Similaridad

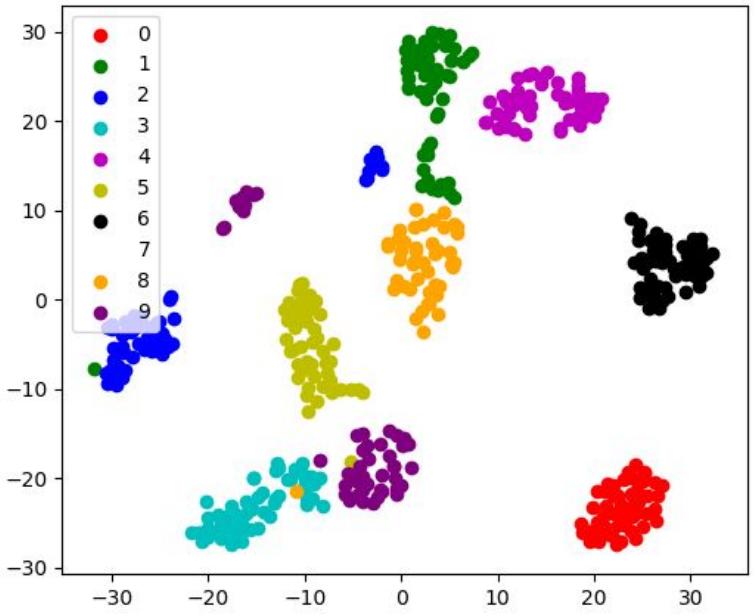
Buscando grupos *naturales* de ~~nodos~~ cosas

v2.2

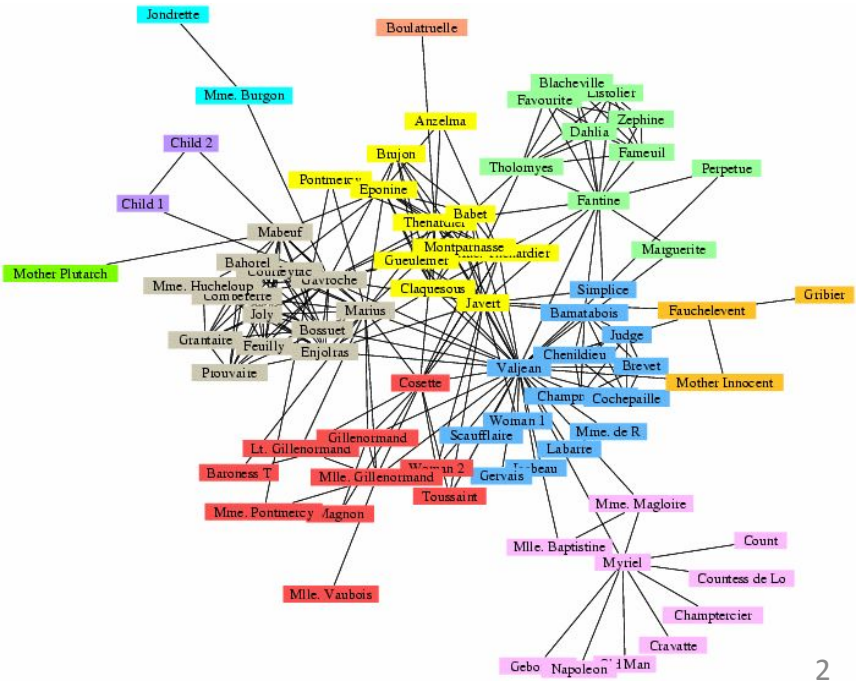
Idea

Para poder **agrupar** entidades necesitamos asumir un criterio, una noción de **similaridad**. Cuando dos cosas son ***parecidas***?

a) Distancias en un espacio métrico



b) Distancias en un espacio topológico

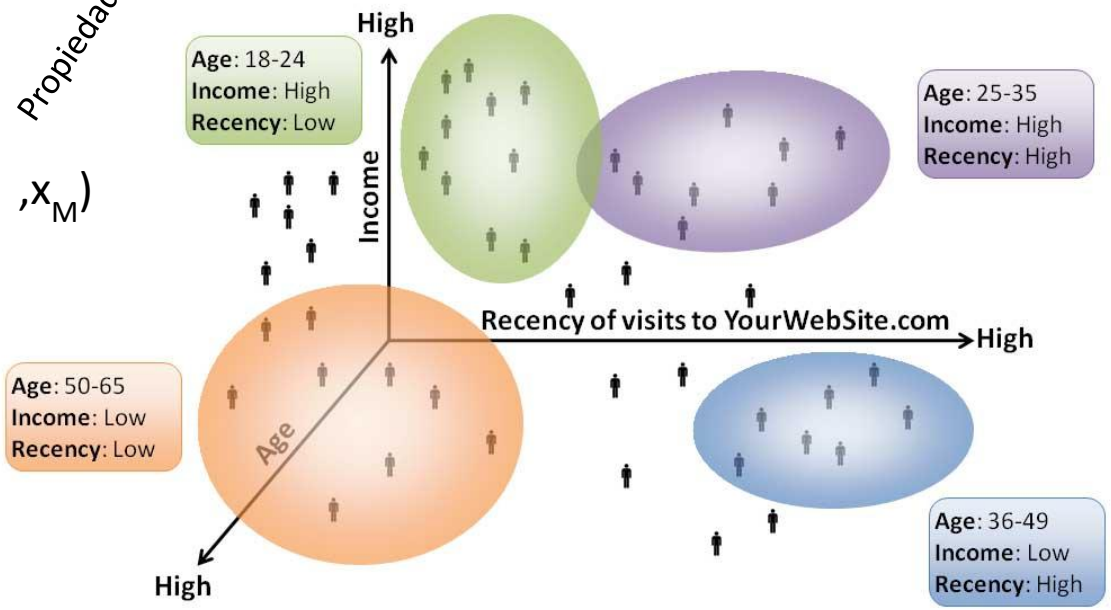


Similaridad en espacio métrico

Similitud a partir de **vectores de características**

Propiedad 1 Propiedad 2 … Propiedad M

objeto → (x_1, x_2, … ,x_M)



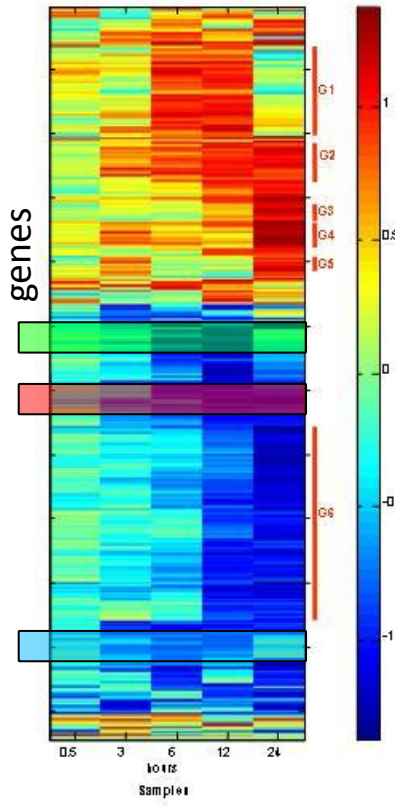
objeto → espacio de características (multidimensional) → Similitud o cercanía (si asumimos espacio métrico)

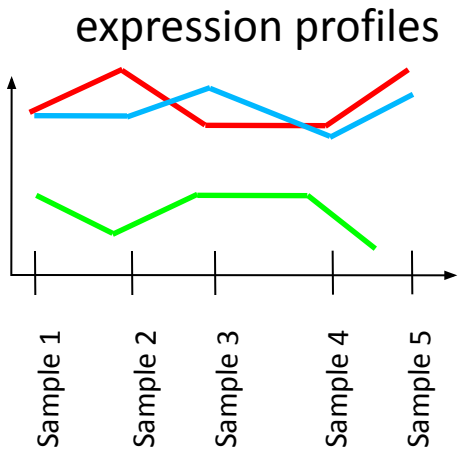
Ejemplo: Patrón de expresión génica

El patrón de expresión del gen X es similar al del gen Y a lo largo de las muestras?

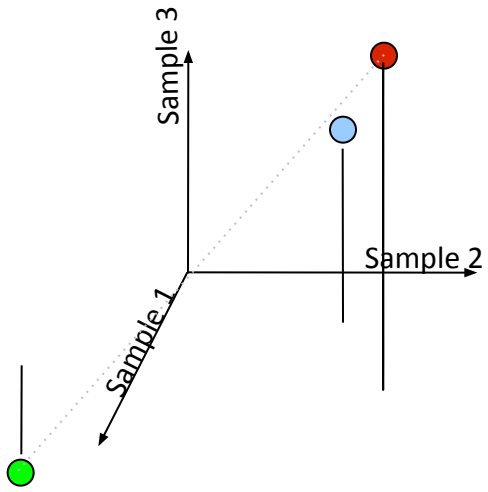
X=(x1,x2,x3,x4,x5), Y=(y1,y2,y3,y4,y5)

muestras(features)

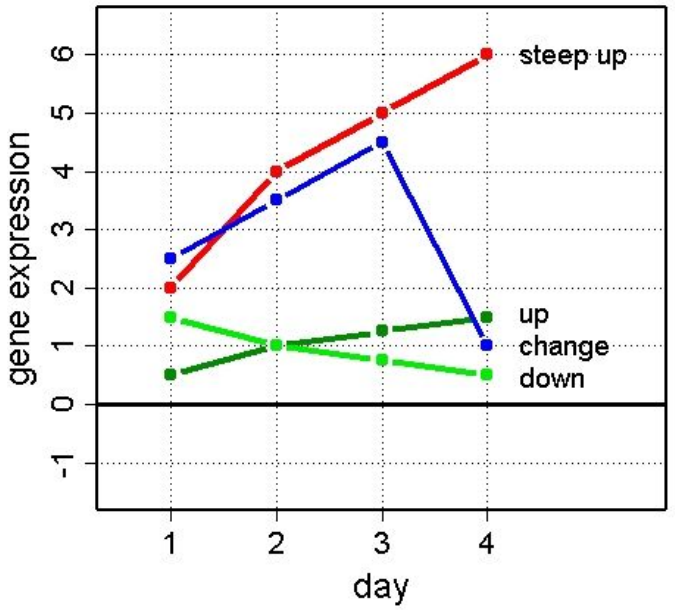




genes as centered points in 5-D (showing only 3D, of course)



Ejemplo:
Patrón de expresión génica



El patrón de expresión del gen X es similar al del gen Y?
X=(x1,x2,x3,x4), Y=(y1,y2,y3,y4)

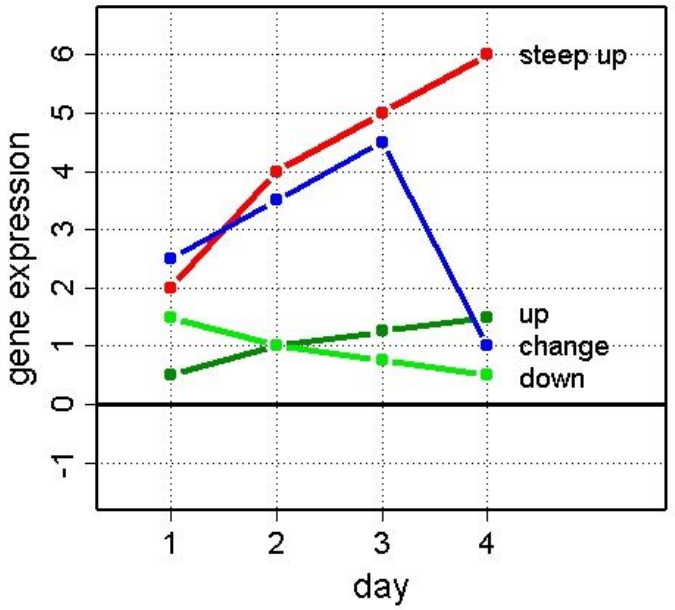
Euclidean Distance

$$d(X,Y)=\sqrt{\sum(x_i-y_i)^2}$$

●	0	6.8	7.6	5.1
●	6.8	0	1.5	4.6
○	7.6	1.5	0	4.6
●	5.1	4.6	4.6	0

0, 1.5, 4.6, 5.1, 6.8,7.6

Ejemplo:
Patrón de expresión génica



El patrón de expresión del gen X es similar al del gen Y?
X=(x1,x2,x3,x4), Y=(y1,y2,y3,y4)

