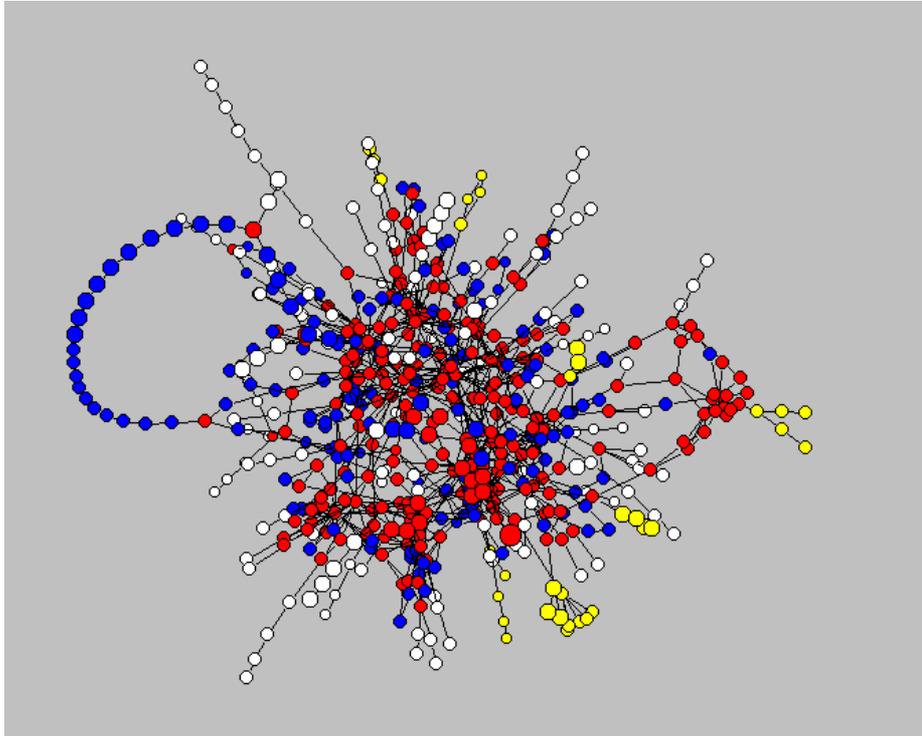
Interacciones de proteínas > no dirigida, sin peso, con auto-interacciones

Redes de colaboración > no dirigida, multi-grafo con peso

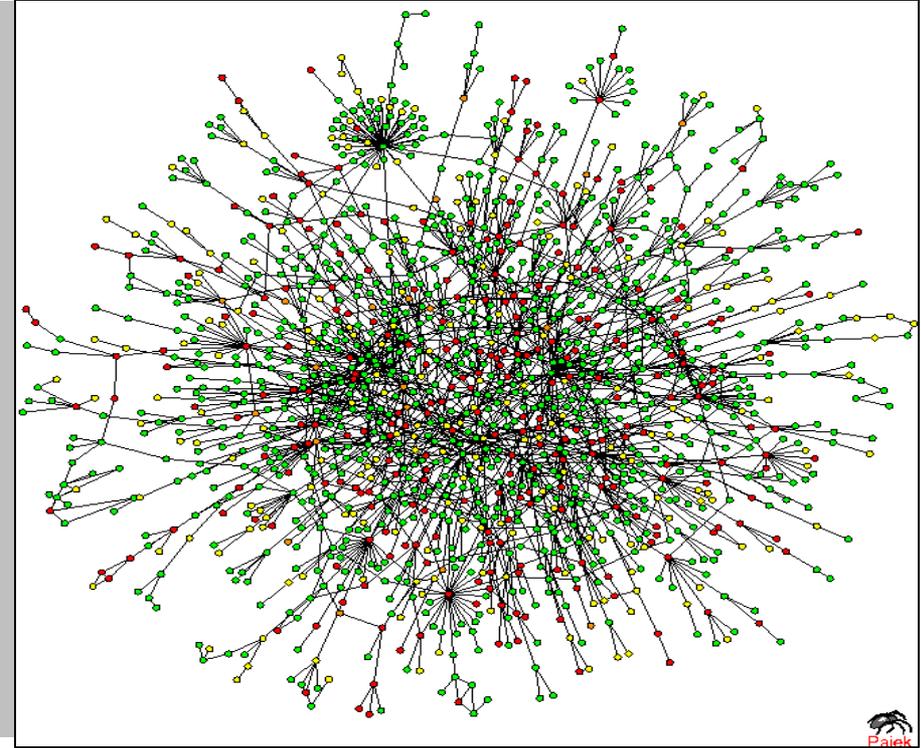
Llamadas de teléfono > dirigida, con peso

