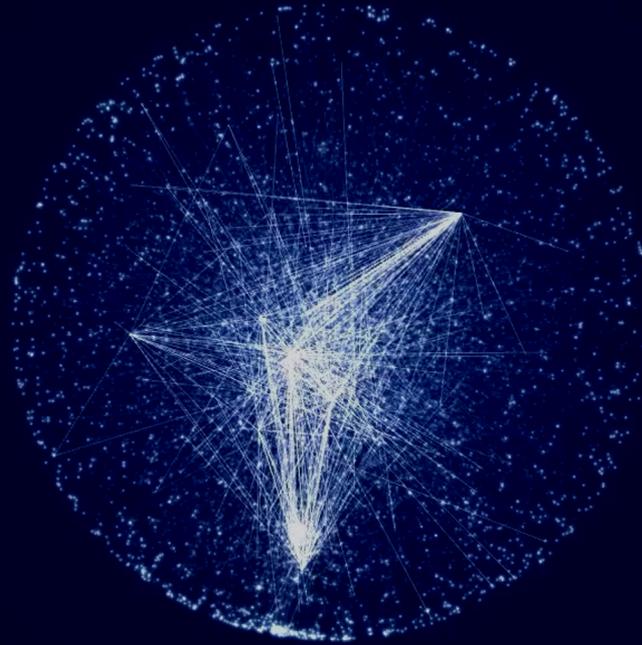


**PARTIMOS EN BREVE**

**MUCHAS GRACIAS**

El test comienza a las 11:55 en Canvas. Sección Evaluaciones

# 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

# Estructuras de Redes

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

- **Dominio: el area específica de estudio o aplicación para la cual la red será analizada.**
- El dominio captura el conjunto de *nodos* y *enlaces* que forman una red en particular, así como también las características y propiedades de esa red.

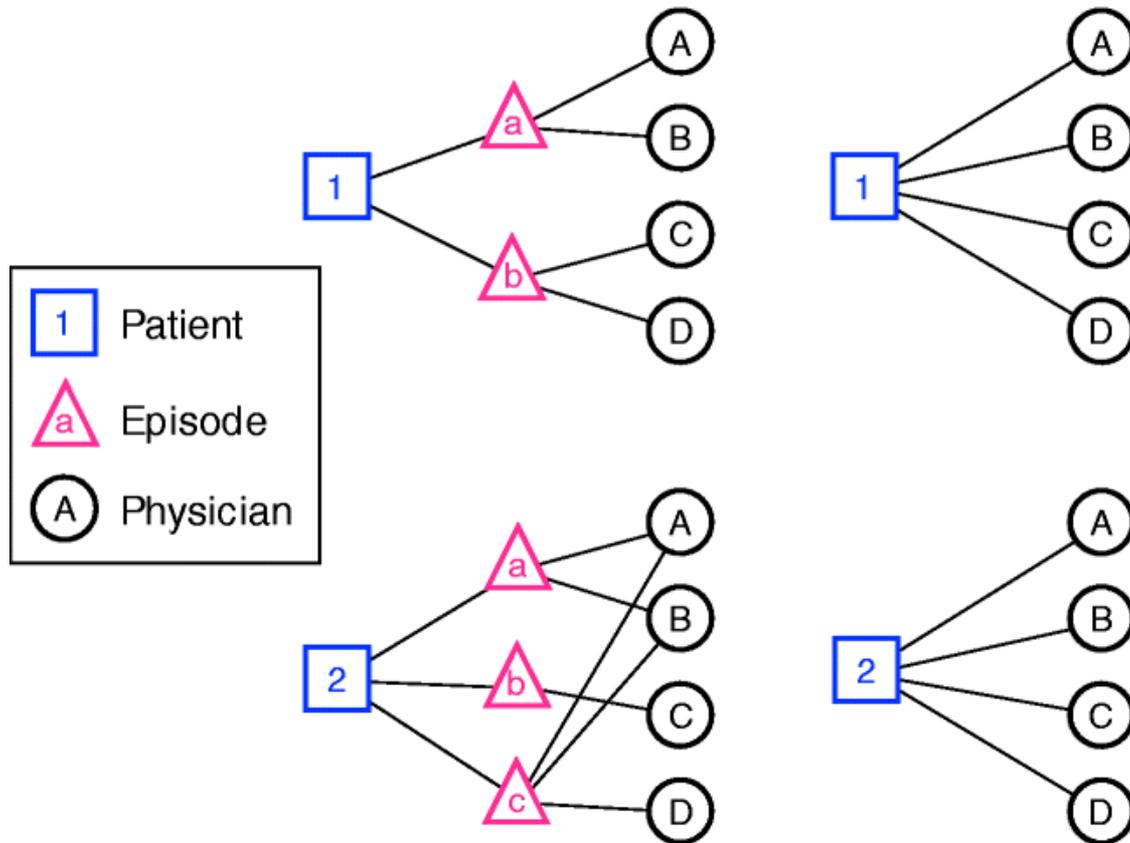
Preguntas a considerar:

- **Qué sistema representa esta red?** E.g. sistema biológico, sistema computacional, sistema eléctrico, sistema social (personas).
- **Cuál es el significado de las relaciones de los nodos?** E.g. patrones de comunicación en una comunidad en particular, interacciones biológicas entre genes o proteínas.
- **Qué medidas de red son útiles para capturar las relaciones entre los nodos?** E.g. in-degree centrality para comunicaciones dirigidas entre individuos, betweenness o eigenvector centrality como indicador de prestigio social, detección de comunidades para capturar grupos de ideología política, flujo y vulnerabilidad en redes de electricidad.

# Niveles de las Redes

Tripartite network

Bipartite network  
(patient-physician)

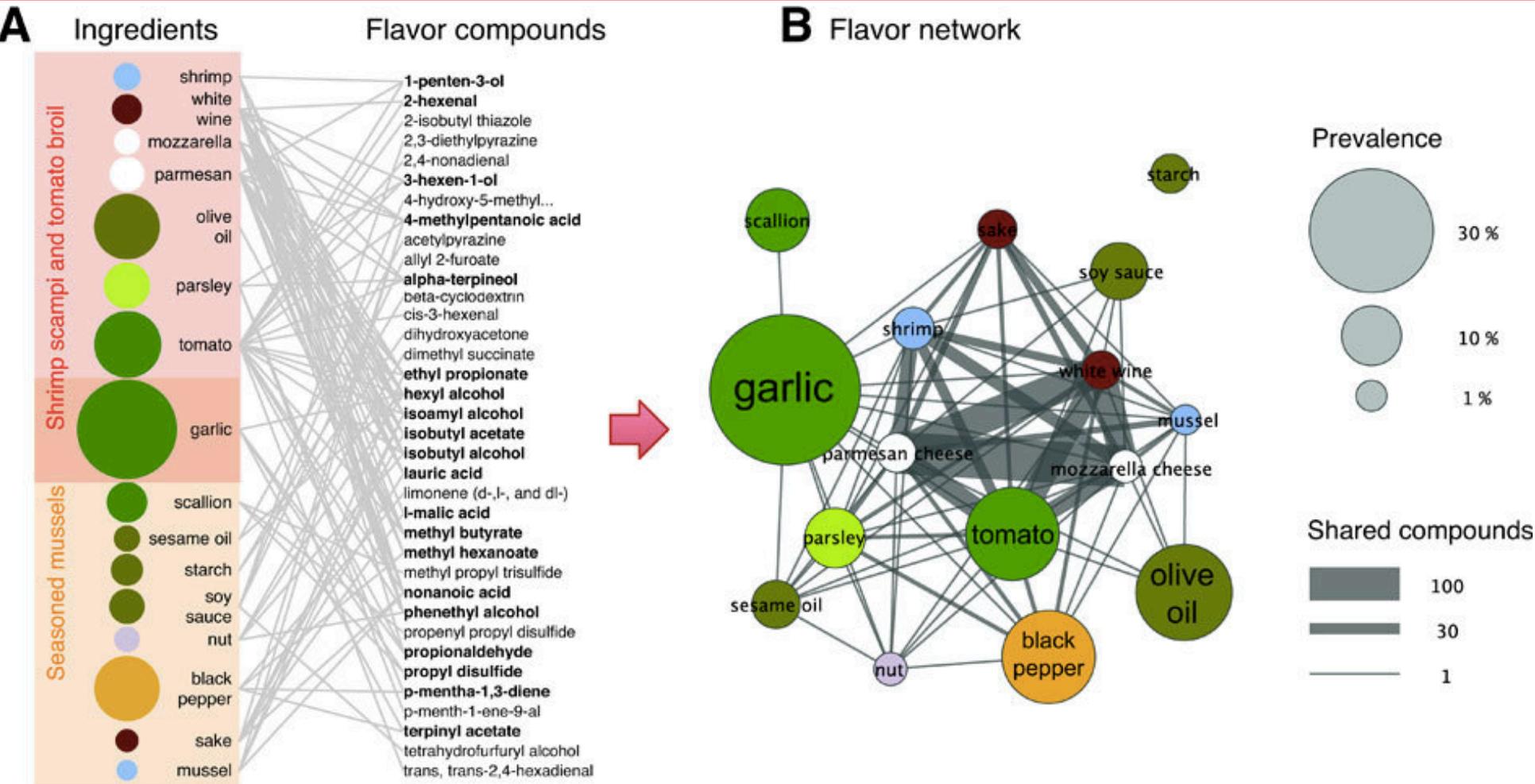


# GRAFOS BIPARTITOS

Un **grafo bipartito** (o bigrafo) es un grafo cuyos nodos se pueden dividir en dos **conjuntos separados**  $U$  y  $V$ , de manera que cada enlace conecta un nodo en  $U$  con uno en  $V$ ; es decir,  $U$  y  $V$  son conjuntos **independientes**.

