

Comunidades

cosas

Buscando grupos *naturales* de ~~nodes~~

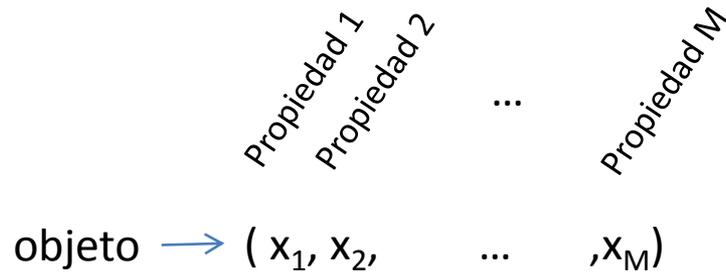
- cliques, k-componentes (Newman 7.8.1, 7.8.2)
- similaridad (N7.12)
 - Geometria: kmeans
 - topologia:
 - equivalencia estructural vs equivalencia regular
 - Clustering jerarquico (Ravasz 9.3)
- Girvan/Newman betweenness
- Modularity
- Overlapping communities
 - clique percolation
 - Link clustering
- Testing/charaterizing communities

Idea

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

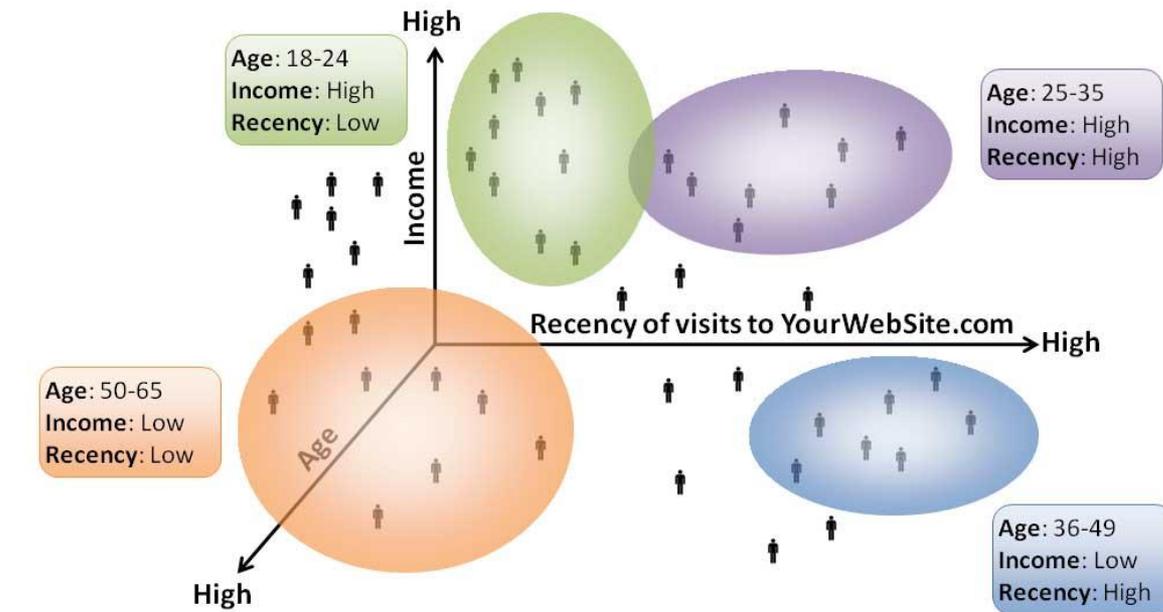
Similaridad

- Cuándo dos cosas son *parecidas*?
 - Similitud a partir de **vectores de características**



Similaridad

- Cuándo dos cosas son *parecidas*?
 - Similitud a partir de **vectores de características**

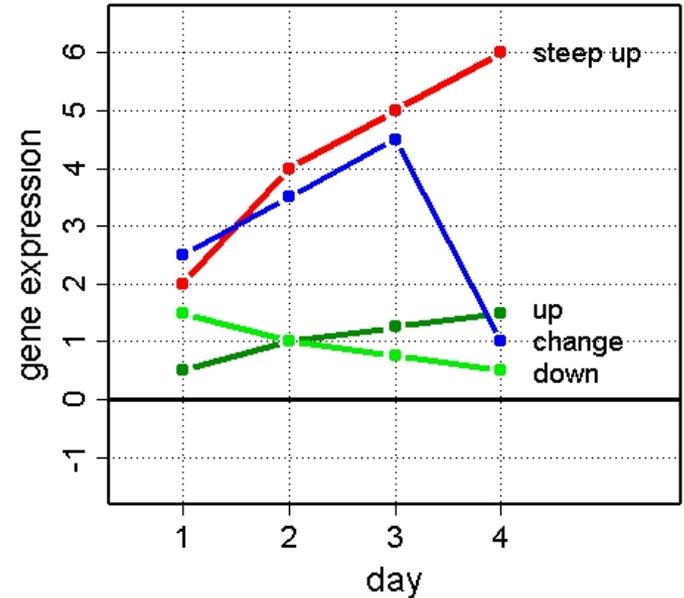


objeto → espacio de características (multidimensional) → Similitud o cercanía (si asumimos espacio métrico)