Amistades en Facebook > no dirigida, sin peso

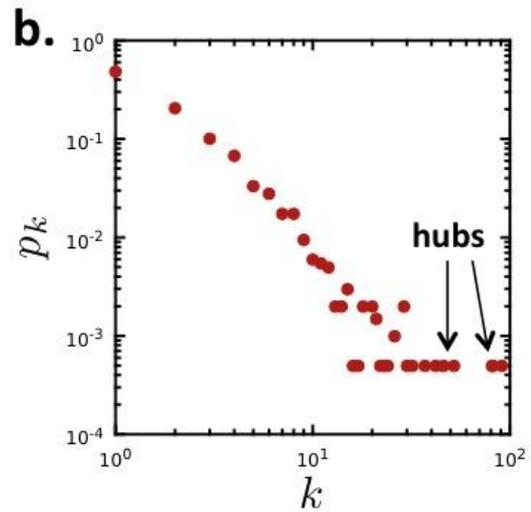
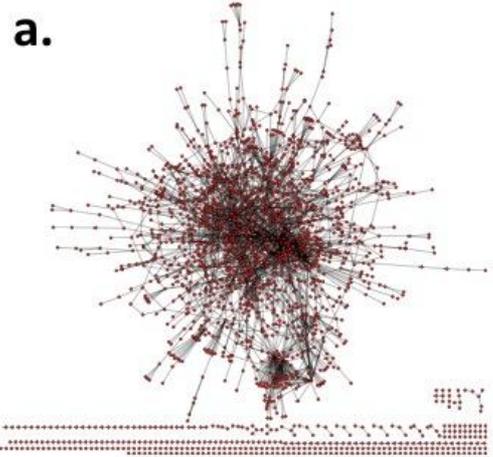
Red metabólica



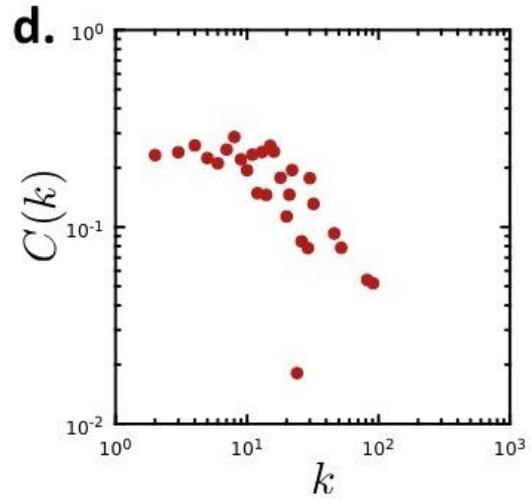
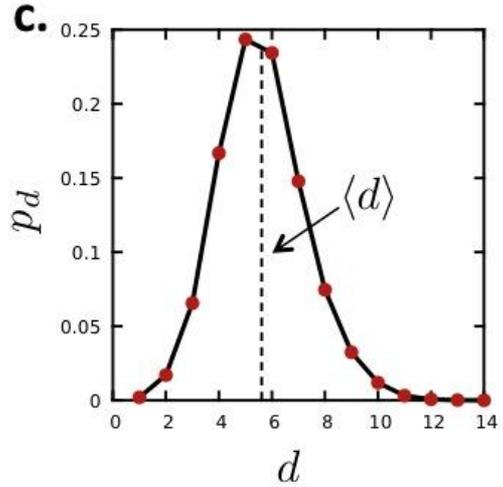
Interacciones de proteínas



