Modelado y Simulación de Sistemas Complejos con Aplicaciones en Economía

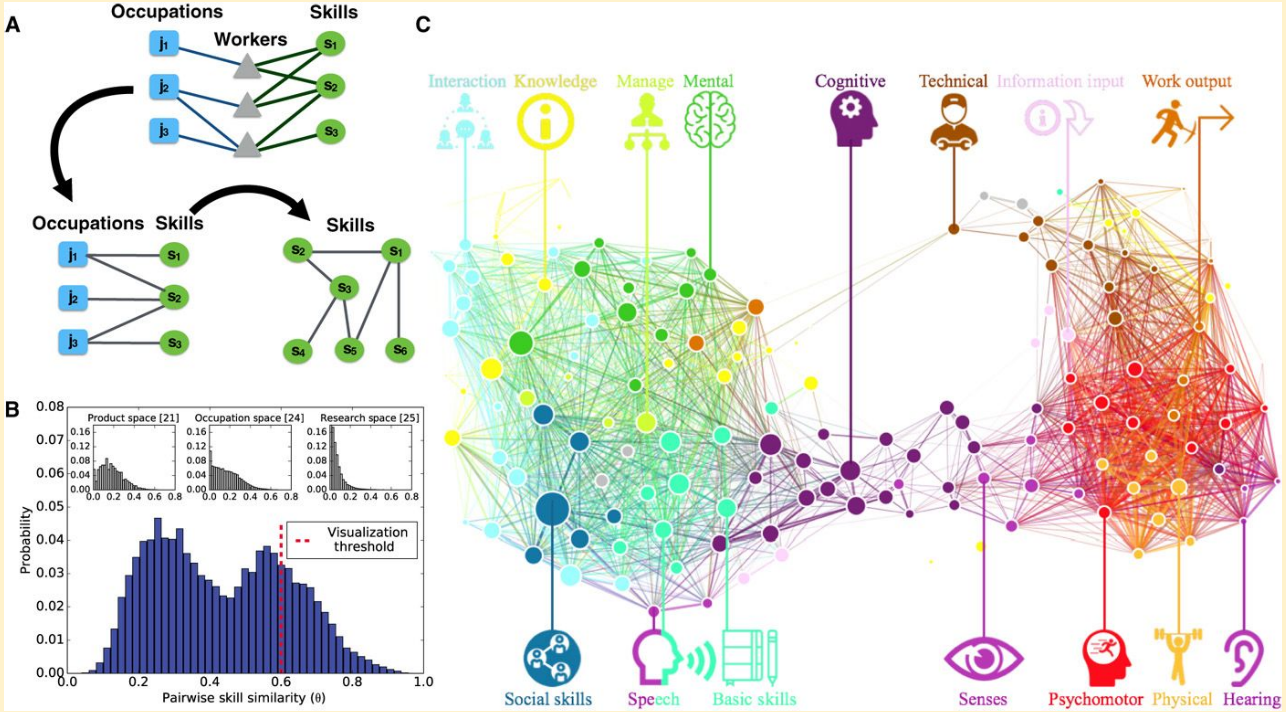
Clase 5

Redes

1C 2024-FCEyN, UBA

Contenido de la clase 5

- Introducción a Redes
- Propiedades, Topología
- Práctica: Product Space

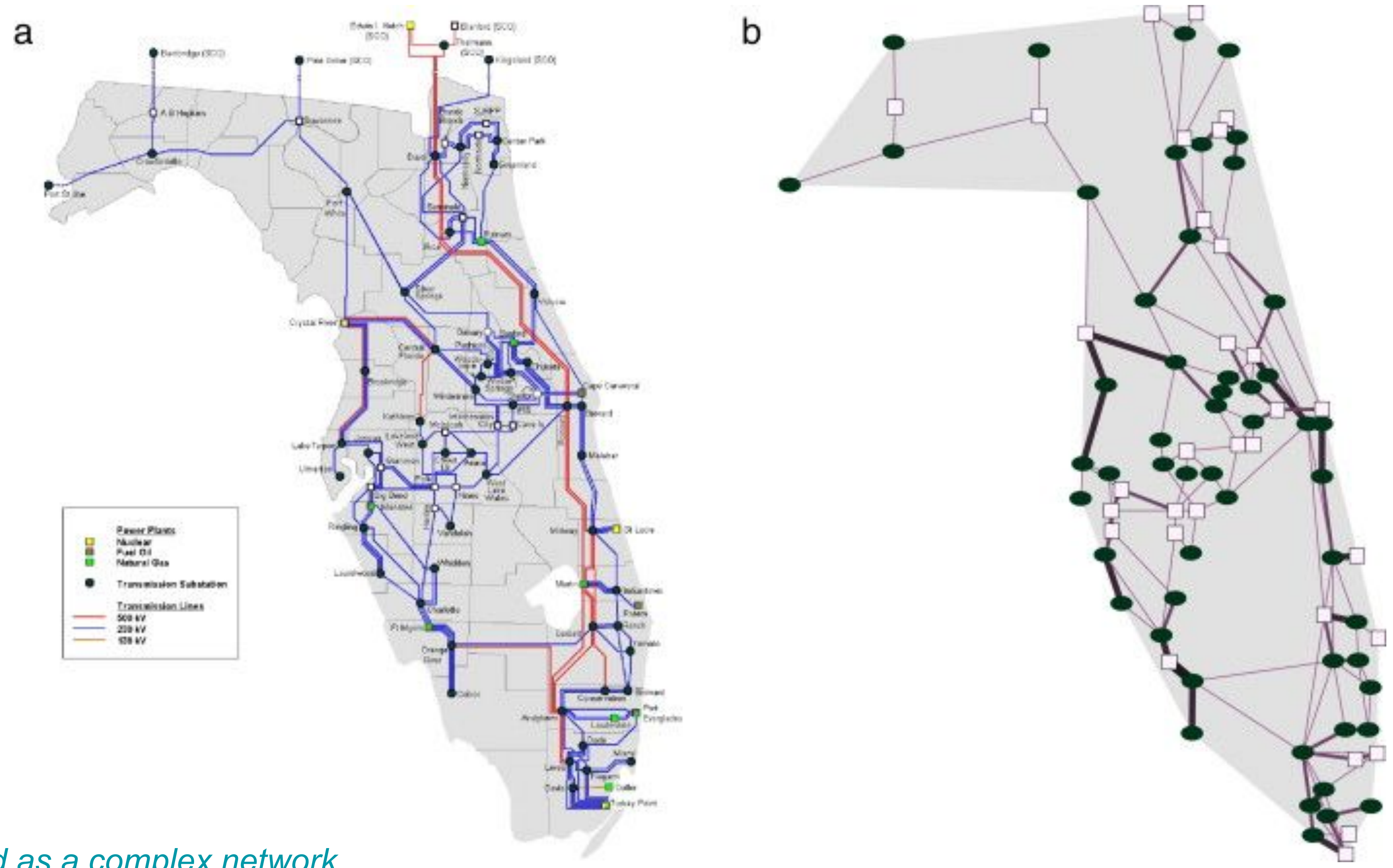


Introducción a Redes

1C 2024-FCEyN, UBA

¿qué son redes?

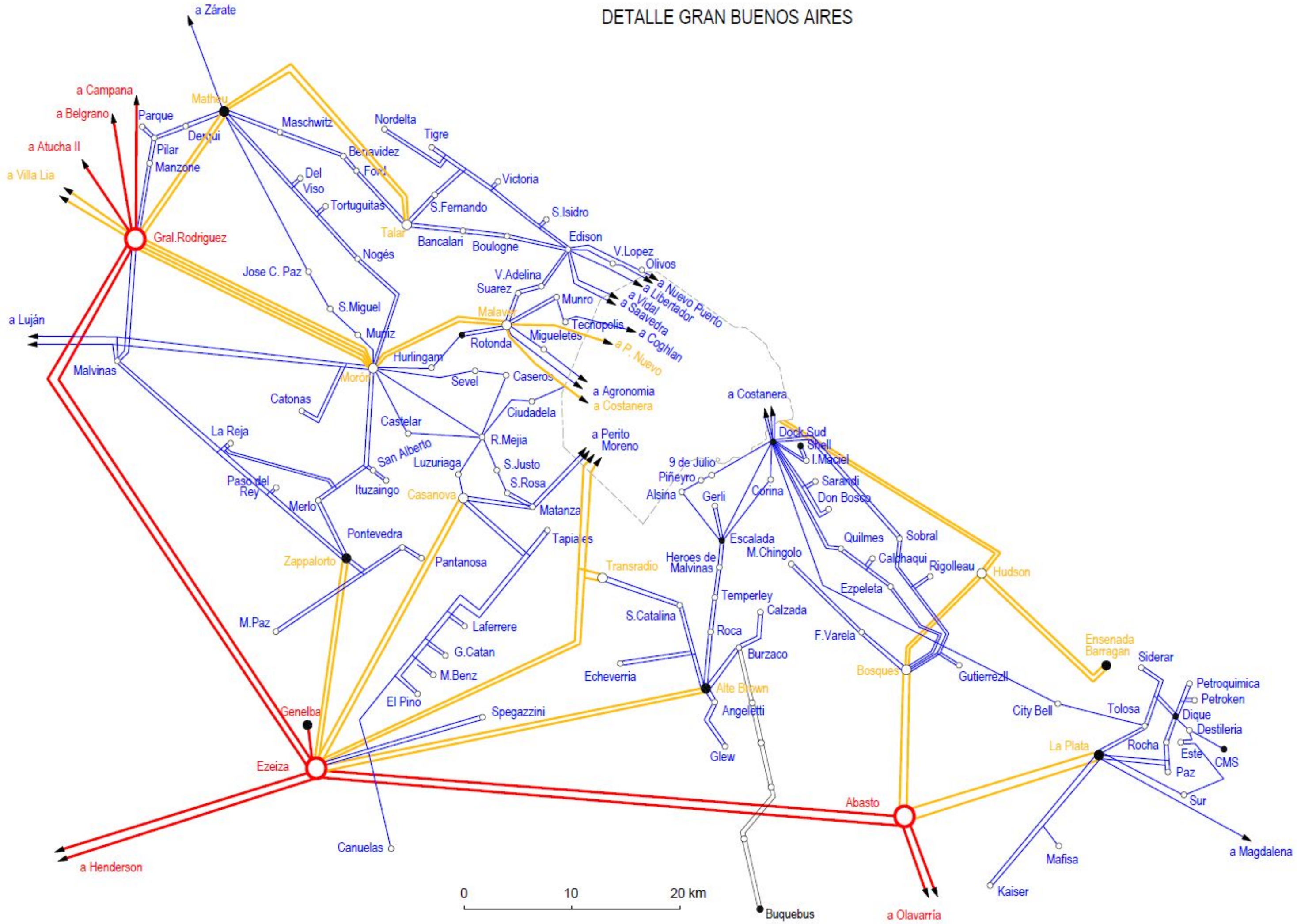
- Redes eléctricas
- Redes de Transporte
- Redes Sociales
- Redes Financieras
- Redes de aeropuerto
- Redes neuronales …

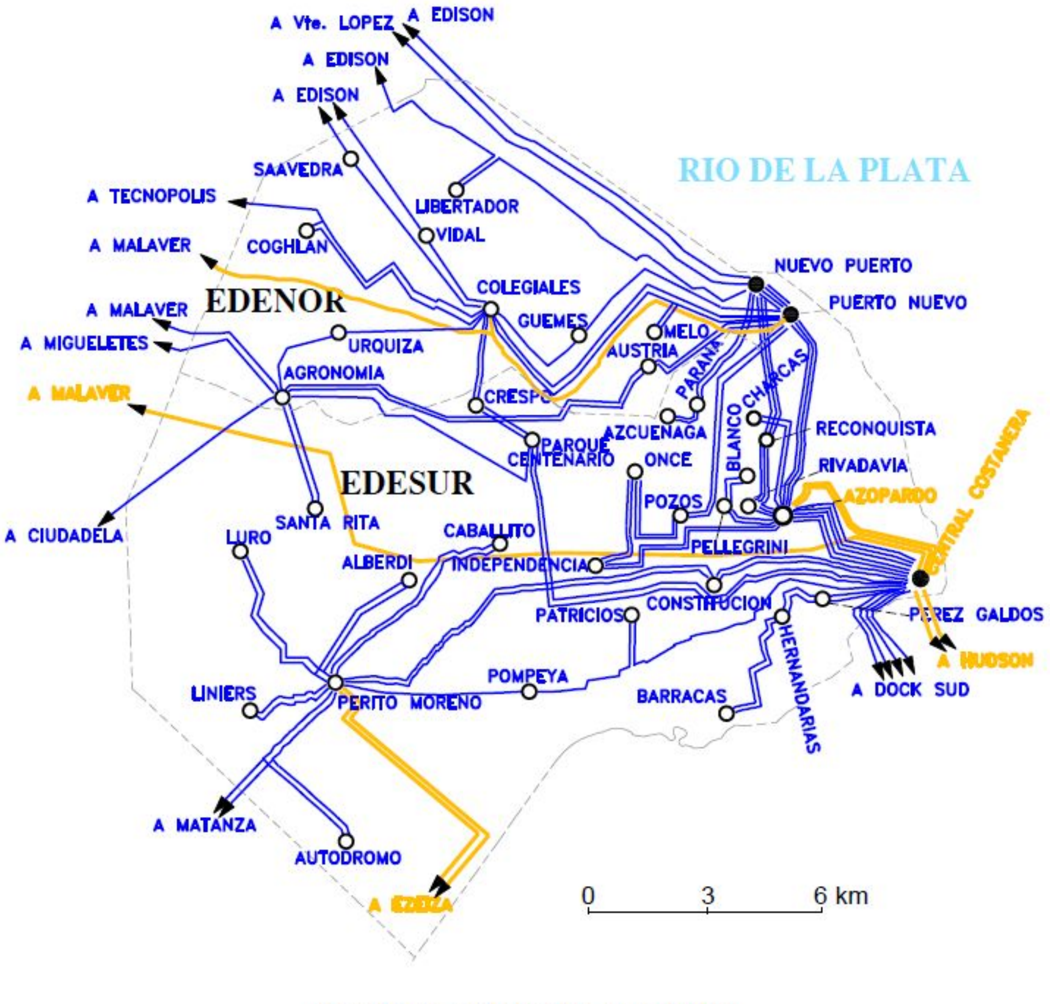


Fuente: [Architecture of the Florida power grid as a complex network]

GCABA

(Fuente: CAMMESA, 2017)