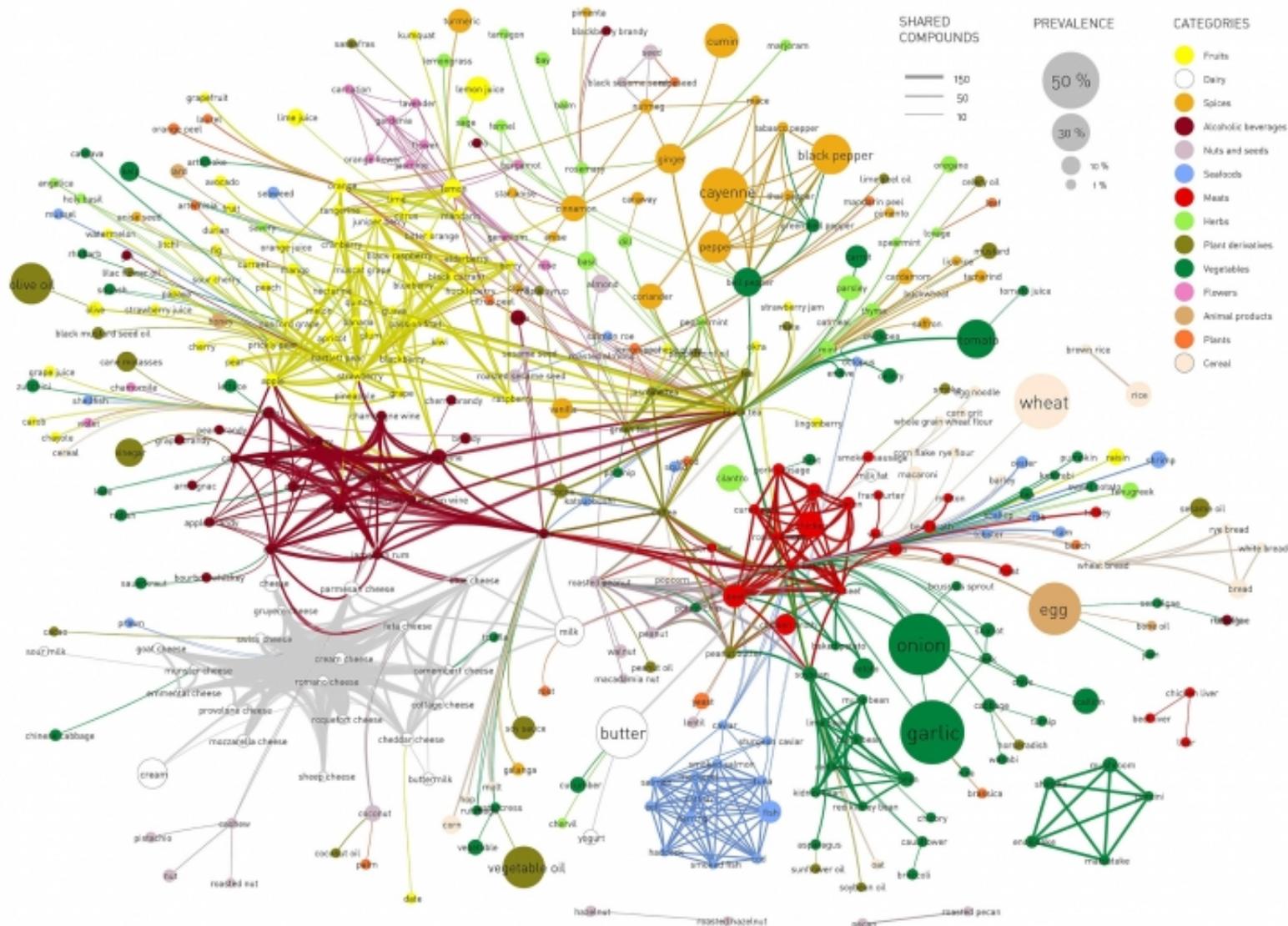
**Ejemplos:**

# Red bipartita ingredient-sabor

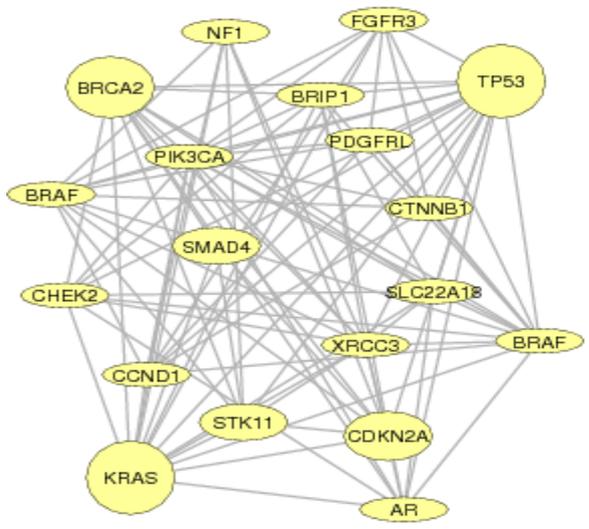


Y.-Y. Ahn, S. E. Ahnert, J. P. Bagrow, A.-L. Barabási  
 Flavor network and the principles of food pairing , *Scientific Reports* 196, (2011).

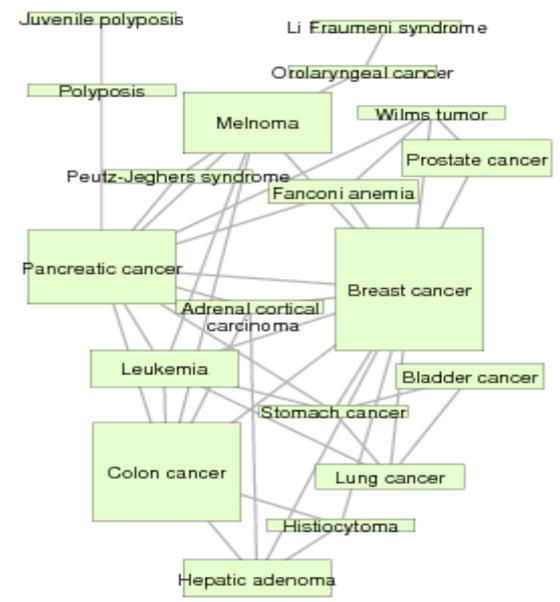
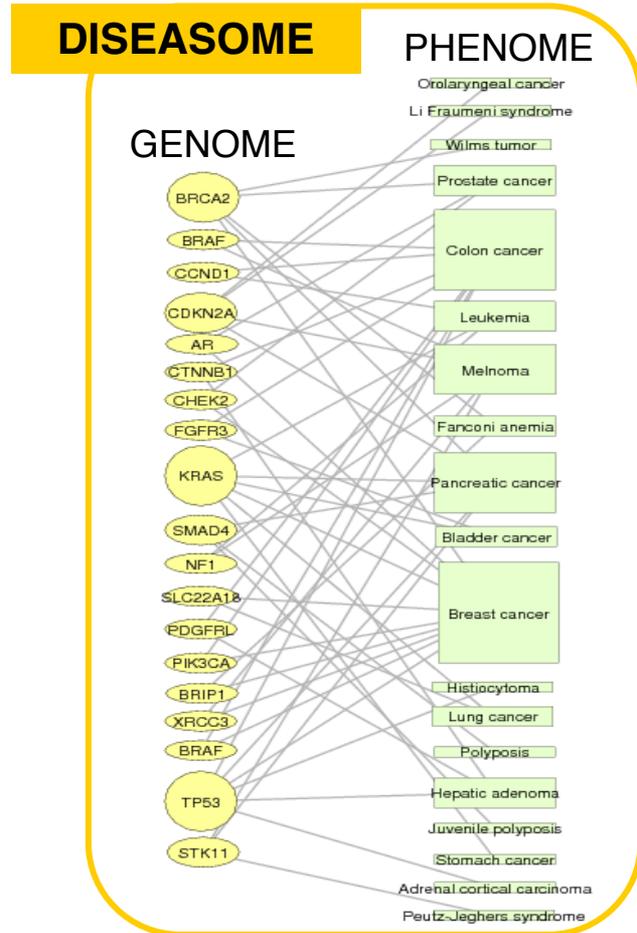
b.



# Red de genes– Red de enfermedades



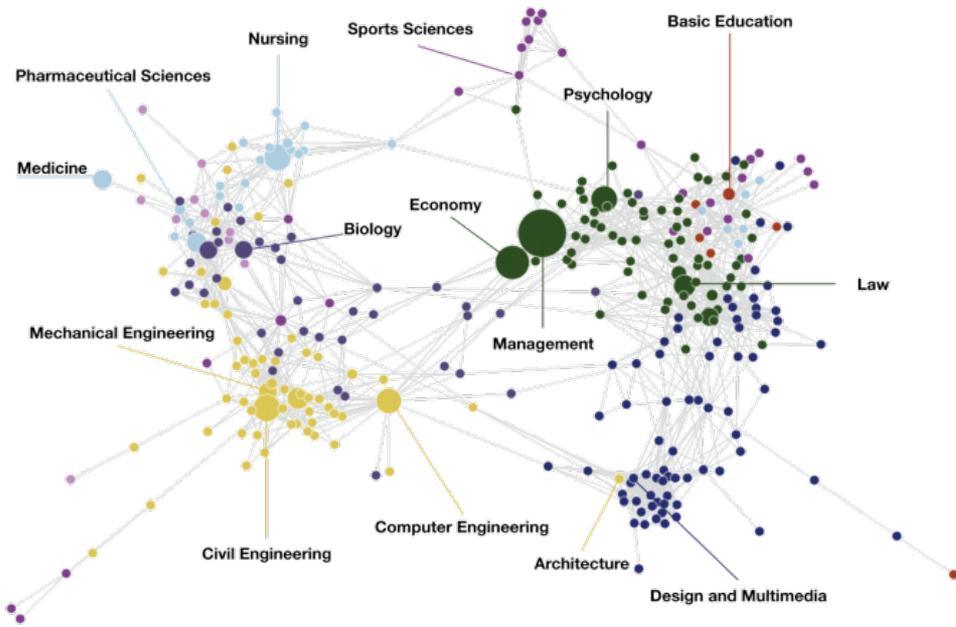
**Gene network**



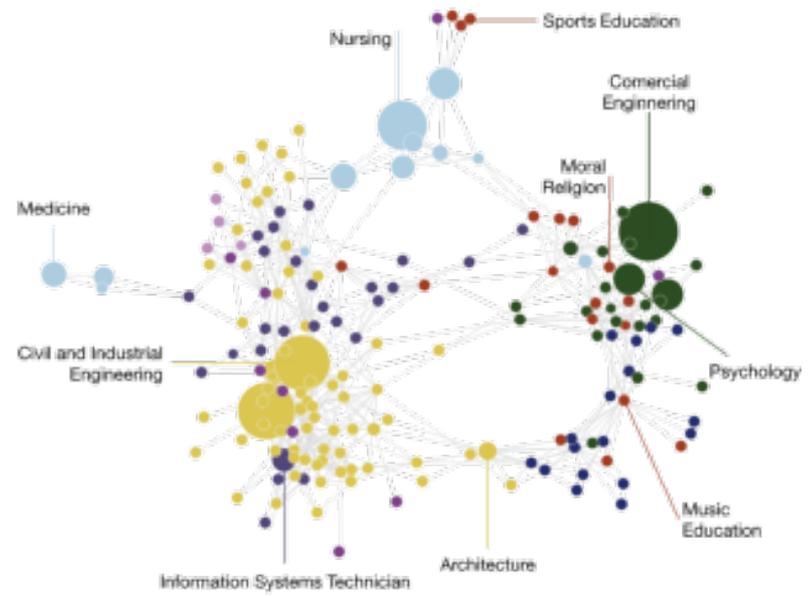
**Disease network**

# Red bipartita carreras-postulantes

a) Portuguese Higher Education System [2008-2015]