Ejemplo: Patrón de expresión génica

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

$X=(x_1,x_2,x_3,x_4)$, $Y=(y_1,y_2,y_3,y_4)$



Euclidean Distance

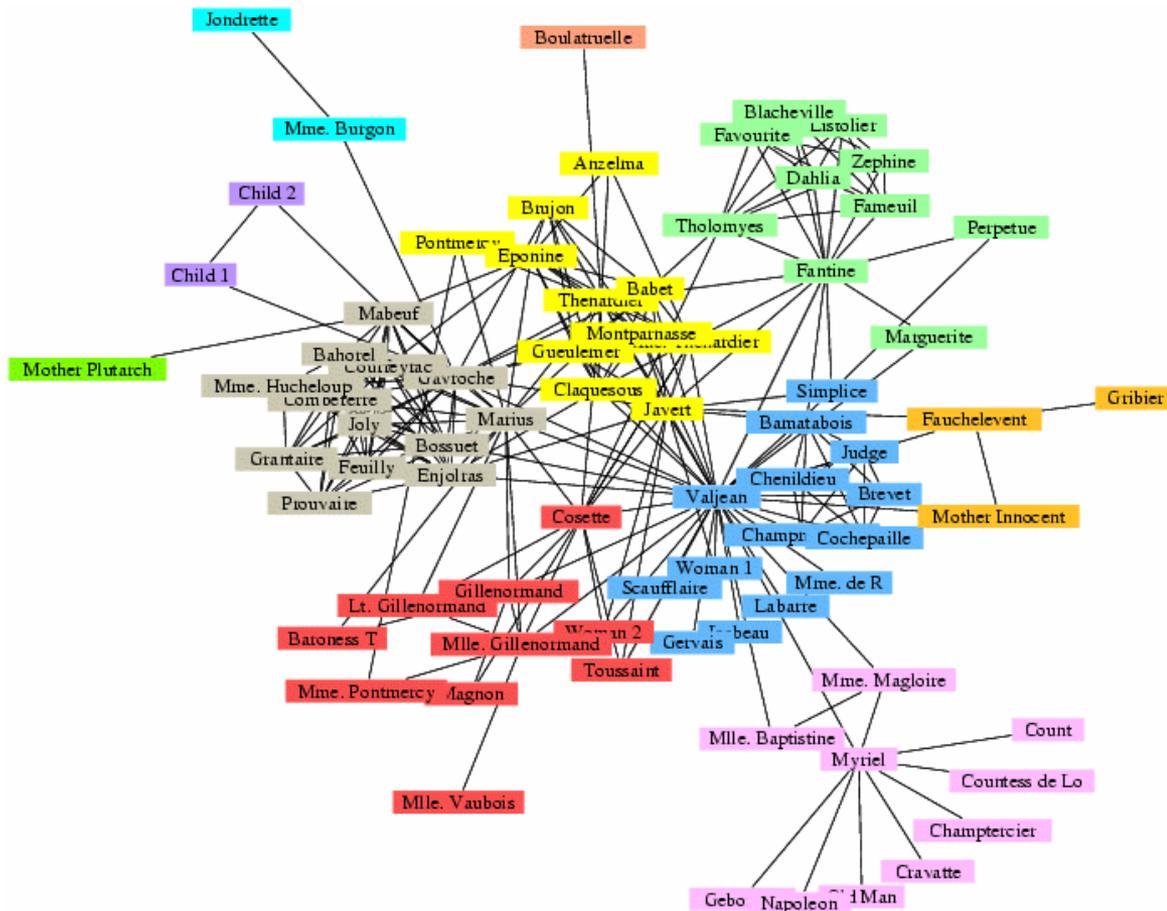
$$d(X, Y) = \sqrt{\sum (x_i - y_i)^2}$$

	●	●	●	●
●	0	2.60	2.75	2.25
●	2.60	0	1.23	2.14
●	2.75	1.23	0	2.15
●	2.25	2.14	2.15	0

0, 1.23, 2.14, 2.15, 2.25, 2.60, 2.75

Similaridad en redes

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

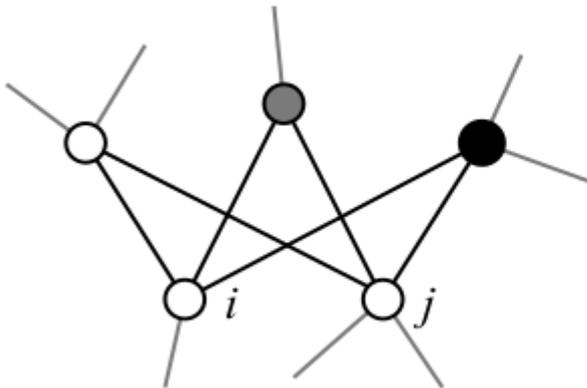


Similaridad en redes

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

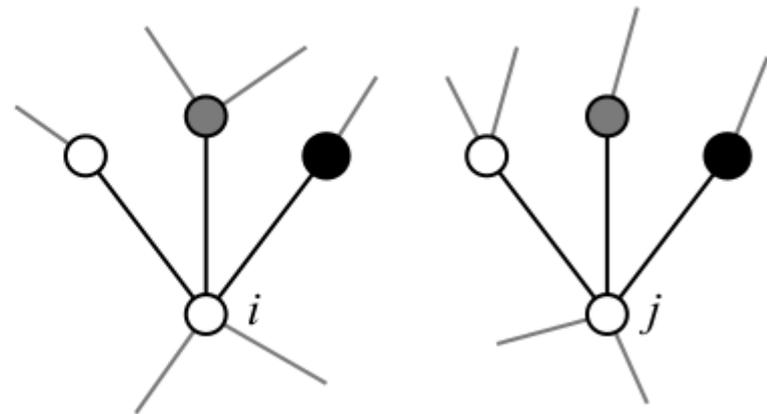
Equivalencia estructural

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**



Similaridad en redes

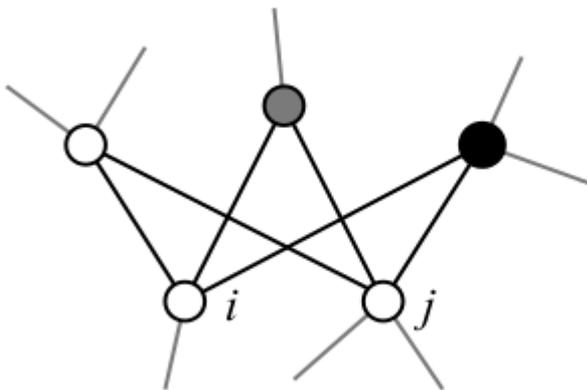


- Cuándo dos vértices de una red son ***parecidos***?
 - Similitud a partir de **propiedades topológicas**