Estudio de caso: red de interacciones de proteínas

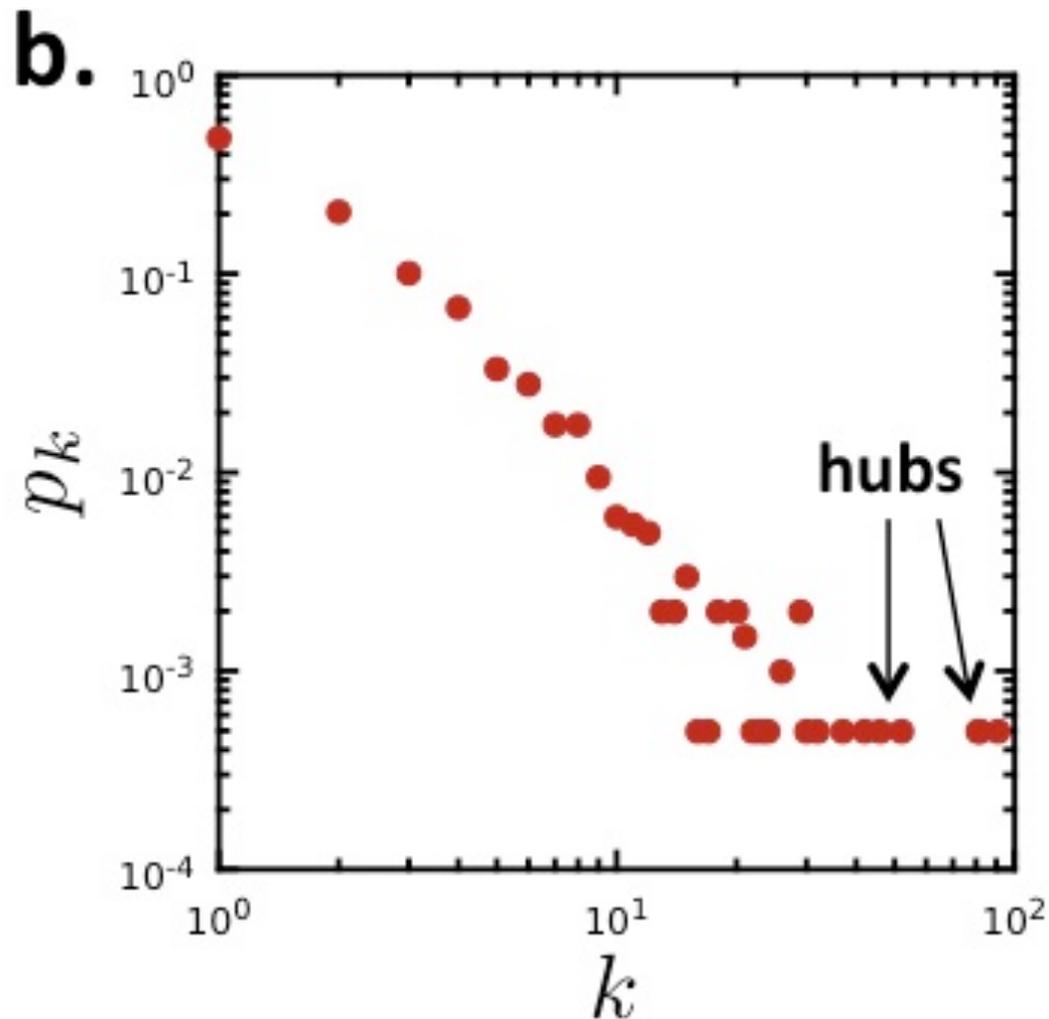


Red no dirigida
N=2,018 las proteínas son nodos
L=2,930 interacciones de union como enlace
Grado promedio $\langle k \rangle = 2.90$.