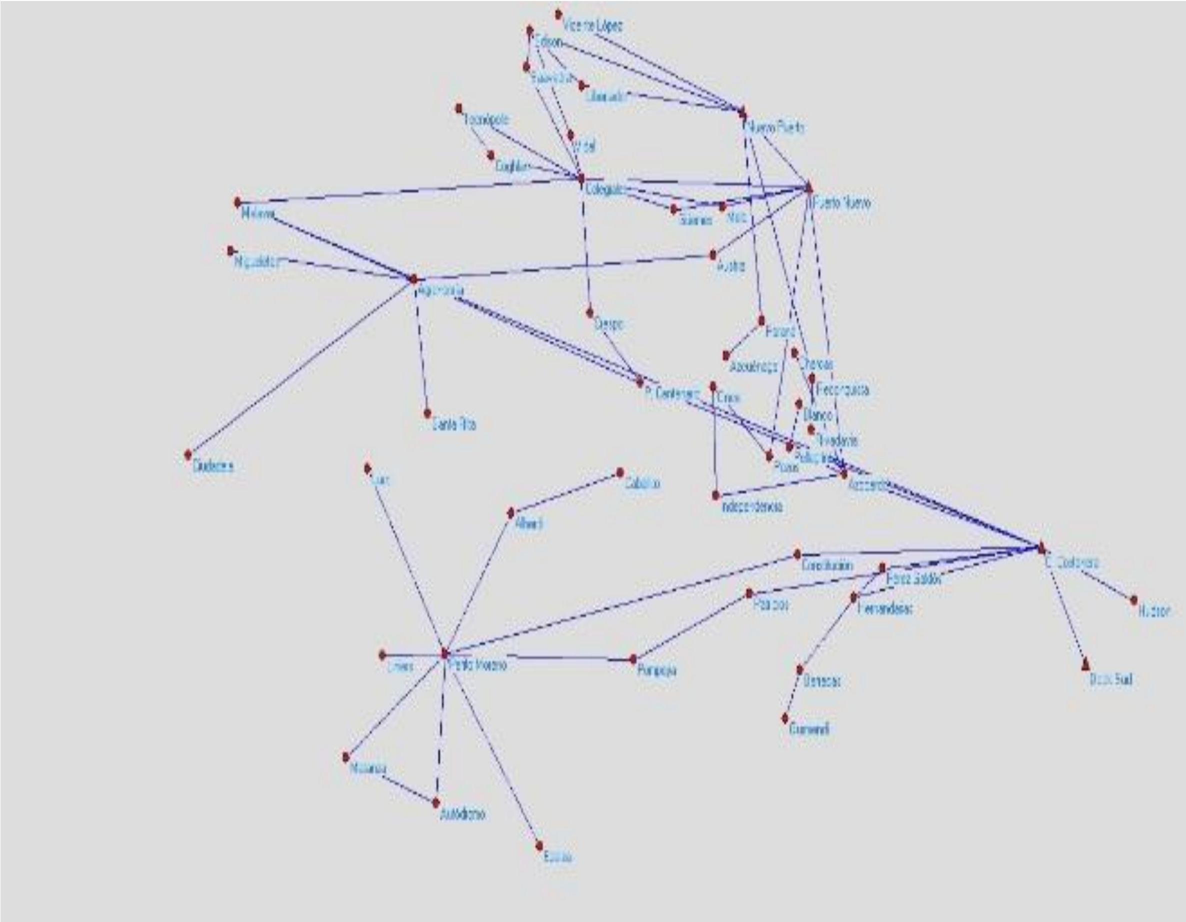




DETALLE CAPITAL FEDERAL

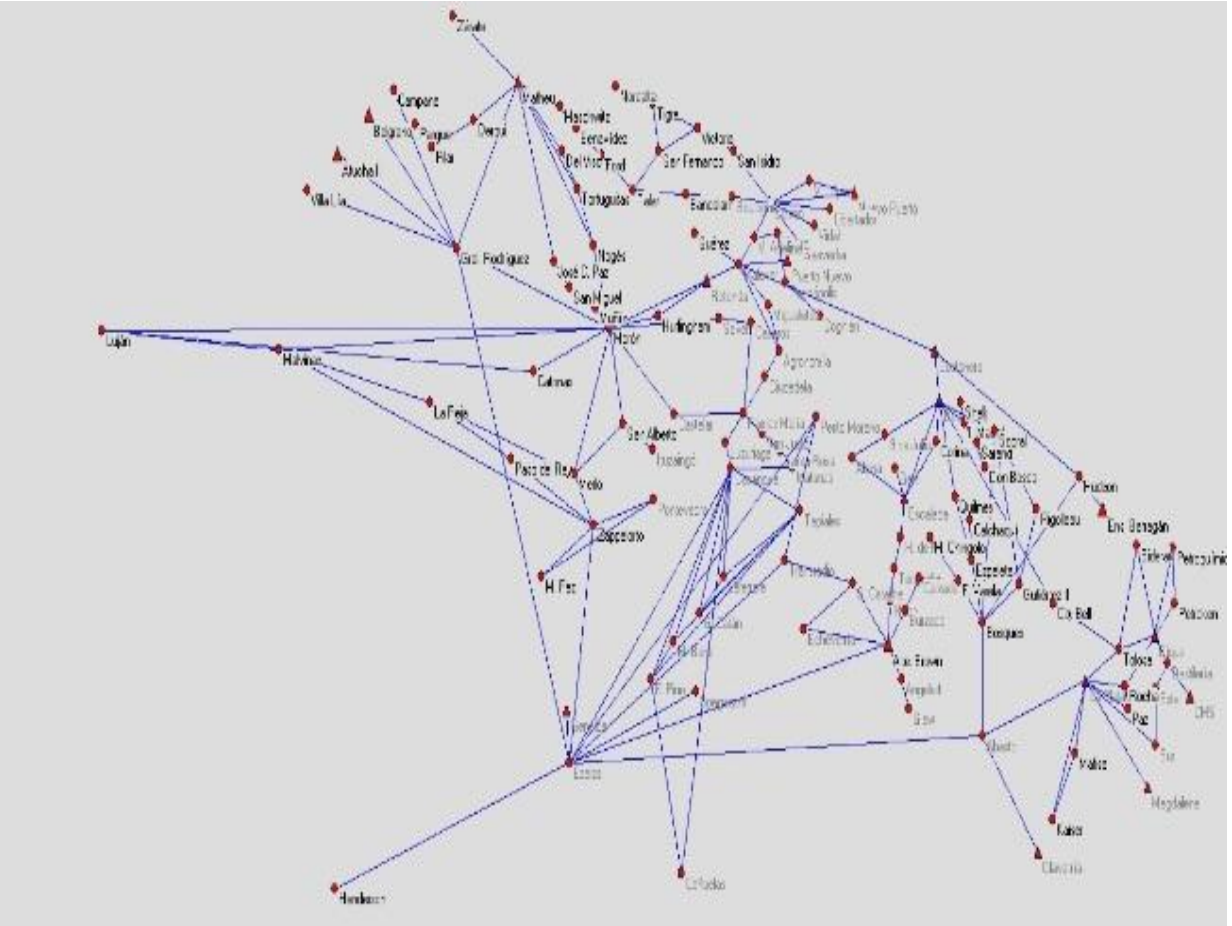
GCABA

Ciudad de Buenos Aires



+

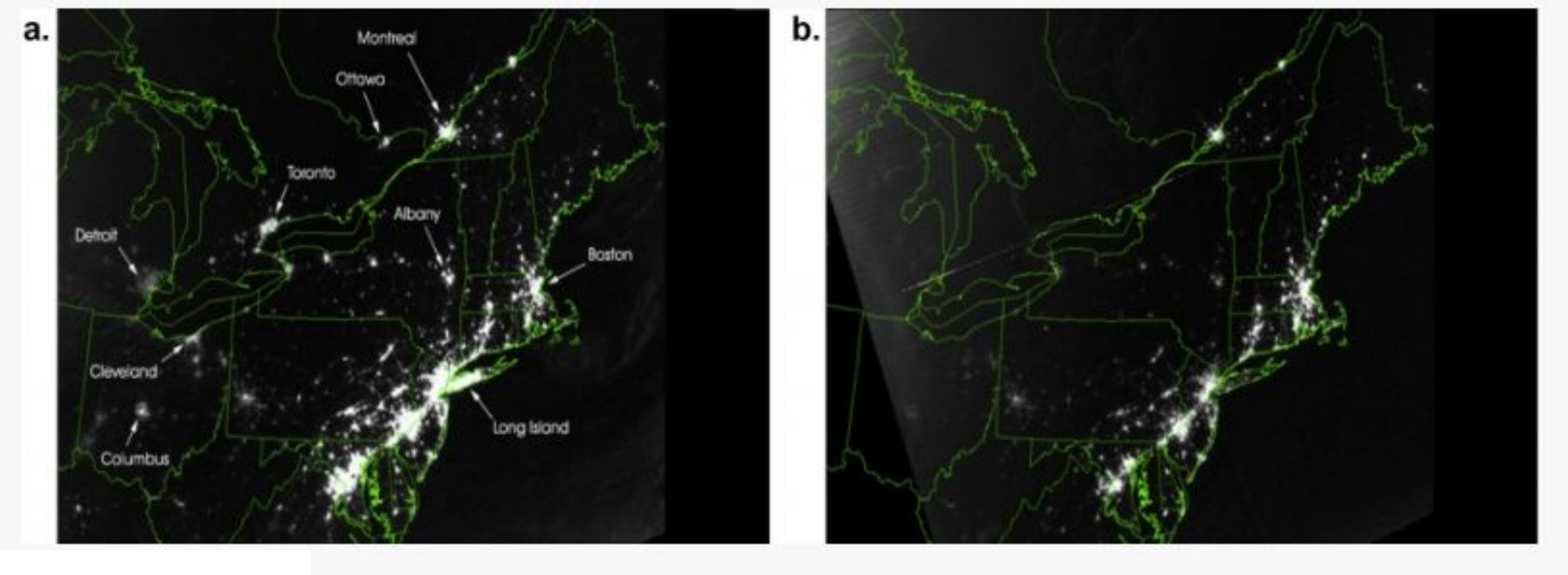
Gran Buenos Aires



Fuente: Elaboración en Pajek en base a CAMMESA, 2015, Ayelen Bargados, Tesis de maestría

Red eléctrica (Power Grid Network)

- A primera vista, las imágenes (a) y (b) son indistinguibles: muestran luces que brillan en zonas muy pobladas y espacios oscuros que marcan vastos bosques y océanos deshabitados.
- En una inspección más detallada, observamos diferencias: Toronto, Detroit, Cleveland, Columbus y Long Island, brillantes en (a), se han oscurecido en (b).
- Imagen real del noreste de EE.UU. el 14 de agosto de 2003, antes y después del apagón que dejó sin electricidad a unos 45 millones de personas en ocho estados de EE.UU. y a otros 10 millones en Ontario.
- EL apagón del 2003 es un ejemplo típico de fallo en cascada. **Vulnerabilidad debida a la interconectividad**



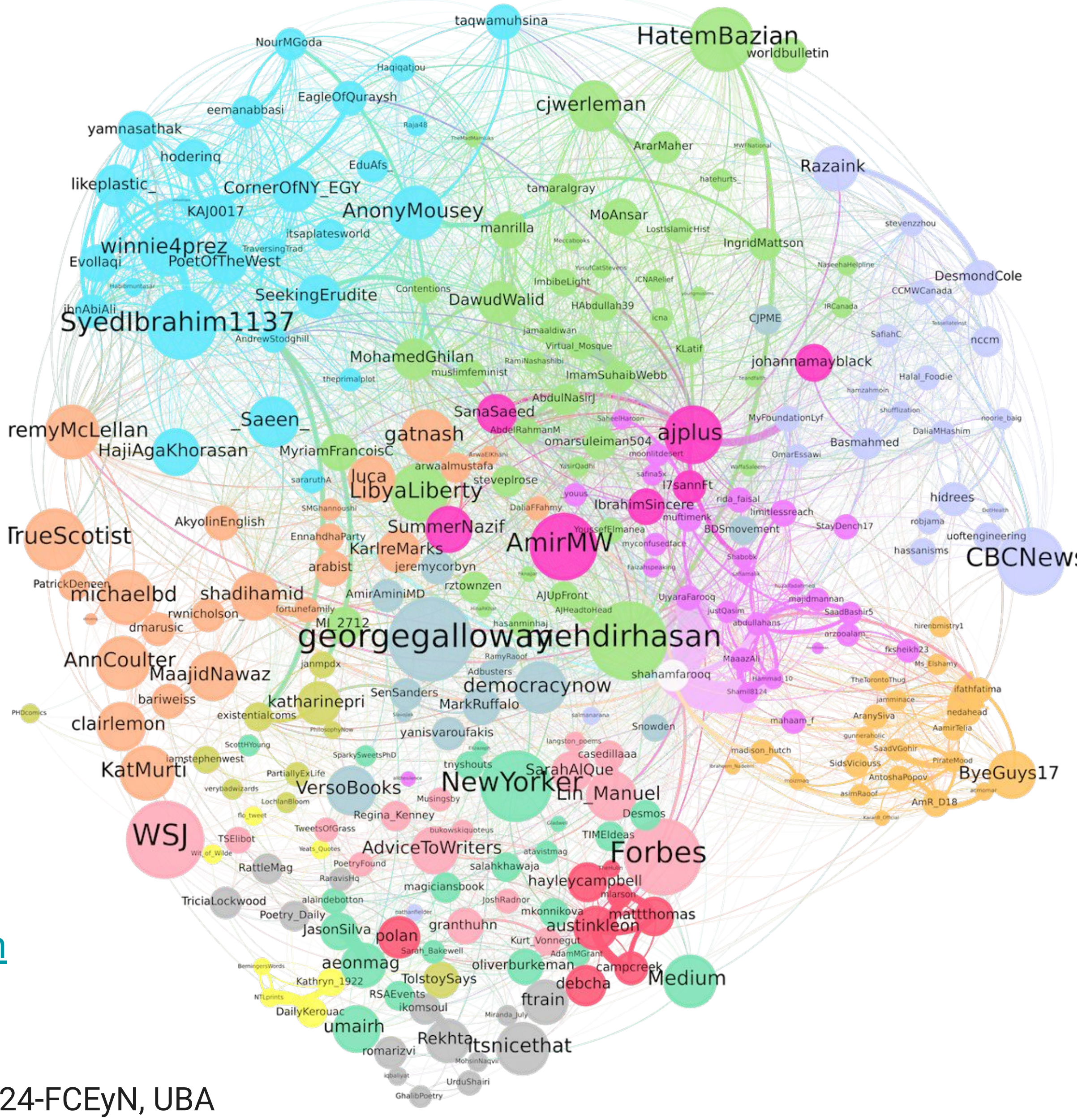
Fuente: Network Science. AL Barabási

Vulnerabilidad debida a la interconectividad: avalanchas

- Cuando una red actúa como un sistema de transporte, un fallo local propaga las cargas a otros nodos. Si la carga adicional es insignificante, el sistema puede absorberla sin problemas y el fallo pasa desapercibido. Sin embargo, si la carga adicional es excesiva para los nodos vecinos, éstos también fallarán y redistribuirán la carga entre sus vecinos.
- Evento de cascada, cuya magnitud depende de la posición y capacidad de los nodos que fallaron inicialmente.
- El ejemplo del apagón del Noreste ilustra varias cuestiones importantes:
  - para evitar los daños de las cascadas, debemos comprender la **estructura** de la red en la que se propaga la cascada
  - necesitamos modelar los procesos dinámicos que tienen lugar en estas redes, como el flujo de electricidad
  - tenemos que averiguar cómo la interacción entre la estructura y la dinámica de la red afecta a la robustez de todo el sistema. Aunque los fallos en cascada puedan parecer aleatorios e impredecibles, siguen leyes reproducibles que pueden cuantificarse e incluso predecirse con herramientas de la ciencia de redes.

¿qué son redes?

- Redes eléctricas
- Redes de Transporte
- Redes de aeropuerto
- **Redes sociales**
- Redes Financieras
- Redes neuronales …

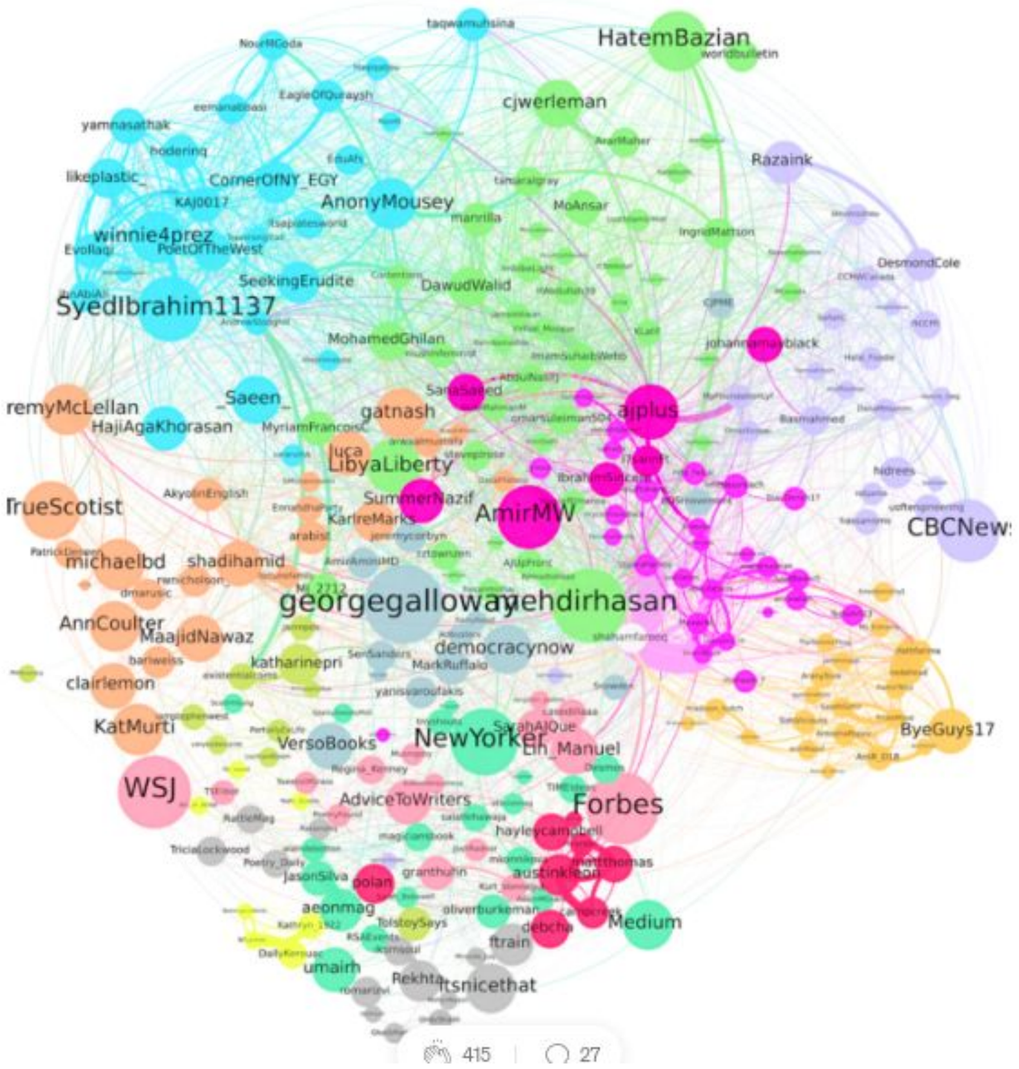


Fuente: [A Graph-based approach to community detection in Twitter Networks]

Una red de egos en Twitter

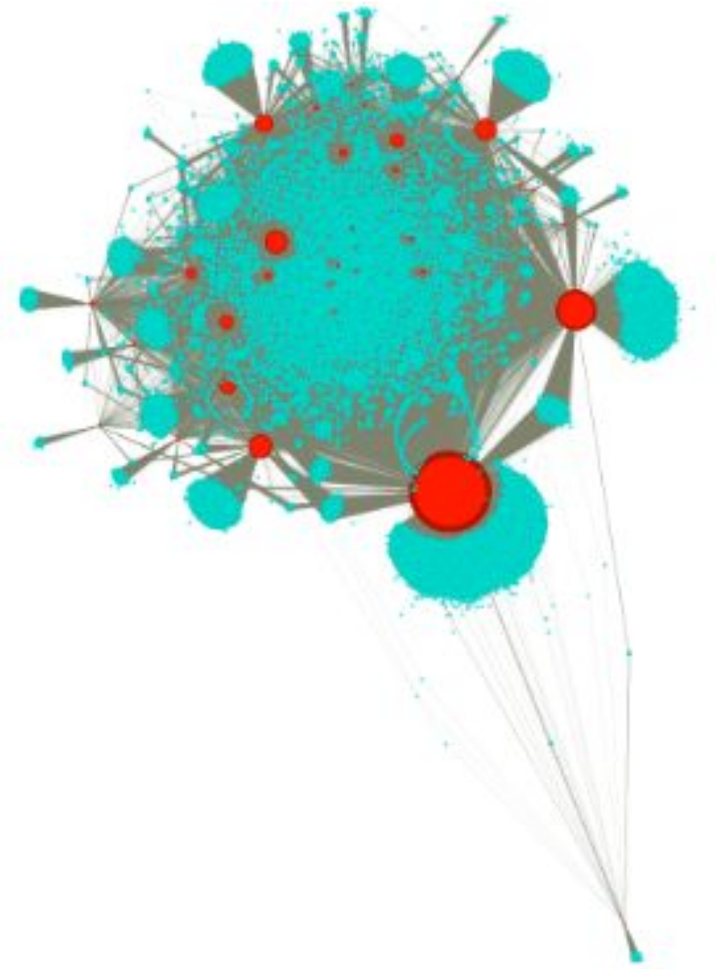
Fuente: [A Graph-based approach to community detection in Twitter Networks]

1. Extracción de datos (API)
2. Construcción de la red
   a. Vecindario de 1 paso
3. Detección de Comunidades
   a. Encontrar comunidades basadas en el "flujo de información”— [Infomap]
   b. Los "usuarios" de las agrupaciones interconectadas tienden a compartir características similares – [Homophily]
   c. Clasificación de las redes temáticas de Twitter mediante el análisis de redes sociales



Redes Financieras: red de bancos y firmas

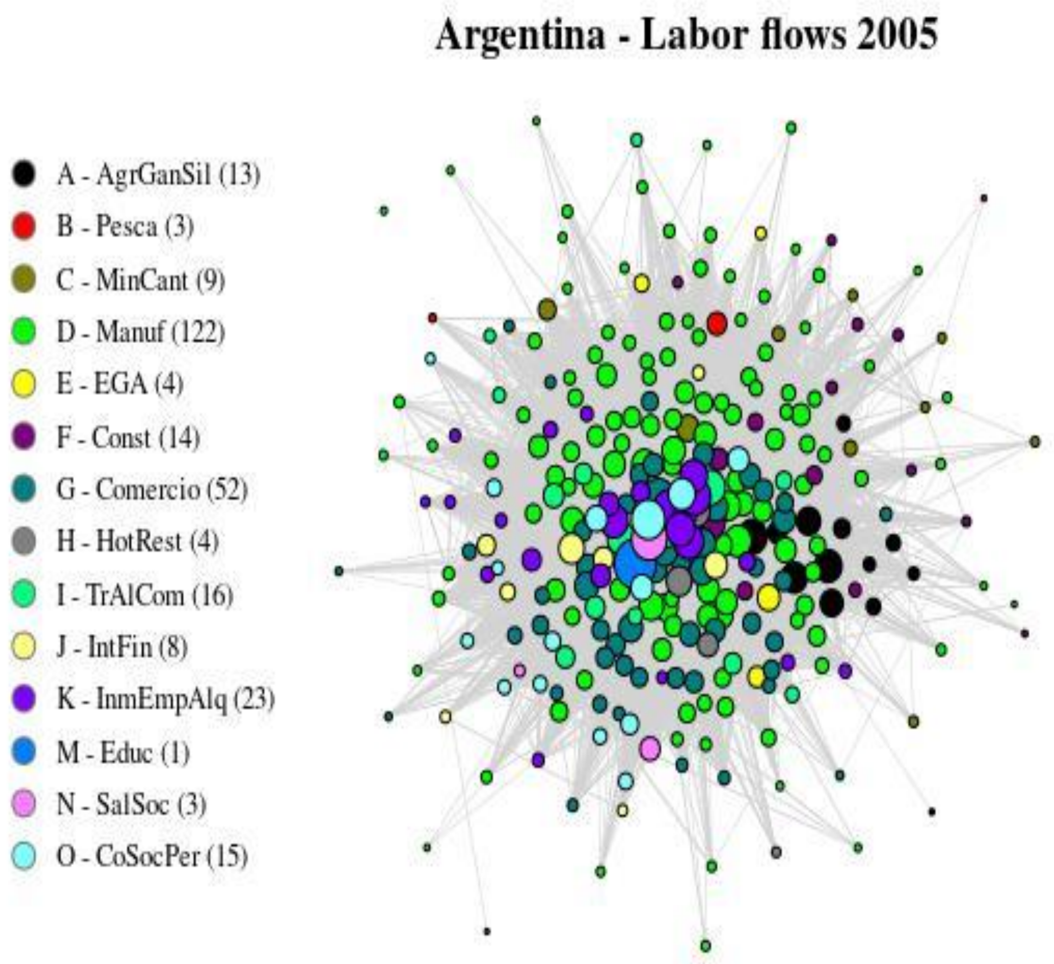
- **Red bipartita**: dos tipos de nodos
    - bancos y empresas
    - Los bancos no se conectan con otros bancos y las empresas no se conectan con otras empresas.
    - los links representan el monto del crédito.
- Los nodos rojos son bancos
- los nodos celestes son empresas
- el tamaño de los nodos es proporcional a su grado



