Detección de Comunidades

en grafos

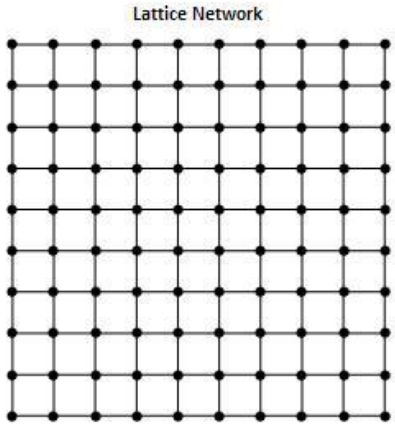
Barabasi Ch9

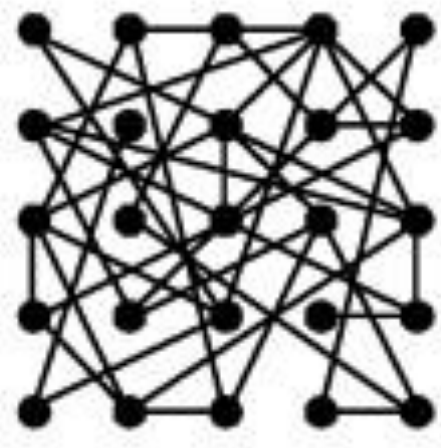
v1.2

2 hipótesis para que funcione

1. **Suponemos** que existe una estructura en comunidades embebida en la red, i.e. en A_{ij}

   Existen realmente grupos? Existen metodologías de búsquedas, no una definición a priori de lo que buscamos. Cómo lo sabemos si no buscamos? Cómo sabemos cuales heterogeneidades son las relevantes? Cómo sabemos cuál es la **escala relevante**?





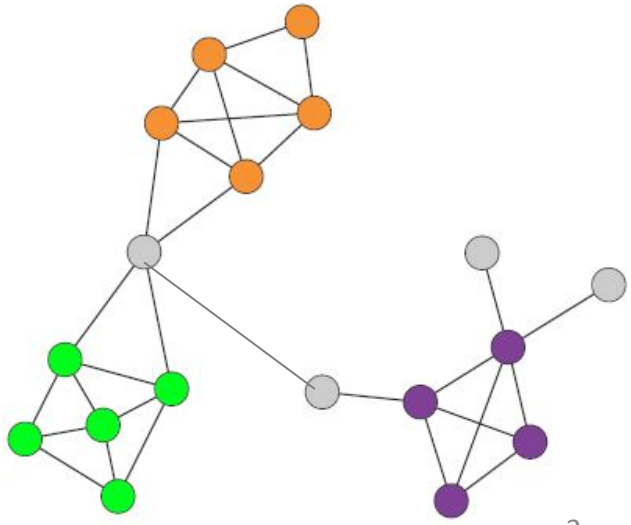
2 hipótesis para que funcione

1. **Suponemos** que existe una estructura en comunidades embebida en la red, i.e. en A_{ij}

2. Qué es una comunidad? Criterio (no definición) de **Conectividad** y **densidad**

   Una comunidad es un subgrafo **conexo, localmente denso**.

   - Desde un nodo de una comunidad puedo alcanzar cualquier otro
   - Un nodo de una comunidad se enlaza con alta probabilidad a nodos de la misma comunidad

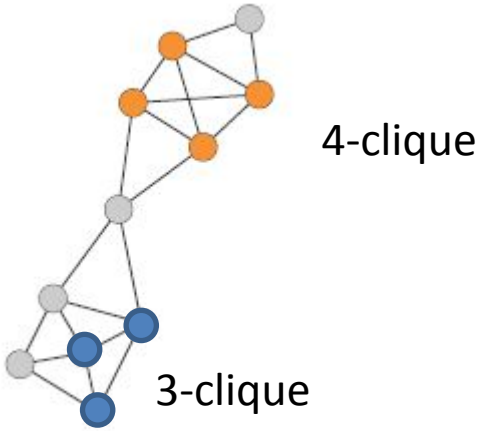


2 hipótesis para que funcione

1. **Suponemos** que existe una estructura en comunidades embebida en la red, i.e. en A_{ij}

2. Qué es una comunidad? Criterio (no definición) de **Conectividad** y **densidad**
   Una comunidad es un subgrafo **conexo, localmente denso**.