No conectado: 185 componentes
El componente gigante tiene 1,647 nodos

Estudio de caso: red de interacciones de proteínas

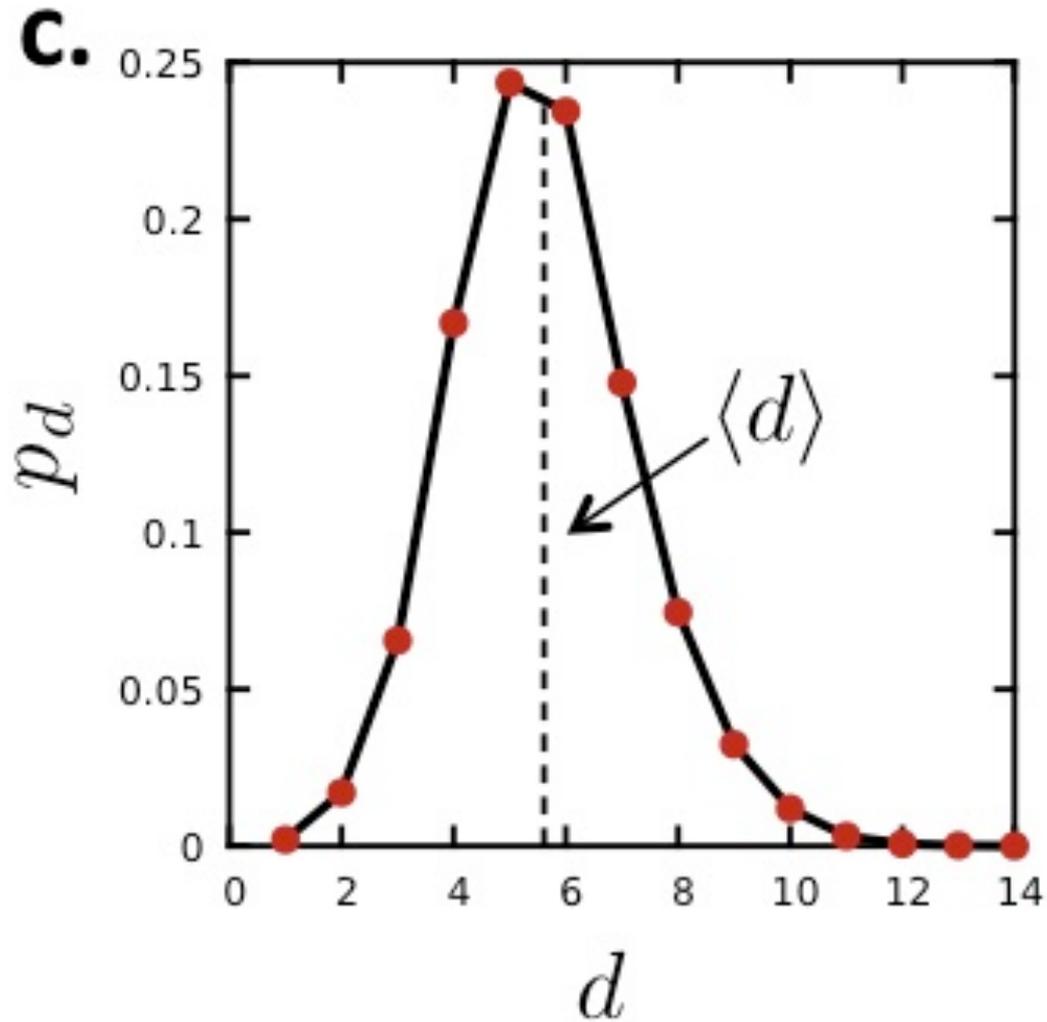


p_k es la probabilidad de que un nodo tenga grado k .