Fuente: Manu Diaz 2023, tesis de maestría

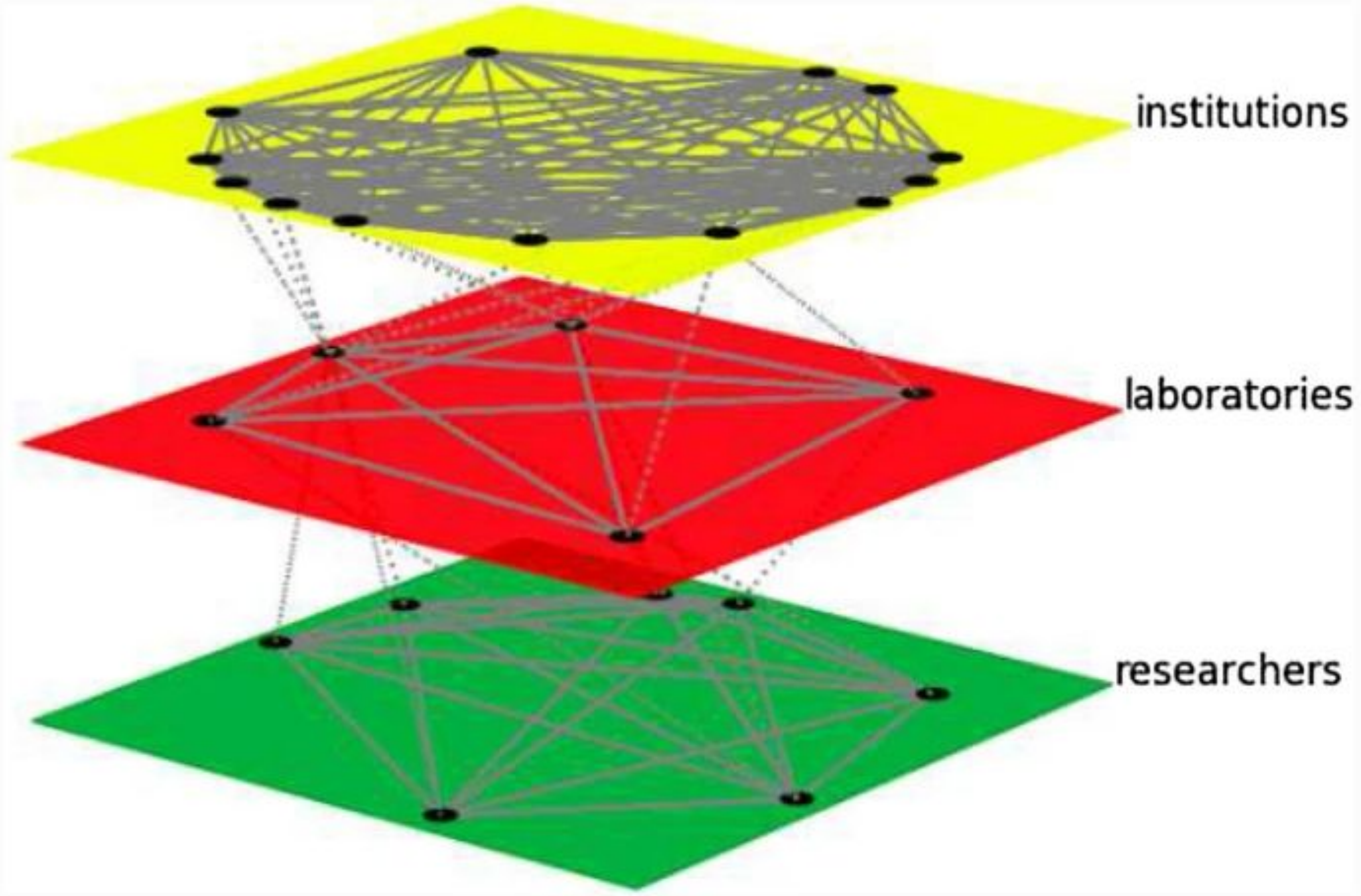
Redes Económicas: red de flujo de trabajo

- Datos: registros administrativos de empleo formal privado AFIP
  - Movilidad laboral: cambios de empleo entre empresas de distintos sectores
- Número de transiciones de empleo formal privado entre empresas de distintas actividades económicas a 4 dígitos
  - Flujos de paneles de individuos
  - matrices de transiciones
  - diferentes periodos



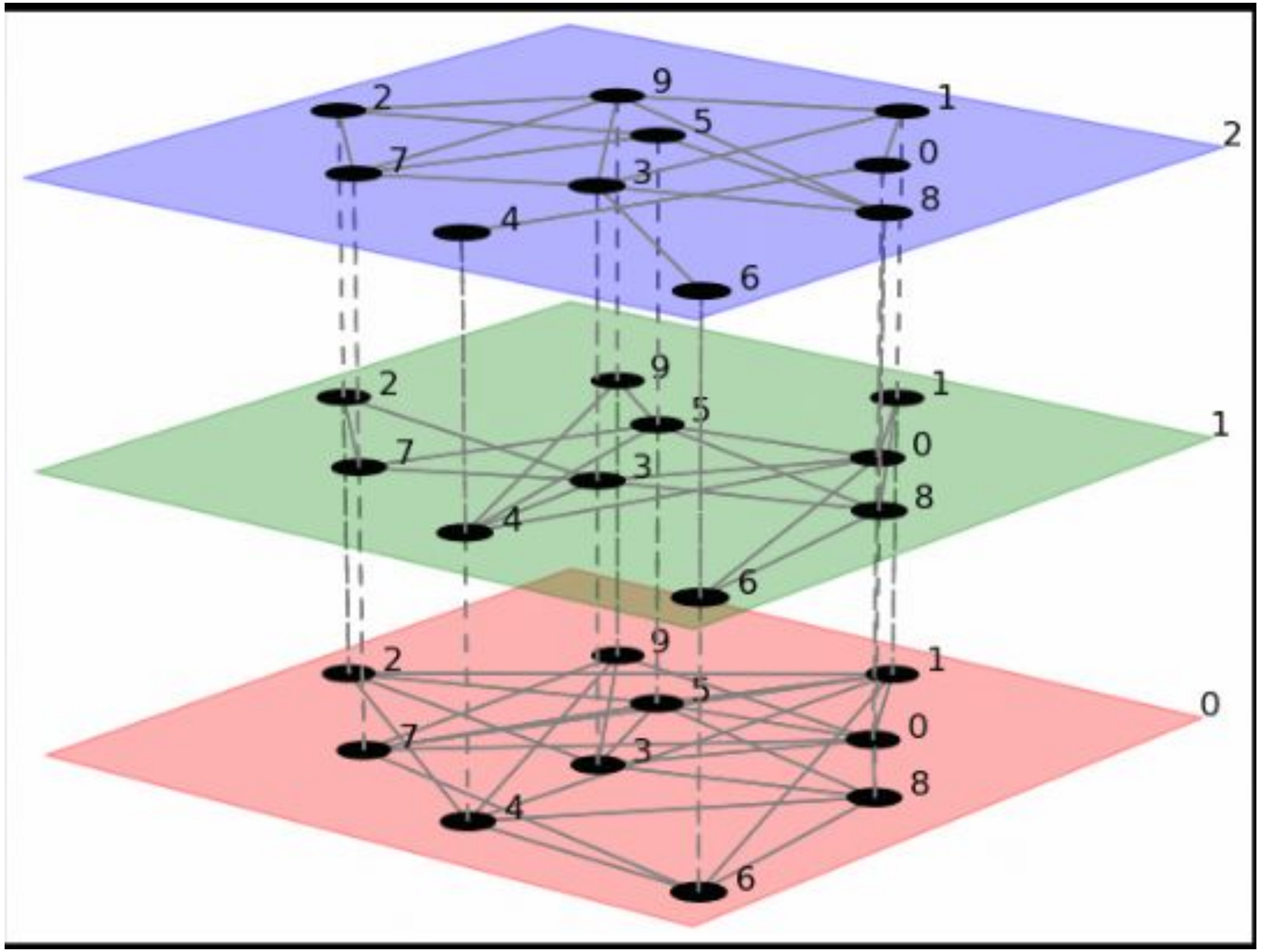
Fuente: Semeshenko, De Raco (IIEP)

Otras dimensiones: redes multicapa



Generated multilayer network by header of a scientific paper

Otras dimensiones: redes múltiplex



otras dimensiones: hipergrafos

- En algunos casos, el uso de grafos simples o dirigidos no proporciona una descripción completa de los sistemas del mundo real que se investigan.
- Por ej., en una red de colaboración representada como un simple grafo sólo sabemos si los científicos han colaborado o no, pero no podemos saber si tres o más autores vinculados en la red fueron coautores del mismo trabajo o no.
- Una posible solución a este problema es representar la red de colaboración como un grafo bipartito en el que un conjunto disjunto de nodos representa los artículos y otro conjunto disjunto representa los autores.
- Sin embargo, en este caso se pierde la "homogeneidad" en la definición de los nodos, ya que tenemos ciertos nodos que representan artículos y otros que representan autores. En el estudio de la conectividad, la agrupación y otras propiedades topológicas, esta distinción entre dos clases de nodos con interpretaciones completamente diferentes puede dar lugar a artefactos en los datos.
- podemos representar la red de colaboración como un hipergrafo en el que los nodos representan a los autores y los hipervínculos a los grupos de autores que han publicado artículos juntos

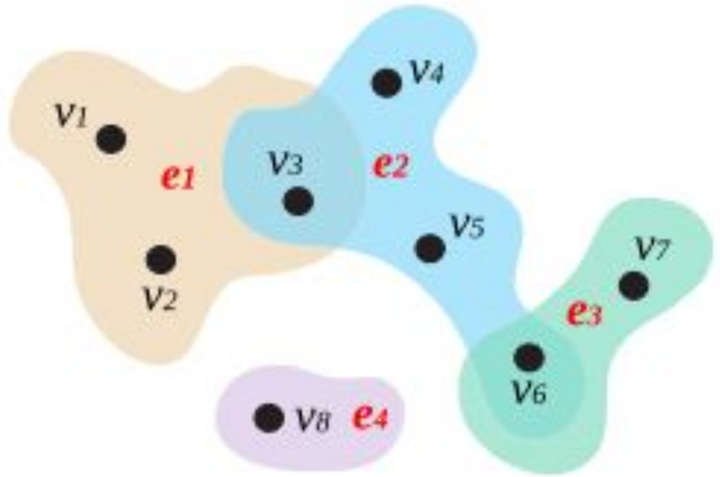
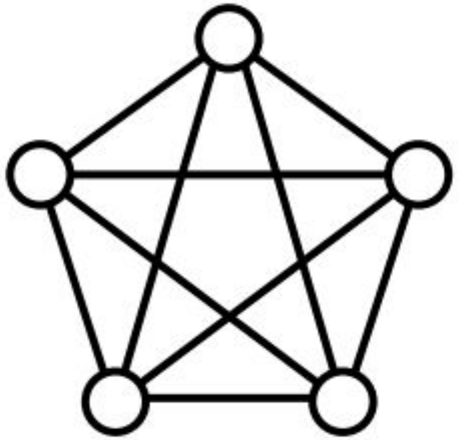


FIG. 1. Graphical representation of a hypergraph. Mathematically, $\mathcal{V} = \{v_1, v_2, v_3, v_4, v_5, v_6, v_7, v_8\}$, $\mathcal{E} = \{e_1, e_2, e_3, e_4\}$, where the hyperedges are $e_1 = \{v_1, v_2, v_3\}$, $e_2 = \{v_3, v_4, v_5, v_6\}$, $e_3 = \{v_6, v_7\}$ and $e_4 = \{v_8\}$.

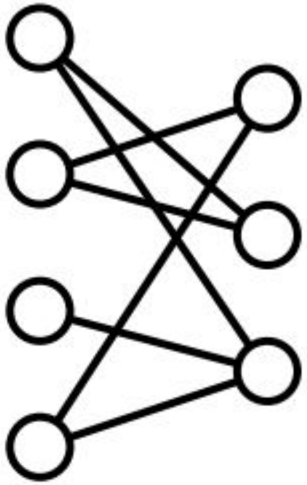
Cada enlace une un par de vértices y representa una interacción entre los individuos correspondientes

¿Qué es una red ?

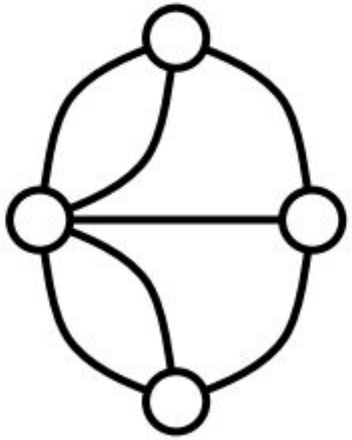
- Estructura matemática que consta de “nodos” (o vértices) y “aristas” (o enlaces) que hacen conexión entre nodos



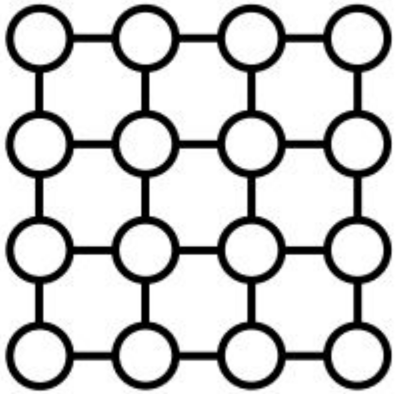
Completa



Bipartita



Aleatoria



Regular
(Grilla)

…

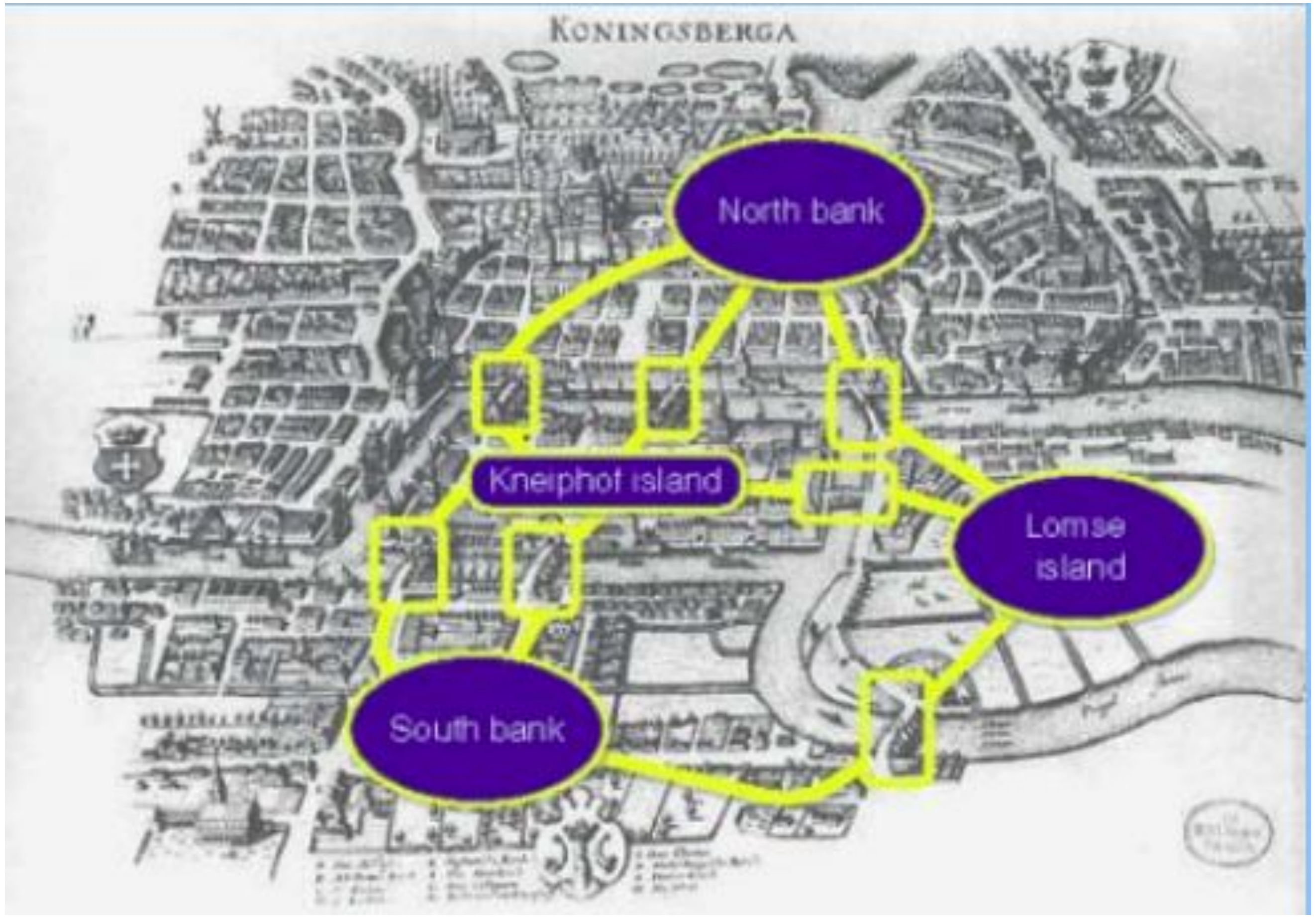
Los puentes de Königsberg

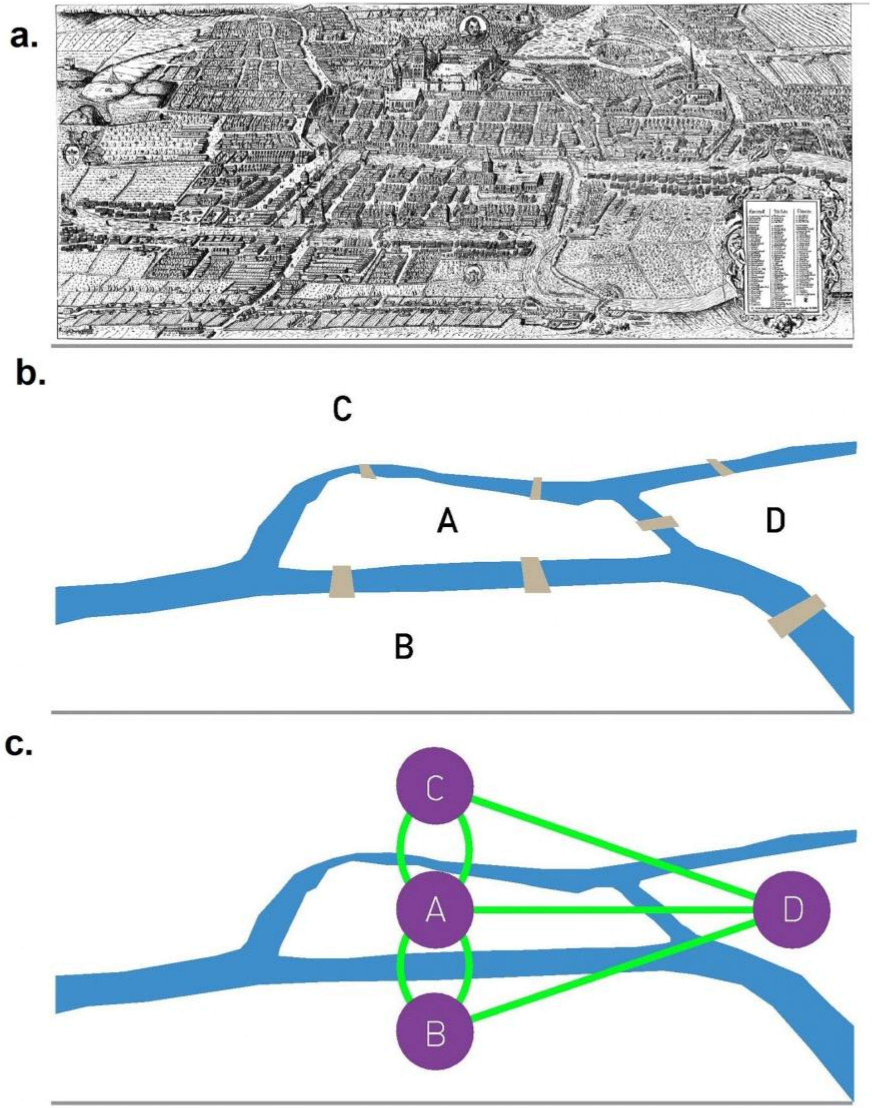
- La teoría de grafos está detrás de la ciencia de las redes. Sus raíces se remontan a 1735 en Königsberg, la capital de Prusia Oriental, una ciudad comercial próspera de su tiempo. El comercio apoyado por su ocupada flota de barcos permitió a los funcionarios de la ciudad construir siete puentes a través del río Pregel que rodeaba la ciudad. Cinco de ellos conectaron con el continente la elegante isla Kneiphof, atrapada entre las dos ramas del Pregel. Los dos restantes cruzaron las dos ramas del río. Este peculiar arreglo dio origen a un rompecabezas contemporáneo:

¿se puede cruzar los siete puentes y nunca cruzar el mismo dos veces?

A pesar de muchos intentos, nadie pudo encontrar ese camino. El problema no se resolvió hasta 1735, cuando Leonard Euler, un matemático nacido en Suiza, ofreció una prueba matemática rigurosa de que tal camino no existe.