Euclidean Distance

$$d(X,Y)=\sqrt{\sum(x_i-y_i)^2}$$

0	6.8	7.6	5.1
6.8	0	1.5	4.6
7.6	1.5	0	4.6
5.1	4.6	4.6	0

0, 1.5, 4.6, 5.1, 6.8,7.6

Manhattan Distance

$$d(X,Y)=\sum|x_i-y_i|$$

0	12.75	13.25	6.50
12.75	0	2.5	8.25
13.25	2.5	0	7.75
6.50	8.25	7.75	0

0, 2.5, 6.5, 7.75, 8.25, 12.75, 13.25

Correlation Distance

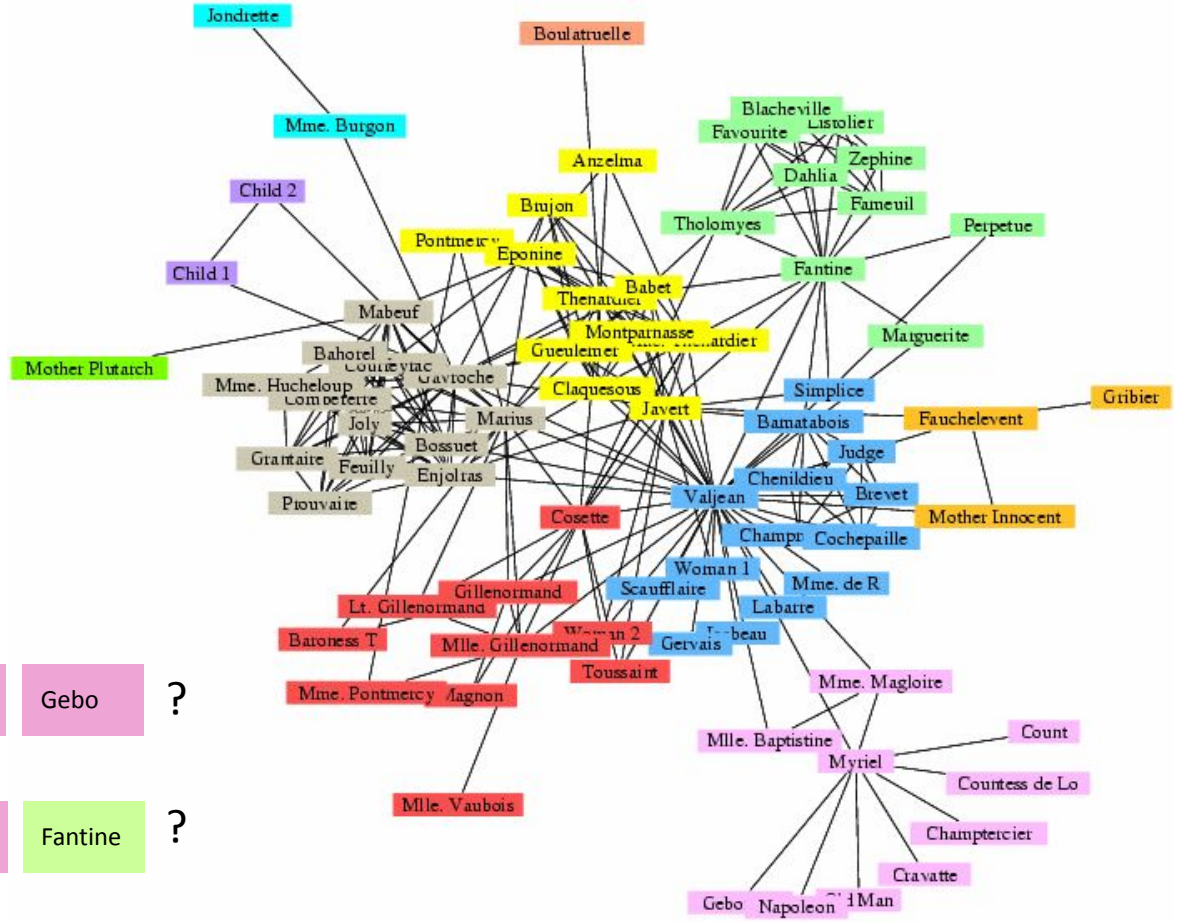
$$d(X,Y)=1-\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2\sum(y_i-\bar{y})^2}$$

0	0	2	1.18
0	0	2	1.18
2	2	0	0.82
1.18	1.18	0.82	0

0, 0.82, 1.18, 2

Similaridad **en redes**

- Cuándo dos vértices de una red son ***similares***?



son similares Myriel Gebo ?

son similares Myriel Fantine ?

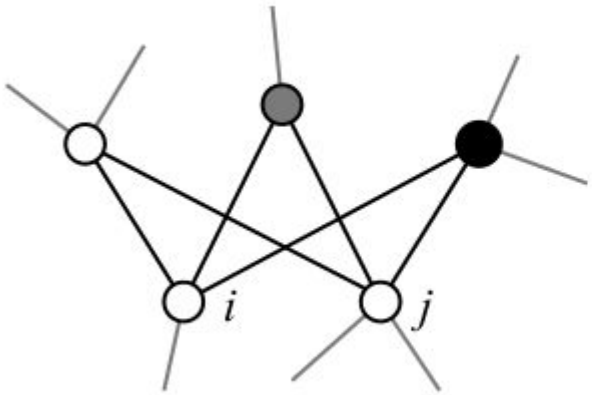
Newman Chap 7.12

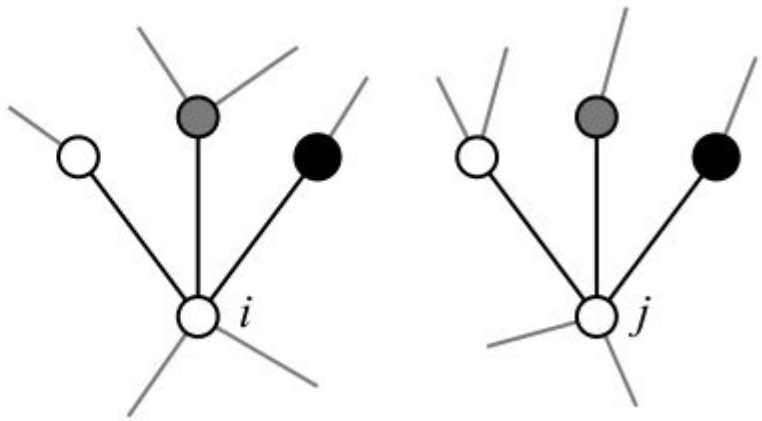
Similaridad **en redes**

- Cuándo dos vértices de una red son ***parecidos***?
  - Similitud a partir de **propiedades topológicas**

Equivalencia estructural
dos nodos de una red son **estructuralmente equivalentes** si **comparten** muchos **vecinos**

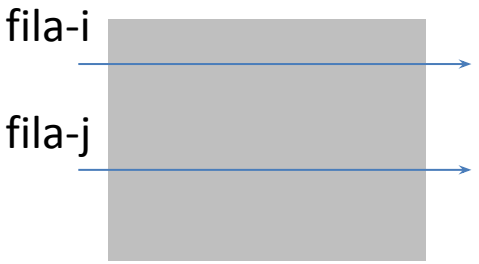
Equivalencia regular
dos nodos de una red son **regularmente equivalentes** si tienen **vecinos** que son **similares**





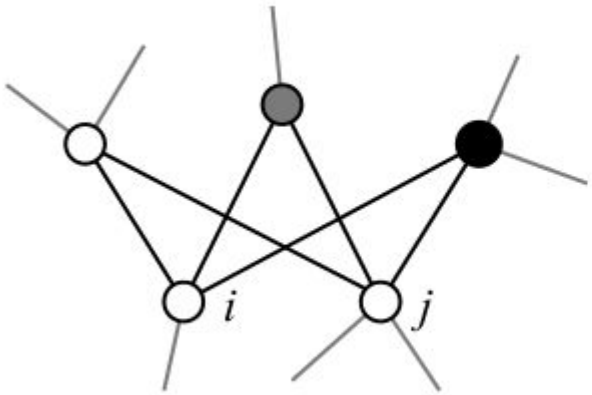
Newman Chap 7.12

Similaridad **en redes**



- Cuándo dos vértices de una red son ***parecidos***?
  - Similitud a partir de **propiedades topológicas**

<u>Equivalencia estructural</u>
dos nodos de una red son
estructuralmente equivalentes
si **comparten** muchos **vecinos**



- Similitud entre nodo-i y nodo-j a partir del número de vecinos compartidos

$$n_{ij} = \sum_k A_{ik} A_{jk} = \sum_k A_{ik} A^T_{kj} = AA^T$$

Producto interno entre los vectores fila-i y fila-j de la matriz de adyacencia

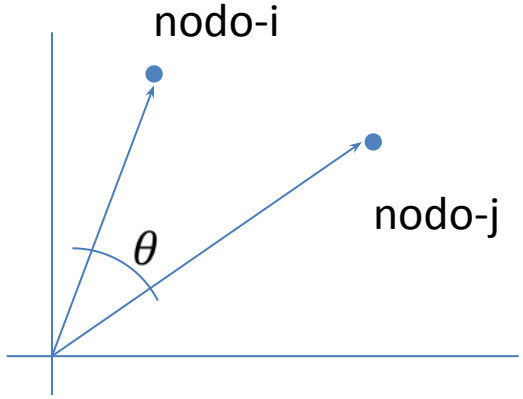
Equivalencia Estructural

Similaridad tipo **Coseno**

- Similitud entre nodo-i y nodo-j a partir del número de vecinos compartidos, **normalizado**:

$$\sigma_{ij} = \frac{\sum_k A_{ik} A_{jk}}{\sqrt{\sum_k A_{ik}{}^2}\sqrt{\sum_k A_{jk}{}^2}} = \cos\theta_{ij}$$

modulo del vector fila-j



$$\sigma_{ij} = \frac{\sum_k A_{ik} A_{jk}}{\sqrt{k_i k_j}} = \frac{n_{ij}}{\sqrt{k_i k_j}}$$

redes no-pesadas

nro vecinos comunes normalizado por la media geometrica de los grados

Equivalencia Estructural

Similaridad tipo **Correlación de Pearson**

- Similaridad como nro de vecinos mútuos respecto a lo que <u>cabría esperar en una **red al azar**</u>:
  - Probabilidad de un dado nodo de conectarse con el nodo-i y además con el nodo-j de manera aleatoria:

$$\frac{k_i}{n}\frac{k_j}{n}$$

  - Ergo, el nro esperado de nodos comunes resulta:

$$\frac{k_i k_j}{n}$$

  - Usamos como medida de similaridad la desviación del nro de vecinos comunes respecto del esperado en el caso aleatorio

$$\sum_k A_{ik}A_{jk} - \frac{k_i k_j}{n}$$

Equivalencia Estructural

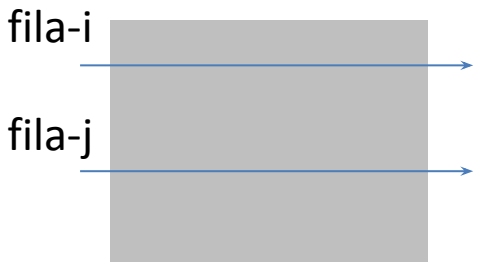
Similaridad tipo **Correlación de Pearson**

$$\sum_k A_{ik}A_{jk} - \frac{k_i k_j}{n} = \sum_k A_{ik}A_{jk} - \frac{1}{n}\sum_k A_{ik}\sum_l A_{jl}$$

$$= \sum_k A_{ik}A_{jk} - n\langle A_i\rangle\langle A_j\rangle$$

$$= \sum_k \left[A_{ik}A_{jk} - \langle A_i\rangle\langle A_j\rangle\right]$$

$$= \sum_k (A_{ik} - \langle A_i\rangle)(A_{jk} - \langle A_j\rangle)$$



Desviación del nro de vecinos respecto del esperado en el caso aleatorio

$n\ cov(A_i, A_j)$

El maximo valor de $cov(A_i,A_j)$ se obtiene cuando ambos vectores son idénticos. Eso ocurre cuando:

$cov(A_i,A_i) = var(A_i) = \sigma_i^2$ o $cov(A_j,A_j) = var(A_j) = \sigma_j^2$

Equivalencia Estructural

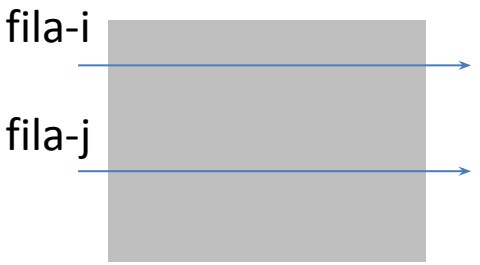
Similaridad tipo **Correlación de Pearson**

$$\sum_k A_{ik}A_{jk} - \frac{k_i k_j}{n} = \sum_k A_{ik}A_{jk} - \frac{1}{n}\sum_k A_{ik}\sum_l A_{jl}$$

$$= \sum_k A_{ik}A_{jk} - n\langle A_i\rangle\langle A_j\rangle$$

$$= \sum_k \left[A_{ik}A_{jk} - \langle A_i\rangle\langle A_j\rangle\right]$$

$$= \sum_k (A_{ik} - \langle A_i\rangle)\left(A_{jk} - \langle A_j\rangle\right)$$

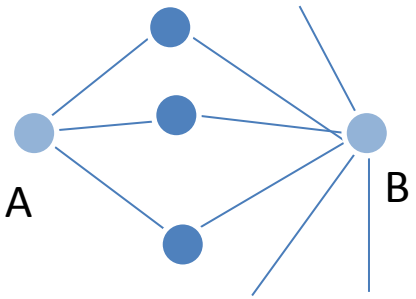


$$r_{ij} = \frac{cov(A_i, A_j)}{\sqrt{\sigma_i^2\sigma_i^2}} = \frac{\sum_k (A_{ik} - \langle A_i\rangle)\left(A_{jk} - \langle A_j\rangle\right)}{\sqrt{\sum_k (A_{ik} - \langle A_i\rangle)^2}\sqrt{\sum_k \left(A_{jk} - \langle A_j\rangle\right)^2}}$$

$-1 \le r_{ij} \le 1$ y resulta nulo si los nodos-i y j tienen el nro de vecinos esperados por azar

Equivalencia Estructural
Superposición topológica *(topological overlap)*

$$TO_{ij} = \frac{n_{ij} + \Theta(A_{ij})}{min(k_i, k_j) + 1 - \Theta(A_{ij})}$$



$$\Theta(x) = \begin{matrix} 0\ si\ x \leq 0 \\ 1\ si\ x > 0 \end{matrix}$$

$$TO_{AB} = \frac{3}{3+1} = 0.75$$

Equivalencia Estructural

Otras medidas de similaridad

- Similaridad: podríamos utilizar el número de vecinos comunes **normalizado** por el número esperado

$$\frac{n_{ij}}{\frac{k_i k_j}{n}} = n\frac{\sum_k A_{ik}A_{jk}}{\sum_k A_{ik}\sum_k A_{jk}}$$

- 1, si el n_{ij} es el numero esperado en una red aleatoria
- >1 si hay más vecinos comunes
- <1 si hay menos
- 0 si no tienen vecinos en comun

- **Distancia** Euclídea (distancia de Hamming en realidad)

$$d_{ij} = \sum_k \left(A_{ik} - A_{jk}\right)^2$$

Nro de vecinos diferentes entre los dos nodos

$$\widetilde{d_{ij}} = \frac{d_{ij}}{k_i + k_j} = \frac{1}{k_i + k_j}\sum_k \left(A_{ik} + A_{jk} - 2A_{ik}A_{jk}\right) = 1 - 2\frac{n_{ij}}{k_i + k_j}$$

Equivalencia Estructural

Otras medidas de similaridad

- Similaridad: podríamos utilizar el número de vecinos comunes **normalizado** por el número esperado

$$\frac{n_{ij}}{\frac{k_i k_j}{n}} = n\frac{\sum_k A_{ik}A_{jk}}{\sum_k A_{ik}\sum_k A_{jk}}$$

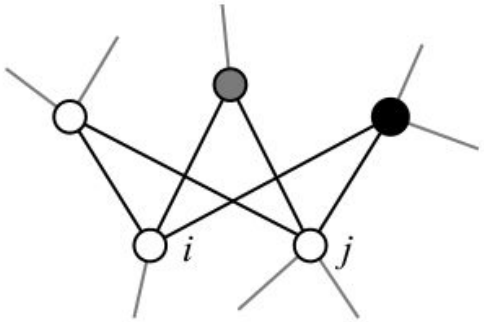
- 1, si el n_{ij} es el numero esperado en una red aleatoria
- >1 si hay más vecinos comunes
- <1 si hay menos
- 0 si no tienen vecinos en comun

- **Distancia** Euclídea (distancia de Hamming en realidad)

$$d_{ij} = \sum_k \left(A_{ik} - A_{jk}\right)^2$$ ← Nro de vecinos diferentes entre los dos nodos

$$\widetilde{d_{ij}} = 1 - 2\frac{n_{ij}}{k_i + k_j} \longrightarrow s_{ij} = 1 - \widetilde{d_{ij}} = 2\frac{n_{ij}}{k_i + k_j}$$

Equivalencia Estructural



- Similaridad tipo coseno

$$\sigma_{ij} = \frac{n_{ij}}{\sqrt{k_i k_j}} = \cos\theta_{ij}$$

- Similaridad de Pearson

$$n_{ij} - \frac{k_i k_j}{n} \rightarrow \frac{cov(A_i, A_j)}{\sigma_i \sigma_j}$$

- Overlap Topologico

$$TO_{ij} = \frac{n_{ij} + \Theta(A_{ij})}{min(k_i, k_j) + 1 - \Theta(A_{ij})}$$

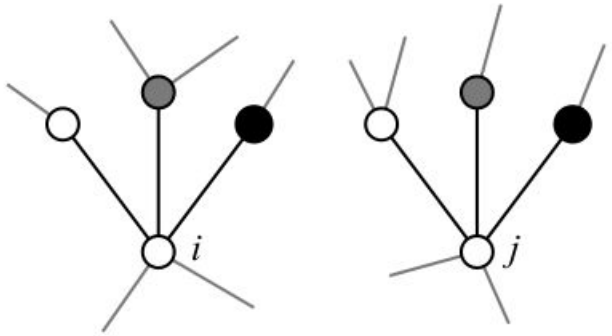
- Similaridad: número de vecinos comunes **normalizado** por el número esperado

$$\frac{n_{ij}}{\frac{k_i k_j}{n}}$$

- Similaridad Euclídea (Hamming)

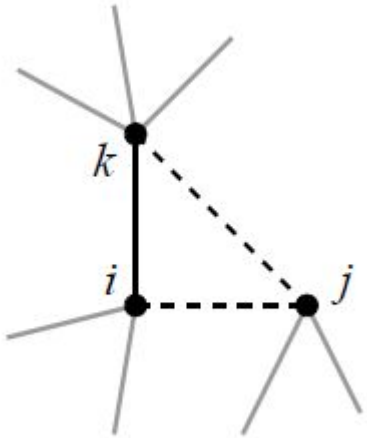
$$s_{ij} = 2\frac{n_{ij}}{k_i + k_j}$$

Equivalencia Regular



Dos nodos son similares si sus vecinos lo son

$$\sigma_{ij} = \alpha \sum_{kl} A_{ik} A_{jl} \sigma_{kl}$$



Versión que usaremos como definición:

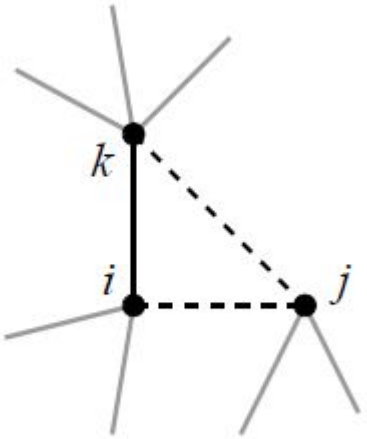
Dos vértices, *i* y *j* son similares si *i* tiene como vecino a *k*, quien a su vez es similar a *j.*

$$\sigma_{ij} = \alpha \sum_{k} A_{ik} \sigma_{kj} + \delta_{ij}$$

σ_{ij} es un campo definido sobre enlaces entre pares que satisface la ec. de consitencia

Newman Chap 7.12.4

Equivalencia Regular



Versión que usaremos como definición:

Dos vértices, *i* y *j* son similares si *i* tiene como vecino a *k*, quien a su vez es similar a *j.*

$$\sigma_{ij} = \alpha \sum_k A_{ik}\sigma_{kj} + \delta_{ij}$$

$$\boldsymbol{\sigma} = \alpha \boldsymbol{A}\boldsymbol{\sigma} + \boldsymbol{I}$$

$$\boldsymbol{\sigma}^{(0)} = 0$$
$$\boldsymbol{\sigma}^{(1)} = \boldsymbol{I}$$
$$\boldsymbol{\sigma}^{(2)} = \alpha\boldsymbol{A} + \boldsymbol{I}$$
$$\boldsymbol{\sigma}^{(3)} = \alpha^2\boldsymbol{A^2} + \alpha\boldsymbol{A} + \boldsymbol{I}$$
$$\dots$$

$$\boldsymbol{\sigma} = \sum_{m=0}^{\infty} (\alpha\boldsymbol{A})^m = (\boldsymbol{I} - \alpha\boldsymbol{A})^{-1}$$

$$\alpha < 1$$
$$\alpha < 1/\kappa_1$$

Autovalor dominante de $\boldsymbol{A}$

Recordemos centralidad de Katz

$$x_i = \alpha \sum_j A_{ij}x_j + 1$$

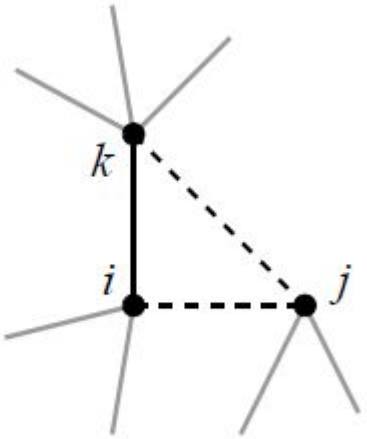
$$\boldsymbol{x} = (\boldsymbol{I} - \alpha\boldsymbol{A})^{-1}\boldsymbol{1}$$

$$x_i = \sum_j \sigma_{ij}$$

Centralidad de Katz de un nodo es su similaridad, en el sentido regular, acumulada