Primeras definiciones utilizaron el concepto de ***clique*** o ***subgrafo completo***:
subgrafo conexo de máxima densidad de enlaces



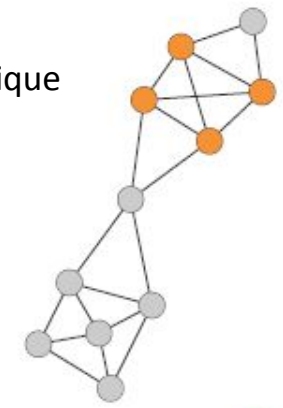
El concepto de clique suele ser demasiado restrictivo en la practica

2 hipótesis para que funcione

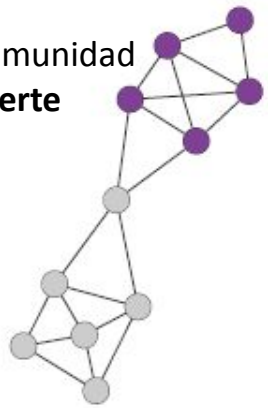
1. **Suponemos** que existe una estructura en comunidades embebida en la red, i.e. en A_{ij}

2. Qué es una comunidad? Criterio (no definición) de **Conectividad** y **densidad**
   Una comunidad es un subgrafo **conexo, localmente denso**.

4-clique

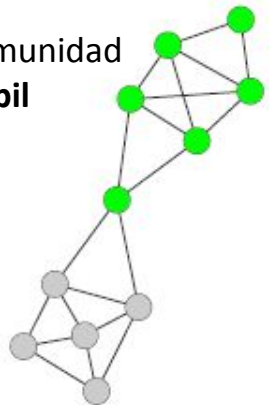


comunidad **fuerte**



$$k_i^{int} > k_i^{ext} \quad \forall i \in C$$

comunidad **débil**



$$\sum_{i \in C} k_i^{int} > \sum_{i \in C} k_i^{ext}$$

2 hipótesis para que funcione

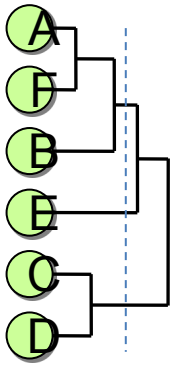
1. **Suponemos** que existe una estructura en comunidades embebida en la red, i.e. en A_{ij}

2. Qué es una comunidad? Criterio de **Conectividad** y **densidad**

   Existen realmente grupos? Existen metodologías de búsquedas, no una definición a priori de lo que buscamos. Cómo lo sabemos si no buscamos? Cómo sabemos cuales heterogeneidades son las relevantes? Cómo sabemos cuál es la **escala relevante**?

Ya vimos algoritmo **aglomerativo**:

I. Noción de distancia (similaridad)
II. Agrupamiento jerárquico. Partiendo de elementos disjuntos, se adosan sucesivamente nodos y comunidades de alta similaridad
III. Definición de grupos a partir del dendrograma



$$d_{a,c} \leq \max(d_{a,b}, d_{bc})$$

2 hipótesis para que funcione

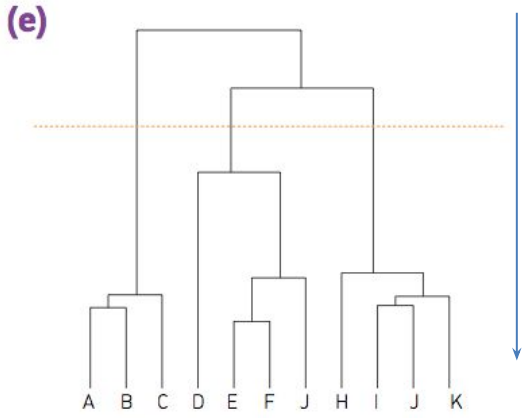
1. **Suponemos** que existe una estructura en comunidades embebida en la red, i.e. en A_{ij}