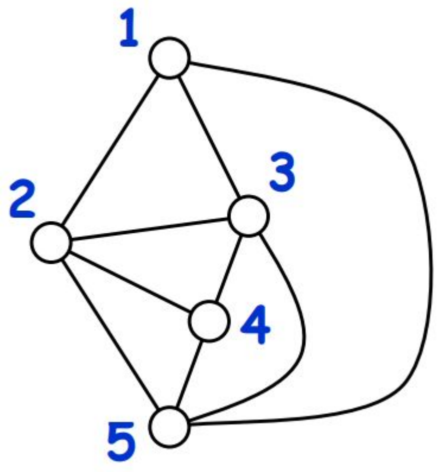
Los puentes de Königsberg



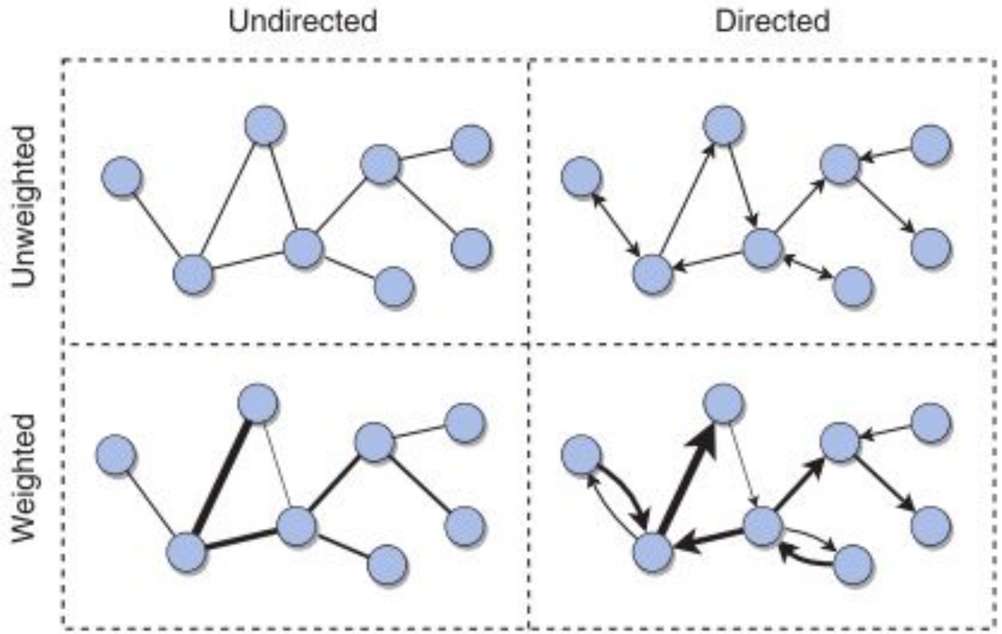


Grafo = Red

- G(V, E) : grafo (red)
- V: vértices (nodos), E: edges (lazos)
- Nodos = 1, 2, 3, 4, 5
- Links =
  - 1<->2, 1<->3, 1<->5
  - 2<->3, 2<->4, 2<->5
  - 3<->4, 3<->5, 4<->5

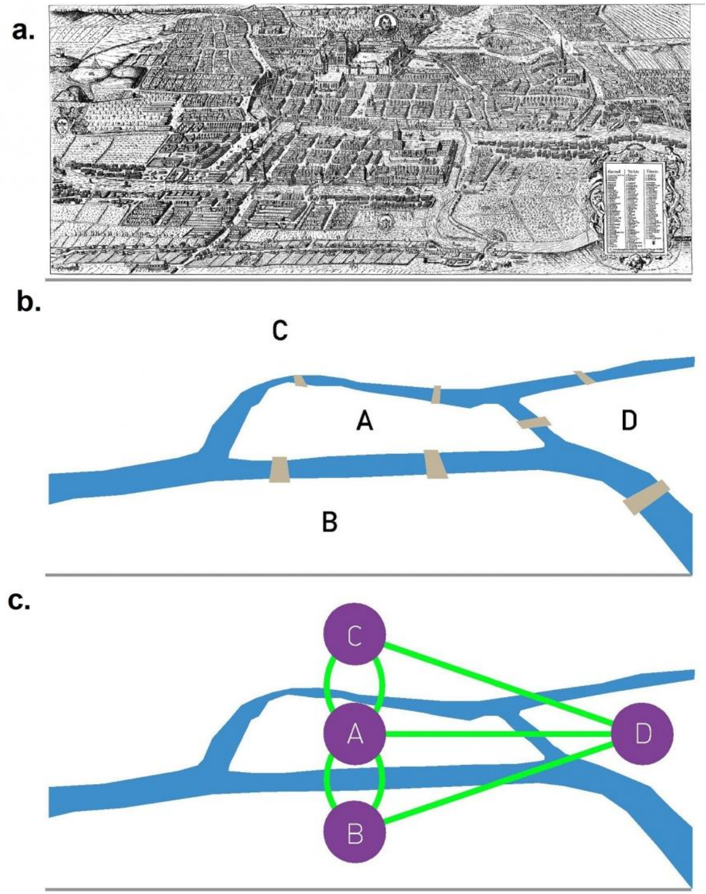


- Los nodos pueden tener estados
- Los lazos pueden tener direcciones y pesos

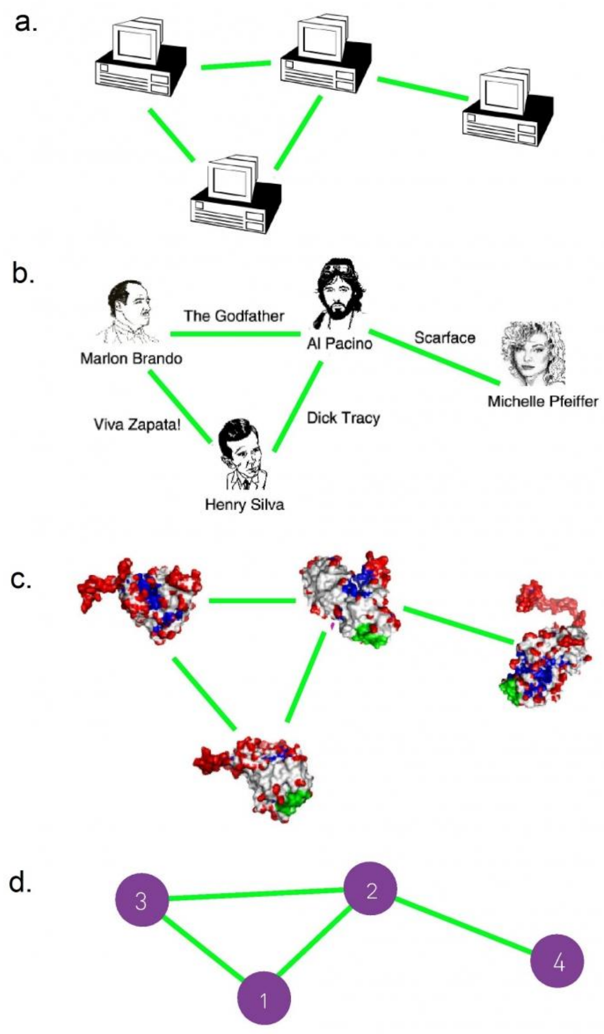


Redes unipartitas

Abstracción



Analogía



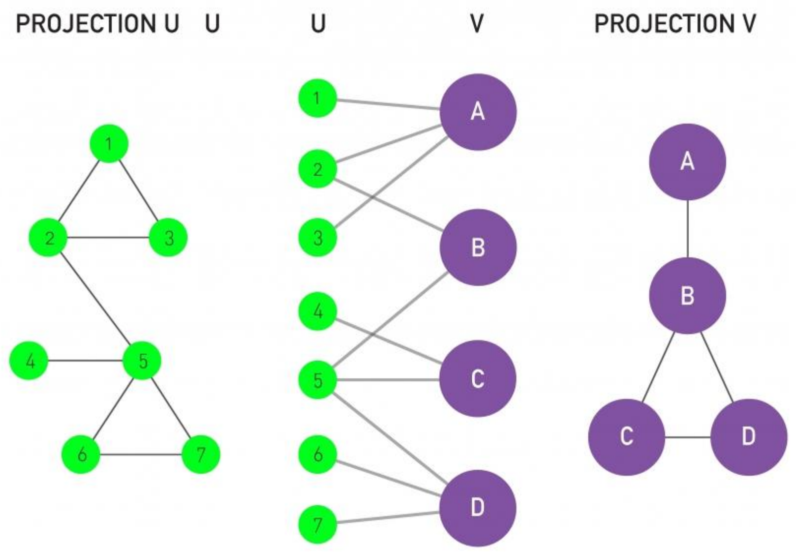
Computadoras / Internet

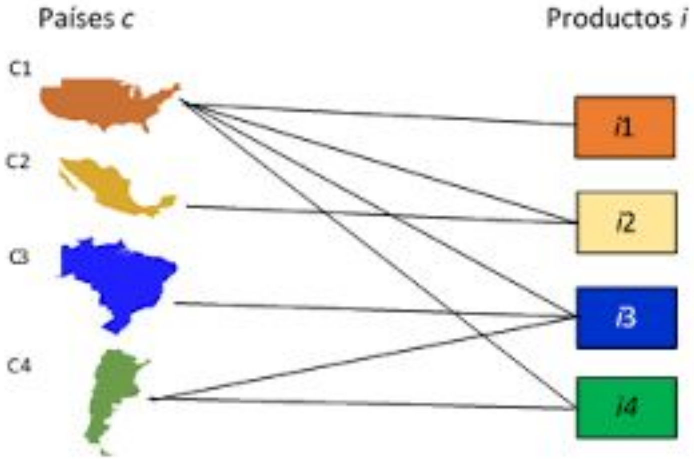
Actores co-actuación

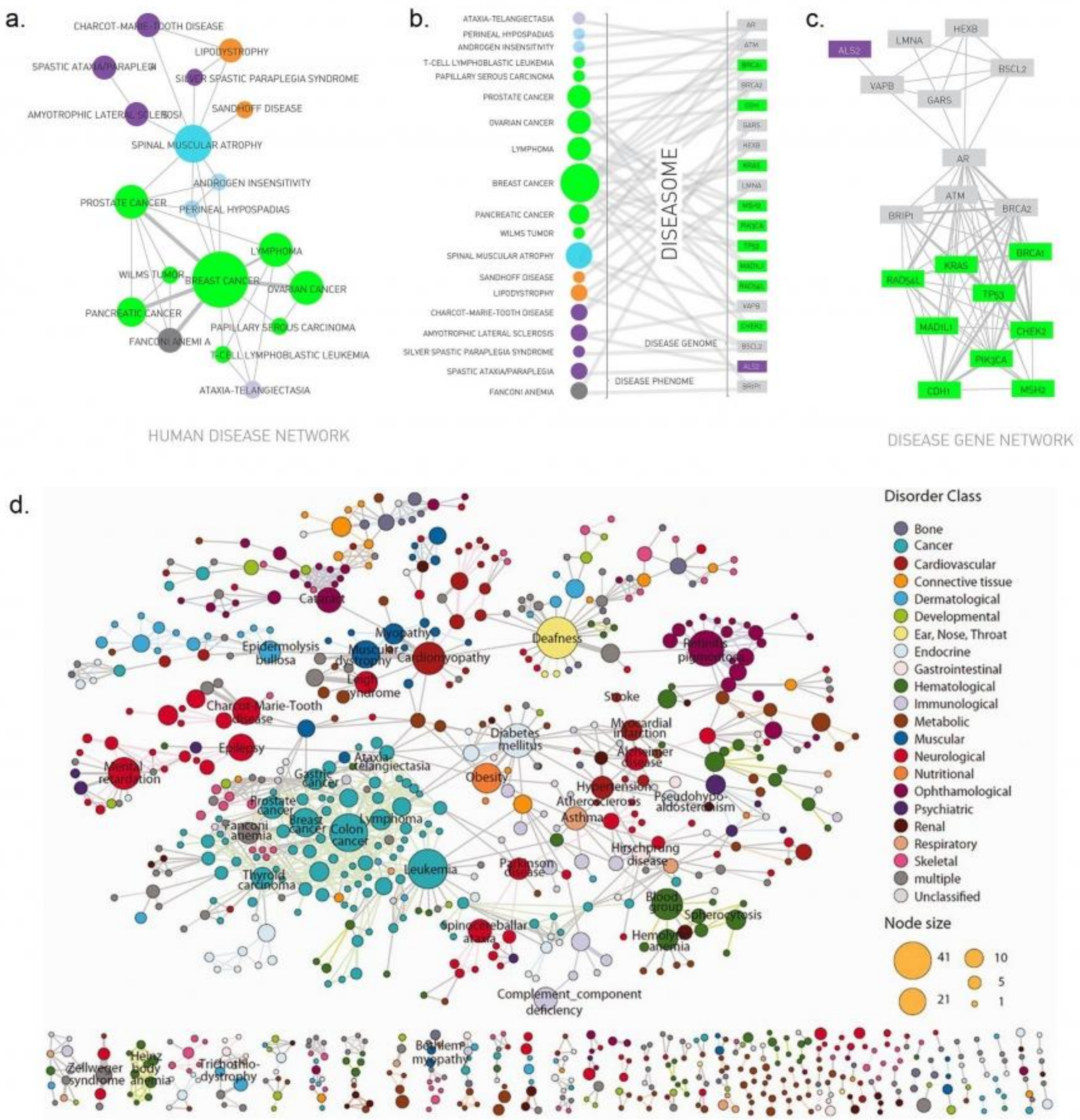
Proteínas

Grafo (misma red…)

Redes bipartitas







Representación de la Red

- Matriz de adyacencia A_{ij}
  - Una matriz con filas y columnas donde cada elemento a_{ij} toma valor 1 cuando haya una arista que une los vértices *i* y *j*. En caso contrario el elemento a_{ij} toma valor 0.
  - La matriz de adyacencia, por tanto, estará formada por ceros y unos.
- Lista de adyacencia
  - Una lista de links cuyo elemento "$i \rightarrow j$" muestra un link que va del nodo *i* al nodo *j*.
  - También representado como "$i \rightarrow \{j_1, j_2, j_3, \ldots\}$"

Grado de la red

- El grado k_i del nodo *i* se puede obtener directamente de los elementos de la matriz de adyacencia:
  - No dirigida: es la suma sobre las filas o las columnas de la matriz (simétrica)

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

  - Dirigida: las sumas sobre las filas y columnas de la matriz de adyacencia proporcionan los grados entrantes y salientes, respectivamente

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

Densidad de la Red

El número máximo de enlaces está limitado por el número posible de conexiones distintas entre los nodos de la red

- **Completa**: si todos los pares posibles de nodos están conectados por enlaces

- No dirigida: el número de pares distintos de nodos $$L_{\max} = \binom{N}{2} = N(N-1)/2$$
- Dirigida: cada par de nodos cuenta, uno para cada dirección $$L_{max} = N(N-1)$$

- **Densidad** $$d = L/L_{\max}$$
  - No dirigida: $$d = L/L_{\max} = \frac{2L}{N(N-1)}$$
  - Dirigida: $$d = L/L_{\max} = \frac{L}{N(N-1)}$$

Redes: Sparsity

Cuantos menos enlaces hay en una red, más esparsa es

- Red *Completa* , $L = L_{max}$, $\rightarrow d = 1$
- Red *Esparsa*, $L \ll L_{max}$, $\rightarrow d \ll 1$
- La red es escasa si el número de enlaces crece proporcionalmente al número de nodos ($L \sim N$), o incluso más lento
- La red es densa si el número de enlaces crece más rápido, ej. ($L \sim N^2$)

To illustrate the importance of network sparsity, let us consider the example of Facebook. At the time of writing, Facebook has around 2 billion users ($N \approx 2 \times 10^9$). If this was a complete network, there would be $L \approx 10^{18}$ links — that is a number with 18 zeros, and there is no way to store so much data! But fortunately, social networks are very sparse and Facebook is no exception. Each user has on average 1000 friends or less, so that the density is approximately $d \approx 10^{-6}$. That is still a lot of data, but Facebook can manage it.

Redes: algunos detalles

Network	Nodes	Links	Directed / Undirected	N	L	‹K›
Internet	Routers	Internet connections	Undirected	192,244	609,066	6.34
WWW	Webpages	Links	Directed	325,729	1,497,134	4.60
Power Grid	Power plants, transformers	Cables	Undirected	4,941	6,594	2.67
Mobile-Phone Calls	Subscribers	Calls	Directed	36,595	91,826	2.51
Email	Email addresses	Emails	Directed	57,194	103,731	1.81
Science Collaboration	Scientists	Co-authorships	Undirected	23,133	93,437	8.08
Actor Network	Actors	Co-acting	Undirected	702,388	29,397,908	83.71
Citation Network	Papers	Citations	Directed	449,673	4,689,479	10.43
E. Coli Metabolism	Metabolites	Chemical reactions	Directed	1,039	5,802	5.58
Protein Interactions	Proteins	Binding	Undirected	2,018	2,930	2.90

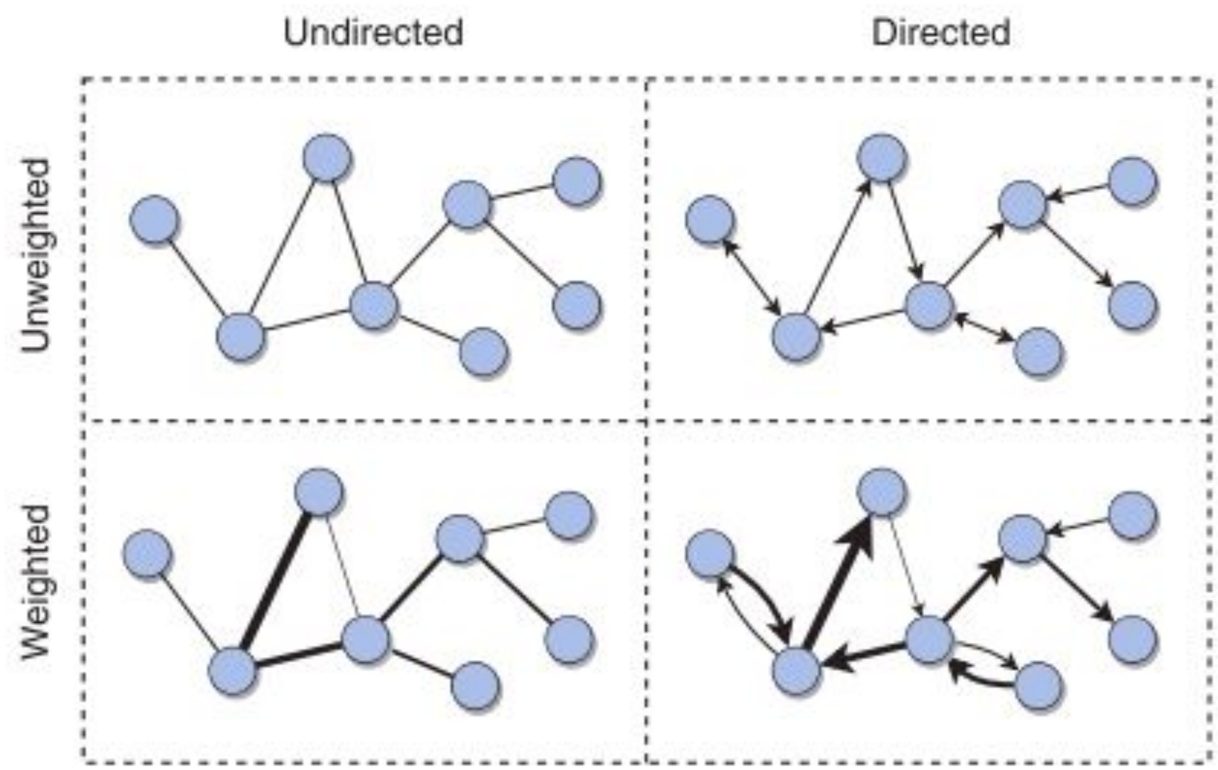
Grado Promedio

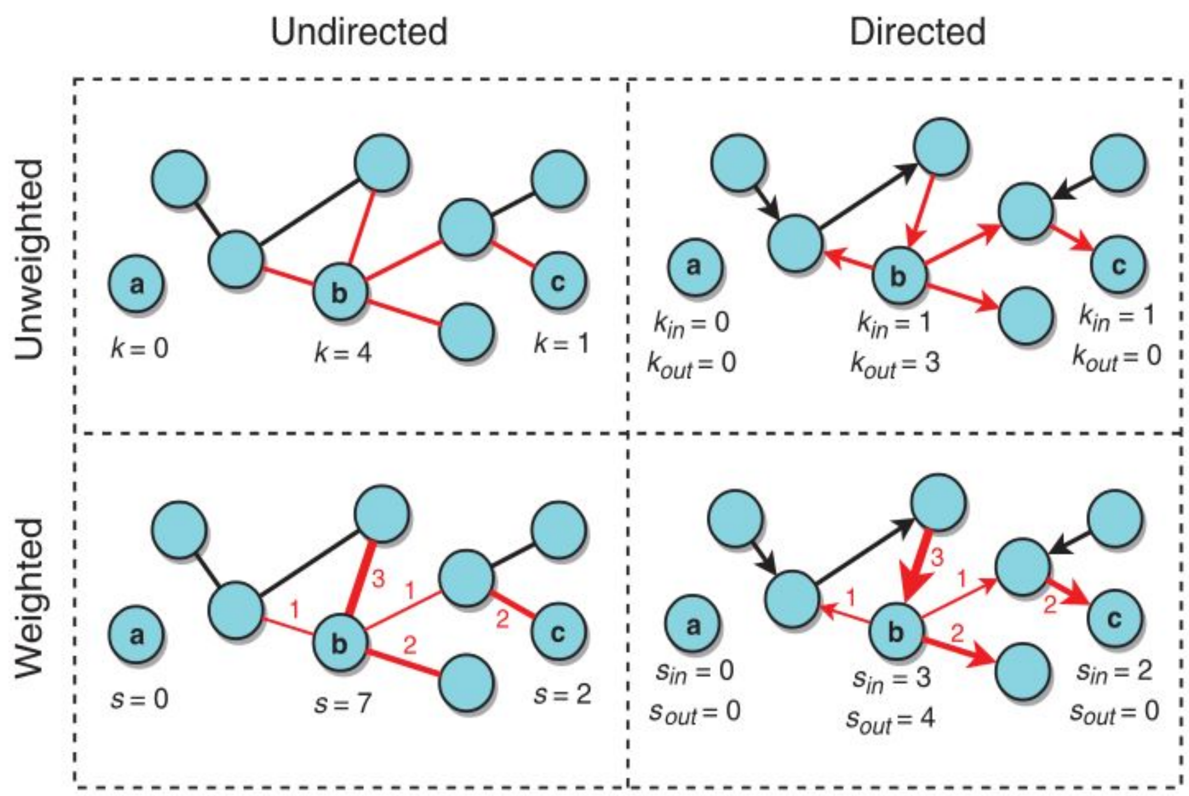
- k_i es el grado del nodo i^{th}: representa el número de enlaces a otros nodos