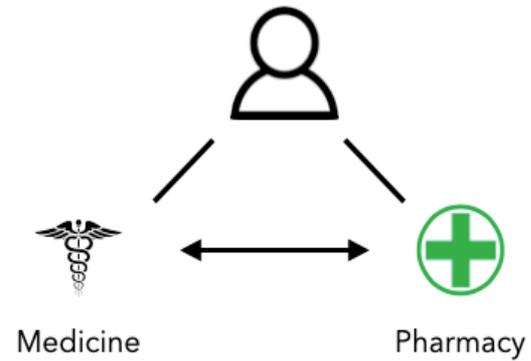
c) Chilean Higher Education System [2012-2017]



ISCED Classification		■ Arts and Humanities	■ Agriculture	■ Social Sciences	# Observations   Node Size min  max
■ Education	■ Sciences	■ Services	■ Engineering	■ Health	

## The Higher Education Space

Connecting Degree Programs



### Lista de Preferencias

1. Medicina
2. Odontología
3. Tecnología Médica
4. Odontología
5. Cs. Físicas y Astronómicas



### Pares de Carreras

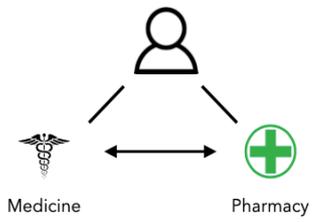
Medicina	—●—	Odontología	Odontología	—●—	Odontología
Medicina	—●—	Tecnología Médica	Odontología	—●—	Cs. Físicas y Astronómicas
Medicina	—●—	Odonotología	Tecnología Médica	—●—	Odontología
Medicina	—●—	Cs. Físicas y Astronómicas	Tecnología Médica	—●—	Cs. Físicas y Astronómicas
Odontología	—●—	Tecnología Médica	Odontología	—●—	Cs. Físicas y Astronómicas

Consideramos todas las preferencias de cada postulantes.  
Las preferencias repetidas indican una postulación a dos instituciones de educación superior distintas.

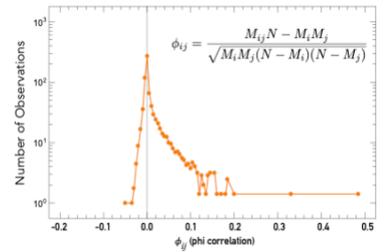
Creamos todos los pares de carreras posibles. Luego, descartamos los que contienen la misma carrera en ambos extremos (color gris).

# The Higher Education Space

Connecting Degree Programs

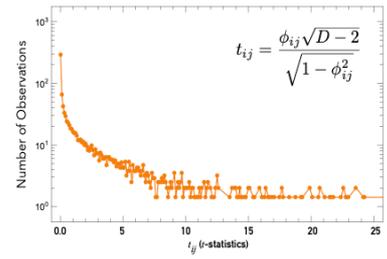


Finding Correlations Between Degrees

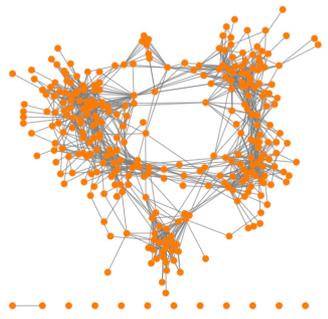


Discard all negative correlations

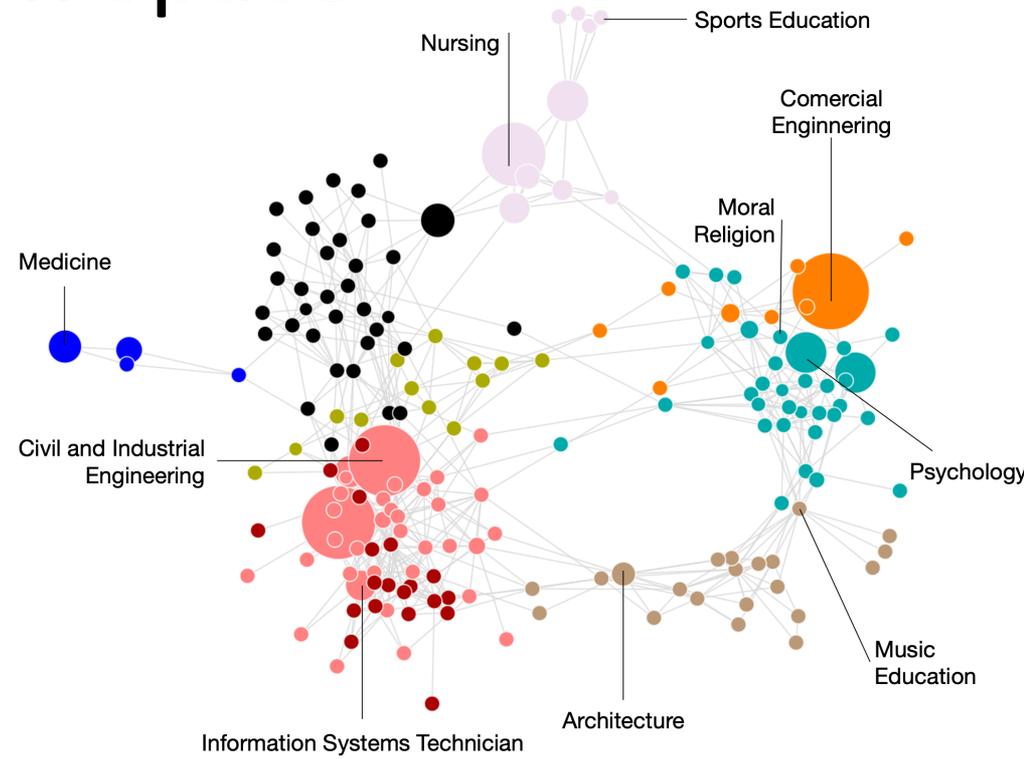
Finding Statistical Significance of Correlations