2. Qué es una comunidad? Criterio de **Conectividad** y **densidad**

   Existen realmente grupos? Existen metodologías de búsquedas, no una definición a priori de lo que buscamos. Cómo lo sabemos si no buscamos? Cómo sabemos cuales heterogeneidades son las relevantes? Cómo sabemos cuál es la **escala relevante**?

   Veamos ahora algoritmo **divisivo**: Girvan-Newman

Clusters à la Newman-Girvan



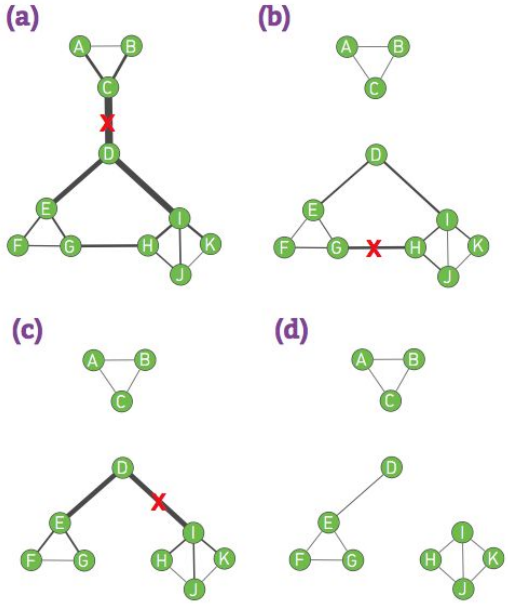
Idea: Partir de un cluster gigante. Ir dividiéndolo **removiendo enlaces** que conecten nodos de baja similaridad

1. Definir una **centralidad de enlaces**

   Intermediatez de enlaces (*link betweenness*):
   nro de caminos más cortos entre todos los pares de nodos, que atraviesen un dado enlace.