Equivalencia estructural

dos nodos de una red son

estructuralmente equivalentes
si **comparten** muchos **vecinos**



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

$$n_{ij} = \sum_k A_{ik}A_{kj} = \mathbf{A}^2 \Big|_{ij}$$

$$A_{ki} = A_{ik} \Rightarrow \sum_k A_{ik}A_{jk}$$

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

Vamos a concentrarnos en redes no-dirigidas

Equivalencia Estructural

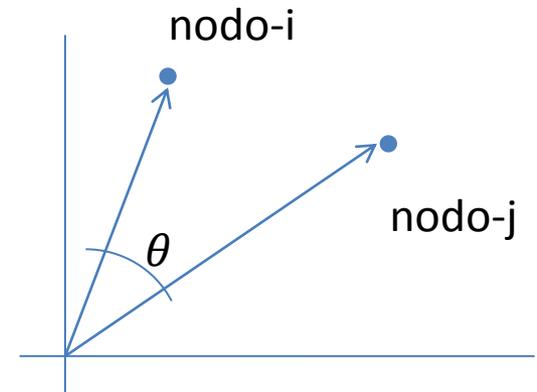
Similaridad tipo **Coseno**



- 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 



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



nro vecinos comunes
normalizado por la media
geometrica de los grados

redes no-pesadas

Equivalencia Estructural

Similaridad tipo **Correlación de Pearson**

- Similaridad como nro de vecinos mútuos respecto a lo que cabría esperar en una red al azar:

- Probabilidad de un dado nodo de conectarse con el nodo- i y además con el nodo- j de manera aleatoria:

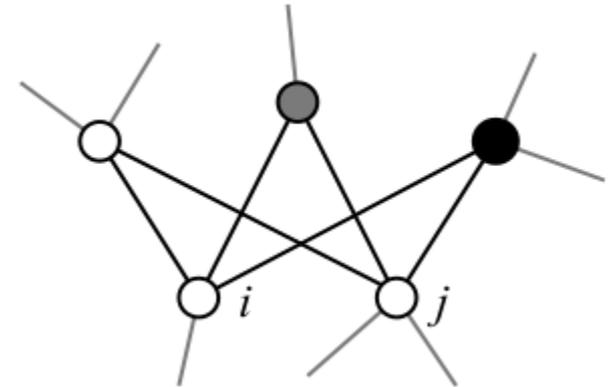
$$\frac{k_i}{n} \frac{k_j}{n}$$

- 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**

$$\begin{aligned}
 \sum_k A_{ik}A_{jk} - \frac{k_i k_j}{n} &= \sum_k A_{ik}A_{jk} - \frac{1}{n} \sum_k A_{ik} \sum_l A_{jl} \\
 &= \sum_k A_{ik}A_{jk} - n \langle A_i \rangle \langle A_j \rangle \\
 &= \sum_k [A_{ik}A_{jk} - \langle A_i \rangle \langle A_j \rangle] \\
 &= \sum_k (A_{ik} - \langle A_i \rangle)(A_{jk} - \langle A_j \rangle)
 \end{aligned}$$



$$n \text{ cov}(A_i, A_j)$$

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

El máximo valor de $\text{cov}(A_i, A_j)$ se obtiene cuando ambos vectores son idénticos. Eso ocurre cuando:

$$\text{cov}(A_i, A_j) = \text{cov}(A_i, A_i) = \text{var}(A_i) \quad \text{o} \quad \text{cov}(A_i, A_j) = \text{cov}(A_j, A_j) = \text{var}(A_j)$$

Equivalencia Estructural

Similaridad tipo **Correlación de Pearson**

$$\begin{aligned}
 \sum_k A_{ik}A_{jk} - \frac{k_i k_j}{n} &= \sum_k A_{ik}A_{jk} - \frac{1}{n} \sum_k A_{ik} \sum_l A_{jl} \\
 &= \sum_k A_{ik}A_{jk} - n \langle A_i \rangle \langle A_j \rangle \\
 &= \sum_k [A_{ik}A_{jk} - \langle A_i \rangle \langle A_j \rangle] \\
 &= \sum_k (A_{ik} - \langle A_i \rangle)(A_{jk} - \langle A_j \rangle)
 \end{aligned}$$



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

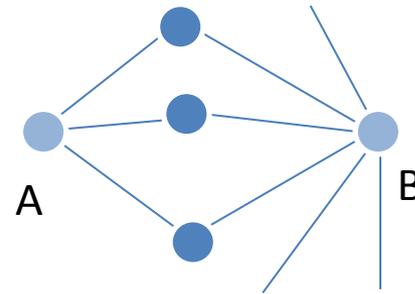
$-1 \leq r_{ij} \leq 1$ y resulta nulo si los nodos-i y j tienen el nro de vecinos esperados por azar

Equivalencia Estructural

Superposición topológica (*topological overlap*)

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

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



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

Equivalencia Estructural

Otras medidas de similaridad



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

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

- 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)

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

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

Equivalencia Estructural

Otras medidas de similaridad



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

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

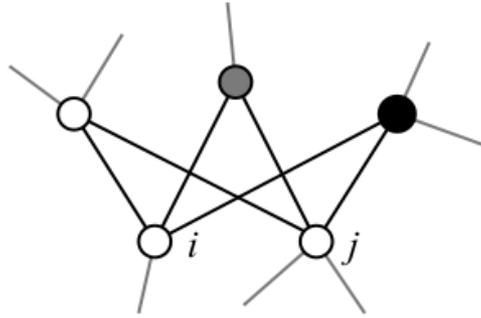
- 1, si el n_{ij} es el número esperado en una red aleatoria
- >1 si hay más vecinos comunes
- <1 si hay menos
- 0 si no tienen vecinos en común