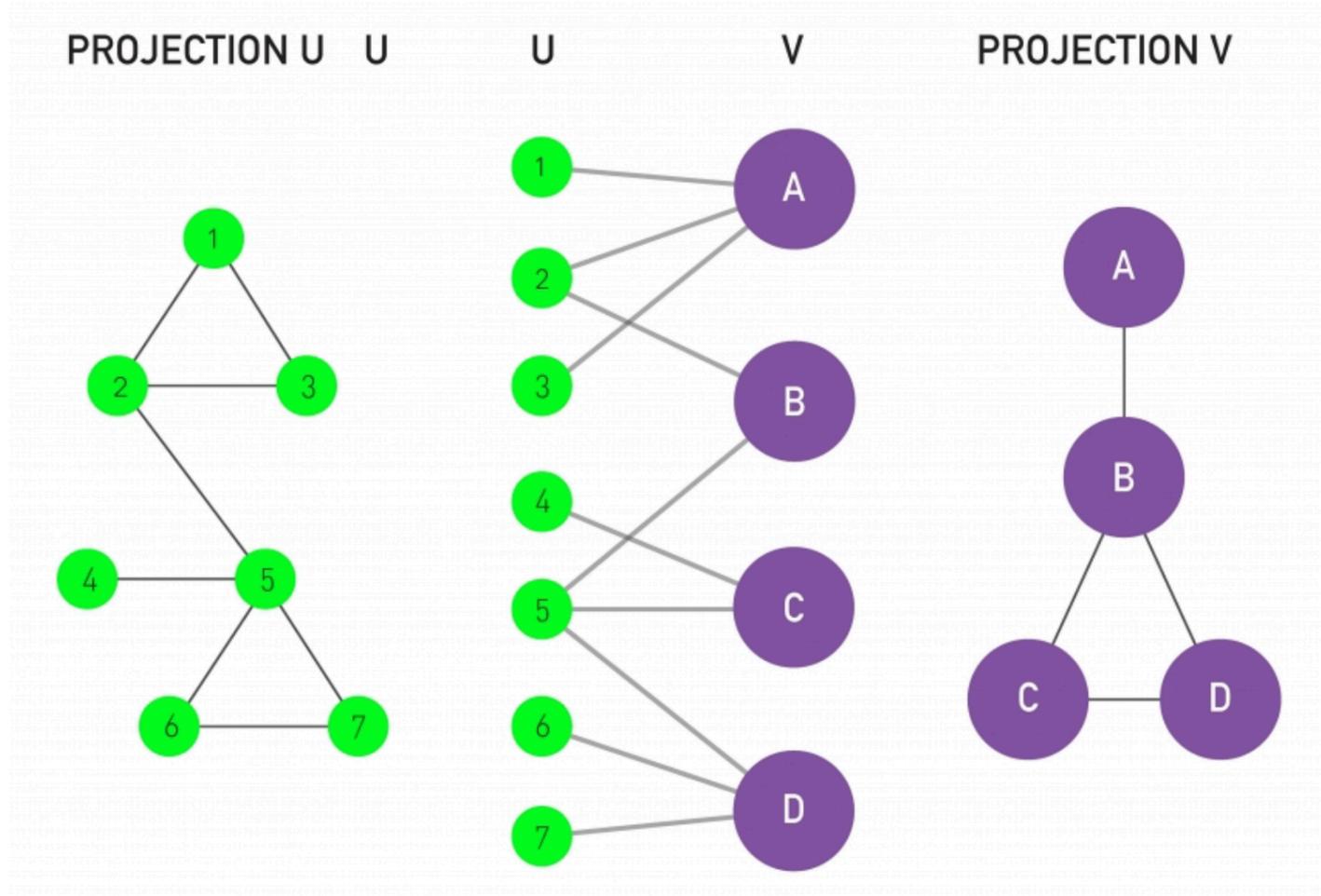
Discard all non-significant links



Discard all loose Nodes



# Proyecciones de una red bipartita

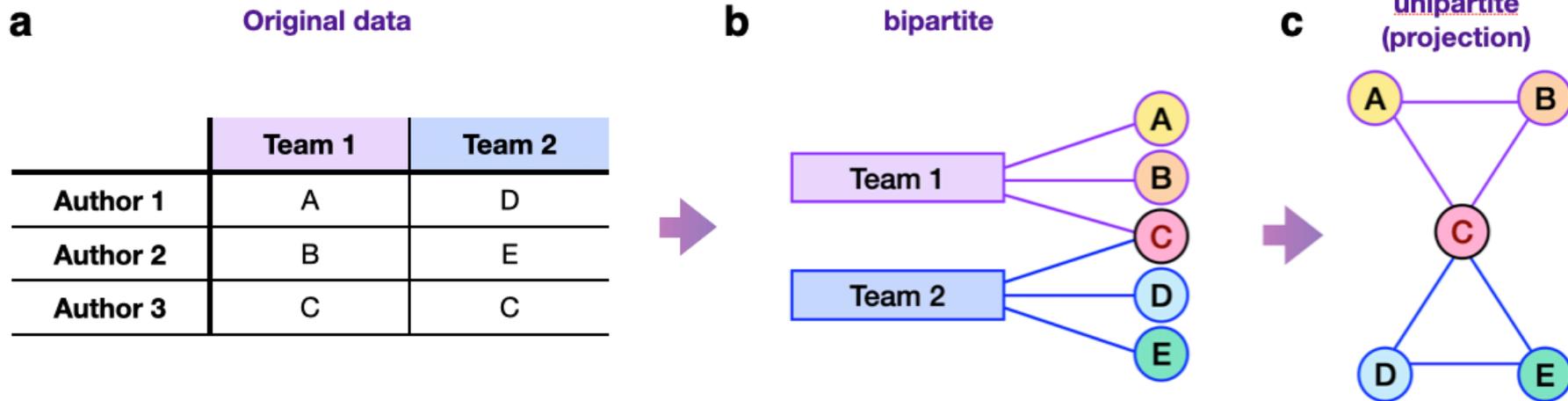


# Ejemplo

Each node represent an *author*  $i = \{1,2,3,\dots,n\}$  where  $n$  is the number of *authors* in the network  $G$ . The network of authors is represented as  $G = (I, V)$ , where  $I$  is the set of authors (nodes,  $k$ ) and  $V$  is the set of connections between authors (edges,  $v$ ).

There is an edge  $v$  between two authors  $i_1, i_2 \in I$  so that  $v(i_1, i_2) \in V$  if they collaborated in the same scientific publication.

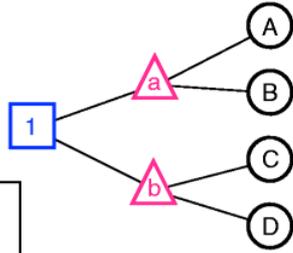
Authors' productivity is measured using their *degree centrality*.



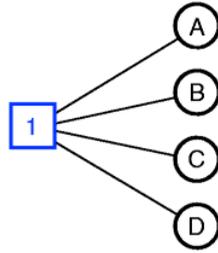
In this example, connections are observable

# Proyecciones de diferentes niveles

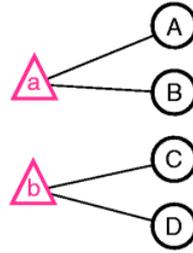
Tripartite network



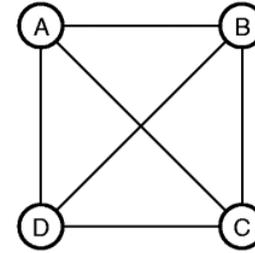
Bipartite network (patient-physician)



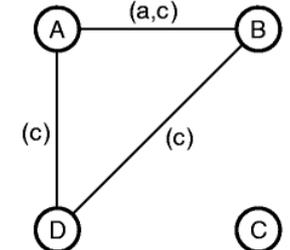
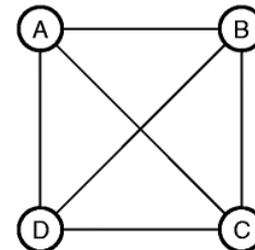
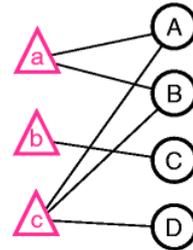
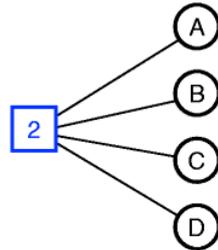
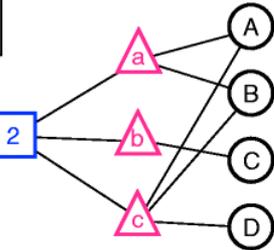
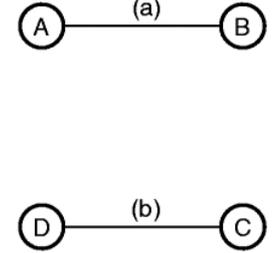
Bipartite network (episode-physician)



Unipartite network (patient projection)



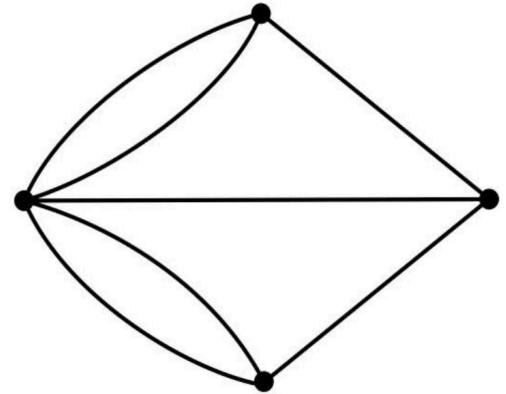
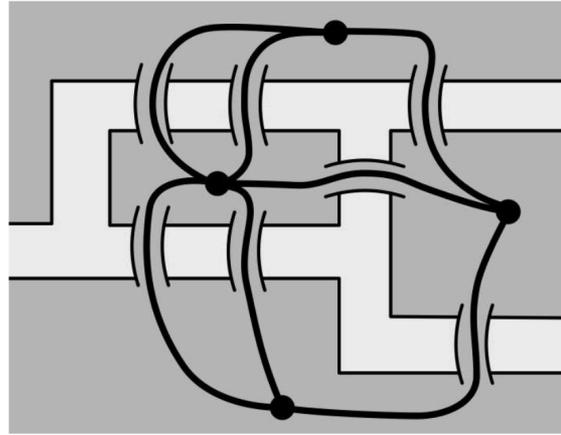
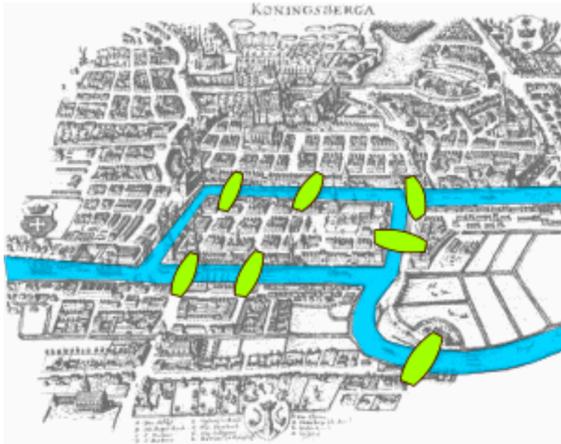
Unipartite network (episode projection)



1 Patient  
a Episode  
A Physician

# Visualización de Estructuras de Red

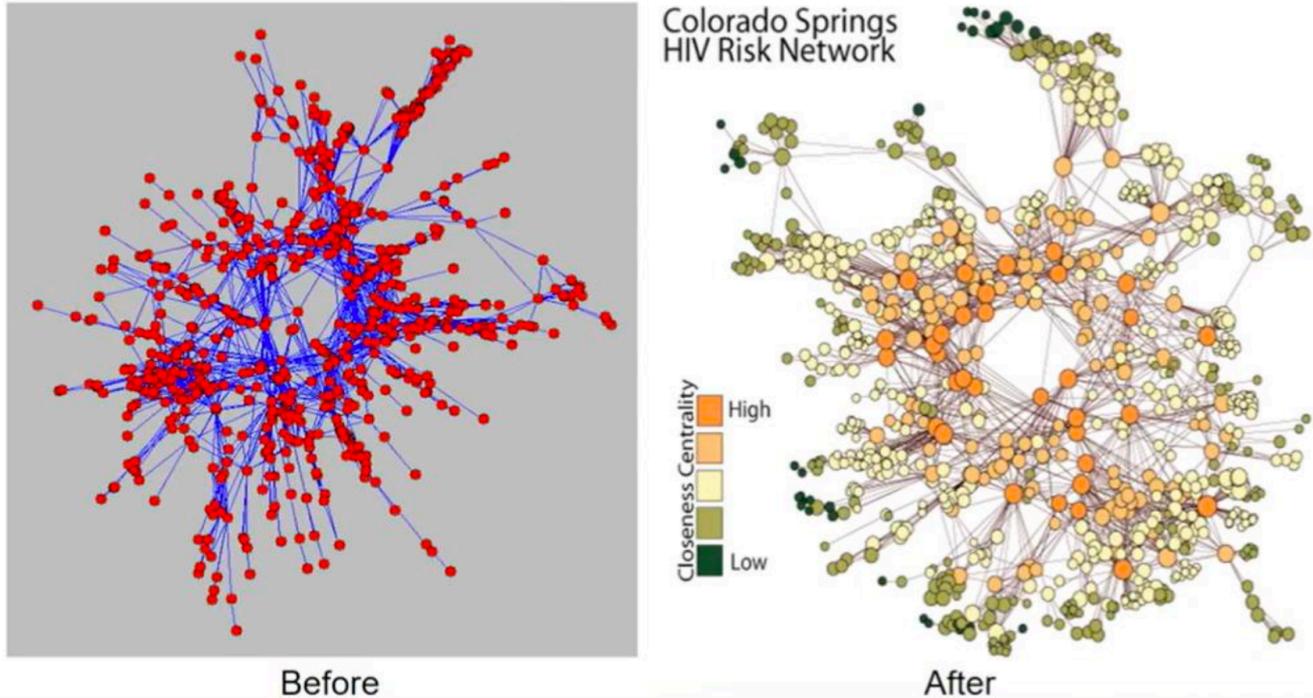
# Convertir una situación del mundo real, a un modelo abstracto con nodos y enlaces