- **No dirigida**:
  - el número total de enlaces, L, $L = \frac{1}{2}\sum_{i=1}^{N} k_i$
  - **Grado Promedio** $$\langle k \rangle = \frac{1}{N}\sum_{i=1}^{N} k_i = \frac{2L}{N}$$

- Dirigida: distinguimos entre grado entrante, k_i^{in}, el número de enlaces que apuntan al nodo *i*, y grado saliente, k_i^{out}, el número de enlaces que apuntan desde el nodo *i* a otros nodos. El grado total $k_i = k_i^{in} + k_i^{out}$
  - número total de enlaces $$L = \sum_{i=1}^{N} k_i^{in} = \sum_{i=1}^{N} k_i^{out}$$
  - El Grado Promedio $$\langle k^{in} \rangle = \frac{1}{N}\sum_{i=1}^{N} k_i^{in} = \langle k^{out} \rangle = \frac{1}{N}\sum_{i=1}^{N} k_i^{out} = \frac{L}{N}$$

Redes: grados y fuerzas

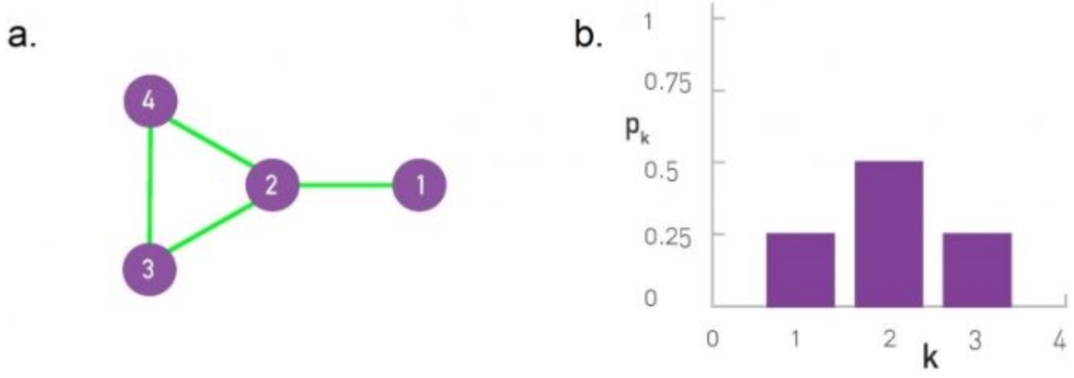




Fuente. Barabási

Distribución de Grado

- La distribución de grados, p_k , es la probabilidad de que un nodo seleccionado al azar en la red tenga un grado k. Dado que p_k es una probabilidad, esa debe normalizarse, es decir: $\sum_{k=1}^{\infty} p_k = 1$
- Para una red con *N* nodos, la distribución de grados es el histograma normalizado dado por $p_k = \frac{N_k}{N}$, donde N_k es el número de nodos de grado *k* Por lo tanto, el número de nodos de grado *k* se puede obtener de la distribución de grados como $N_k = Np_k$.



Caminos y Distancias

- En redes la distancia es un concepto desafiante. ¿Cuál es la distancia entre dos páginas web o entre dos personas que no se conocen? La distancia física no es relevante aquí.
- La distancia física es reemplazada por la *longitud del camino (path length)*. La longitud es un camino que recorre los enlaces de la red. La longitud de un camino representa el número de enlaces que contiene el camino

Caminos y Distancias

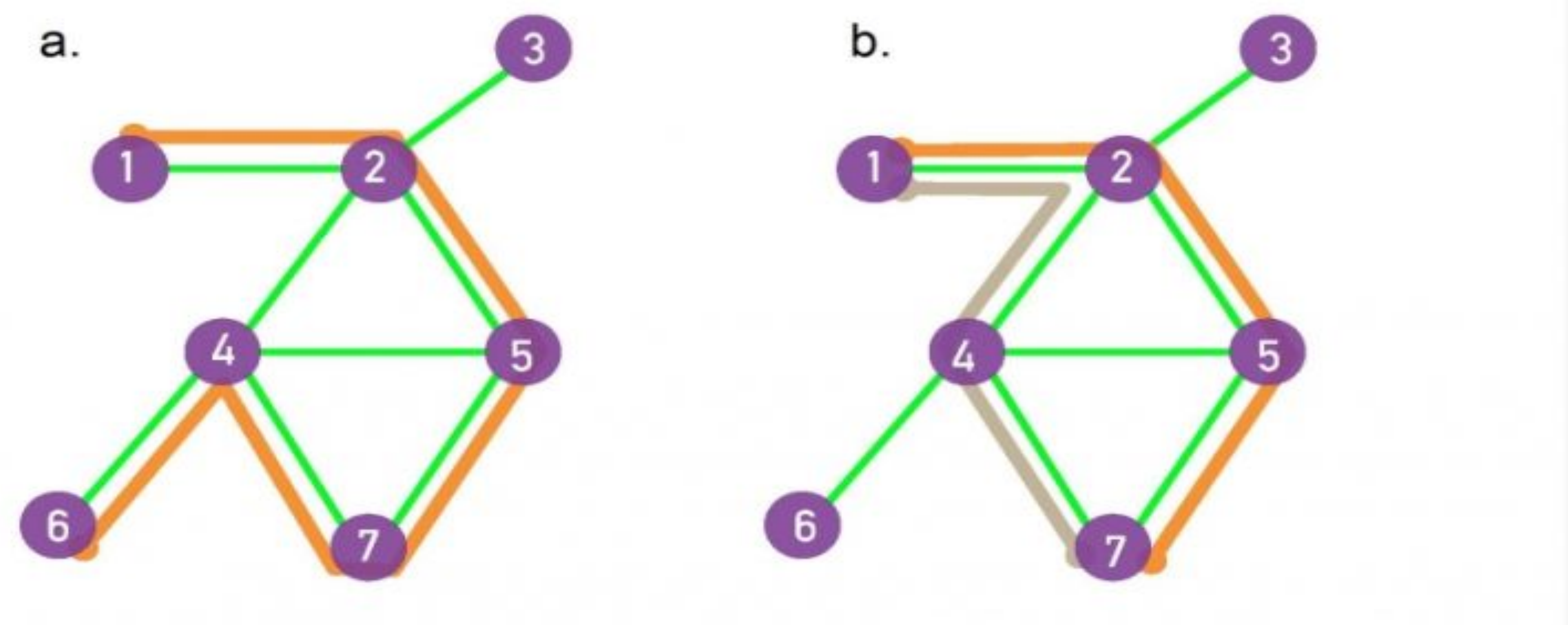


Image 2.12

Paths

- A path between nodes i_0 and i_n is an ordered list of n links P = $\{(i_0, i_1), (i_1, i_2), (i_2, i_3), \ldots ,(i_{n-1}, i_n)\}$. The length of this path is n. The path shown in orange in (a) follows the route 1→2→5→7→4→6, hence its length is n = 5.
- The shortest paths between nodes 1 and 7, or the distance d_{17}, correspond to the path with the fewest number of links that connect nodes 1 to 7. There can be multiple paths of the same length, as illustrated by the two paths shown in orange and grey. The network diameter is the largest distance in the network, being d_{max} = 3 here.

Camino más corto (Shortest Path)

- El camino más corto entre los nodos *i* y *j* es el camino con el menor número de enlaces
- El camino más corto a menudo se llama la distancia entre los nodos *i* y *j*, y se denota por d_{ij}, o simplemente d. Podemos tener múltiples caminos más cortos de la misma longitud d entre un par de nodos. El camino más corto nunca contiene bucles o se interseca a sí mismo.
- En una red no dirigida $d_{ij} = d_{ji}$, es decir, la distancia entre los nodos *i* y *j* es la misma que la distancia entre los nodos *j* y *i*. En una red dirigida a menudo dijimos $d_{ij} \neq d_{ji}$. Además, en una red dirigida, la existencia de un camino desde el nodo *i* al nodo *j* no garantiza la existencia de un camino desde *j* hasta *i*.

Camino más corto (Shortest Path)

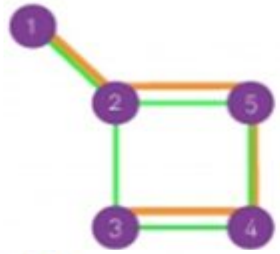
- En una red ponderada, el camino corto es un camino con la suma mínima de pesos de borde

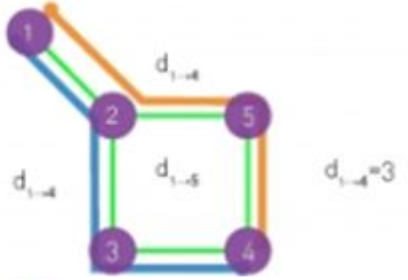
Problema del viajante

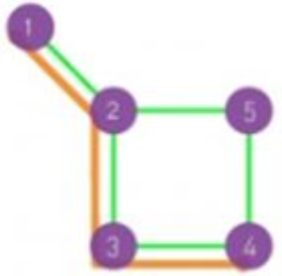
Travelling Salesman Problem (1930)

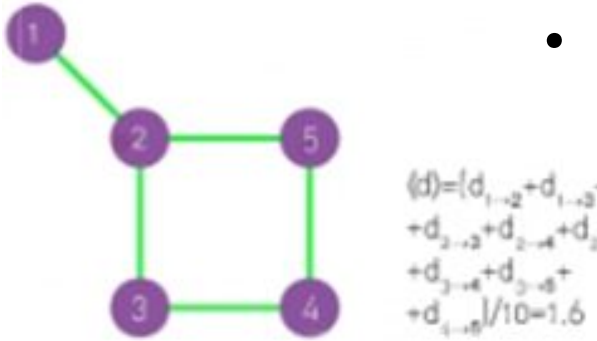
- dada una lista de ciudades y las distancias entre cada par de ellas, ¿cuál es la ruta más corta posible que visita cada ciudad exactamente una vez y al finalizar regresa a la ciudad origen?
- problema NP-Hard dentro en la optimización combinatoria
- En el problema se presentan N! rutas posibles, aunque se puede simplificar ya que dada una ruta nos da igual el punto de partida y esto reduce el número de rutas a examinar en un factor N quedando (N-1)!. Como no importa la dirección en que se desplace el viajante, el número de rutas a examinar se reduce nuevamente en un factor 2. Por lo tanto, hay que considerar (N-1)!/2 rutas posibles
  - 5 ciudades hay (5-1)!/2=12 rutas diferentes y no necesitamos un ordenador para encontrar la mejor ruta
  - 10 ciudades hay (10-1)!/2=181.440 rutas diferentes
  - 30 ciudades hay más de 4·10^30 rutas posibles. Un ordenador que calcule un millón de rutas por segundo necesitaría 10^17 años para resolverlo. Dicho de otra forma, si se hubiera comenzado a calcular al comienzo de la creación del universo (hace unos 13.400 millones de años) todavía no se habría terminado.

Elementos









- Camino
  - Una secuencia de nodos de modo que cada nodo esté conectado al siguiente nodo a lo largo del camino mediante un enlace. Cada ruta consta de n + 1 nodos y n enlaces. La longitud de un camino es el número de sus enlaces, contando múltiples enlaces varias veces. (Ej: $1 \rightarrow 2 \rightarrow 5 \rightarrow 4 \rightarrow 3$ covers a path of length four)
- Camino corto
  - El camino con la distancia más corta entre dos nodos (puede ser no único).
  - Ej: between nodes 1 and 4 we have two shortest paths, $1\rightarrow 2\rightarrow 3\rightarrow 4$ (blue) and $1\rightarrow 2\rightarrow 5\rightarrow 4$ (orange), having the same length $d_{1,4}$ =3.
- Diámetro
  - El camino más corto máximo (the longest shortest path) en un gráfico, o la distancia entre los dos nodos más lejanos
  - Ej: the diameter is between nodes 1 and 4, hence d_{max}=3.
- Average path length $\langle d \rangle$
  - The average of the shortest paths between all pairs of nodes.
    (Ej: $\langle d \rangle$=1.6)

Clustering Coefficient

- Captures the degree to which the neighbours of a given node link to each other
- For a node *i* with degree k_i, the local clustering coefficient :

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

where L_i represents the number of links between the k_i neighbours of node *i*

- C_i *is between 0 and 1*
- C_i*=0* if none of the neighbours of node *i* link to each other
- C_i*=1* if the neighbours of node i form a complete graph
- C_i is the probability that two neighbours of a node link to each other.
- C_i*=0.5* → 50% proba that two neighbours of a node are linked.
- *Thus, the more densely interconnected the neighbourhood of node i, the higher is the local clustering coefficient*

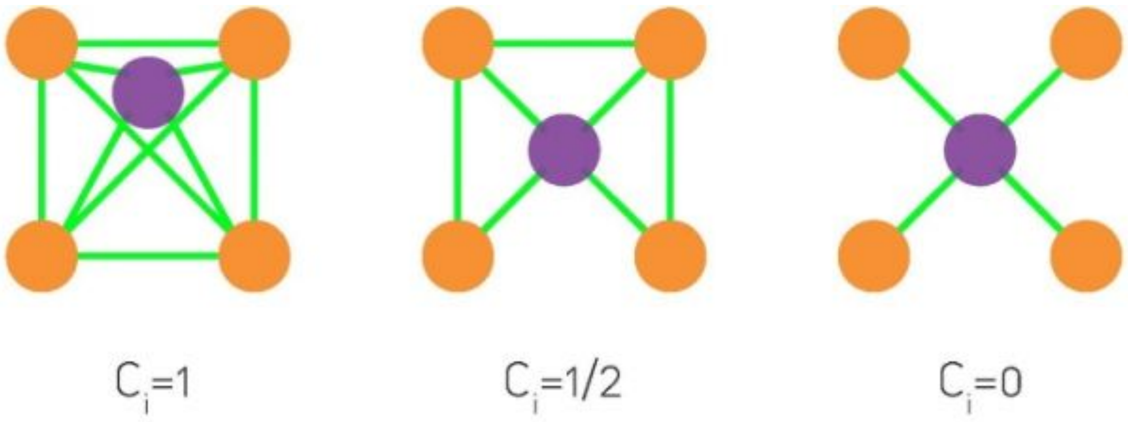
Clustering

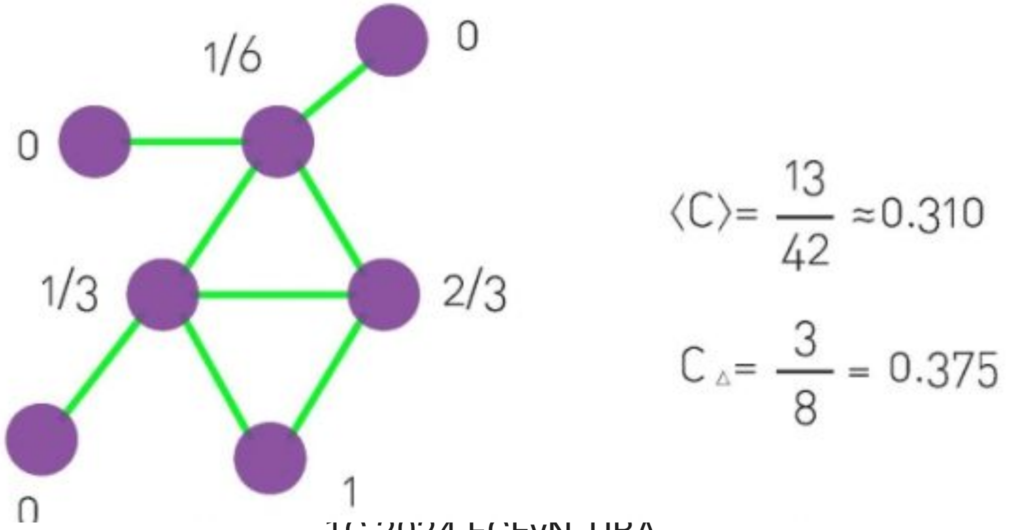
- Average clustering coefficient $\langle C \rangle$, represents the average of C_i over all nodes

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

- $\langle C \rangle$ is the probability that two neigbours of a randomly selected node link to each other