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

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

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

Equivalencia Estructural



- Similaridad tipo coseno
- Similaridad de Pearson
- Overlap Topologico
- Similaridad: número de vecinos comunes **normalizado** por el número esperado
- Similaridad Euclídea (Hamming)

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

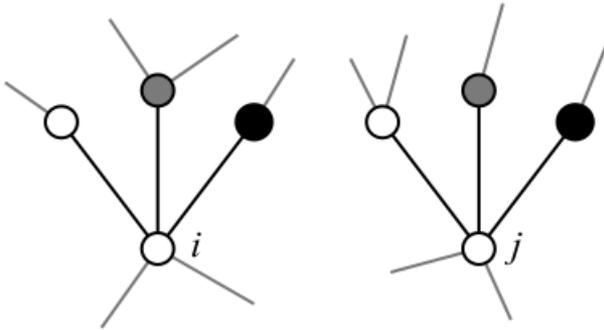
$$n_{ij} - \frac{k_i k_j}{n} \rightarrow \frac{\text{cov}(A_i, A_j)}{\sigma_i \sigma_j}$$

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

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

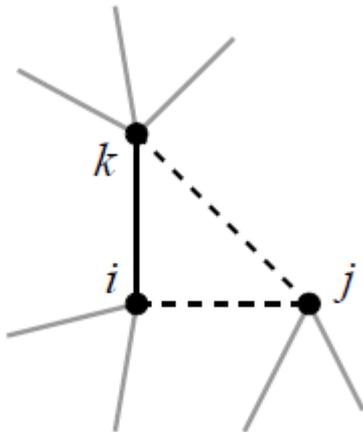
$$s_{ij} = 2 \frac{n_{ij}}{k_i + k_j}$$

Equivalencia Regular



Dos nodos son similares si sus vecinos lo son

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

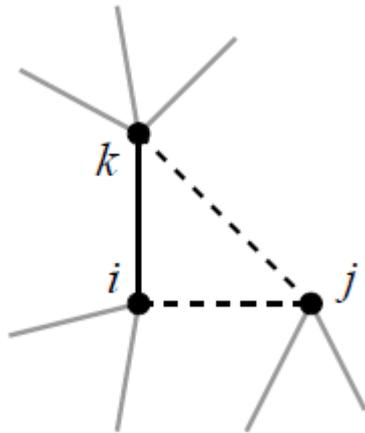


Versión que usaremos como definición:

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

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

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 .

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

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

$$\begin{aligned} \boldsymbol{\sigma}^{(0)} &= \mathbf{0} \\ \boldsymbol{\sigma}^{(1)} &= \mathbf{I} \\ \boldsymbol{\sigma}^{(2)} &= \alpha \mathbf{A} + \mathbf{I} \\ \boldsymbol{\sigma}^{(3)} &= \alpha^2 \mathbf{A}^2 + \alpha \mathbf{A} + \mathbf{I} \\ &\dots \end{aligned}$$

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

$$\alpha < 1$$

$$\alpha < 1/\kappa_1$$

Autovalor dominante de \mathbf{A}

Recordemos centralidad de Katz

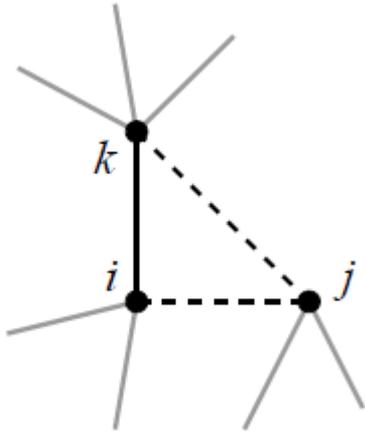
$$x_i = \alpha \sum_j A_{ij} x_j + 1$$

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

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

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

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 más chances de aumentar su nivel de similitud general. Si se desea mitigar este efecto...

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

Variante:

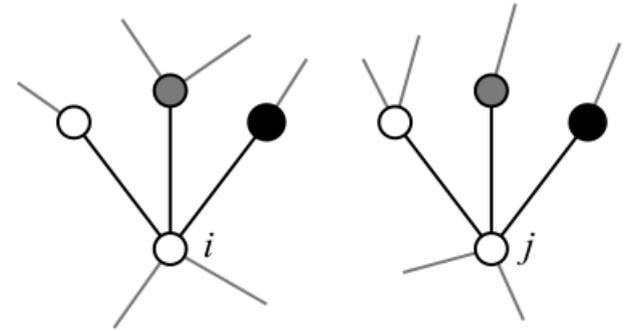
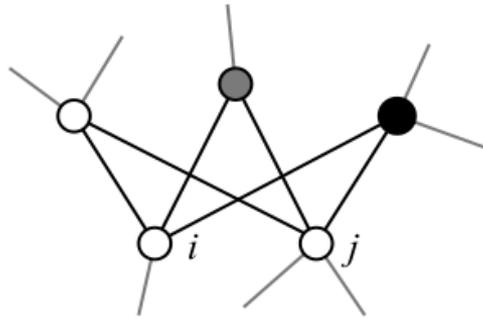
$$\sigma_{ij} = \frac{\alpha}{k_i} \sum_k A_{ik} \sigma_{kj} + \delta_{ij}$$

Variante:

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

$$\boldsymbol{\sigma} = (\mathbf{D} - \alpha \mathbf{A})^{-1} \mathbf{D} \quad \text{Con } \mathbf{D} \text{ la matriz diagonal}$$
$$D_{ii} = k_i$$