2. Agrupamiento jerárquico divisivo

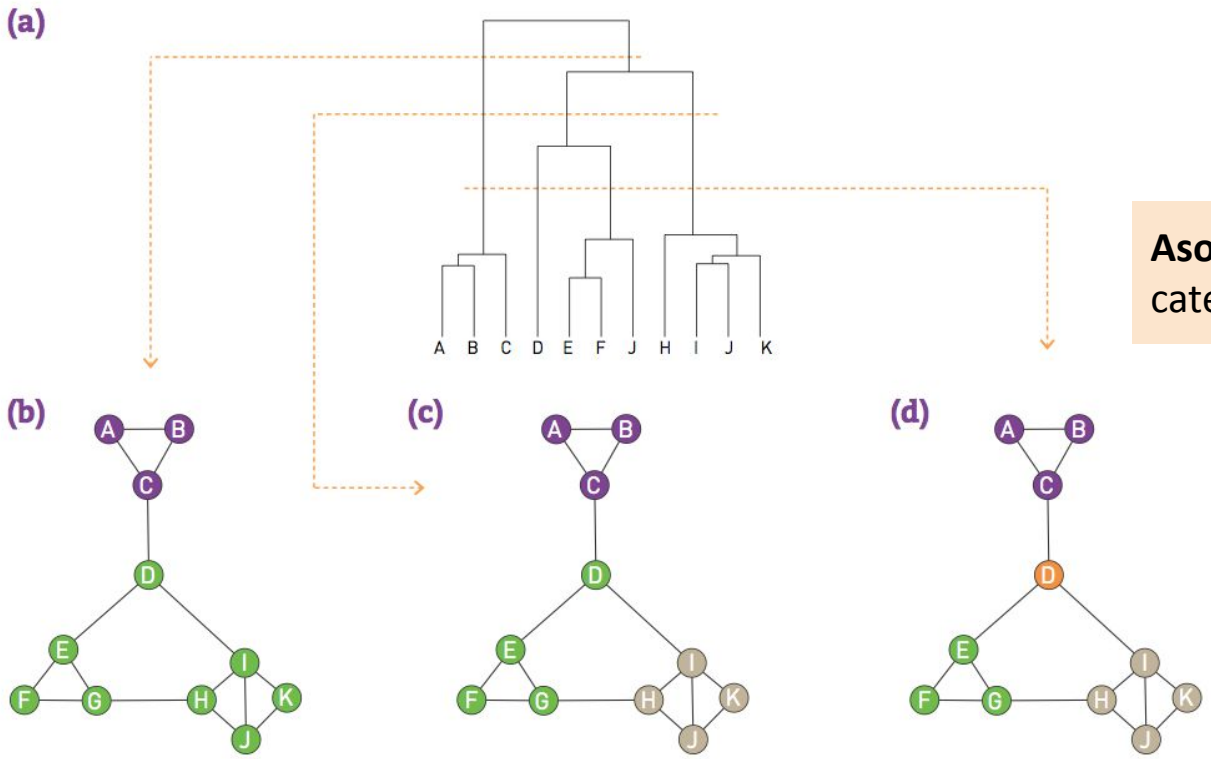
   I. Computar centralidad (betweenness) de enlaces
   II. Remover el enlace de mayor centralidad
   III. Recalcular centralidad de enlaces
   IV. Repetir hasta descartar el último enlace



M. Girvan & M.E.J. Newman, *PNAS* 99 (2002).
Barabási, *Ch9*

Clusters à la Newman-Girvan

Donde **cortar** el dendrograma para definir los clusters?



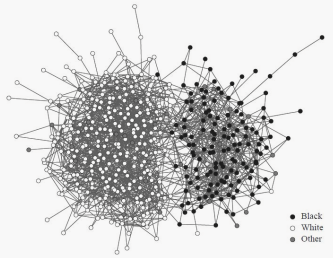
pocos enlaces internos

demasiados enlaces externos

Como cuantificar el **acuerdo** entre **cableado** y **partición** en grupos?

Asortatividad en la variable categórica: “pertenencia al cluster C”

Assortative mixing: **caracerísticas categóricas**



Supongamos que existen n_c clases diferentes para los n nodos de una red de m enlaces.
Sea c_i la clase del nodo-i. El número de enlaces entre mismo tipo de nodos resulta:

Red real

$$\sum_{\text{edges } (i,j)} \delta(c_i, c_j) = \frac{1}{2} \sum_{ij} A_{ij}\, \delta(c_i, c_j)$$

$$\delta(c_i, c_j) = 1 \text{ sii } c_i = c_j$$

Red aleatoria (recableado)

$$\frac{1}{2} \sum_{ij} \frac{k_i k_j}{2m}\, \delta(c_i, c_j)$$

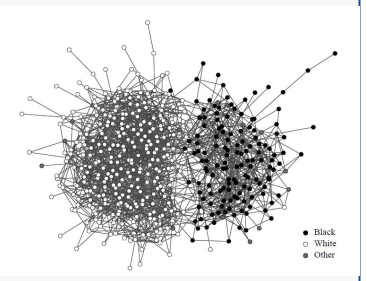
Consideramos la diferencia: $\frac{1}{2} \sum_{ij} A_{ij}\, \delta(c_i, c_j) - \frac{1}{2} \sum_{ij} \frac{k_i k_j}{2m}\, \delta(c_i, c_j)$

modularidad

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j)$$

- Q > 0 si hay más enlaces entre vertices del mismo tipo que los esperados por azar
- Q < 0 si hay menos enlaces entre vertices del mismo tipo que los esperados por azar