Newman Chap 7.12.4

Equivalencia Regular



Versión que usaremos como definición:

Dos vértices, *i* y *j* son similares si *i* tiene como vecino a *k*, quien a su vez es similar a *j.*

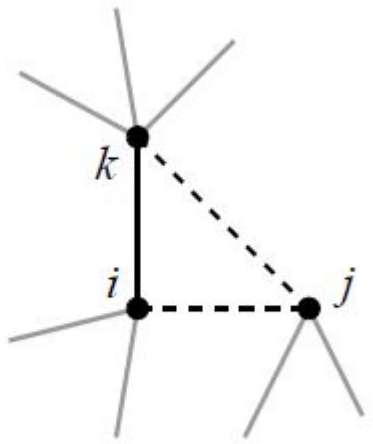
$$\boldsymbol{\sigma} = \sum_{m=0}^{\infty} (\alpha \boldsymbol{A})^m = \boldsymbol{I} + \alpha \boldsymbol{A} + \alpha^2 \boldsymbol{A^2} + \cdots$$

$$\alpha < 1/\kappa_1$$

$$\sigma_{ij} = \delta_{ij} + \alpha A_{ij} + \alpha^2 A^2{}_{ij} + \cdots$$

Newman Chap 7.12.4

Equivalencia Regular



Versión que usaremos como definición:

Dos vértices, *i* y *j* son similares si *i* tiene como vecino a *k*, quien a su vez es similar a *j.*

$$\sigma_{ij} = \alpha \sum_k A_{ik}\sigma_{kj} + \delta_{ij}$$

Nodos con alto grado tienen más vecinos, que pueden ser similares a terceros nodos, por lo que tienen mas chances de aumentar su nivel de similaridad general. Si se desea mitigar este efecto…

Si se desea incluir alguna noción de similaridad **externa: S^{ext}**

Variante:

$$\sigma_{ij} = \frac{\alpha}{k_i} \sum_k A_{ik}\sigma_{kj} + \delta_{ij}$$

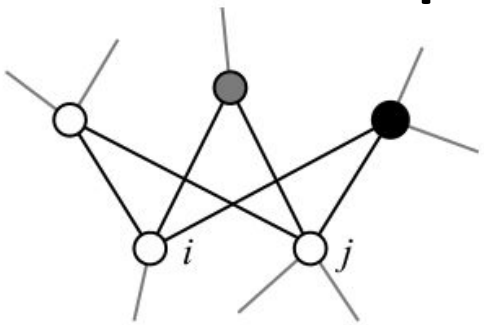
$$\boldsymbol{\sigma} = (\boldsymbol{D} - \alpha \boldsymbol{A})^{-1} \boldsymbol{D}$$

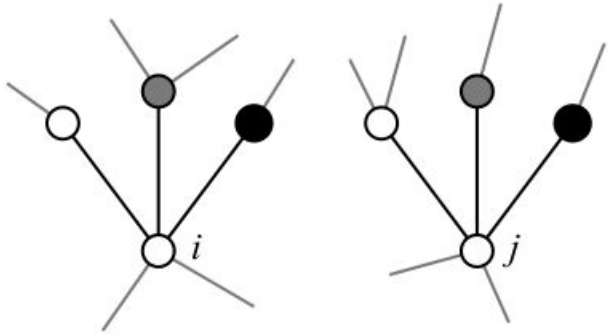
Con ***D*** la matriz diagonal $D_{ii}=k_i$

Variante:

$$\sigma_{ij} = \alpha \sum_k A_{ik}\sigma_{kj} + s^{ext}{}_{ij}$$

Similaridades topológicas





- Similaridad tipo coseno

$$\sigma_{ij} = \frac{n_{ij}}{\sqrt{k_i k_j}} = \cos\theta_{ij}$$

$$\sigma_{ij} = \alpha \sum_k A_{ik}\sigma_{kj} + \delta_{ij}$$

- Similaridad de Pearson

$$n_{ij} - \frac{k_i k_j}{n} \rightarrow \frac{cov(A_i, A_j)}{\sigma_i \sigma_j}$$

$$\sigma_{ij} = \frac{\alpha}{k_i} \sum_k A_{ik}\sigma_{kj} + \delta_{ij}$$

- Overlap topologico: $TO_{ij} = \frac{n_{ij} + \Theta(A_{ij})}{min(k_i, k_j) + 1 - \Theta(A_{ij})}$

$$\sigma_{ij} = \alpha \sum_k A_{ik}\sigma_{kj} + s^{ext}{}_{ij}$$

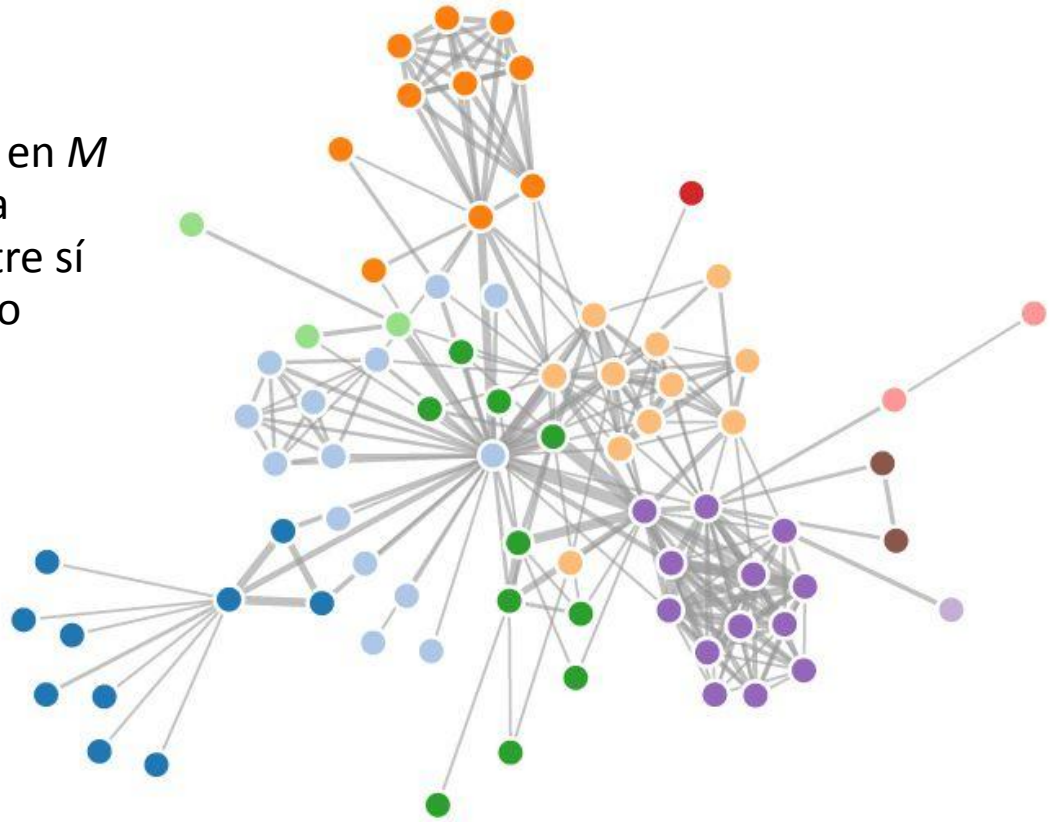
- Similaridad: **normalizado** por el número esperado $\frac{n_{ij}}{\frac{k_i k_j}{n}}$

- Similaridad Euclídea (Hamming) $s_{ij} = 2\frac{n_{ij}}{k_i + k_j}$

Detectando comunidades en redes

Noción de comunidades:

Partición de los *N* vértices de la red en *M* grupos tales que nodos de la misma comunidad sean **más parecidos** entre sí que respecto a nodos fuera de dicho grupo



- Similaridad topológica de vértices
- Heurística

Hacia una descripción mesoscópica

comunidades en redes

impacts

From molecular to modular cell biology

Leland H. Hartwell, John J. Hopfield, Stanislas Leibler and Andrew W. Murray

Although living systems obey the laws of physics and chemistry, the notion of function or purpose differentiates biology from other natural sciences. Organisms exist to reproduce, whereas, outside religious belief, rocks and stars have no purpose. Selection for function has produced the living cell, with a unique set of properties that distinguish it from inanimate systems of interacting molecules. Cells exist far from thermal equilibrium by harvesting energy from their environment. They are composed of thousands of different types of molecule. They contain information for their survival and reproduction, in the form of their DNA. Their interactions with the environment depend in a byzantine fashion on this information and the information and the many components. For example, in the signal transduction system in yeast that converts the detection of a pheromone into the act of mating, there is no single protein responsible for amplifying the input signal provided by the pheromone molecule.

To describe biological functions, we need a vocabulary that contains concepts such as amplification, adaptation, robustness, insulation, error correction and coincidence detection. For example, to decipher how the binding of a few molecules of an attractant to receptors on the surface of a bacterium can make the bacterium move towards the attractant (chemotaxis) will require understanding how cells robustly detect and amplify signals in a noisy environment. Having described such concepts, we need to explain how they arise from interactions among components in the cell.

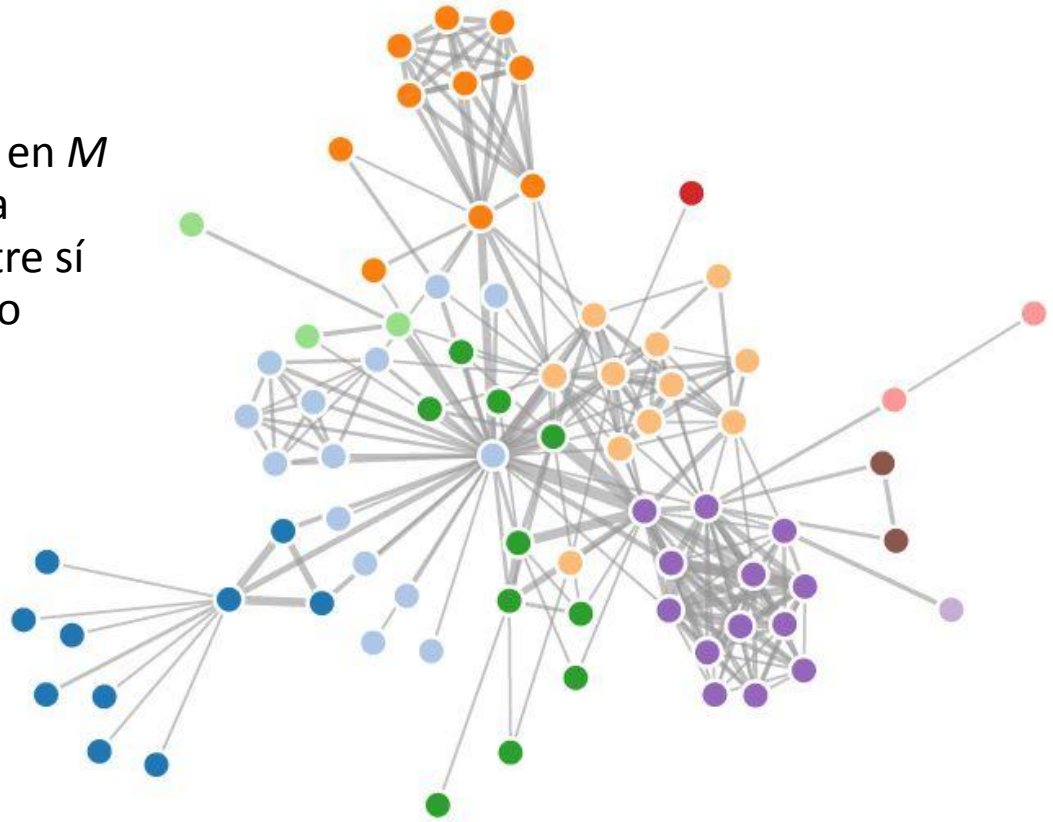
We argue here for the recognition of functional 'modules' as a critical level of biological organization. Modules are composed of many types of molecule. They have discrete functions that arise from interactions among their components (proteins, DNA, RNA and small molecules), but these functions cannot easily be predicted by studying the properties of the isolated components. We believe that general 'design principles' — profoundly shaped by the constraints of evolution — govern the structure and function of modules. Finally, the notion of function and functional properties separates biology