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

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

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

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

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

- Similaridad: **normalizado** por el número esperado

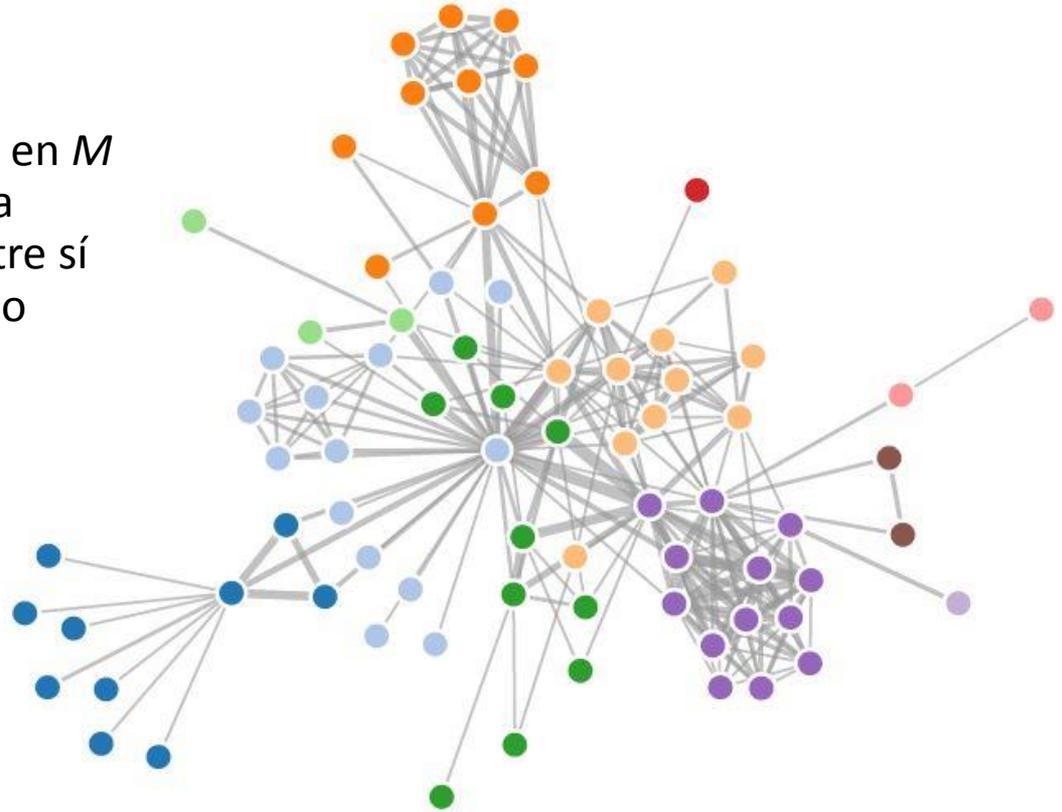
$$\frac{n_{ij}}{k_i k_j}$$

- 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.

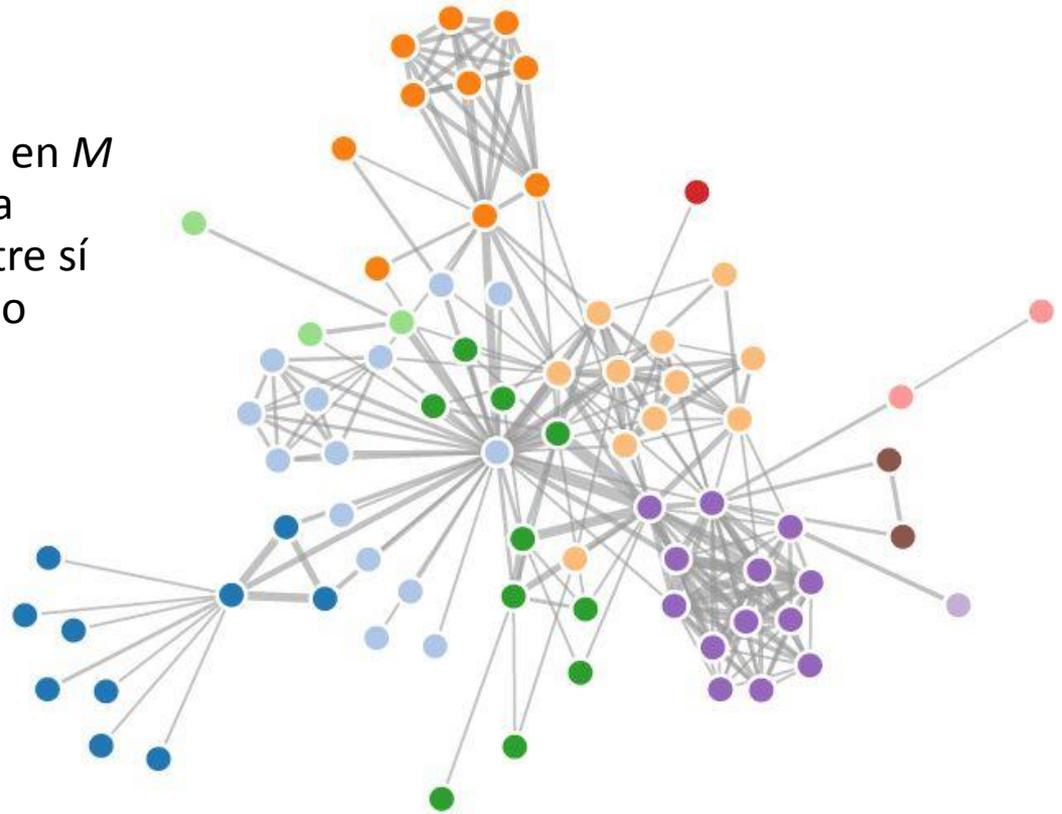
Having described such concepts, we need to explain how they arise from interactions among components in the cell.