$N_k = \#$ nodos con grado k

$$p_k = N_k / N$$

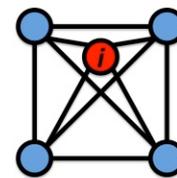
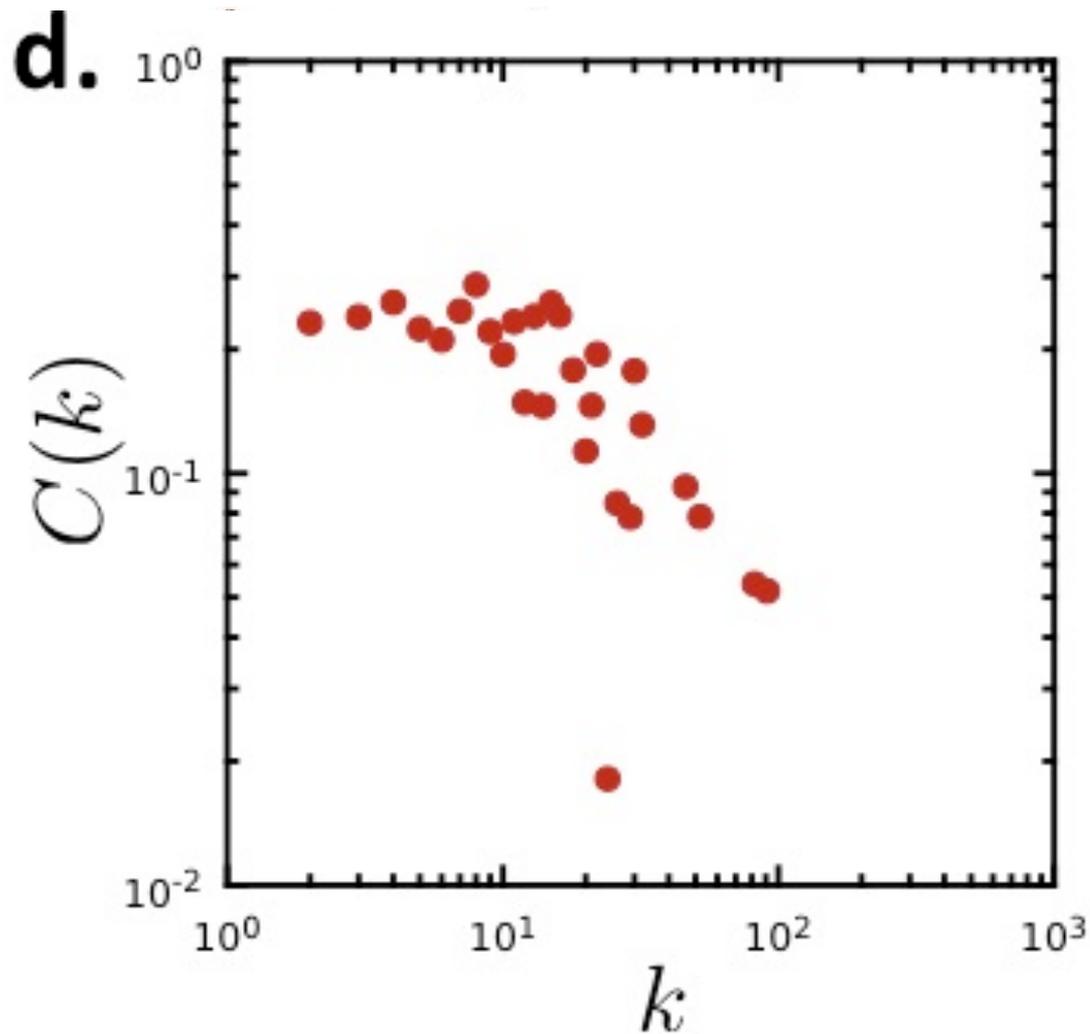
Estudio de caso: red de interacciones de proteínas



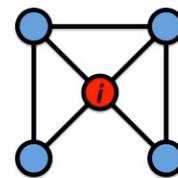
$$d_{\max}=14$$

$$\langle d \rangle = 5.61$$

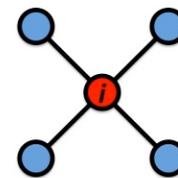
Estudio de caso: red de interacciones de proteínas



$$C_i = 1$$



$$C_i = 1/2$$



$$C_i = 0$$

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

$$\langle C \rangle = 0.12$$

Preguntas básicas

- Cómo son los diámetros de tres redes de 500 nodos si cada red se formó de manera distinta: aleatoria, mundo pequeño, y libre de escala?
- Tienes una red aleatoria de N nodos. Qué pasa con las propiedades estructurales de esta red cuando cambias el modelo de formación a una estructura tipo árbol?
- Cuál es la utilidad de la medida de coeficiente de agrupamiento?
- Describe la relación entre clustering coeficiente, longitud de camino, y la estructura de una red en términos de propiedades generales.
- Proporciona un ejemplo de red real y discute sus propiedades. Qué utilidad tienen las propiedades de red para entender el contexto real de la red? Puedes sacar inferencia a partir de estas medidas? Explica.