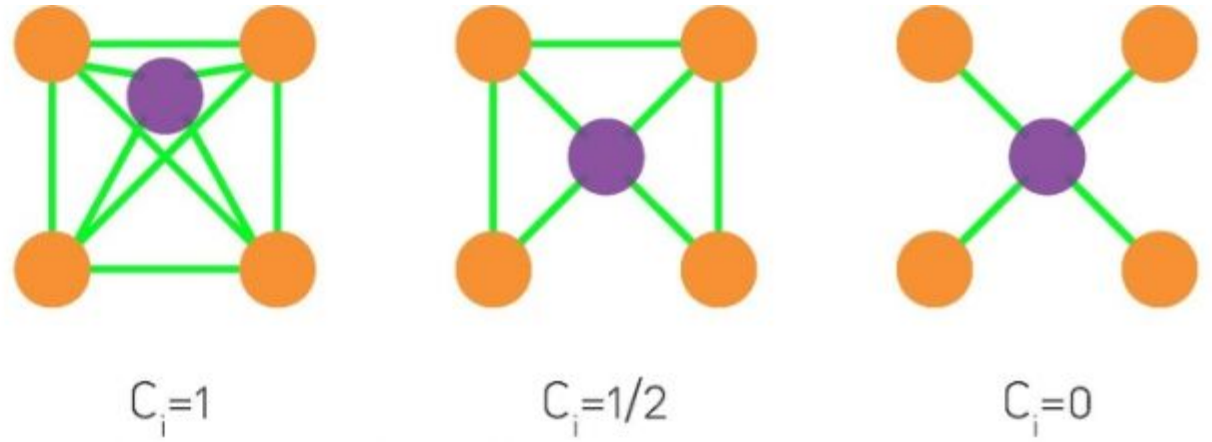
Clustering coefficient



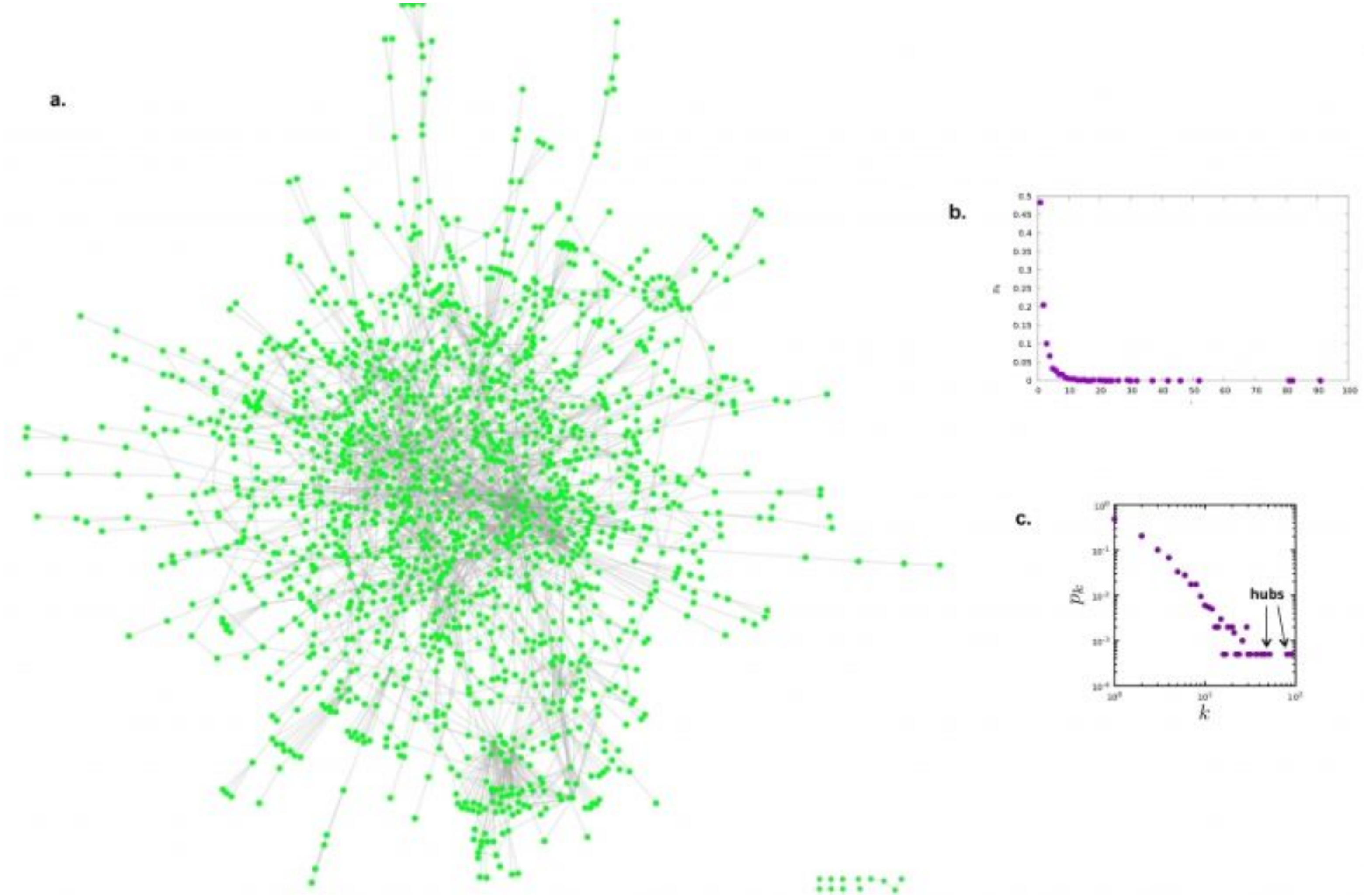


Coeficiente de agrupamiento global

- Measures the total number of closed triangles in a network.
- Going back to $C_i = \frac{2L_i}{k_i(k_i-1)}$, L_i is the number of triangles that a node *i* participates in, as each link between two neighbors of node *i* closes a triangle.



Real network example: The protein-protein interaction (PPI) network



a) PPI□ undirected network
N= 2,018 proteins
L= 2,930 links
<k>=2.90 □ a typical protein interacts with approx. Two to three other proteins.
b) Degree distribution of PPI: the vast majority Of nodes have only a few links.
These numerous nodes with few links coexist with very highly connected nodes (scale-free)

The network has a large component that connects 81% of the proteins and several smaller components.

The av clustering coef <C>=0.12, indicates significant degree of local clustering.

Centrality Measures

- **Node centrality** : degree
- **Closeness:** What if it's not so important to have many direct friends? Or be "between" others. But one still wants to be in the "middle" of things, not too far from the center.
  - Closeness is based on the length of the average shortest path between a vertex and all vertices in the graph
- **Betweenness**: how many pairs of individuals would have to go through you in order to reach one another in the minimum number of hops?

Closeness Centrality

- The closeness of the node *i* in an undirected network *G* is defined as:

$$CC(i) = \frac{n-1}{S(i)}$$

where $S(i)$ is the distance sum, calculated from the shortest path distances:

$$S(i) = \sum_{j \in V(G)} d(i,j)$$

- In a directed network a node has in- and out-closeness centrality.
- In-closeness corresponds to how close this node is to nodes it is receiving information from.
- Out-closeness centrality indicates how close the node is from those it is sending information to.

Betweenness Centrality

- The betweenness of the node *i* in an undirected network *G* is defined as:

$$BC(i) = \sum_i \sum_k \frac{\rho(j,i,k)}{\rho(j,k)}, i \neq j \neq k$$

- where $\rho(j,k)$ is the number of shortest paths connecting the node *j* to the node *k*
- $\rho(j,i,k)$ is the number of these shortest paths that pass through node *i* in the network.

Topologías de Redes

- Aleatorias (ER)
- Pequeños Mundos (SW)
- Libre Escala (SF)

Random Network (RN)

- From a Cocktail Party to Random Networks
- A random network consists of N nodes where each node pair is connected with probability p
- There are two definitions of a random network:
  - G(N, L) Model: N labeled nodes are connected with L randomly placed links (Erdős and Rényi, papers on random networks). Model fixes the total number of links L.
  - G(N,p) Model: Each pair of N labeled nodes is connected with probability p, a model introduced by Gilbert. The model fixes the probability p that two nodes are connected.

Rapoport (1951) demonstrated that if we increase the average degree of a network, we observe an abrupt transition from disconnected nodes to a graph with a giant component.

Random Networks (RN)

- The interesting thing about the $G(N, p)$ model is that even though edges are chosen independently with no “collusion”, certain global properties of the graph emerge from the independent choices
  - For small p, with $p = d/N$, $d < 1$, each connected component in the graph is small.
  - For $d > 1$, there is a giant component consisting of a constant fraction of the vertices.
  - As d increases there is a rapid transition in p of a giant component at $d = 1$.

Suppose the vertices of the graph represents people and an edge means the two people it connects have met and became friends. The probability **p** that two people meet and become friends is statistically independent of all other friendships. The value of **d** can be interpreted as the expected number of friends a person knows. The question arises as to how large are the components in this friendship graph?

If the expected number of friends each person has is more than one, then a giant component will be present consisting of a constant fraction of all the people. On the other hand, if in expectation, each person has less than one friend, the largest component is a vanishingly small fraction of the whole. Furthermore, the transition from the vanishing fraction to a constant fraction of the whole happens abruptly between d slightly less than one to d slightly more than one. There is no global coordination of friendships. Each pair of individuals becomes friends independently.

RN: Number of links expected

- Each RN generated with the same parameters N, p looks slightly different
- How many links we expect for a particular realization of a RN with fixed N and p?

The average degree of a random network is

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

Average Distance:

$$\langle d \rangle \propto \frac{\ln N}{\ln \langle k \rangle}$$

Clustering Coefficient:

$$\langle C \rangle = \frac{\langle k \rangle}{N}$$

in a random network the chance of observing a hub decreases faster than exponentially.

The probability that a random network has exactly L links is the product of three terms:

- The probability that L of the attempts to connect the $N(N-1)/2$ pairs of nodes have resulted in a link, which is p^L.
- The probability that the remaining $N(N-1)/2 - L$ attempts have not resulted in a link, which is $(1-p)^{N(N-1)/2-L}$.
- A combinational factor,

$$\binom{\frac{N(N-1)}{2}}{L} \qquad (3.0)$$

counting the number of different ways we can place L links among $N(N-1)/2$ node pairs.

We can therefore write the probability that a particular realization of a random network has exactly L links as

$$p_L = \binom{\frac{N(N-1)}{2}}{L} p^L (1-p)^{\frac{N(N-1)}{2}-L} \qquad (3.1)$$

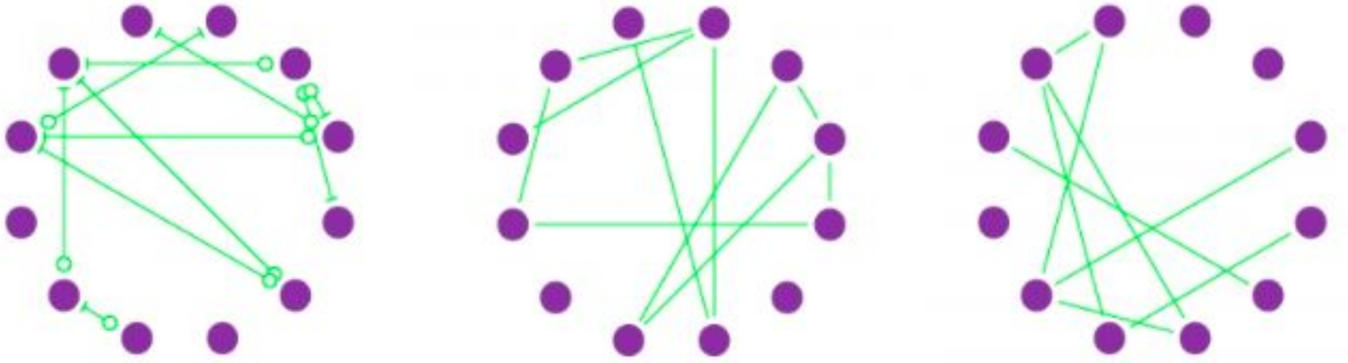
As (3.1) is a binomial distribution (BOX 3.3), the expected number of links in a random graph is

\

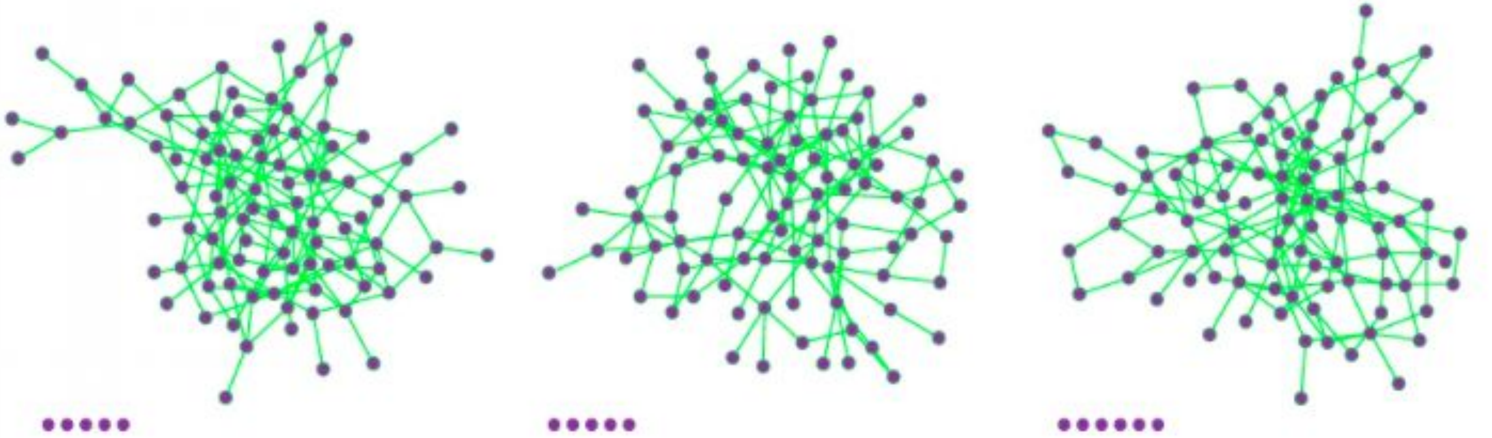
$$\langle L \rangle = \sum_{L=0}^{\frac{N(N-1)}{2}} L p_L = p\frac{N(N-1)}{2} \qquad (3.2)$$

Random Networks (RN)

3 realizations of a RNs generated with the same parameters: p=1/6 and N=12. Despite the identical parameters, the networks not only look different, but they have a different number of links:
(L=10, 10, 8).

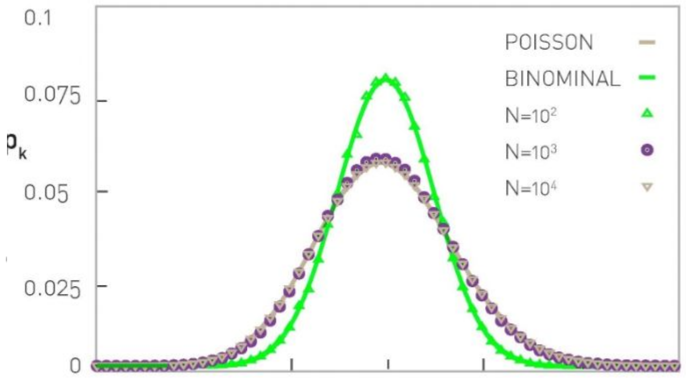


3 realizations of a RN with p=0.03 and N=100. Several nodes have degree k=0, shown as isolated nodes at the bottom



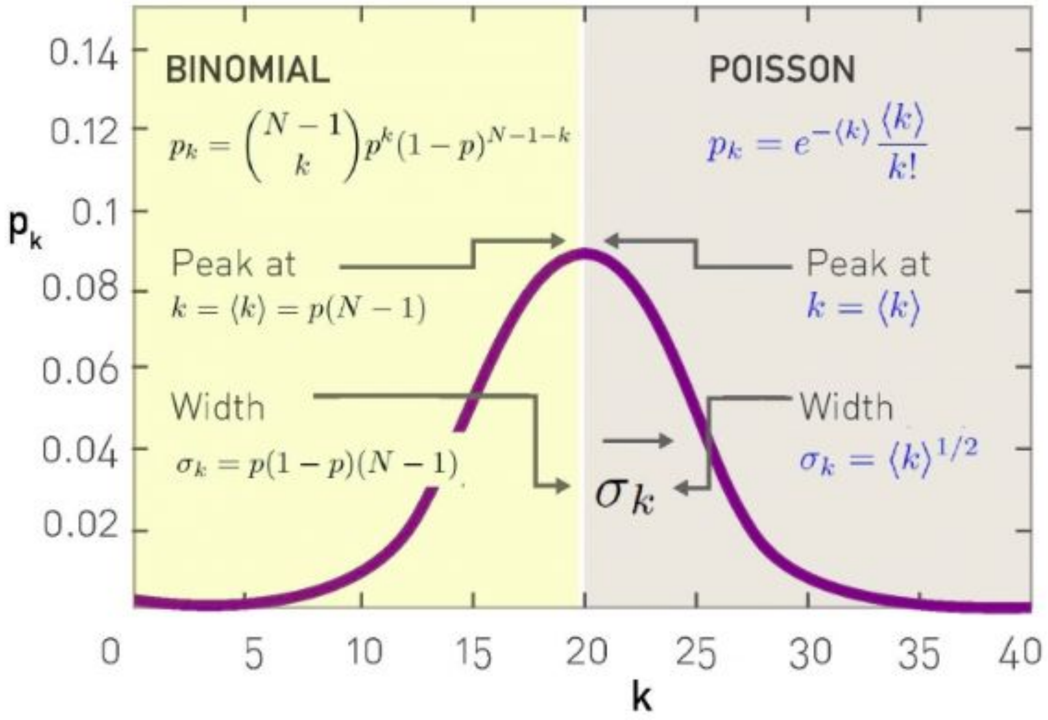
Poisson distribution

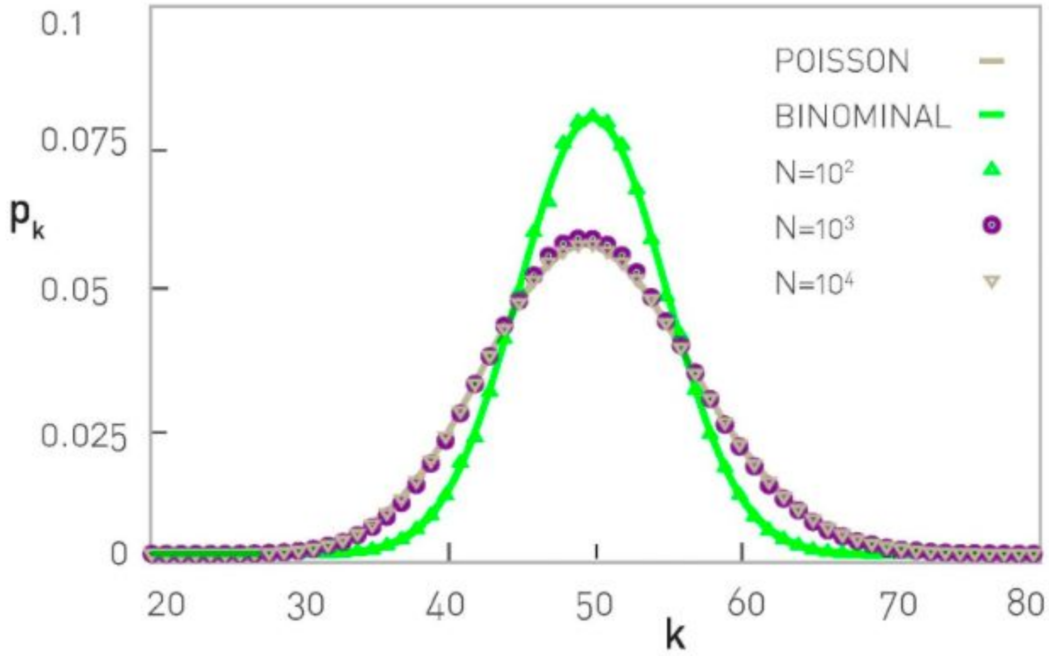
- The degree distribution of a random network with $\langle k \rangle = 50$, and N= 100, 1000, 10^4
- Small Networks: BINOMIAL
  - For N=100, the degree distribution deviates significantly from the Poisson form, as the condition for the poisson aprox N>> $\langle k \rangle$ is not satisfied. Hence for small networks, one needs use the exact binomial form.



- Large Networks: POISSON
  - For large networks (N=10^3,10^4) the degree distribution becomes indistinguishable from the Poisson prediction. Therefore, for large N the degree distribution is independent of the network size.