## Elementos de visualización de la red:

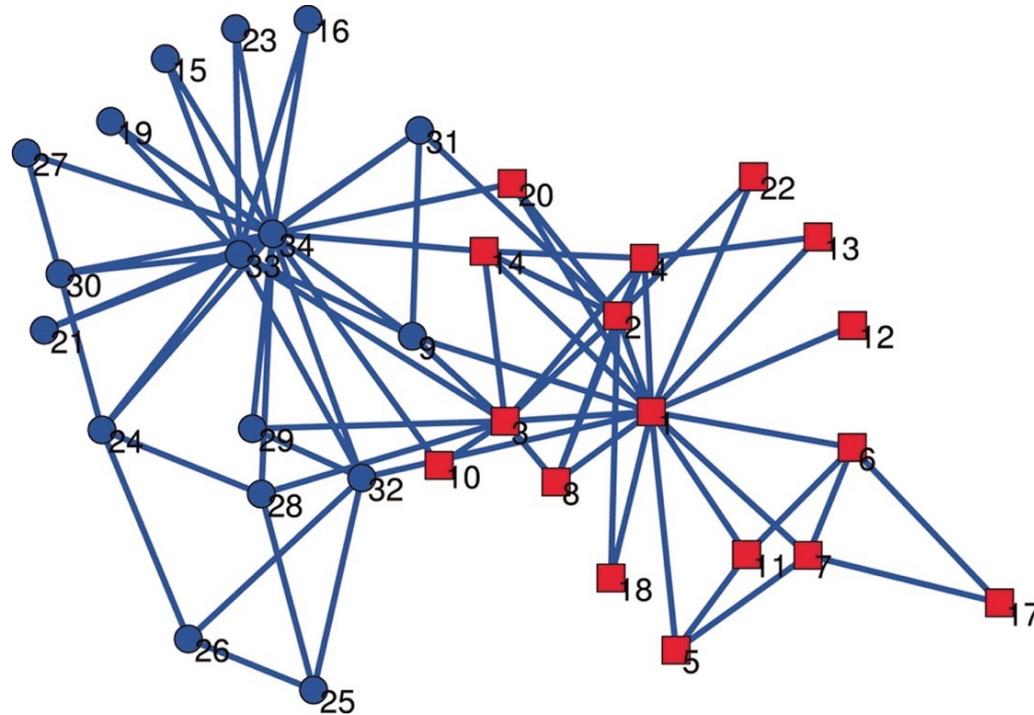
- **Nodos:** color, tamaño, forma, leyenda
- **Enlaces:** color, grosor, curvatura
- **Otros:** comunidades, distribución de la estructura (layout), leyendas, filtros

# Colores: HIV risk



Source: Moody J. slides from Social Networks and Health Workshop 2016

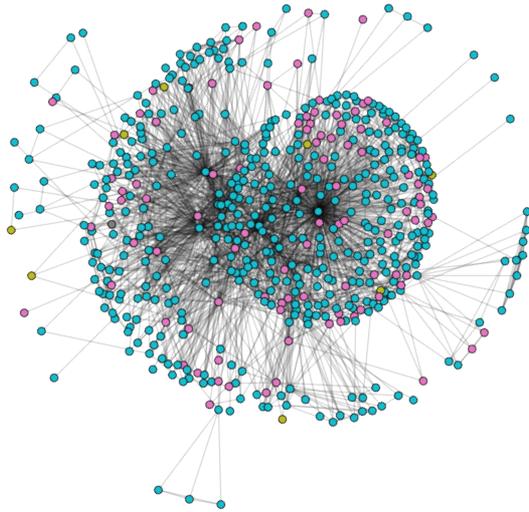
# Comunidades: Zachary Karate Club



Zachary, W. W. (1977). *Journal of anthropological research*, 452-473.

# Layout: Colaboraciones por género

Spring

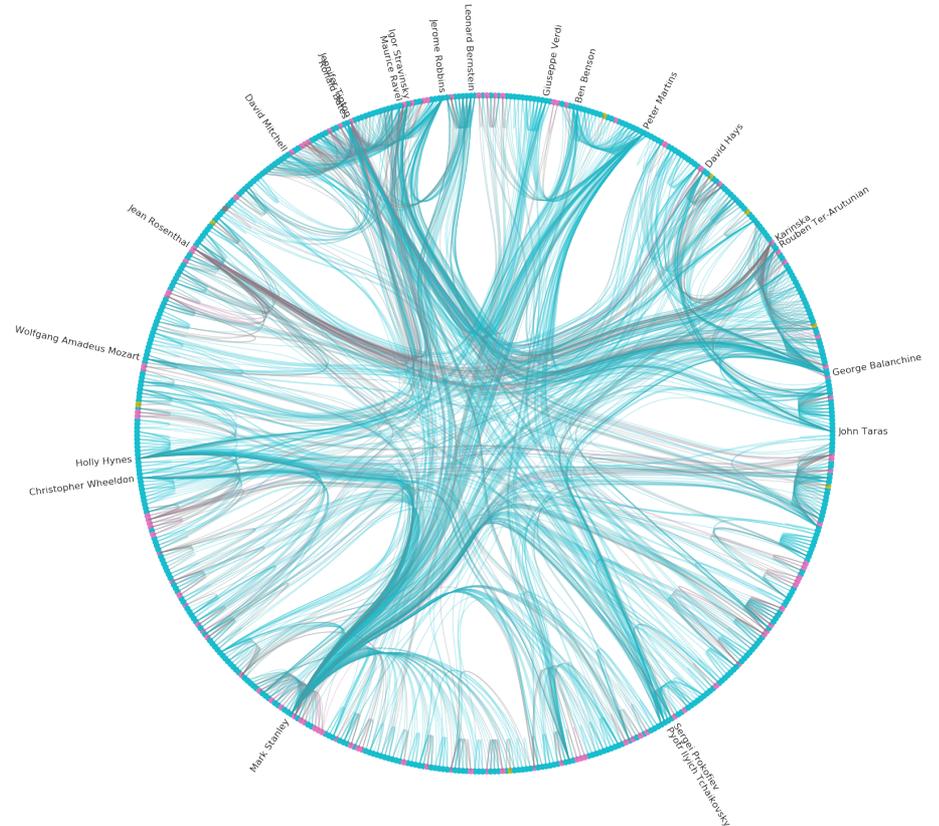


gender

- female
- male
- unknown
- various



Hierarchical Edge Bundling



Source: Yessica Herrera

- **Comunidades, importancia de nodos y tipos de enlaces para monitorear patrones de movilidad urbana**

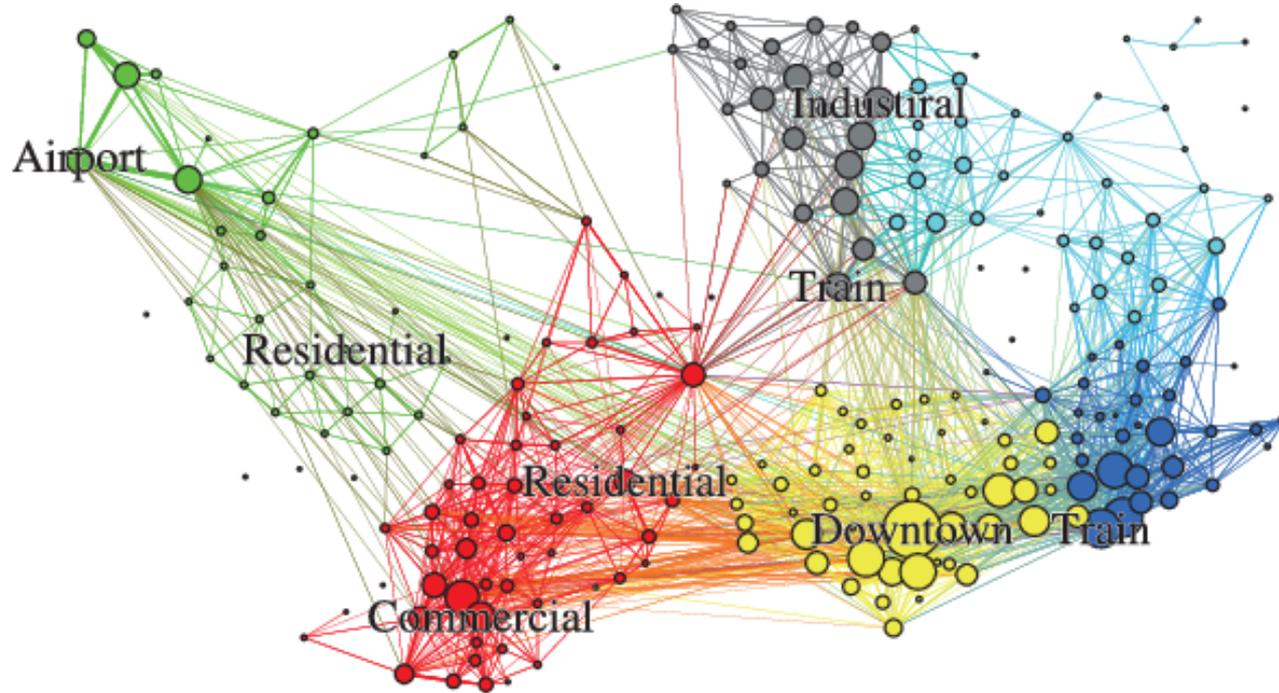


Fig. 2. Human mobility from cellphone data.