Nature 1999

Detectando comunidades en redes

Noción de comunidades:

Partición de los *N* vértices de la red en *M* grupos tales que nodos de la misma comunidad sean **más parecidos** entre sí que respecto a nodos fuera de dicho grupo



- Similaridad topológica de vértices
- Heurística: **Agrupamiento Jerárquico**

Agrupamiento Jerárquico

1. Se construye una matriz de similaridad de vértices a partir de la matriz de adyacencia

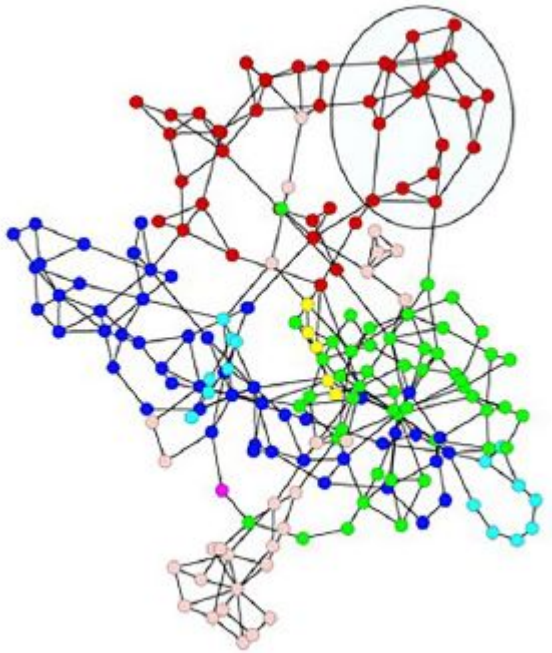
2. Iterativamente se identifican grupos de nodos de alta o baja similaridad dependiendo de alguna de estas dos estrategias:
    1. Estrategia **aglomerativa**: se adosan sucesivamente nodos y comunidades de alta similaridad
    2. Estrategia **divisiva**: se dividen sucesivamente comunidades, removiendo enlaces que conectan nodos de baja similaridad.

3. El resultado del ordenamiento producido se puede visualizar en una estructura llamada **dendrograma** o árbol jerárquico.

4. Es posible definir la partición en comunidades buscada a partir del dendrograma.

Agrupamiento Jerárquico:
Ravasz

1. Se construye una matriz de similaridad de vértices a partir de la matriz de adyacencia



Red metabolica E.Coli

$$TO_{ij} = \frac{n_{ij} + \Theta(A_{ij})}{min(k_i, k_j) + 1 - \Theta(A_{ij})}$$

Agrupamiento Jerárquico:
Ravasz

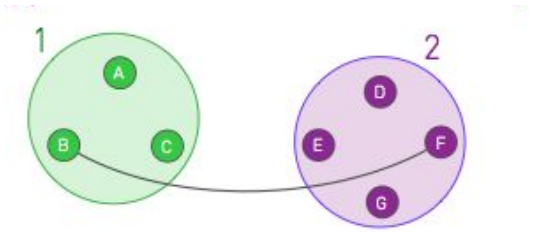
1. Se construye una matriz de similaridad de vértices a partir de la matriz de adyacencia

$$TO_{ij} = \frac{n_{ij} + \Theta(A_{ij})}{min(k_i, k_j) + 1 - \Theta(A_{ij})}$$

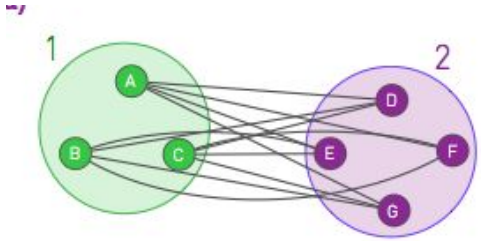
2. Criterio para determinar distancias entre **grupos**

$$\frac{1}{TO_{ij}} = r_{ij} =$$

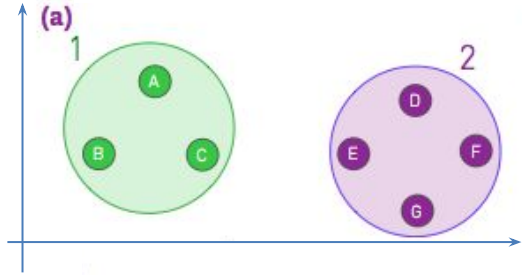
	D	E	F	G
A	2.75	2.22	3.46	3.08
B	3.38	2.68	3.97	3.40
C	2.31	1.59	2.88	2.34

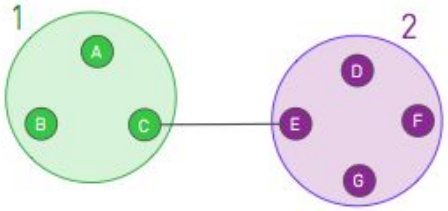


Complete Linkage: $x_{12} = 3.97$



Average Linkage: $x_{12} = 2.84$





Single Linkage: $x_{12} = 1.59$

Agrupamiento Jerárquico:
Ravasz

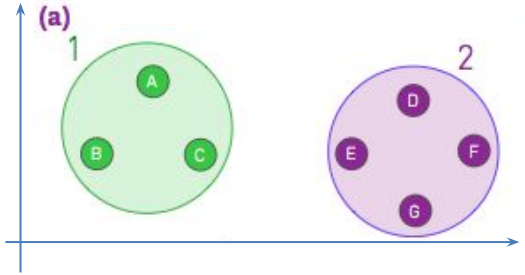
1. Se construye una matriz de similaridad de vértices a partir de la matriz de adyacencia

$$TO_{ij} = \frac{n_{ij} + \Theta(A_{ij})}{min(k_i, k_j) + 1 - \Theta(A_{ij})}$$

2. Criterio para determinar distancias entre **grupos**

$$\frac{1}{TO_{ij}} = r_{ij} =$$

	D	E	F	G
A	2.75	2.22	3.46	3.08
B	3.38	2.68	3.97	3.40
C	2.31	1.59	2.88	2.34



Distancia de Ward:

Da cuenta del incremento de la desviación cuadrática respecto al CM si se juntan los dos clusters examinados

$$\begin{aligned}\Delta(A,B) &= \sum_{i\in A\cup B} \|\vec{x}_i - \vec{m}_{A\cup B}\|^2 - \sum_{i\in A} \|\vec{x}_i - \vec{m}_A\|^2 - \sum_{i\in B} \|\vec{x}_i - \vec{m}_B\|^2 \\ &= \frac{n_A n_B}{n_A + n_B} \|\vec{m}_A - \vec{m}_B\|^2\end{aligned}$$

- Dados dos pares de clusters con CMs separados a la misma distancia, el algoritmo favorecerá unir a los clusters más chicos
- Utilizar Ward mantiene este *costo* lo más chico posible…es apropiado si pensamos que los grupos en nuestros datos poseen una estructura compacta

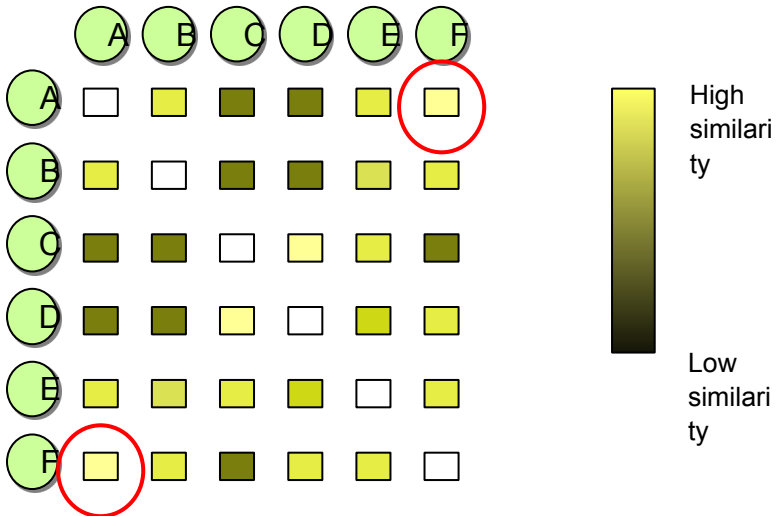
Agrupamiento Jerárquico:
Ravasz

1. Se construye una matriz de similaridad de vértices a partir de la matriz de adyacencia

2. Criterio para determinar distancias entre **grupos**
3. Estrategia **aglomerativa**: se adosan sucesivamente nodos y comunidades de alta similaridad

$$TO_{ij} = \frac{n_{ij} + \Theta(A_{ij})}{min(k_i, k_j) + 1 - \Theta(A_{ij})}$$

- Single linkage
- Complete linkage
- Average Linkage
- Ward distance
- …



1. Se asigna cada nodo a su propia comunidad y se evalua la similaridad entre todos los pares de nodos.
2. Se identifica el par con la similaridad más alta y se los combina en un único grupo
3. Se calculan las similaridades del nuevo grupo con todo el resto
4. Se repite desde el punto 2 hasta que todos los nodos se encuentran en un mismo grupo

Agrupamiento Jerárquico:
Ravasz

1. Se construye una matriz de similaridad de vértices a partir de la matriz de adyacencia

2. Estrategia **aglomerativa**: se adosan sucesivamente nodos y comunidades de alta similaridad

3. El resultado del ordenamiento producido se puede visualizar en una estructura llamada **dendrograma** o árbol jerárquico.