Random Networks: Degree Distribution

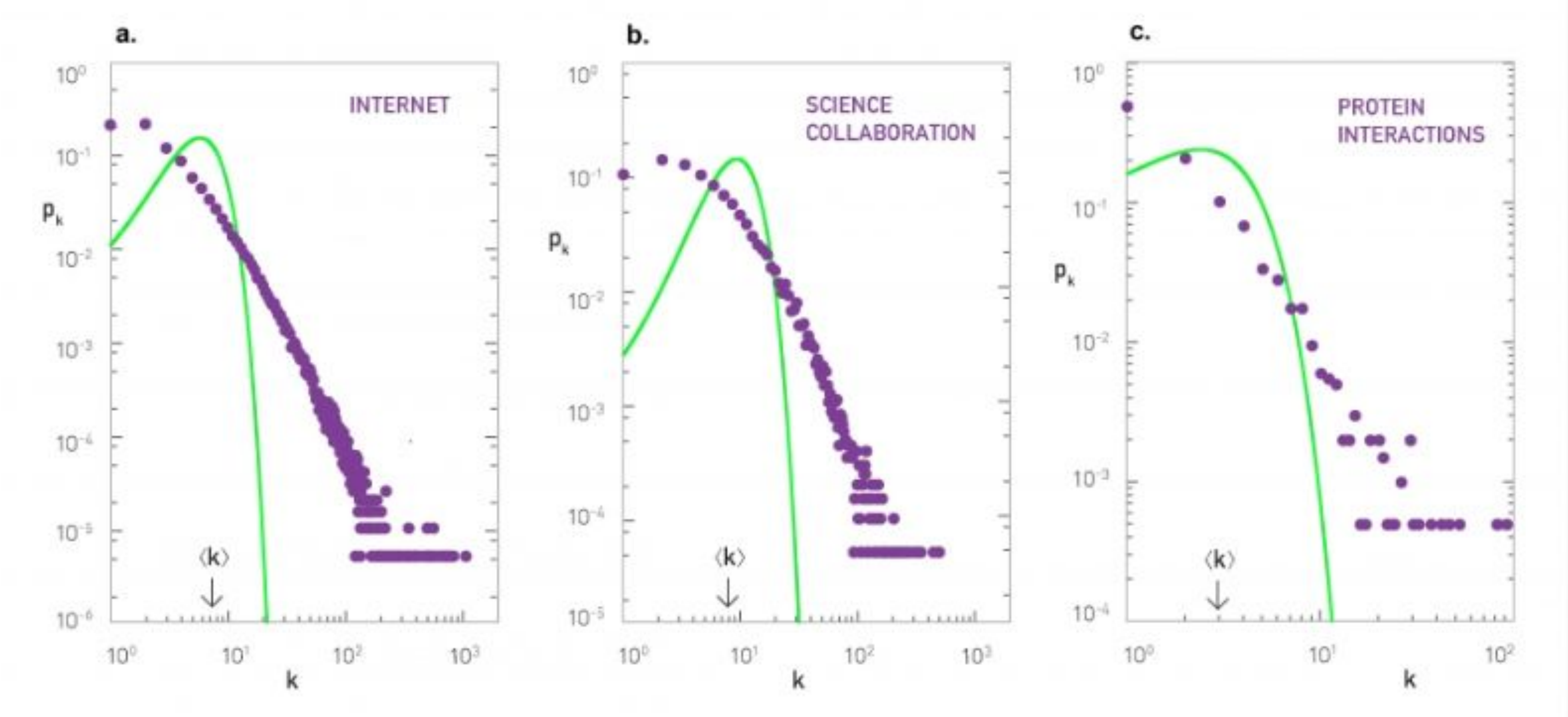




<u>Binomial vs. Poisson Degree Distribution</u>

The exact form of the degree distribution of a random network is the binomial distribution. For N ›› ‹k› the binomial is well approximated by a Poisson distribution. Both describe the same distribution, they have the identical properties, but they are expressed in terms of different parameters: The binomial distribution depends on p and N, while the Poisson distribution depends only on ‹k›.

Random Networks (RN)



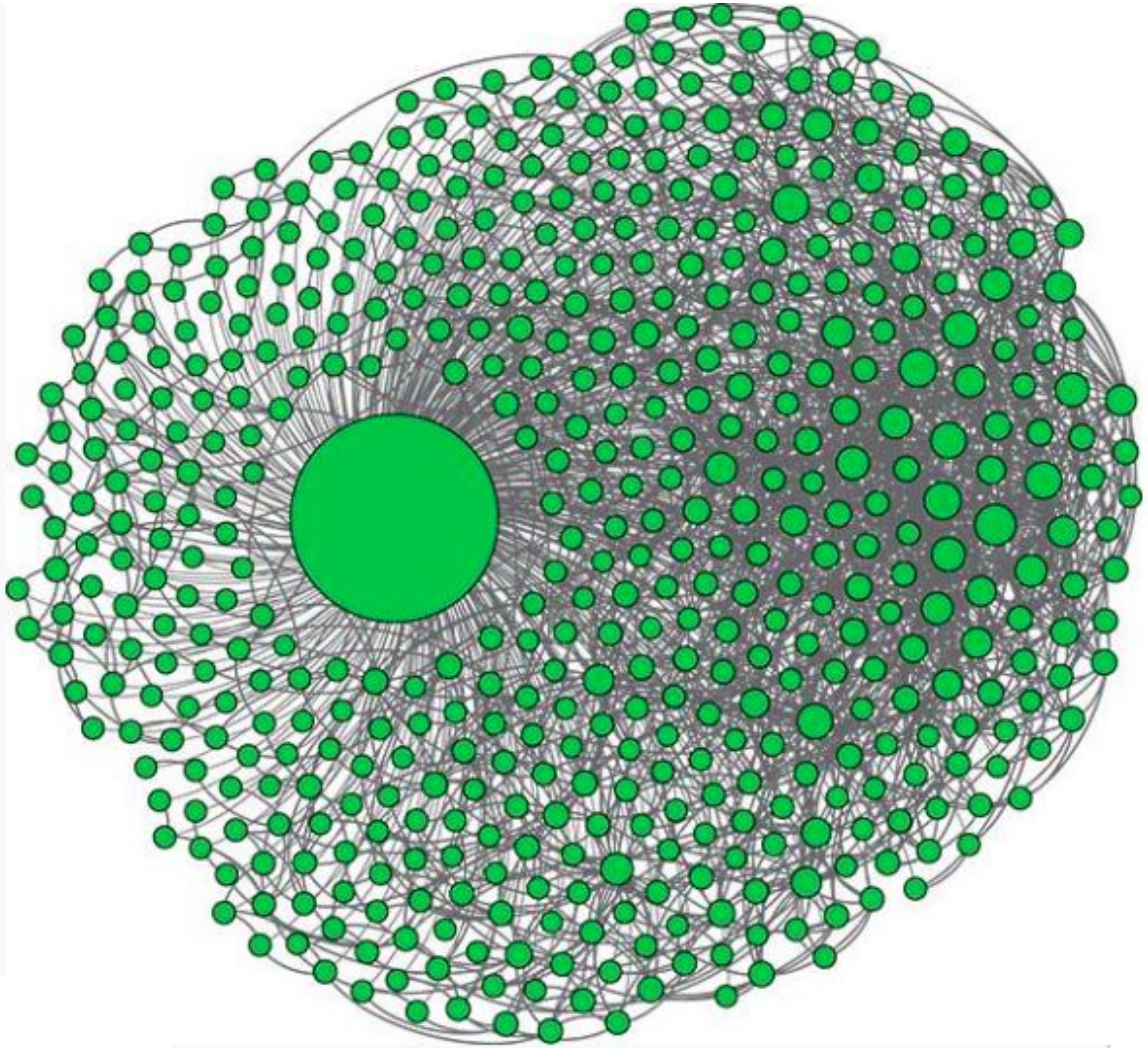
Coauthorship networks

- Nodes are scholars, when a publication is co-authored by two or more scholars, we can infer links between them in the network.
- Paul Erdős, a mathematician, made critical contributions to network science.
- Mathematicians are fond of studying their distance in the coauthorship network from the particular node corresponding to Erdős. They call this distance their *Erdős number.*
- Many mathematicians have a very small Erdős number.
- There is even an online tool to compute the Erdős number for mathematicians (www.ams.org/mathscinet/collaborationDistance.html)

Erdős collaboration network

Image: illustrates the network of collaborations involving Erdős and his over 500 coauthors. In reality, scholars are not just close to Erdős; they are close to everyone. This is typical of collaboration networks: there are short paths among all pairs of nodes. Pick any two scholars and they will not be very far from each other.

Source: Menczer et al: Filippo Menczer, Santo Fortunato, Clayton A. Davis, [A First Course in Network Science. Cambridge University Press 2020]



19 Degrees of Separation

How many clicks do we need to reach a randomly chosen document on the Web? The difficulty in addressing this question is rooted in the fact that we lack a complete map of the WWW—we only have access to small samples of the full map. We can start, however, by measuring the WWW's average path length in samples of increasing sizes, a procedure called *finite size scaling*. The measurements indicate that the average path length of the WWW increases with the size of the network as [21]

$$\langle d \rangle \approx 0.35 + 0.89 \ln N$$

In 1999 the WWW was estimated to have about 800 million documents [22], in which case the above equation predicts $\langle d \rangle \approx 18.69$. In other words in 1999 two randomly chosen documents were on average 19 clicks from each other, a result that became known as *19 degrees of separation*. Subsequent measurements on a sample of 200 million documents found $\langle d \rangle \approx 16$ [23], in good agreement with the $\langle d \rangle \approx 17$ prediction. Currently the WWW is estimated to have about trillion nodes ($N \sim 10^{12}$), in which case the formula predicts $\langle d \rangle \approx 25$. Hence $\langle d \rangle$ is not fixed but as the network grows, so does the distance between two documents.

The average path length of 25 is much larger than the proverbial six degrees (BOX 3.7). The difference is easy to understand: The WWW has smaller average degree and larger size than the social network. According to (3.19) both of these differences increase the Web's diameter.

Six degrees of separation: Small World

- Pretty much all social networks have very short paths among nodes
- It is likely to know someone who knows someone who knows someone… and in a few steps get to anyone on the planet!
- *Six Degrees of Kevin Bacon* is a fun game that originates from such a network
  - Play this game online at The Oracle of Bacon (oracleofbacon.org)
  - The website pulls data to build the network from the Internet Movie Database (IMDB.com)
  - The Wiki game. You will be amazed at how quickly you can reach any target with a bit of practice. Wikipedia has short paths.
- Milgram experiment: small world

Milgram Experiment

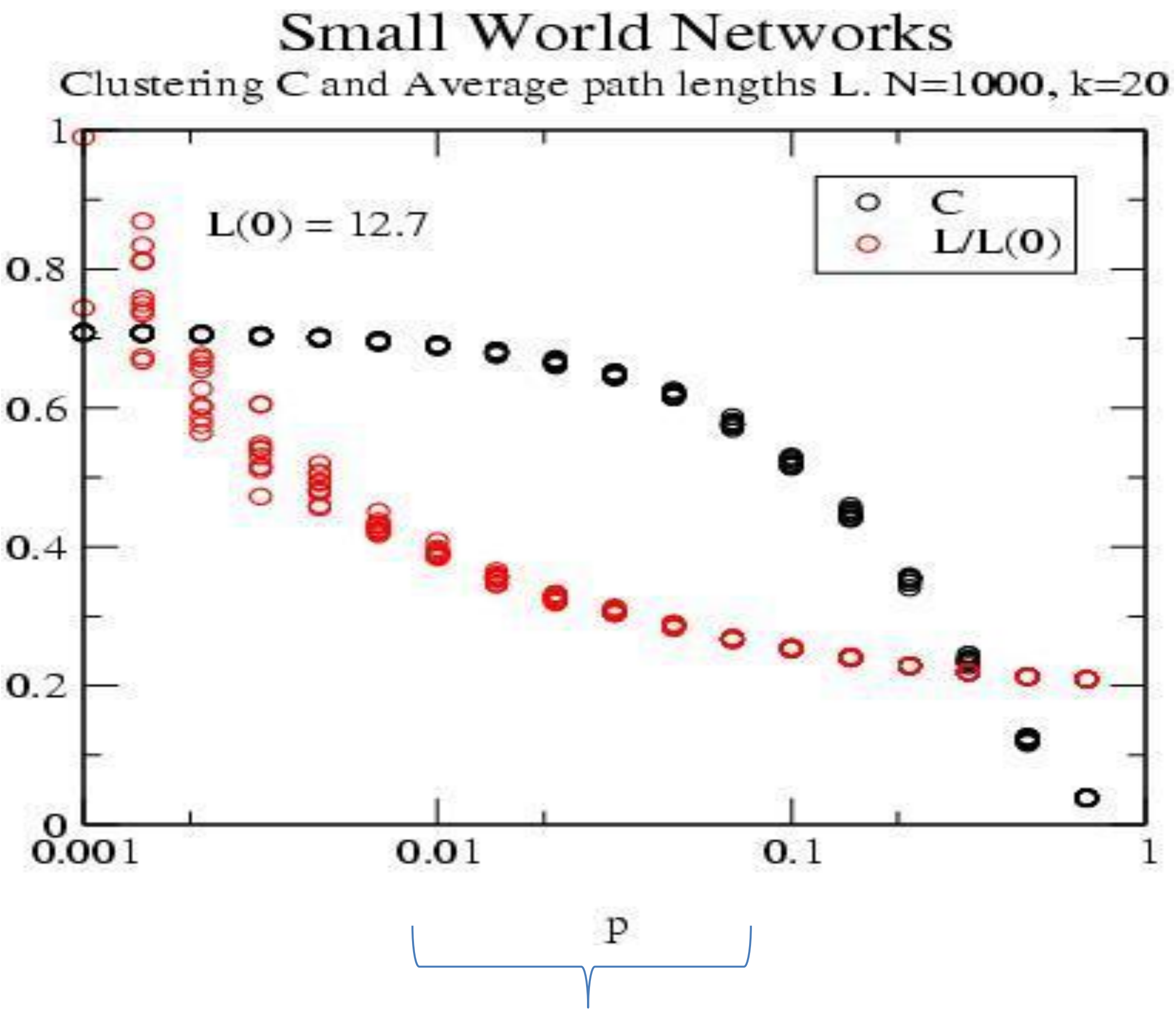
1. Individuals in the U.S. cities of [Omaha, Nebraska](), and [Wichita, Kansas](), to be the starting points and [Boston, Massachusetts](), to be the end point of a chain of correspondence. These cities were selected because they were thought to represent a great distance in the United States, both socially and geographically.
2. Information packets were initially sent to 160 "randomly" selected individuals in Omaha or Wichita. They included letters, which detailed the study's purpose, and basic information about a target contact person in Boston.
3. Upon receiving the invitation to participate, the recipient was asked whether he or she personally knew the contact person described in the letter. If so, the person was to forward the letter directly to that person.
4. In the more likely case that the person did not personally know the target, then the person was to think of a friend or relative who was more likely to know the target.

64 letters eventually reached Boston out of 296, requiring close to a dozen intermediates

Mundos pequeños

- Experimento de Milgram (1960): hay 6 grados de separación entre dos individuos cualesquiera elegidos al azar.
- Si se considera la población mundial del orden N=6*10^9, y la cantidad promedio de personas conocidas por un individuo $\langle k \rangle = 200$, se obtendrá $L_{rand} \approx 4.2$ < 6.
- La cuestión clave: es diferente de la red aleatoria?

MUNDOS PEQUEÑOS



Se han estudiado varios ejemplos de sistemas dinámicos cuyas partes están conectadas con un red del tipo *mundo pequeño*. En ellas se observa un aumento de la velocidad en la propagación de señales y una mejor sincronización entre nodos. Ej: infecciones.

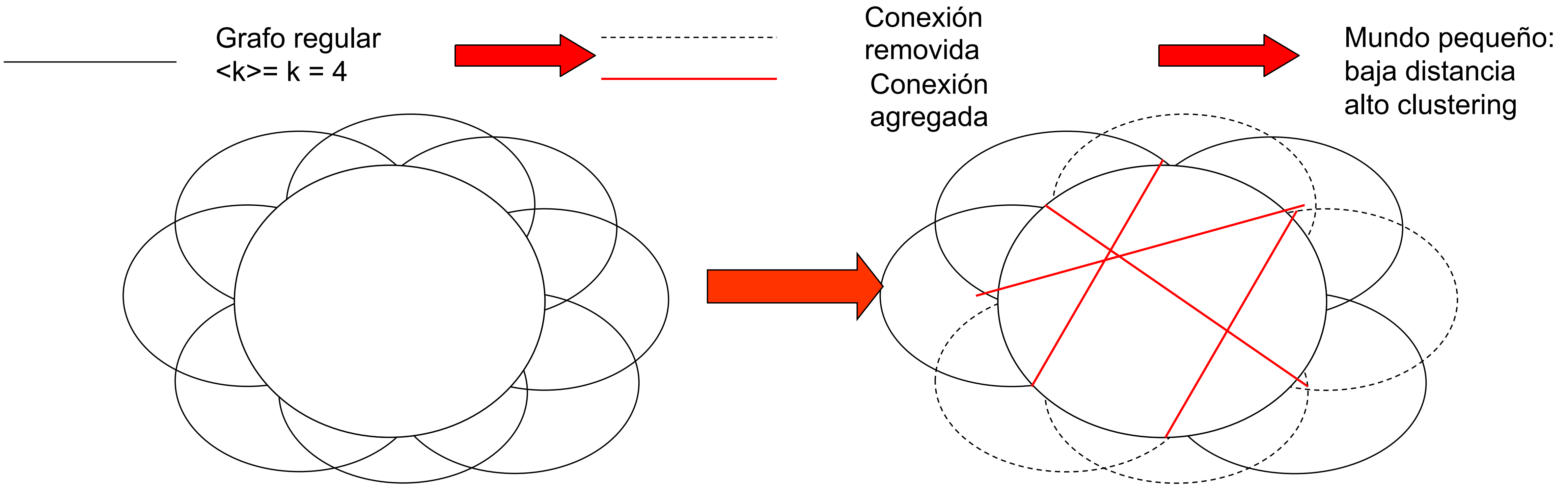
Una red nerviosa propia de un mundo pequeño ofrece entonces ventajas evolutivas sobre una red aleatoria (C. Elegans)

Régimen del mundo pequeño: clustering C es mucho más elevado que para el grafo aleatorio

MUNDOS PEQUEÑOS

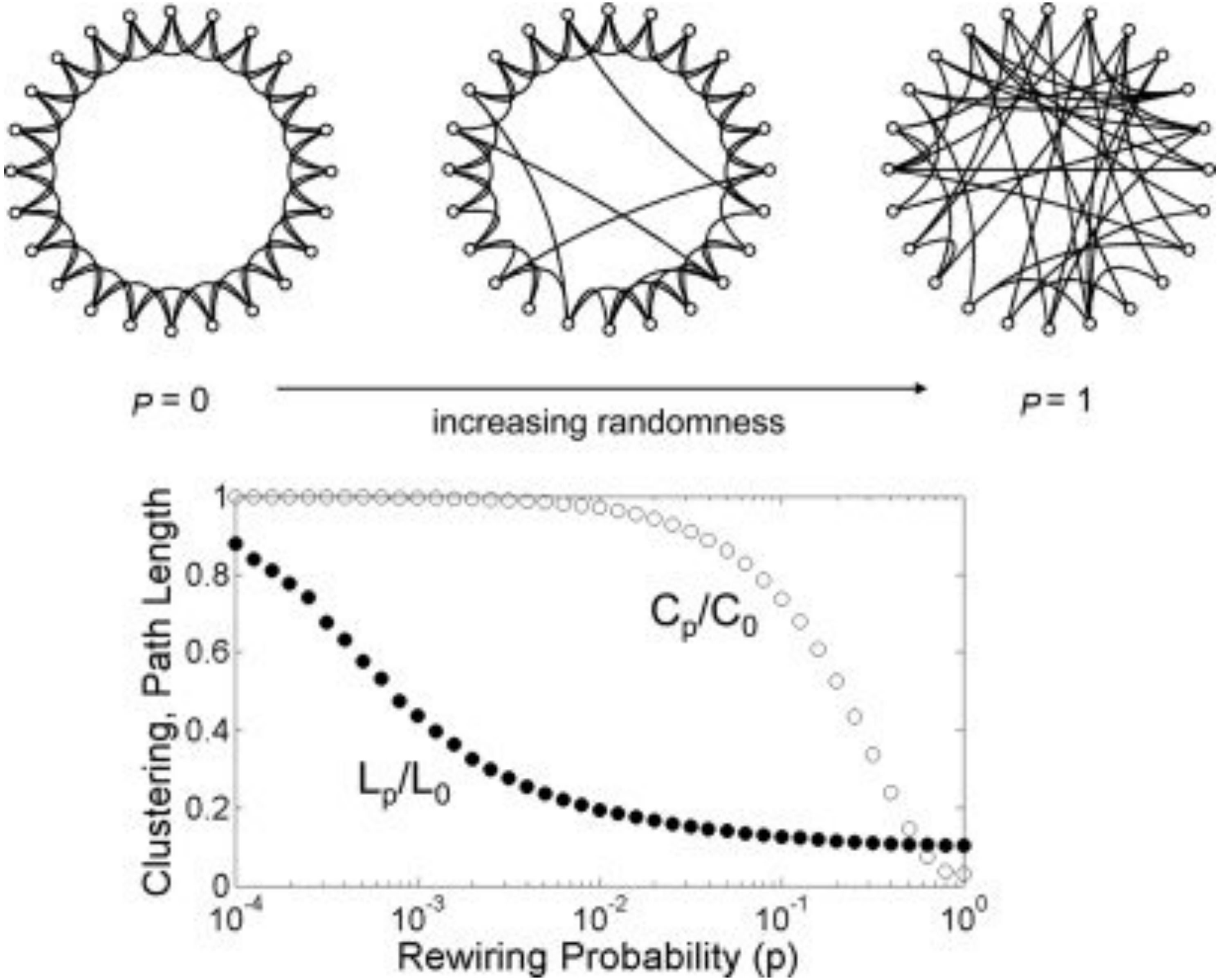
Watts y Strogatz observan que la mayoría de las redes sociales tienen un *elevado clustering* (mayor que en un grafo aleatorio) y *corta distancia* entre nodos (como en un grafo aleatorio).