# Resumen y Conceptos Importantes

# TRES PRINCIPALES CANTIDADES EN NETWORK SCIENCE

**Distribución de grado:**

**$P(k)$**

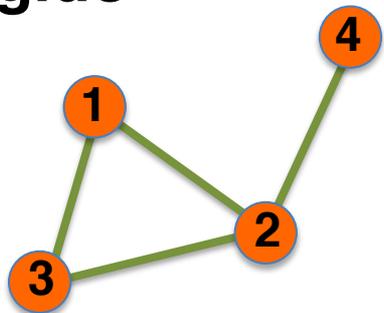
**Longitud de camino:**

**$\langle d \rangle$**

**Coeficiente de clustering:**

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

## No-dirigido



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

$$A_{ii} = 0$$

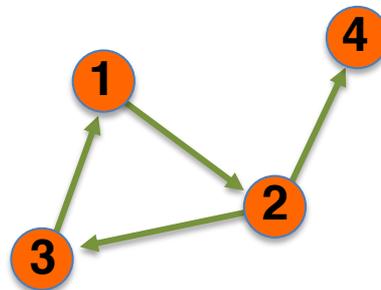
$$A_{ij} = A_{ji}$$

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

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

*Red de actores, interacciones proteina-proteina*

## Dirigido



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

$$A_{ii} = 0$$

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

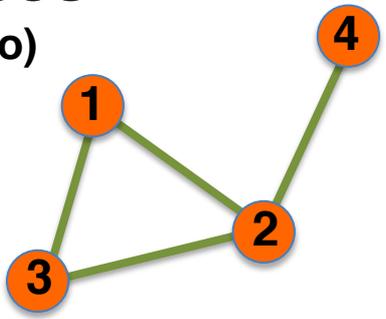
$$L = \sum_{i,j=1}^N A_{ij}$$

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

*WWW, red de citasiones*

In=columnas  
Out=filas

**Sin pesos**  
(no-dirigido)



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

$$A_{ii} = 0$$

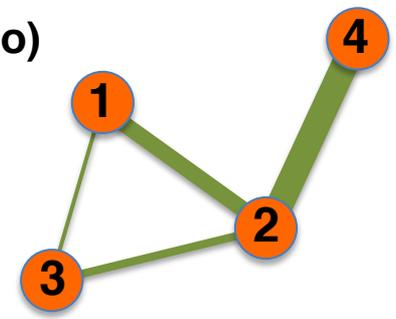
$$A_{ij} = A_{ji}$$

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

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

*Interacciones proteina-proteina, www*

**Con pesos**  
(no-dirigido)



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

$$A_{ii} = 0$$

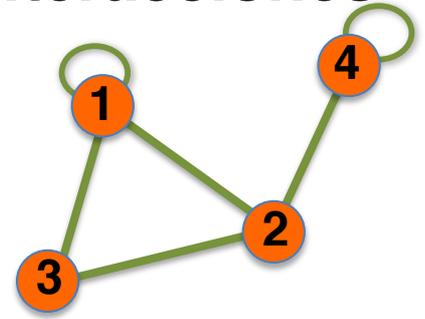
$$A_{ij} = A_{ji}$$

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

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

*Grafo de llamadas, red metabólica*

# Auto-interacciones



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

$$A_{ii} \neq 0$$

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

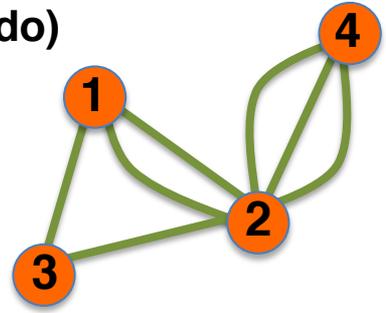
$$L = \frac{1}{2} \sum_{i,j=1, i \neq j}^N A_{ij} + \sum_{i=1}^N A_{ii}$$



Red de interacción de proteínas, [www](http://www)

# Multigrafo

(no-dirigido)



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

$$A_{ii} = 0$$

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

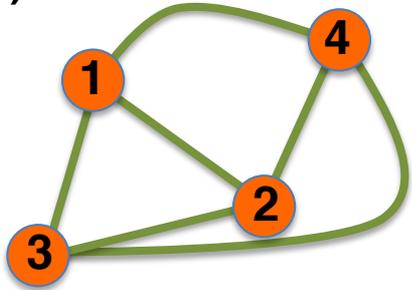
$$L = \frac{1}{2} \sum_{i,j=1}^N \text{nonzero}(A_{ij})$$

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

Redes sociales, Red de colaboraciones

# Grafo completo

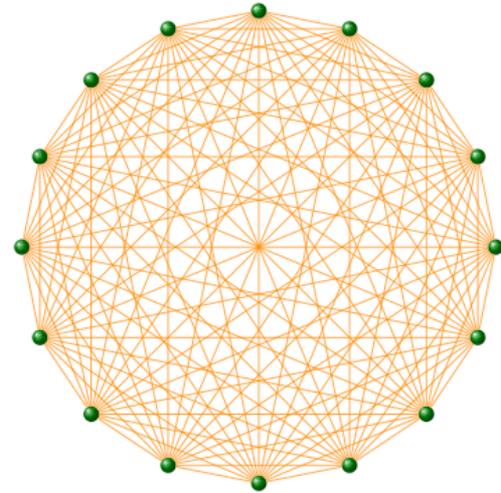
(no-dirigido)



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

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

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



*Red de actores, Interacciones proteina-proteina*

# GRAPHOLOGY: Las redes reales pueden tener multiples características

WWW > Multigrafo dirigido con auto-interacciones

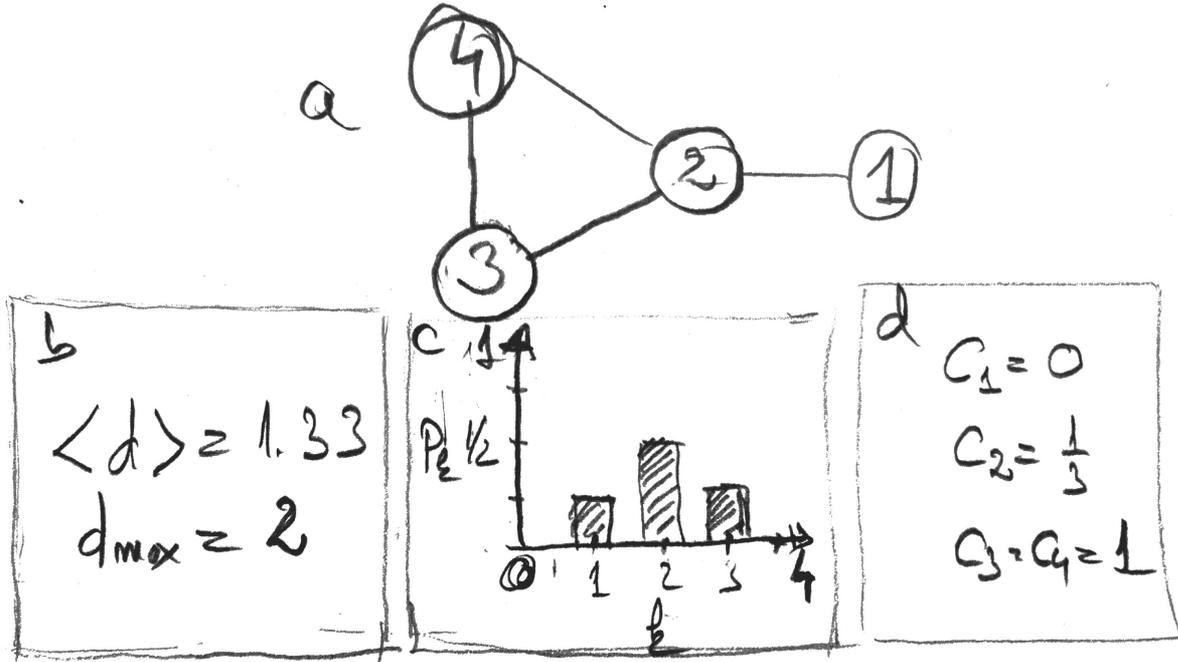
Protein Interactions > no-dirigido, sin pesos, con auto-interacciones

Collaboration network > no-dirigido, multigrafo or con pesos

Mobile phone calls > dirigido, con pesos.

Facebook Friendship links > no-dirigido, sin pesos.