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

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 similaridad siguiendo 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:ejemplo

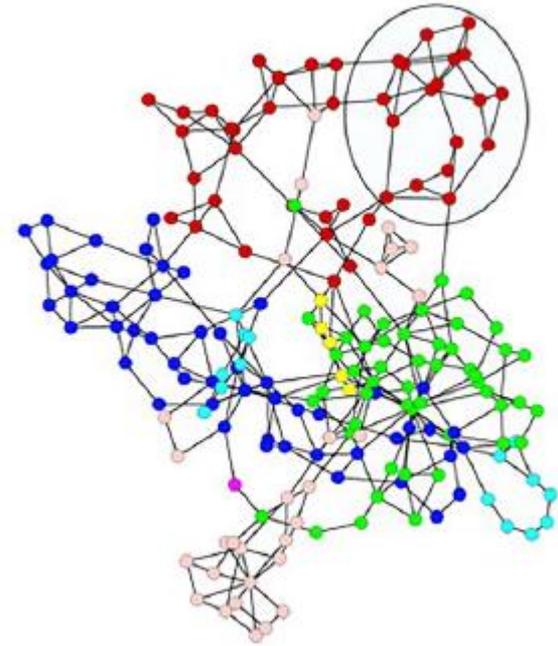
Hierarchical Organization of Modularity in Metabolic Networks

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.

Agrupamiento Jerárquico: Ravasaz

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:

Ravas

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

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

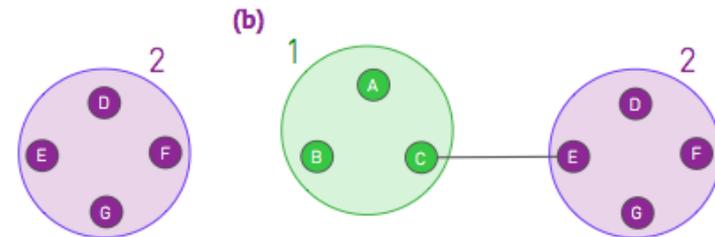
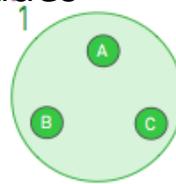
2. Estrategia **aglomerativa**: se adosan sucesivamente nodos y comunidades de **alta similaridad**

Distancia **entre grupos**

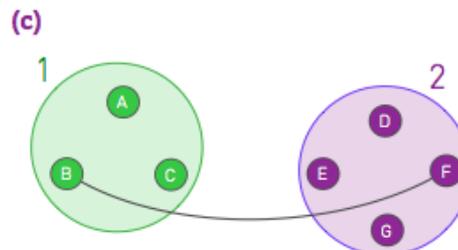


$$\frac{1}{TO_{ij}} = r_{ij}$$

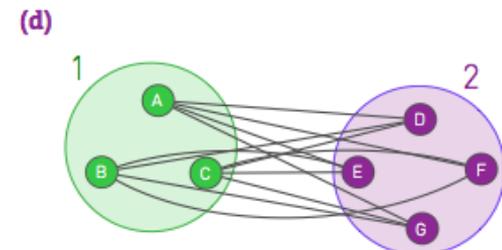
	D	E	F	G
A	2.75	2.22	3.46	3.08
B	3.38	2.68	3.97	3.40
C	2.31	1.59	2.88	2.34



Single Linkage: $x_{12} = 1.59$



Complete Linkage: $x_{12} = 3.97$



Average Linkage: $x_{12} = 2.84$

- Single linkage
- Complete linkage
- Average Linkage

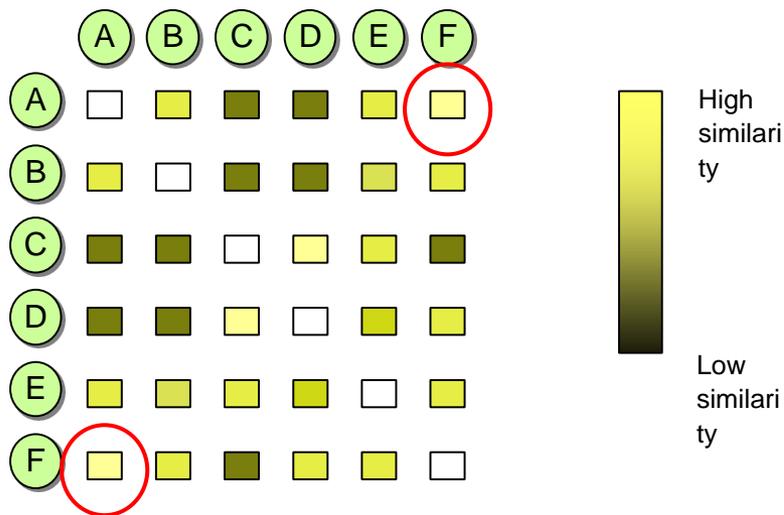
Agrupamiento Jerárquico:

Ravasaz

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. Agrupamiento jerárquico



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

- Single linkage
- Complete linkage
- Average Linkage

1. Se asigna cada nodo a su propia comunidad y se evalúa 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:

Ravas

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

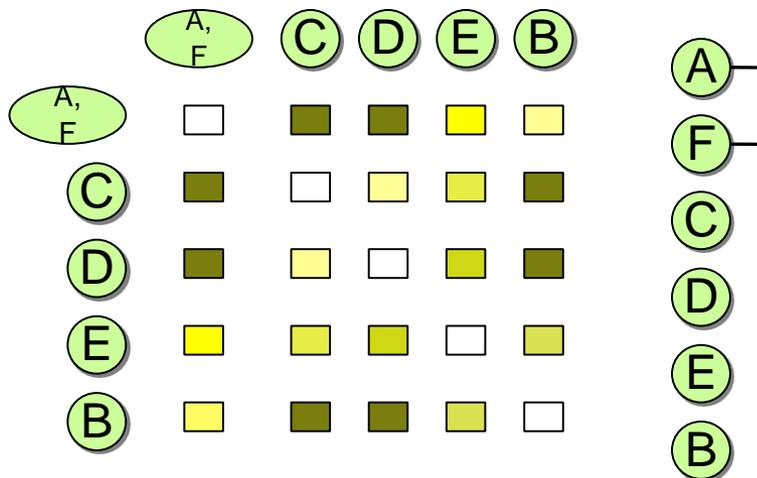
2. Estrategia **aglomerativa**: se adosan sucesivamente nodos y comunidades de alta similaridad

3. Agrupamiento jerárquico

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

- Single linkage
- Complete linkage
- Average Linkage

1. Se asigna cada nodo a su propia comunidad y se evalúa 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



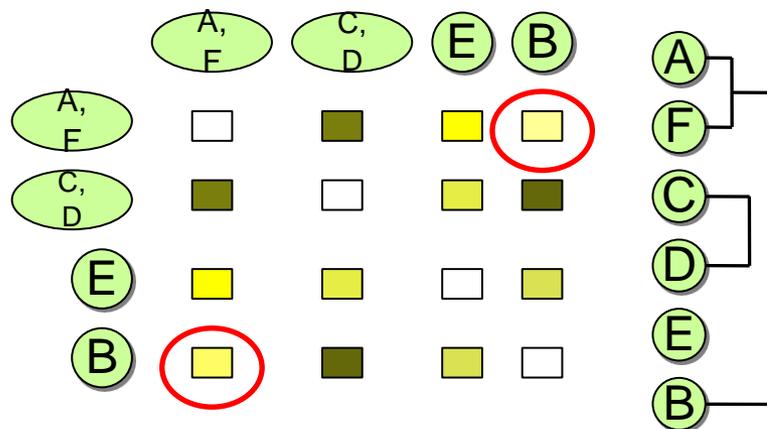
Agrupamiento Jerárquico:

Ravas

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. Agrupamiento jerárquico



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

- Single linkage
- Complete linkage
- Average Linkage

1. Se asigna cada nodo a su propia comunidad y se evalúa 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

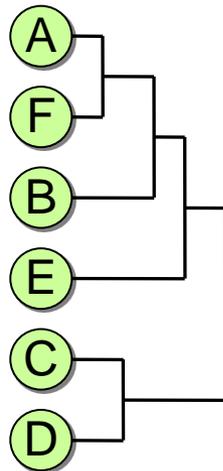
1. Se construye una matriz de similitud de vértices a partir de la matriz de adyacencia

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

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

- Single linkage
- Complete linkage
- Average Linkage

3. Agrupamiento jerárquico

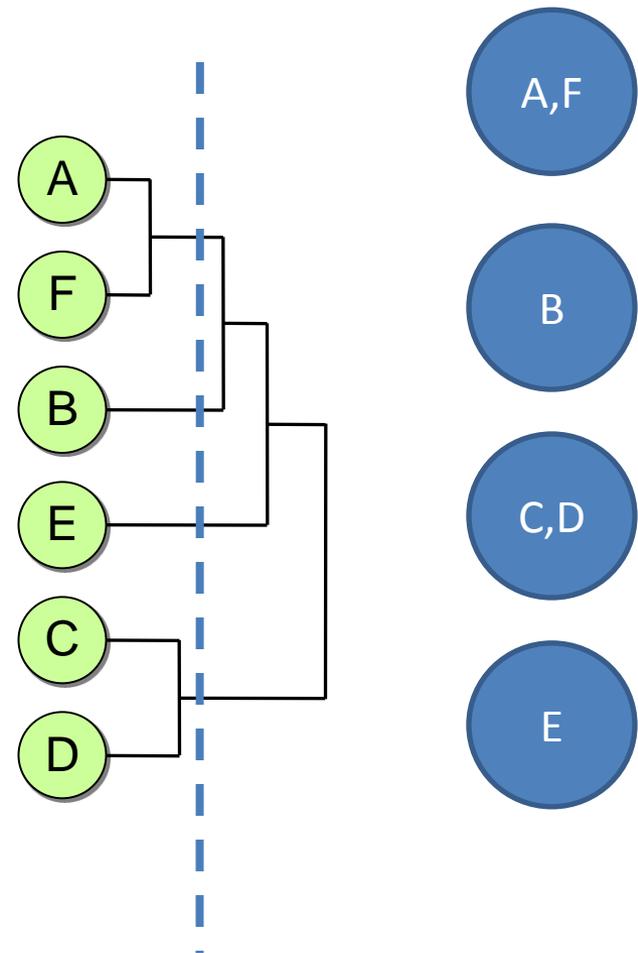


1. Se asigna cada nodo a su propia comunidad y se evalúa la similitud entre todos los pares de nodos.
2. Se identifica el par con la similitud más alta y se los combina en un único grupo
3. Se calculan las similitudes del nuevo grupo con todo el resto
4. Se repite desde el punto 2 hasta que todos los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

Agrupamiento Jerárquico:

Ravasaz

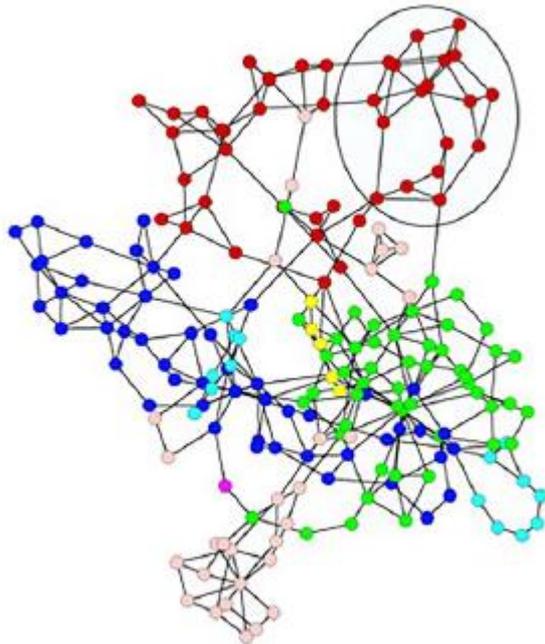
1. Se construye una matriz de similitud de vértices a partir de la matriz de adyacencia
2. Estrategia **aglomerativa**: se adosan sucesivamente nodos y comunidades de alta similitud
3. Agrupamiento jerárquico
4. El dendrograma describe el orden en el que nodos fueron agregados a comunidades. Es posible definir la partición en **comunidades** buscada a partir del dendrograma_



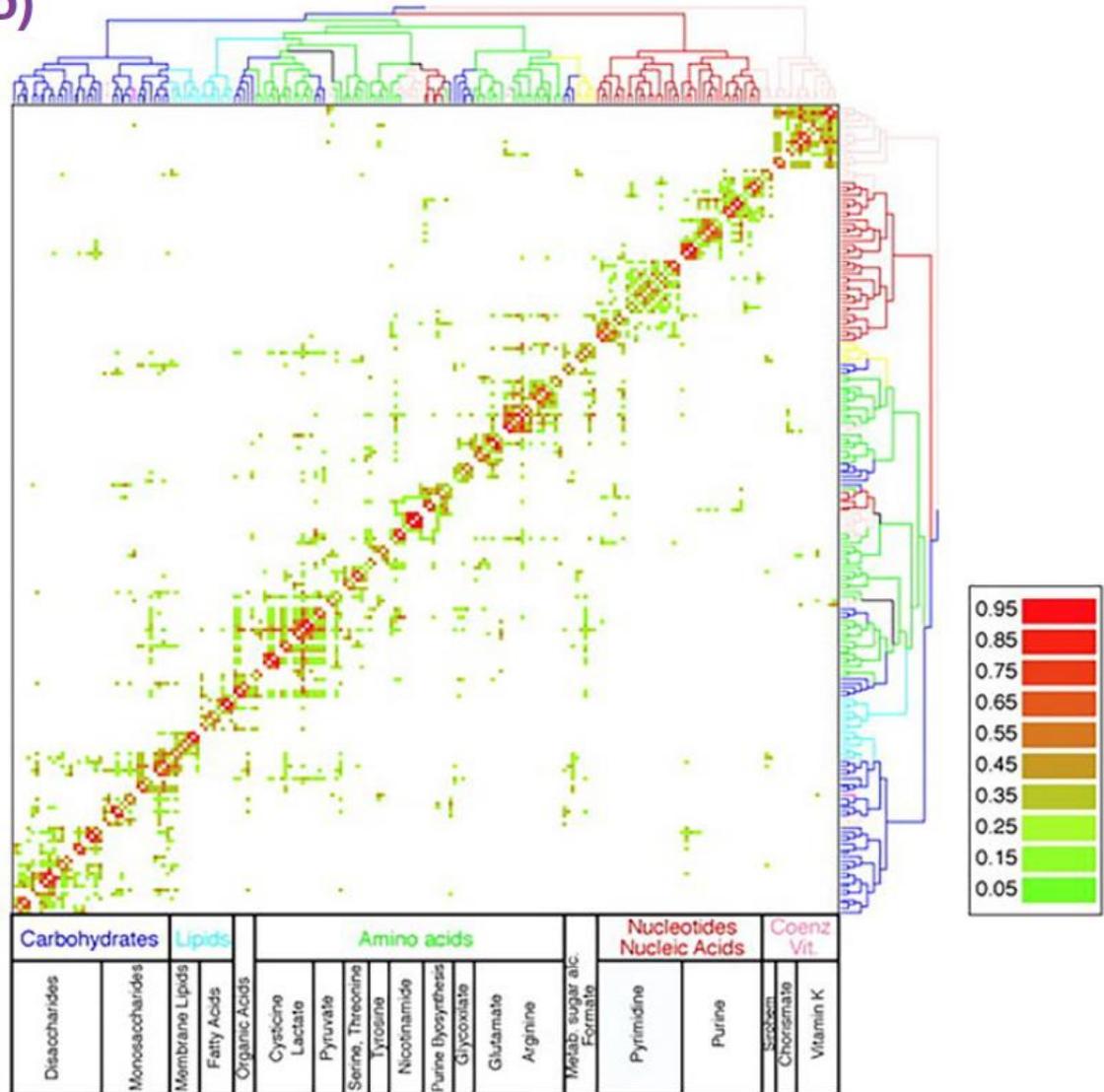
Hierarchical Organization of Modularity in Metabolic Networks

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.



(D)



Agrupamiento Jerárquico

Se utiliza muchísimo para encontrar clusters pero...

- Utiliza una organización jerárquica de los elementos que puede no existir en los datos reales.
- Grados de libertad asociados a la matriz de similaridad: $N(N-1)/2$. Los dendrogramas generados en cambio poseen sólo $N-1$ merges. Asumir orden jerárquico reduce la cantidad de información!
- Dendrogramas satisfacen propiedad ultramétrica :
$$\text{dist}(a,c) \leq \max(\text{dist}(a,b), \text{dist}(b,c))$$
- Entonces el procedimiento produce en realidad un mapeo entre un espacio métrico (o semi métrico) y uno ultramétrico