Introducen la idea de redes que llaman "mundos pequeños". *La red de un **mundo pequeño*** se construye partiendo de una red regular y cambiando una arista por otra (eligiendo al azar el otro nodo) con una probabilidad *p* en cada nodo



Grafo regular

Mundo pequeño

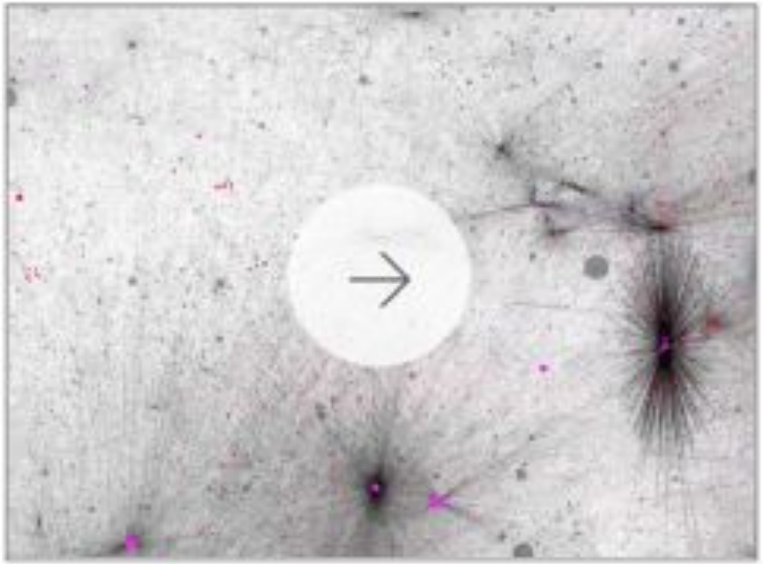


La Plata, Uruguay

Small World Networks

- With p edges are disconnected and attached to a randomly chosen other node
- For $p=1$: all edges are reconnected and a random network is obtained
  - Low clustering, short path length
- For $0<p<1$: only small fraction of rewired links
  - Path length already drops to very low values
  - Clustering coefficient still maintains high
  - □ small world network
  - A very powerful model of many real networks (that display high clustering and short path length)
  - However! Fails to explain important properties of natural networks, s.t. modularity and broad degree distributions with hubs

WWW network: perfect example of SF



- Nodes are documents
- Links are URL (Uniform Resource Locations)
- Web is the largest network, aprox $N \approx 10^{12}$
- First map obtained in order to understand the structure of the network behind WWW (Hawoong Jeong, Univ. Notre Dame)
  - 300 000 documents
  - 1.5 million of links
- The purpose: compare the properties of the WEB graph with the RN model
- Close inspection reveals interesting differences:
  - Numerous small-degree nodes coexist with a few hubs, nodes an exceptionally large number of links
  - In RN, highly connected nodes , or hubs, are effectively forbidden

WWW network: cont.

- If the WWW were to be a random network, the degrees of the Web documents should follow a Poisson distribution. Yet, the Poisson form offers a poor fit for the WWW's degree distribution.

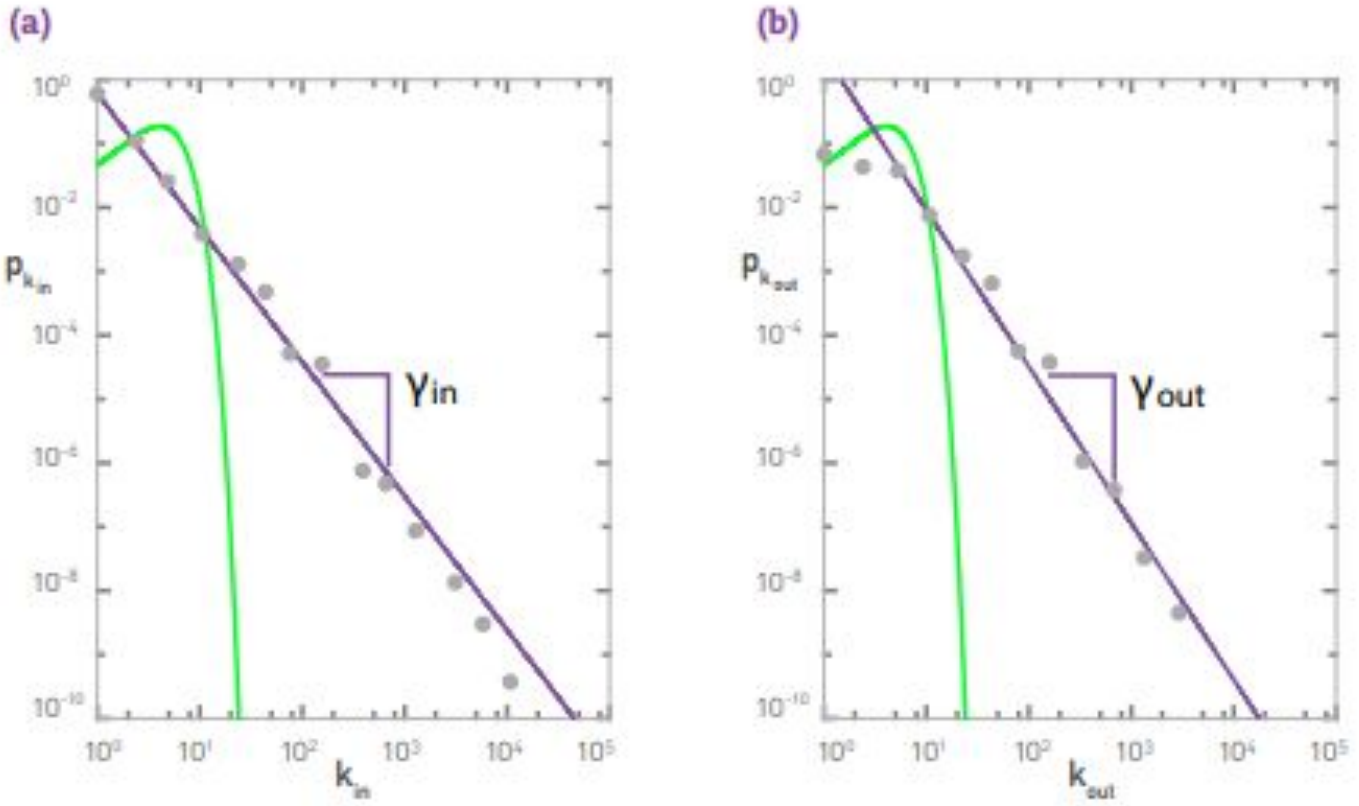
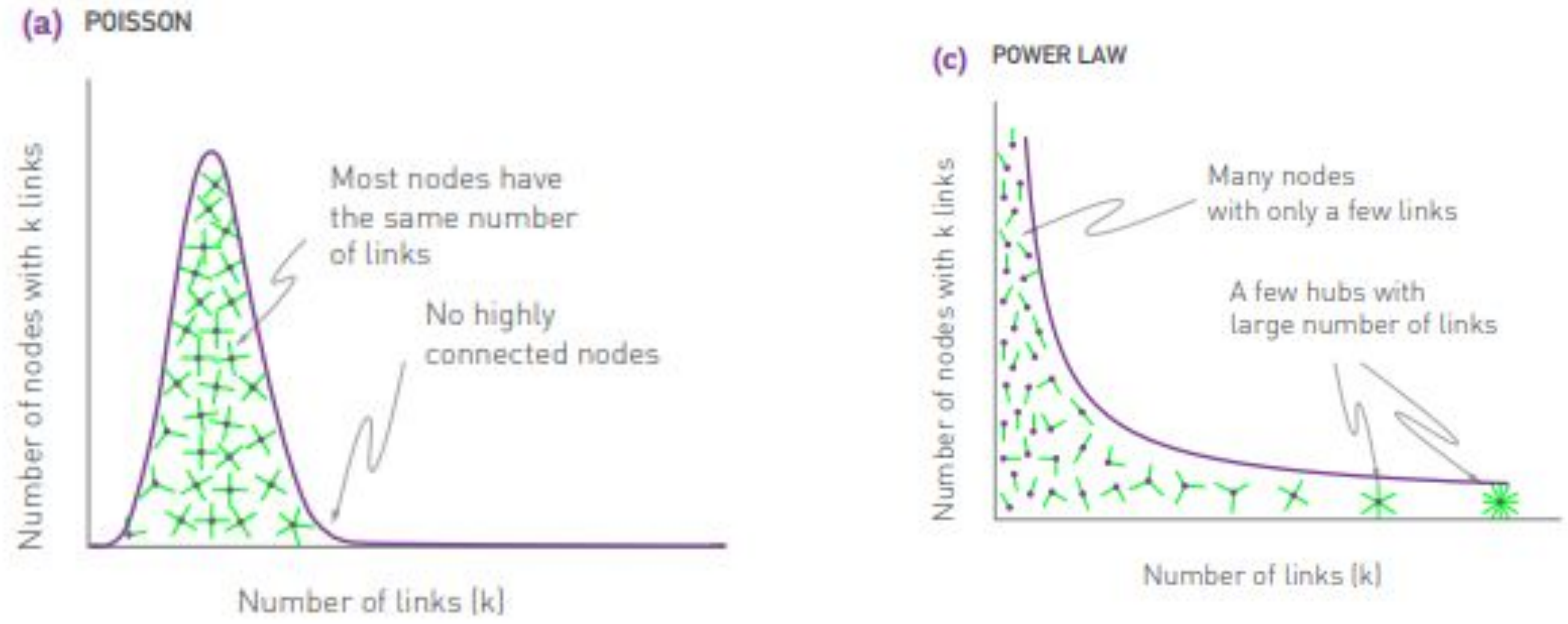


Figure 4.2

The Degree Distribution of the WWW

The incoming **(a)** and outgoing **(b)** degree distribution of the WWW sample mapped in the 1999 study of Albert *et al.* [1]. The degree distribution is shown on double logarithmic axis (log-log plot), in which a power law follows a straight line. The symbols correspond to the empirical data and the line corresponds to the power-law fit, with degree exponents $\gamma_{in} = 2.1$ and $\gamma_{out} = 2.45$. We also show as a green line the degree distribution predicted by a Poisson function with the average degree $\langle k_{in} \rangle = \langle k_{out} \rangle = 4.60$ of the WWW sample.

Poisson vs Power law



WWW network: cont.

- The main difference between a random and a scale-free network comes in the tail of the degree distribution, representing the high-k region of p_k
- For small k the power law is above the Poisson function, indicating that a scale-free network has a large number of small degree nodes, most of which are absent in a random network
- For k in the vicinity of ⟨k⟩ the Poisson distribution is above the power law, indicating that in a random network there is an excess of nodes with degree k≈⟨k⟩
- For large k the power law is again above the Poisson curve. The difference is particularly visible if we show p_k on a *log-log* plot, indicating that the probability of observing a high-degree node, or hub, is several orders of magnitude higher in a scale-free than in a random network.

WWW network: cont.

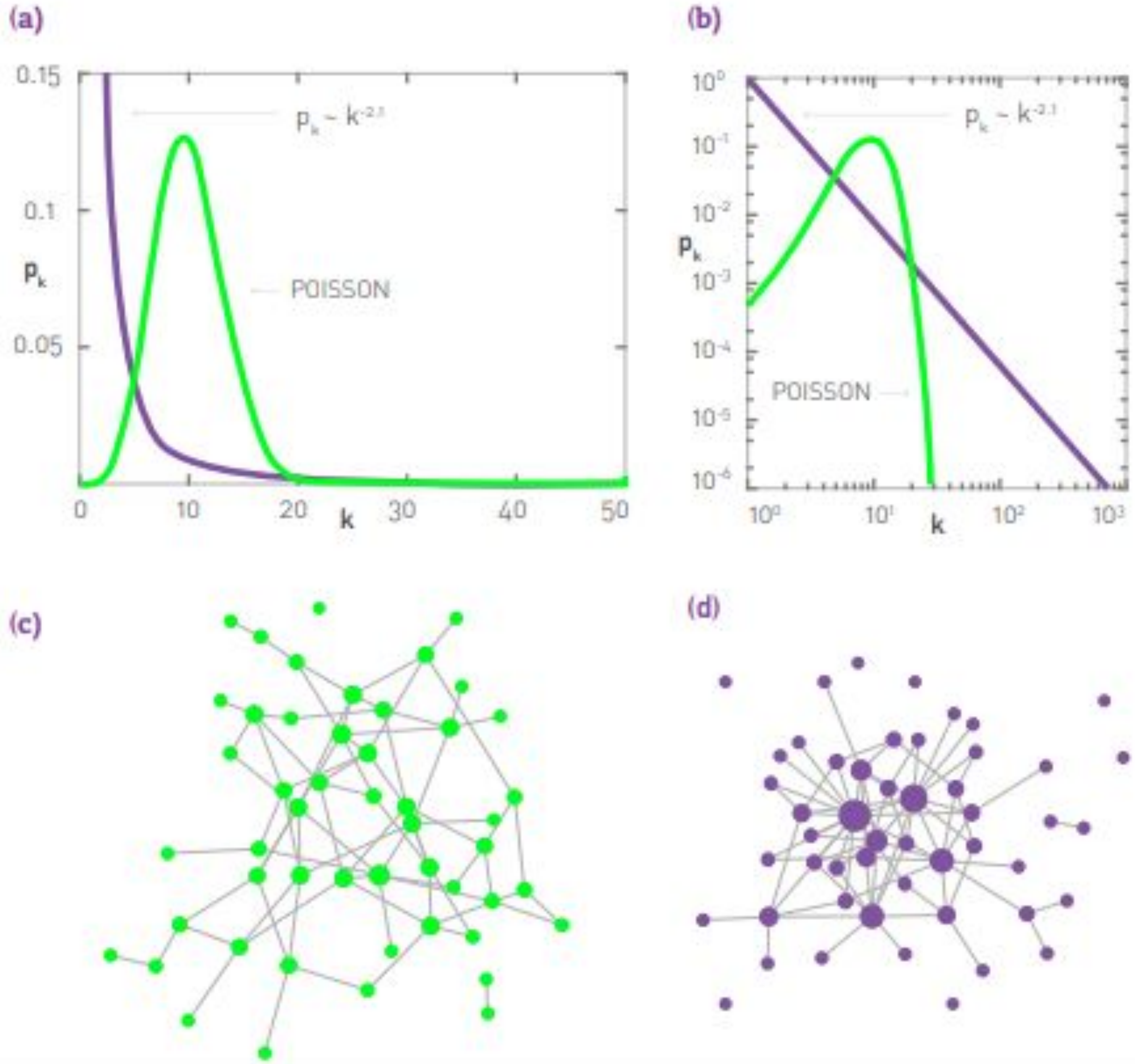


Figure 4.4
Poisson vs. Power-law Distributions

(a) Comparing a Poisson function with a power-law function ($\gamma = 2.1$) on a linear plot. Both distributions have $\langle k \rangle = 11$.

(b) The same curves as in (a), but shown on a log-log plot, allowing us to inspect the difference between the two functions in the high-k regime.

(c) A random network with $\langle k \rangle = 3$ and $N = 50$, illustrating that most nodes have comparable degree $k \approx \langle k \rangle$.

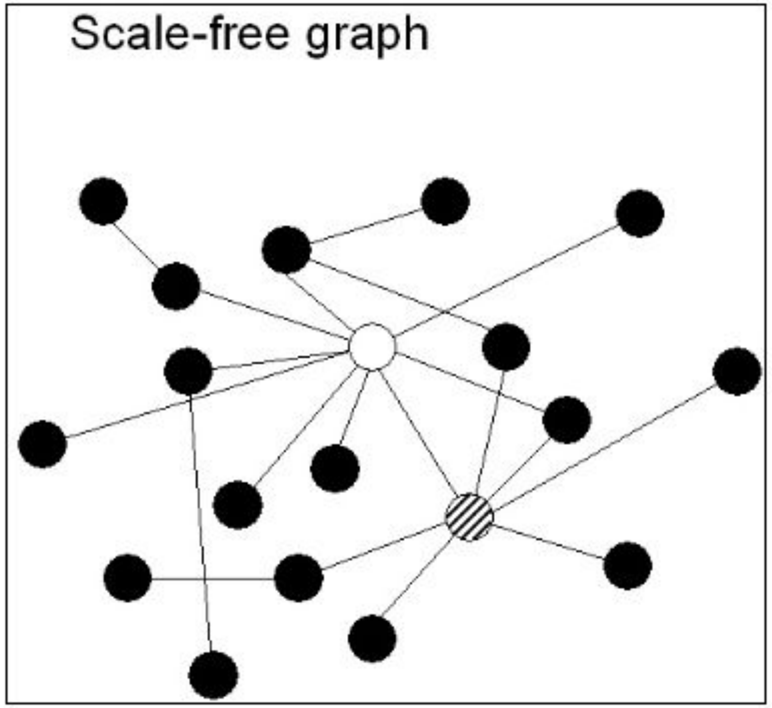
(d) A scale-free network with $\gamma = 2.1$ and $\langle k \rangle = 3$, illustrating that numerous small-degree nodes coexist with a few highly connected hubs. The size of each node is proportional to its degree.

Redes libres de escala

- Barabási y Albert (1999) mostraron que cuando las redes grandes están formadas por las reglas de vinculación preferencial, la red resultante muestra una distribución de la ley de potencia de los grados del nodo: $P(k) \approx Ak^{-\gamma}$ con $2<\gamma<3$.
- Estas son redes libres de escala: los nodos no tienen una cantidad de vecinos que sea característica

Scale-free Networks

- Barabasi and Albert proposed a model of a growing network. At each iteration, a new vertex is added, and it is connected to existing vertices with a probability that depends upon the degree of that node. As a consequence, nodes with a high degree are more likely to receive more connections, increasing their degree even further. This is an example of positive feedback or preferential attachment.
- The most interesting feature of the model is the shape of its degree distribution. After a sufficient number of iterations the degree distribution becomes a power law.
- While the goal of the other models (random graphs and small - world models) is to construct a graph with correct topological features, modelling scale - free networks puts the emphasis on capturing the network dynamics.



Scale-Free Networks: mechanism

- RN and SW: assume that we start with a fixed number N of nodes that are then randomly connected or rewired, without modifying N, the actual number of nodes.
- In contrast, most real world networks describe open systems, which grow by the continuous addition of new nodes. Starting from a small number of nodes, the number of nodes increases through
- Most real networks exhibit preferential attachment, such that the likelihood of connecting to a node depends on the node's degree
  - For example, a webpage will more likely include hyperlinks to popular documents with already high degree, because such highly connected documents are easy to find, etc.
- These two ingredients, growth and preferential attachment, inspired the introduction of the scale-free (SF) model that has a *power-law* degree distribution.

Scale-Free Network: Modelo Barabási-Albert

- Growth: Starting with a small number (m_0) of nodes, at every timestep we add a new node with $m \leq m_0$ edges that link the new node to m different nodes already present in the system.
- Preferential attachment: When choosing the nodes to which the new node connects, we assume that the probability p that a new node will be connected to node i depends on the degree k_i of node i, s.t.: $p(k_i) = \frac{k_i}{\sum_j k_i}$
- After *t* timesteps this algorithm results in a network with $N = t + m_0$ nodes and mt edges.
- Numerical simulations indicated that this network evolves into a scale-free state with the probability that a node has *k* edges following a power-law with an exponent 3, where the scaling exponent is independent of m, the only parameter in the model
- This power law distribution reflects the presence of large number of highly connected nodes or hubs
- However! even SF have their limitations: they do not explain clustering very well, they are not assortative, and have no real modules.

SF: metrics

- Average Path Length:
  - the average path length is smaller in the SF network than in a RN for any N, indicating that the heterogeneous SF topology is more efficient in bringing the nodes close than the homogeneous topology of random graphs. Average path length of the SF network increases approximately logarithmically with N.
  - Apart from the empirical fit there is no theoretical expression that would give a good approximation for the path length in the scale - free model.
  - The dynamical process that generates the network introduces nontrivial correlations that affect all topological properties.
- Clustering coefficient
  - no analytical prediction for the SF model.

¡Nos vemos en la próxima clase!