# TRES CANTIDADES CENTRALES EN NETWORK SCIENCE



**A. Distribución de grado:**

$P_k$

**B. Longitud de camino:**

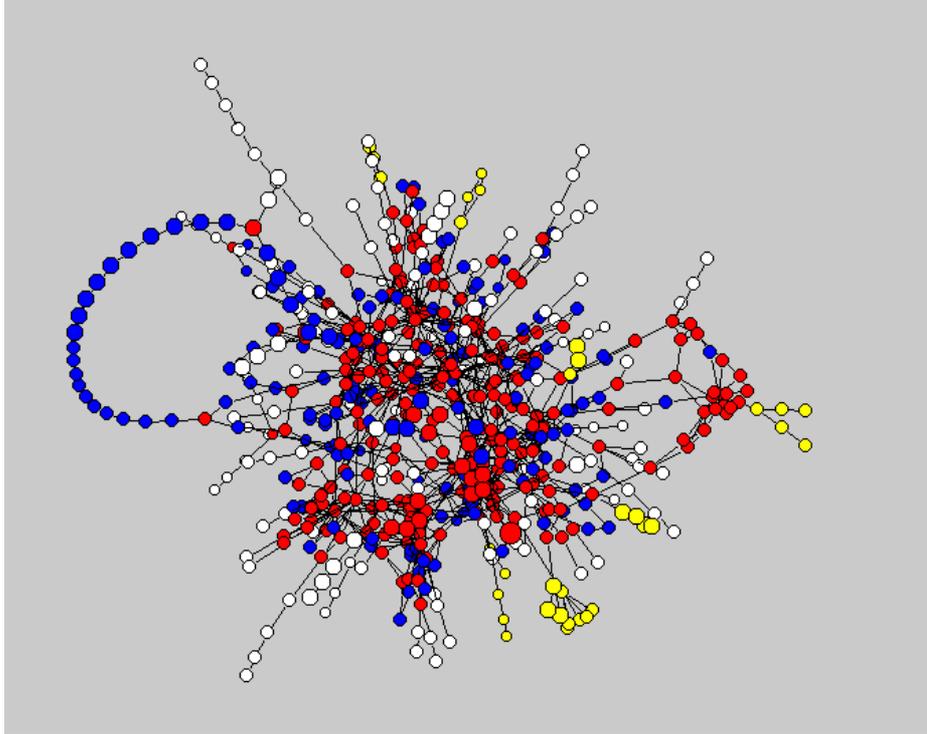
$\langle d \rangle$

**C. Coeficiente de clustering:**

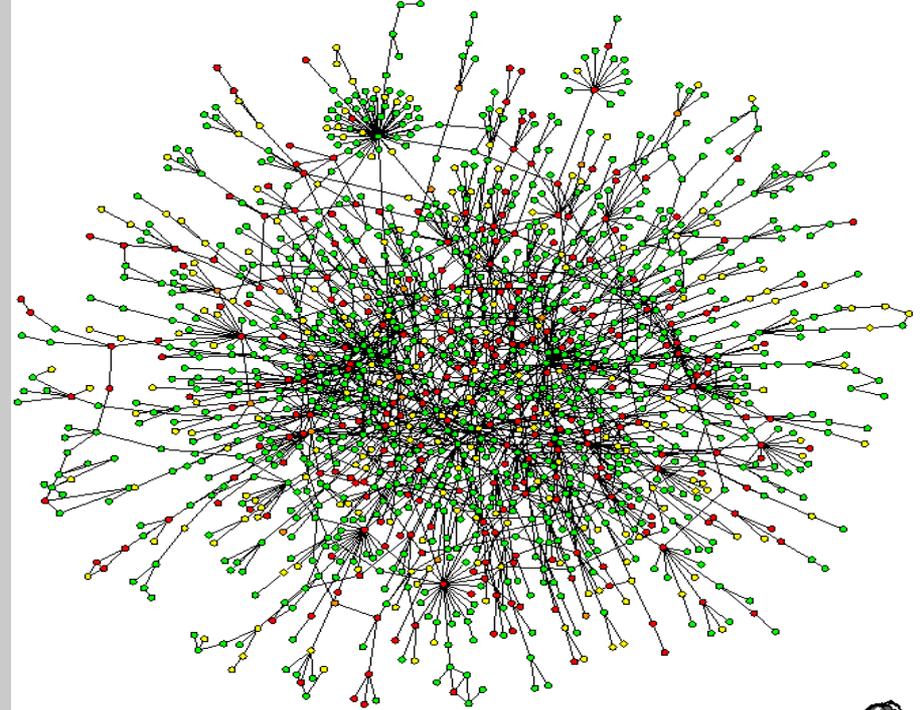
$$C_i = \frac{2e_i}{k_i(k_i - 1)}$$



# Red metabólica



# Interacciones entre Proteínas

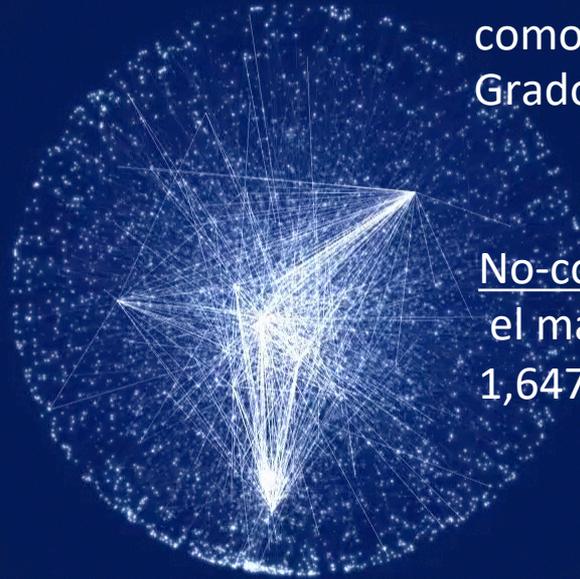


# A CASE STUDY: PROTEIN-PROTEIN INTERACTION NETWORK

Red no-dirigida

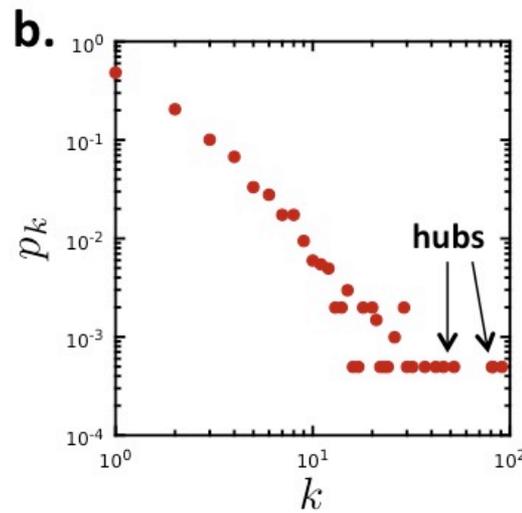
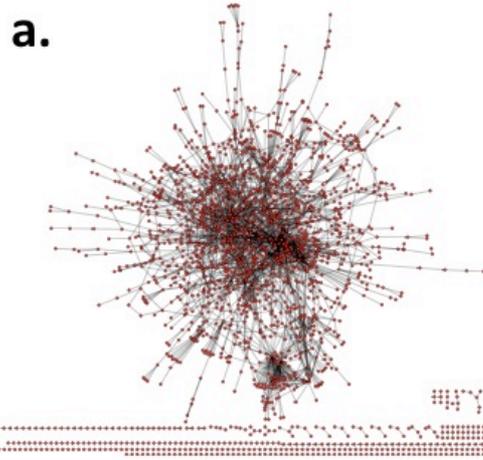
N=2,018 proteínas como nodos  
L=2,930 interacciones vinculantes  
como links.

Grado promedio  $\langle k \rangle = 2.90$ .



No-conectado: 185 componentes  
el más grande (componente gigante)  
1,647/2018 nodos

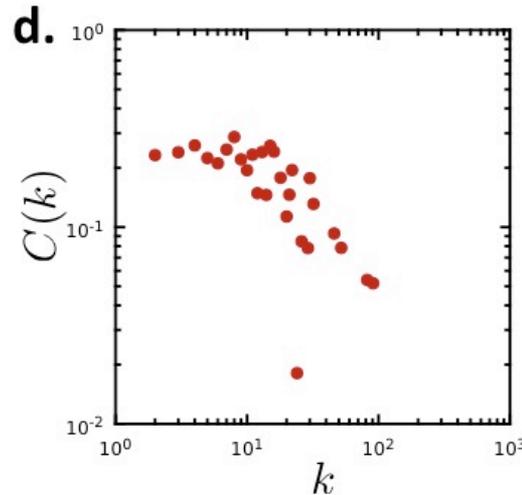
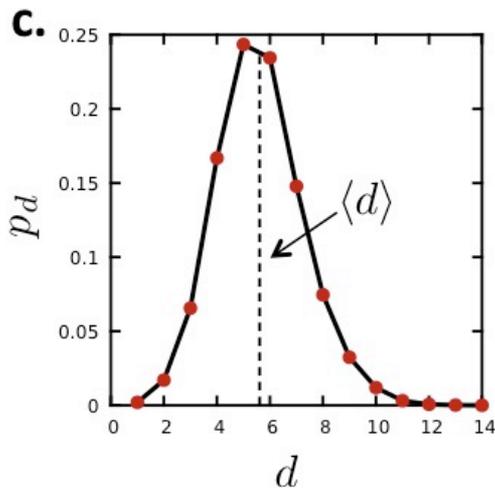
# UN CASO DE ESTUDIO: INTERACCION PROTEINA-PROTEINA



Red no-dirigida

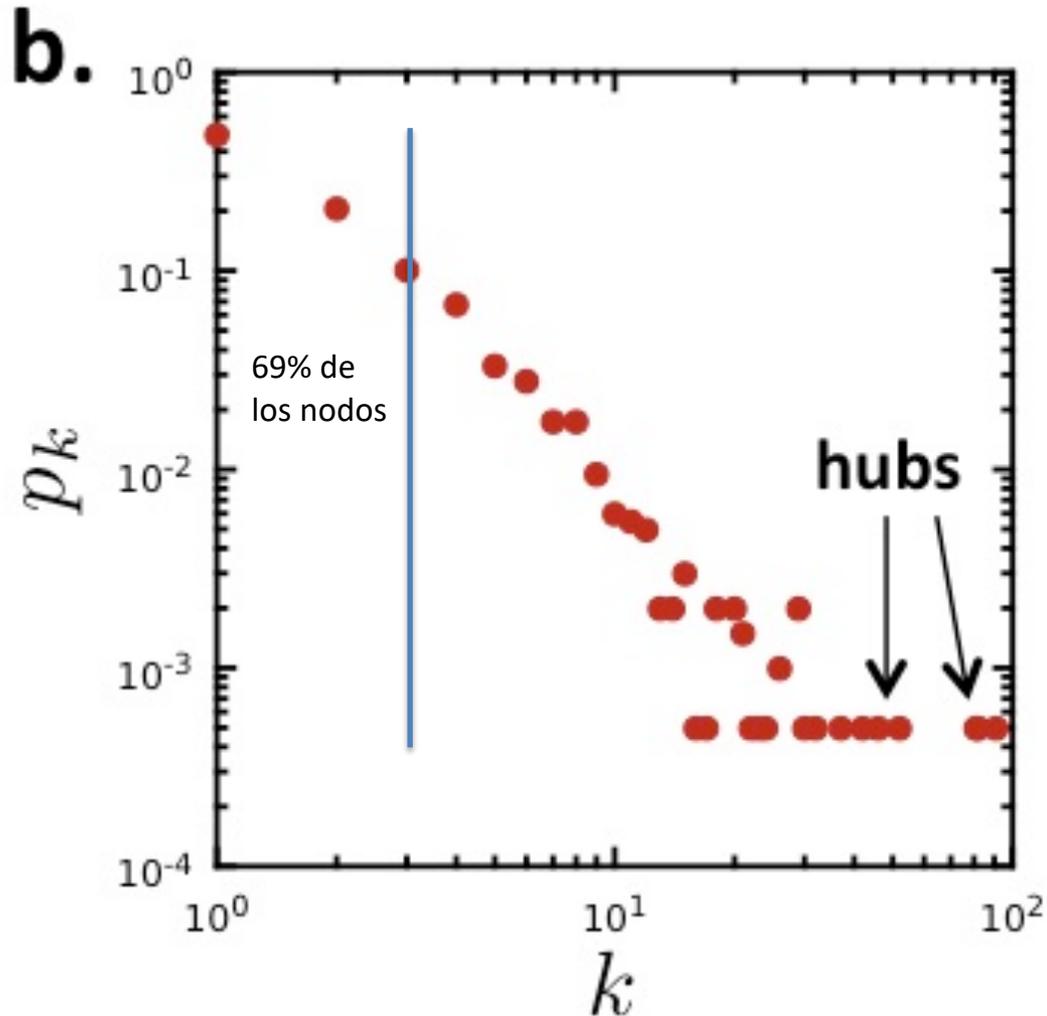
$N=2,018$  proteínas como nodos  
 $L=2,930$  interacciones vinculantes  
como links.