4. Es posible definir la partición en comunidades buscada a partir del dendrograma.

$$TO_{ij} = \frac{n_{ij} + \Theta(A_{ij})}{min(k_i, k_j) + 1 - \Theta(A_{ij})}$$

- Single linkage
- Complete linkage
- Average Linkage

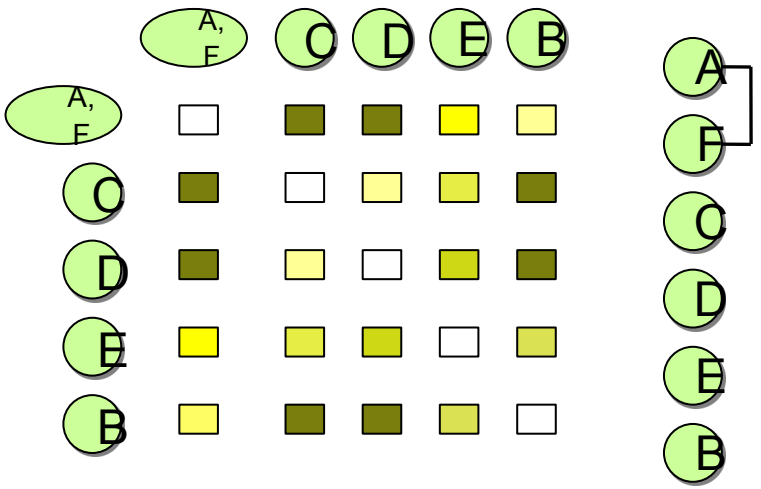
Agrupamiento Jerárquico:
Ravasz

1. Se construye una matriz de similaridad de vértices a partir de la matriz de adyacencia

2. Criterio para determinar distancias entre **grupos**
3. Estrategia **aglomerativa**: se adosan sucesivamente nodos y comunidades de alta similaridad

$$TO_{ij} = \frac{n_{ij} + \Theta(A_{ij})}{min(k_i, k_j) + 1 - \Theta(A_{ij})}$$

- Single linkage
- Complete linkage
- Average Linkage



1. Se asigna cada nodo a su propia comunidad y se evalua la similaridad entre todos los pares de nodos.
2. Se identifica el par con la similaridad más alta y se los combina en un único grupo
3. Se calculan las similaridades del nuevo grupo con todo el resto
4. Se repite desde el punto 2 hasta que todos los nodos se encuentran en un mismo grupo

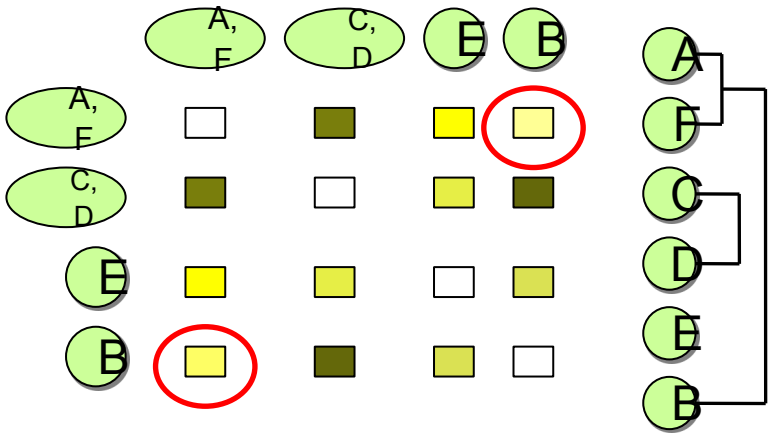
Agrupamiento Jerárquico:
Ravasz

1. Se construye una matriz de similaridad de vértices a partir de la matriz de adyacencia

2. Criterio para determinar distancias entre **grupos**
3. Estrategia **aglomerativa**: se adosan sucesivamente nodos y comunidades de alta similaridad

$$TO_{ij} = \frac{n_{ij} + \Theta(A_{ij})}{min(k_i, k_j) + 1 - \Theta(A_{ij})}$$

- Single linkage
- Complete linkage
- Average Linkage



1. Se asigna cada nodo a su propia comunidad y se evalua la similaridad entre todos los pares de nodos.
2. Se identifica el par con la similaridad más alta y se los combina en un único grupo
3. Se calculan las similaridades del nuevo grupo con todo el resto
4. Se repite desde el punto 2 hasta que todos los nodos se encuentran en un mismo grupo

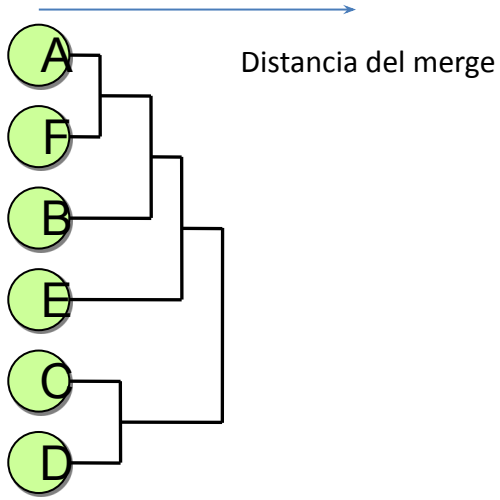
Agrupamiento Jerárquico:
Ravasz

1. Se construye una matriz de similaridad de vértices a partir de la matriz de adyacencia

$$TO_{ij} = \frac{n_{ij} + \Theta(A_{ij})}{min(k_i, k_j) + 1 - \Theta(A_{ij})}$$

2. Criterio para determinar distancias entre **grupos**
3. Estrategia **aglomerativa**: se adosan sucesivamente nodos y comunidades de alta similaridad

- Single linkage
- Complete linkage
- Average Linkage



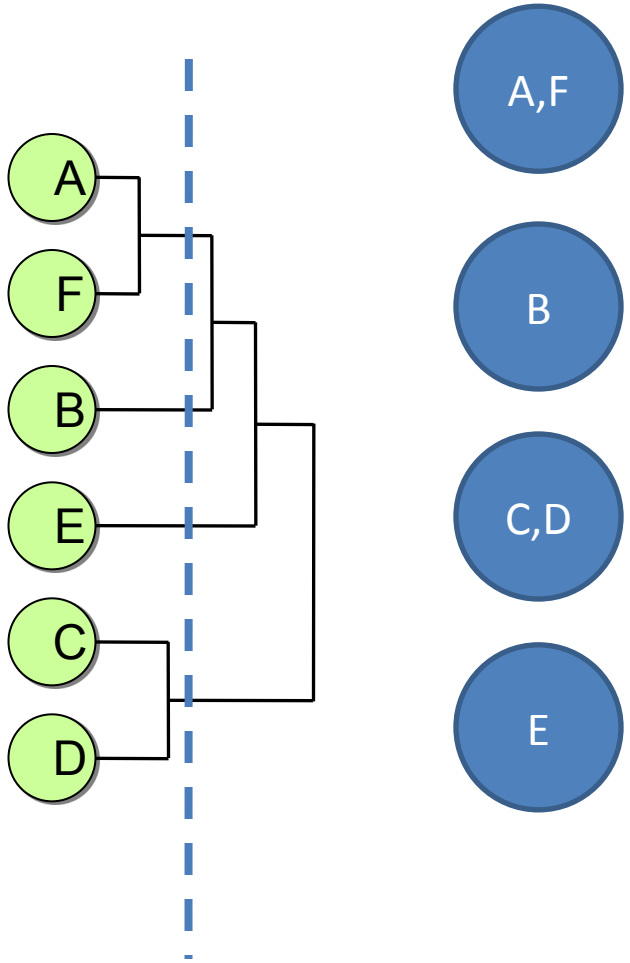
1. Se asigna cada nodo a su propia comunidad y se evalua la similaridad entre todos los pares de nodos.
2. Se identifica el par con la similaridad más alta y se los combina en un único grupo
3. Se calculan las similaridades del nuevo grupo con todo el resto
4. Se repite desde el punto 2 hasta que todos los nodos se encuentran en un mismo grupo

Agrupamiento Jerárquico:
Ravasz

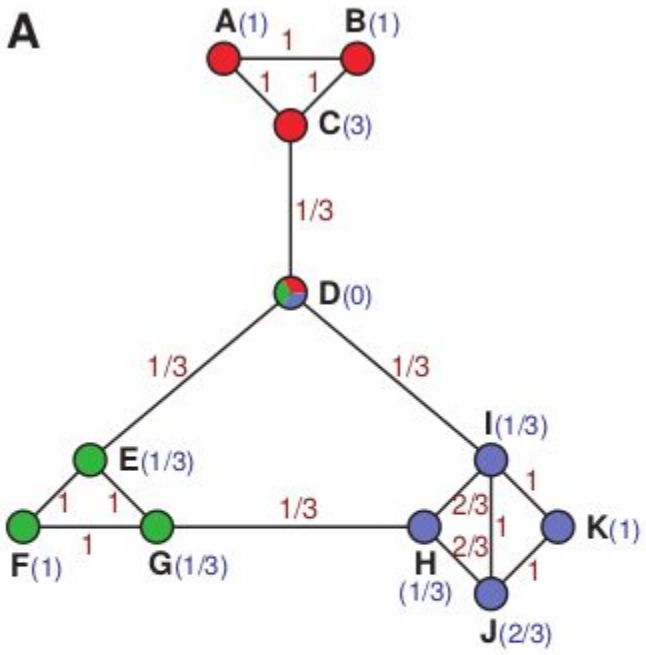
1. Se construye una matriz de similaridad de vértices a partir de la matriz de adyacencia

2. Criterio para determinar distancias entre **grupos**
3. Estrategia **aglomerativa**: se adosan sucesivamente nodos y comunidades de alta similaridad

4. El dendrograma describe el orden en el que nodos fueron agregados a comunidades. Es posible definir la partición en **comunidades** buscada a partir del dendrograma_



Topological Overlap y Dendrogramas



Matrix de TO con filas/columnas reordenadas segun dendrograma →

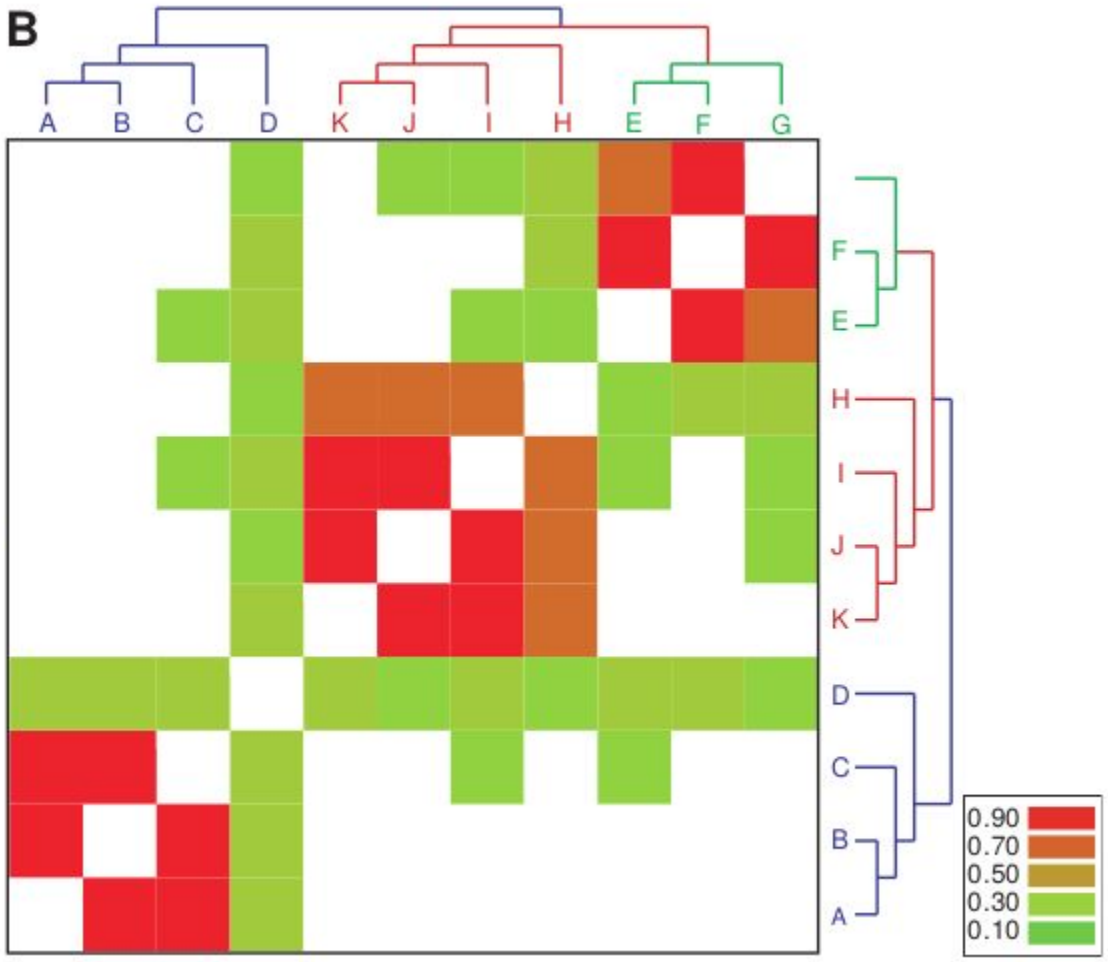


Fig 3 Ravasz 2002

Agrupamiento Jerárquico:ejemplo

Hierarchical Organization of Modularity in Metabolic Networks

E. Ravasz,[1] A. L. Somera,[2] D. A. Mongru,[2] Z. N. Oltvai,[2]*
A.-L. Barabási[1]*

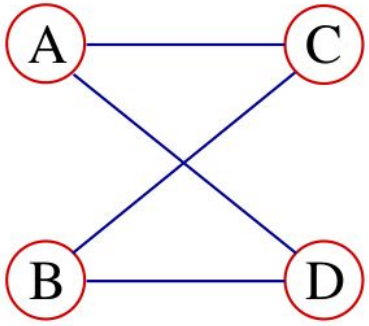
Spatially or chemically isolated functional modules composed of several cellular components and carrying discrete functions are considered fundamental building blocks of cellular organization, but their presence in highly integrated biochemical networks lacks quantitative support. Here, we show that the metabolic networks of 43 distinct organisms are organized into many small, highly connected topologic modules that combine in a hierarchical manner into larger, less cohesive units, with their number and degree of clustering following a power law. Within *Escherichia coli*, the uncovered hierarchical modularity closely overlaps with known metabolic functions. The identified network architecture may be generic to system-level cellular organization.

Science 2002

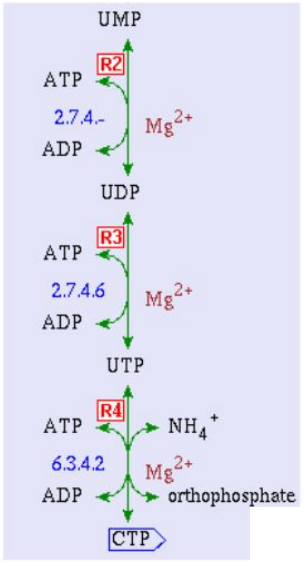
Aproximando la red metabolica

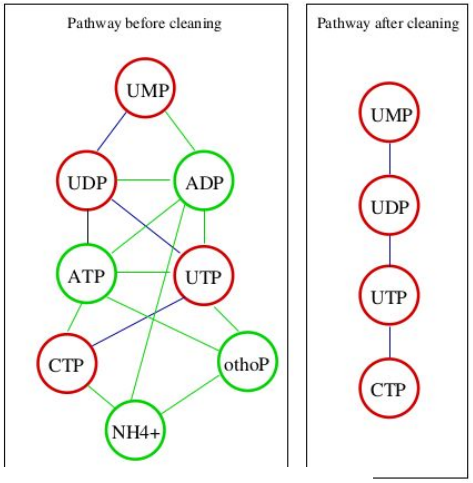
Proyeccion en red de sustratos

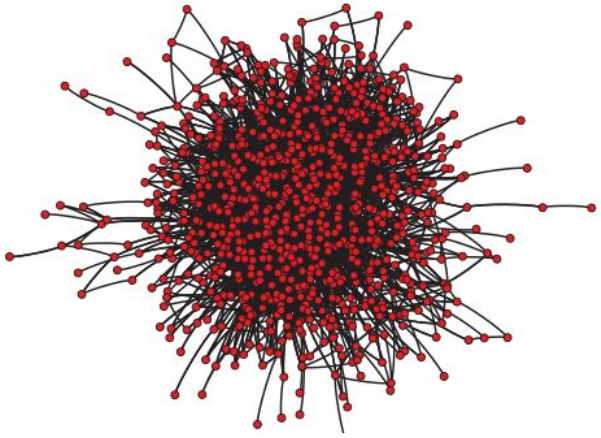
$$A + B \longrightarrow C + D$$

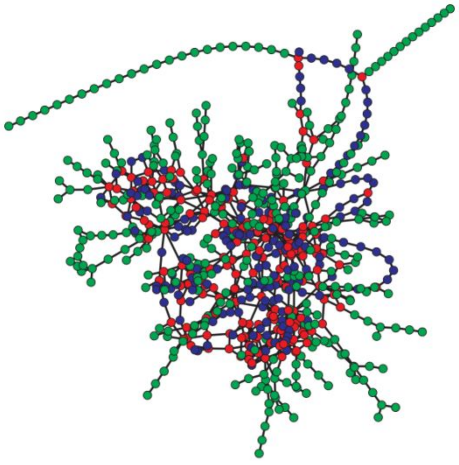


Reduccion bioquimica



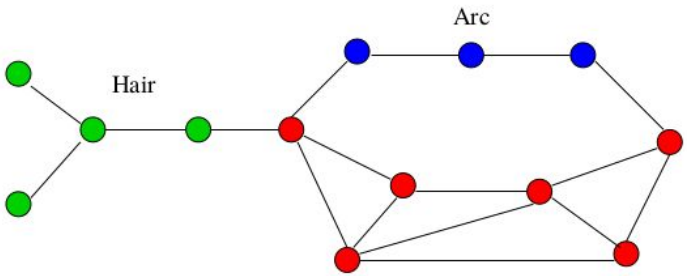


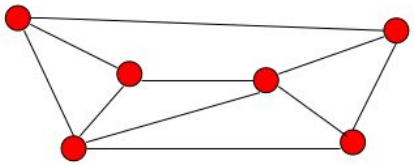


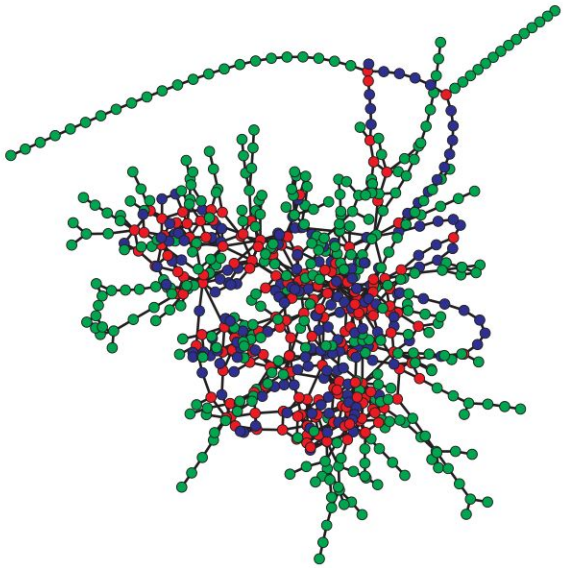


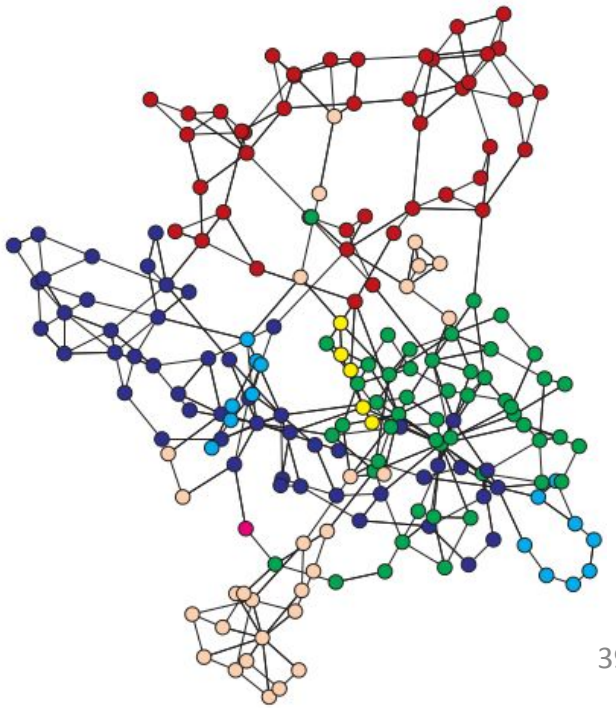
Aproximando la red metabolica

Reduccion topologica

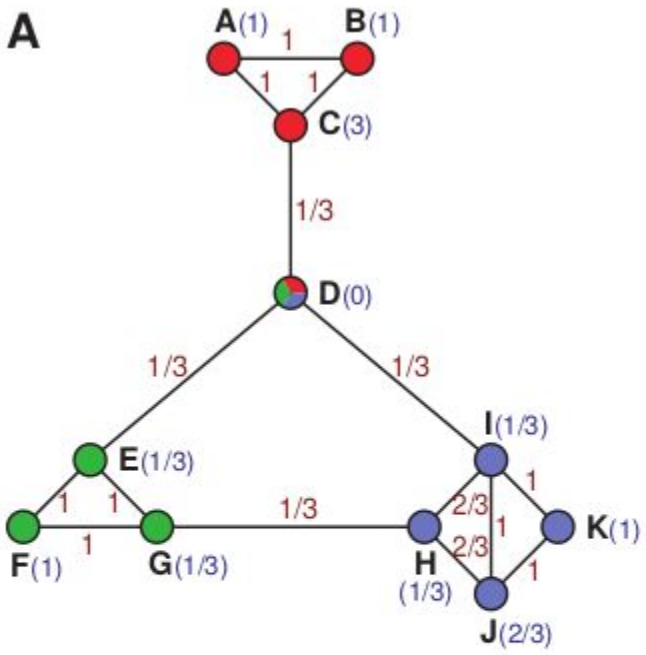








Topological Overlap y Dendrogramas



Matrix de TO con filas/columnas reordenadas segun dendrograma →

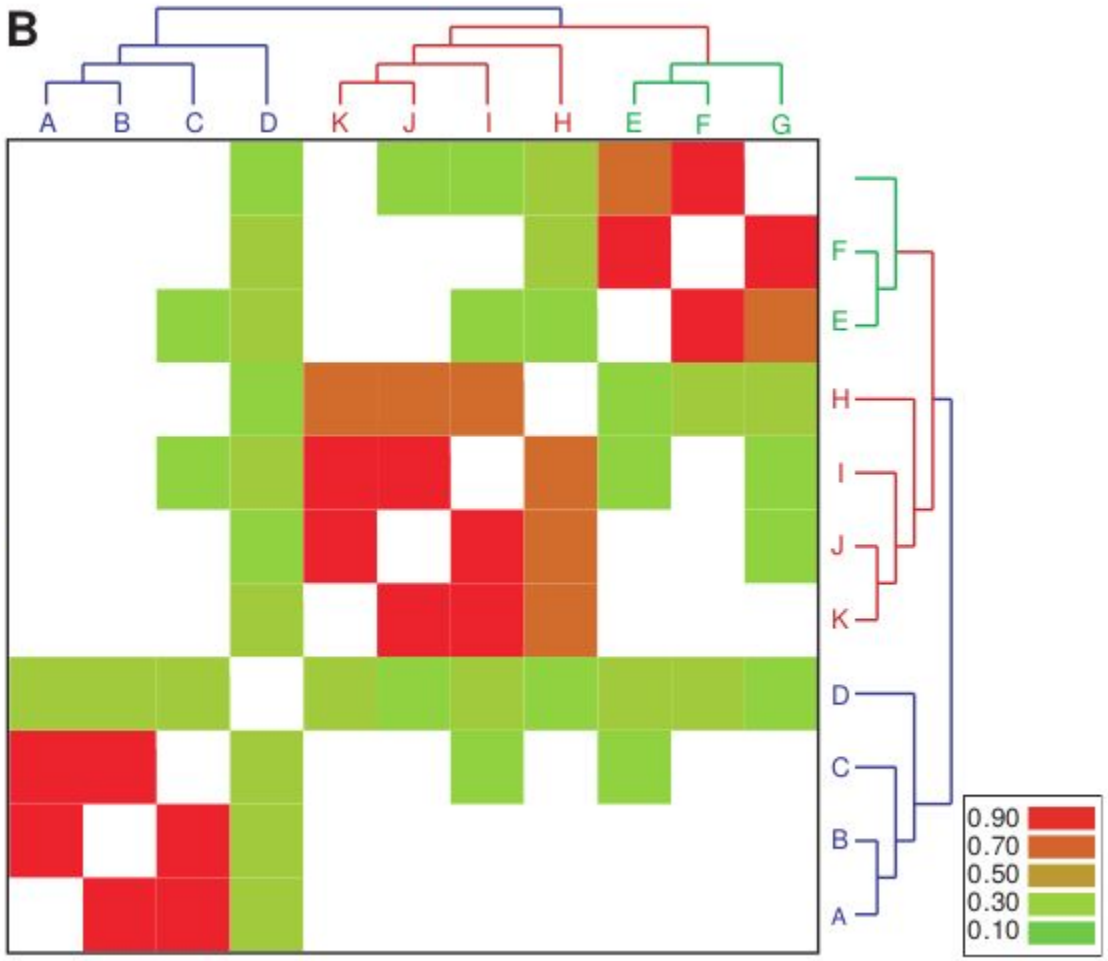
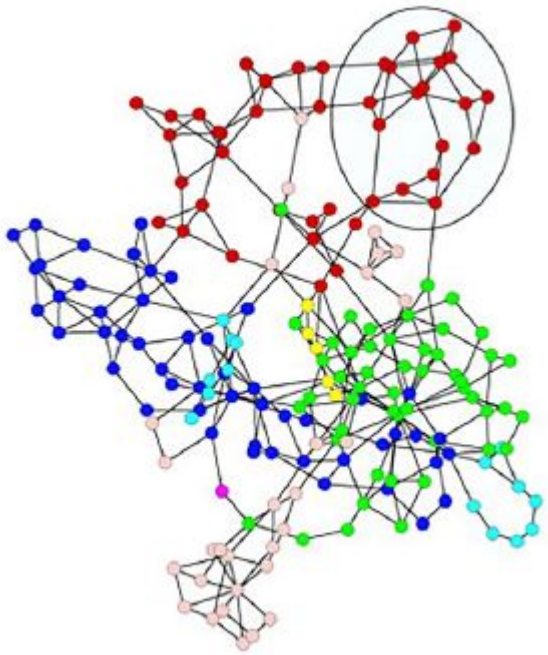


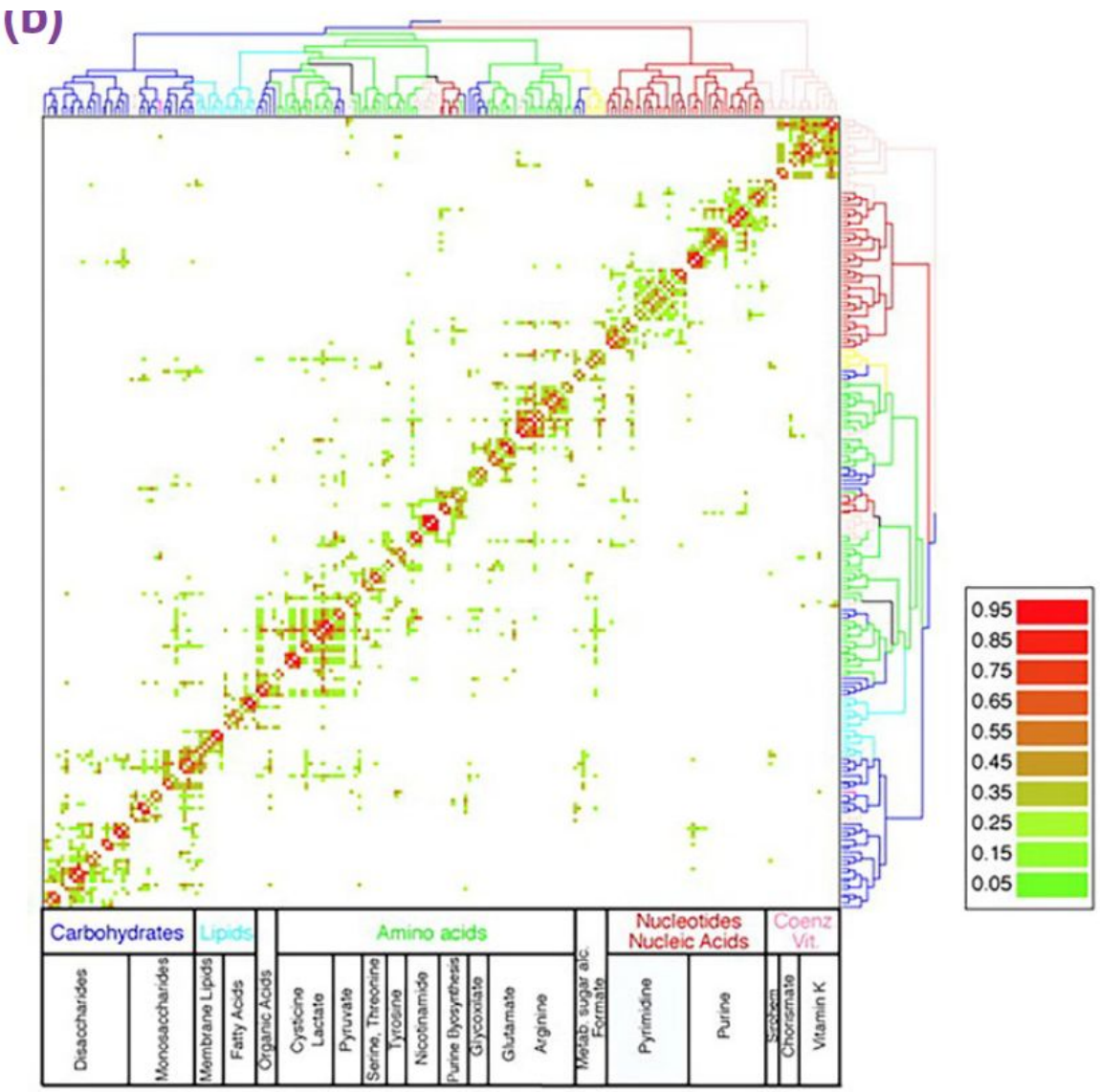
Fig 3 Ravasz 2002

Hierarchical Organization of Modularity in Metabolic Networks

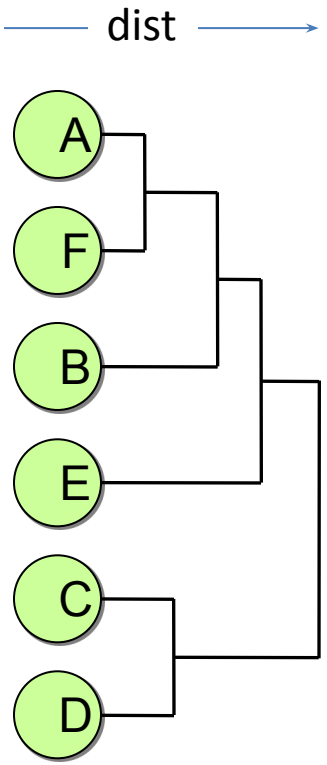
E. Ravasz,[1] A. L. Somera,[2] D. A. Mongru,[2] Z. N. Oltvai,[2]* A.-L. Barabási[1]*

Spatially or chemically isolated functional modules composed of several cellular components and carrying discrete functions are considered fundamental building blocks of cellular organization, but their presence in highly integrated biochemical networks lacks quantitative support. Here, we show that the metabolic networks of 43 distinct organisms are organized into many small, highly connected topologic modules that combine in a hierarchical manner into larger, less cohesive units, with their number and degree of clustering following a power law. Within *Escherichia coli*, the uncovered hierarchical modularity closely overlaps with known metabolic functions. The identified network architecture may be generic to system-level cellular organization.





Agrupamiento Jerárquico



Se utiliza muchisimo para encontrar clusters pero…