Grado promedio  $\langle k \rangle = 2.90$ .



No conectado: 185 componentes  
las más grande (componente gigante)  
1,647 nodos

# UN CASO DE ESTUDIO: INTERACCION PROTEINA-PROTEINA



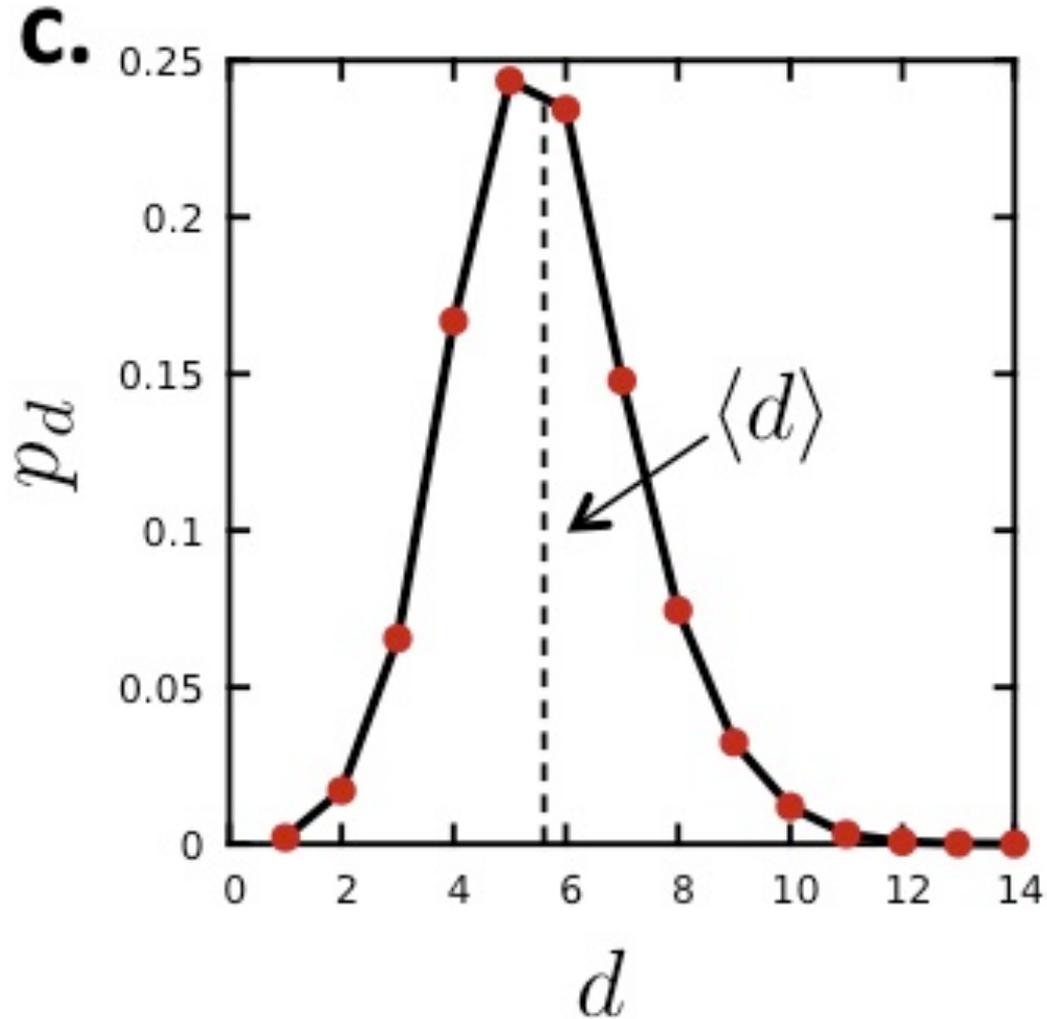
$p_k$  es la probabilidad de que un nodo tenga grado  $k$ .

$N_k = \#$  nodos con grado  $k$

$$p_k = N_k / N$$

Propiedad libre de escala (scale free)

# UN CASO DE ESTUDIO: INTERACCION PROTEINA-PROTEINA

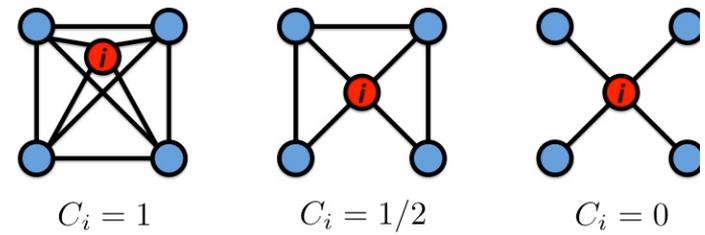
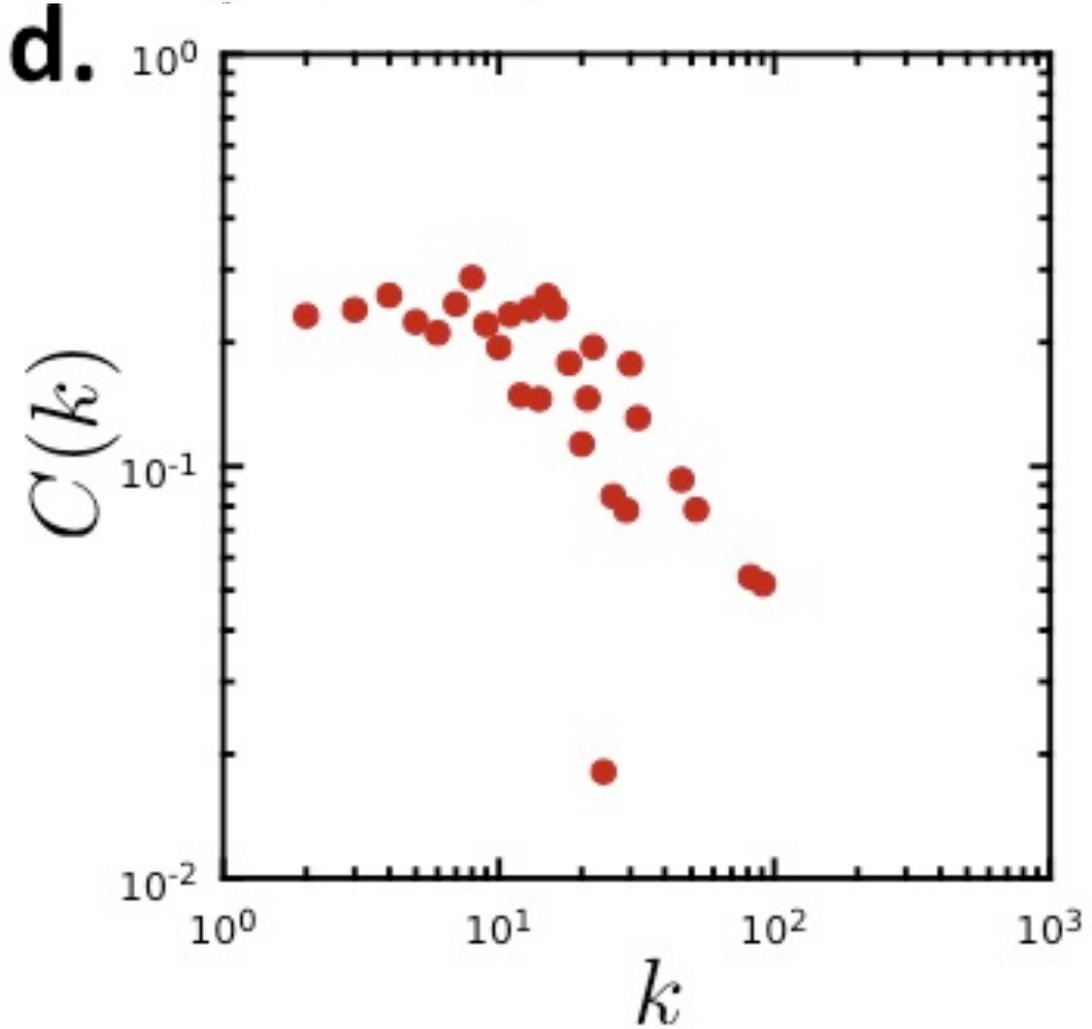


$$d_{\max}=14$$

$$\langle d \rangle = 5.61$$

Propiedad de mundo pequeño

# UN CASO DE ESTUDIO: INTERACCION PROTEINA-PROTEINA



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

$\langle C \rangle = 0.12$

Propiedad de jerarquía

## Preguntas básicas

- Cuáles son las aplicaciones prácticas de redes multi-modos y en qué difieren de las redes unipartita?
- Cuál es la utilidad de usar las proyecciones de una red?
- Cuáles son las diferencias estructurales (visibles) en redes de distinta distribución de grado (aleatorias, mundo pequeño, escala libre)?
- Que información *no* se puede obtener de una visualización de red?

# Lineamientos finales del proyecto

Medida:  $N(t)$ ,  $L(t)$  [ $t$ -tiempo si tiene un sistema dependiente del tiempo];  $P(k)$  (distribución en grados);  $\langle l \rangle$  longitud de camino promedio;  $C$  (coeficiente de agrupamiento),  $C_{\text{rand}}$ ,  $C(k)$ ; Visualización / comunidades;  $P(w)$  si tiene una red ponderada; robustez de la red (si corresponde); propagación (si es apropiado).

No es suficiente medir las cosas, es necesario discutir las ideas que ellas ofrecen (significado):

¿Qué aprendiste de cada cantidad que mediste?

¿Cuáles fueron tus expectativas?

¿Cómo se comparan los resultados con tus expectativas?

La restricción de tiempo serán estrictas. Aproximadamente 7min + 3 min preguntas;

No es necesario escribir un informe, basta con entregar la presentación en formato PDF (se recibirán solo PDF).

Deben enviar un email con nombres / título / nombre curso con **24 horas** antes de la presentación.

La primera diapositiva debe contener nombres y título .

Prueba tus diapositivas con el proyector de antemano (es tu responsabilidad comunicar la información de la mejor manera)

## **Criterio de evaluación:**

Uso de herramientas de red (integridad / uso correcto);

Capacidad para extraer información/perspectivas de sus datos utilizando las herramientas de red;

**(data != información)**

Calidad general del proyecto / presentación.