- Utiliza una organización jerárquica de los elementos que puede no existir en los datos reales.
- **N(N-1)/2** grados de libertad asociados a la matriz de similaridad. Los dendrogramas generados en cambio poseen sólo **N-1** merges. Asumir orden jerárquico reduce la cantidad de información!

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
iris1	5.1	3.5	1.4	0.2
iris2	4.9	3.0	1.4	0.2
iris3	4.7	3.2	1.3	0.2
iris4	4.6	3.1	1.5	0.2
iris5	5.0	3.6	1.4	0.2
iris6	5.4	3.9	1.7	0.4
iris7	4.6	3.4	1.4	0.3

Euclidean distance

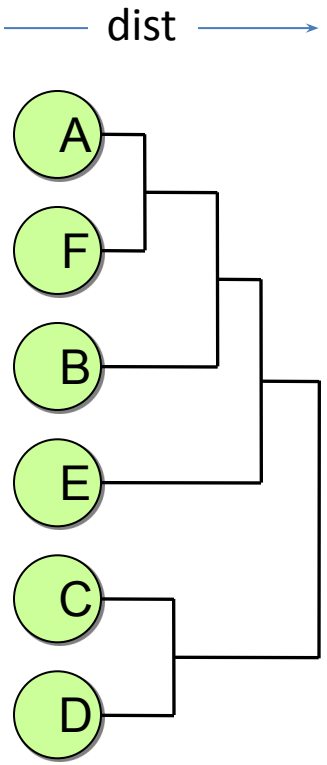
	1	2	3	4	5	6	7
1	0.00	0.54	0.51	0.65	0.14	0.62	0.52
2	0.54	0.00	0.30	0.33	0.61	1.10	0.51
3	0.51	0.30	0.00	0.24	0.51	1.10	0.26
4	0.65	0.33	0.24	0.00	0.65	1.20	0.33
5	0.14	0.61	0.51	0.65	0.00	0.62	0.46
6	0.62	1.10	1.10	1.20	0.62	0.00	0.99
7	0.52	0.51	0.26	0.33	0.46	0.99	0.00

Ultrametric approx

	iris1	iris2	iris3	iris4	iris5	iris6	iris7
iris1	0.00	1.27	1.27	1.27	0.14	0.71	1.27
iris2	1.27	0.00	0.44	0.44	1.27	1.27	0.44
iris3	1.27	0.44	0.00	0.24	1.27	1.27	0.32
iris4	1.27	0.44	0.24	0.00	1.27	1.27	0.32
iris5	0.14	1.27	1.27	1.27	0.00	0.71	1.27
iris6	0.71	1.27	1.27	1.27	0.71	0.00	1.27
iris7	1.27	0.44	0.32	0.32	1.27	1.27	0.00

0
0.14
0.32
0.44
0.71
1.27

Agrupamiento Jerárquico



Se utiliza muchisimo para encontrar clusters pero…

- Utiliza una organización jerárquica de los elementos que puede no existir en los datos reales.
- **N(N-1)/2** grados de libertad asociados a la matriz de similaridad. Los dendrogramas generados en cambio poseen sólo **N-1** merges. Asumir orden jerárquico reduce la cantidad de información!
- Dendrogramas satisfacen propiedad ultrametrica :
  Para cualquier triplete:
  $dist(a,f) \leq \max(dist(a,b), dist(b,f))$

  $dist(a,f) \leq dist(a,b) + dist(b,f)$ desigualdad triangular

- Entonces el procedimiento produce en realidad un mapeo entre un espacio métrico (o semi métrico) y uno ultramétrico