Assortative mixing: **caracerísticas categóricas**



Otra manera de computar asortatividad/modularidad:

$$Q = \frac{1}{2m}\sum_{ij}\left(A_{ij} - \frac{k_i k_j}{2m}\right)\overbrace{\sum_r \delta(c_i,r)\delta(c_j,r)}^{\delta(c_i,c_j)}$$

$$= \sum_r \left[\frac{1}{2m}\sum_{ij} A_{ij}\,\delta(c_i,r)\delta(c_j,r) - \frac{1}{2m}\sum_i k_i\,\delta(c_i,r)\frac{1}{2m}\sum_j k_j\delta(c_j,r)\right]$$

$$e_{rs} = \frac{1}{2m}\sum_{ij} A_{ij}\delta(c_i,r)\delta(c_j,s)$$

fracción de enlaces entre nodos de la clases r y s (o sea, probabilidad de encontrar un nodo de clase *r* y otro de clase *s* en los extremos de un enlace).

$$a_r = \frac{1}{2m}\sum_{ij} A_{ij}\delta(c_i,r)$$

fracción de enlaces adyacente a nodos del tipo r (o sea, probabilidad de encontrar un nodo de tipo *r* en un extremo de un enlace)

Entonces…

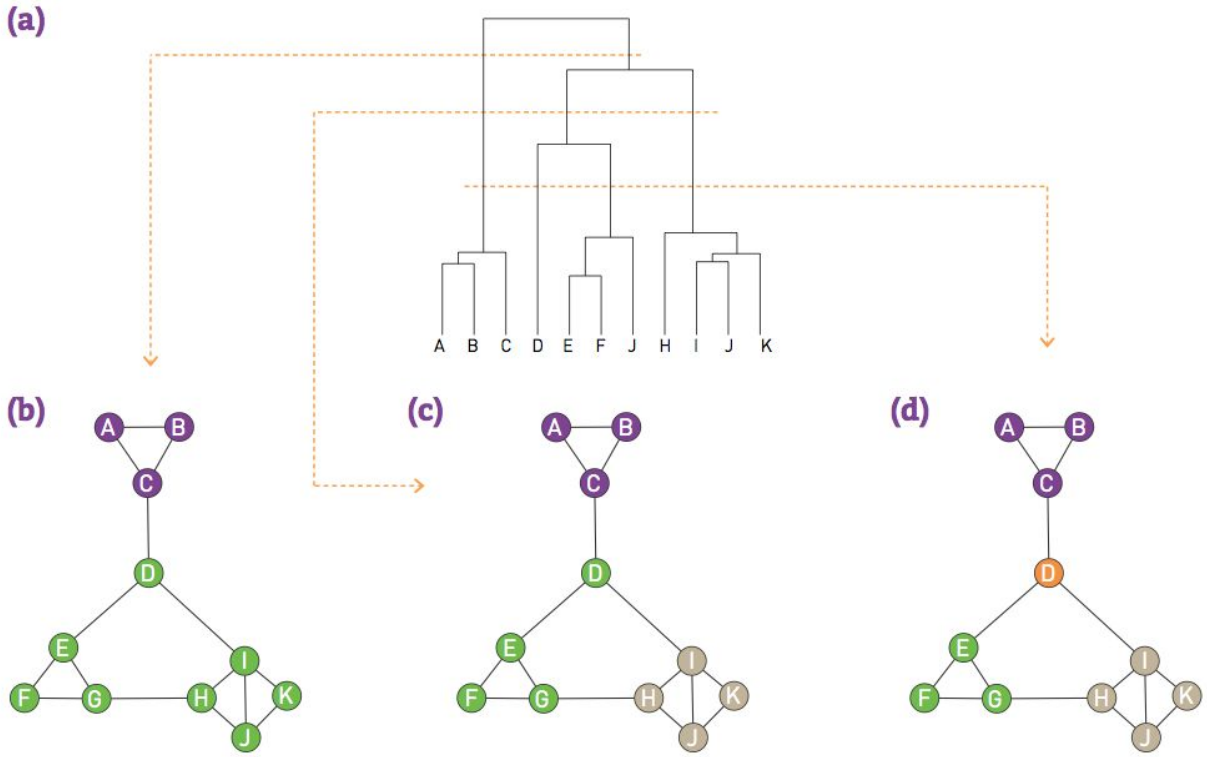
$$Q = \sum_r (e_{rr} - {a_r}^2)$$

Notar que en una red sin correlaciones

$$e_{rs} = a_r\, a_s \rightarrow Q = 0$$

Clusters à la Newman-Girvan

Donde **cortar** el dendrograma para definir los clusters?



pocos enlaces internos

demasiados enlaces externos

Como cuantificar el **acuerdo** entre **cableado** y **partición** en grupos?

fracción enlaces adyacentes a nodos de cluster r

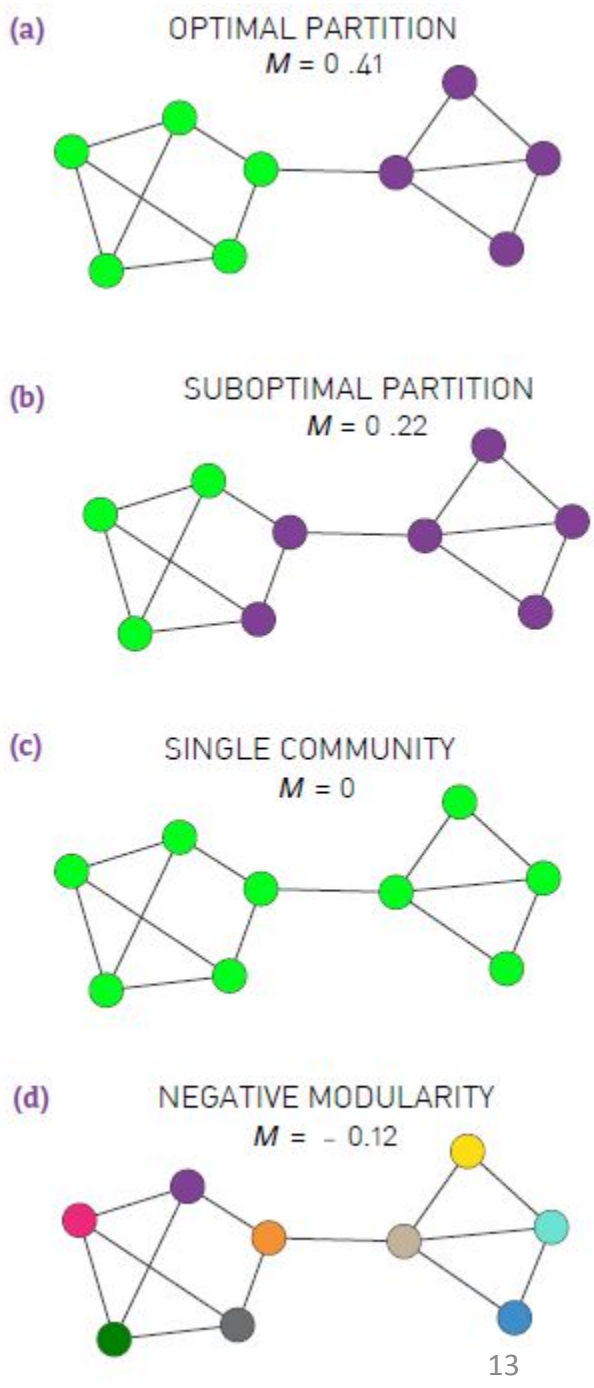
$$Q = \sum_r \left(\frac{L_r}{L} - \left(\frac{k_r}{2L} \right)^2 \right)$$

fracción enlaces entre nodos de cluster r

Propiedades de la modularidad

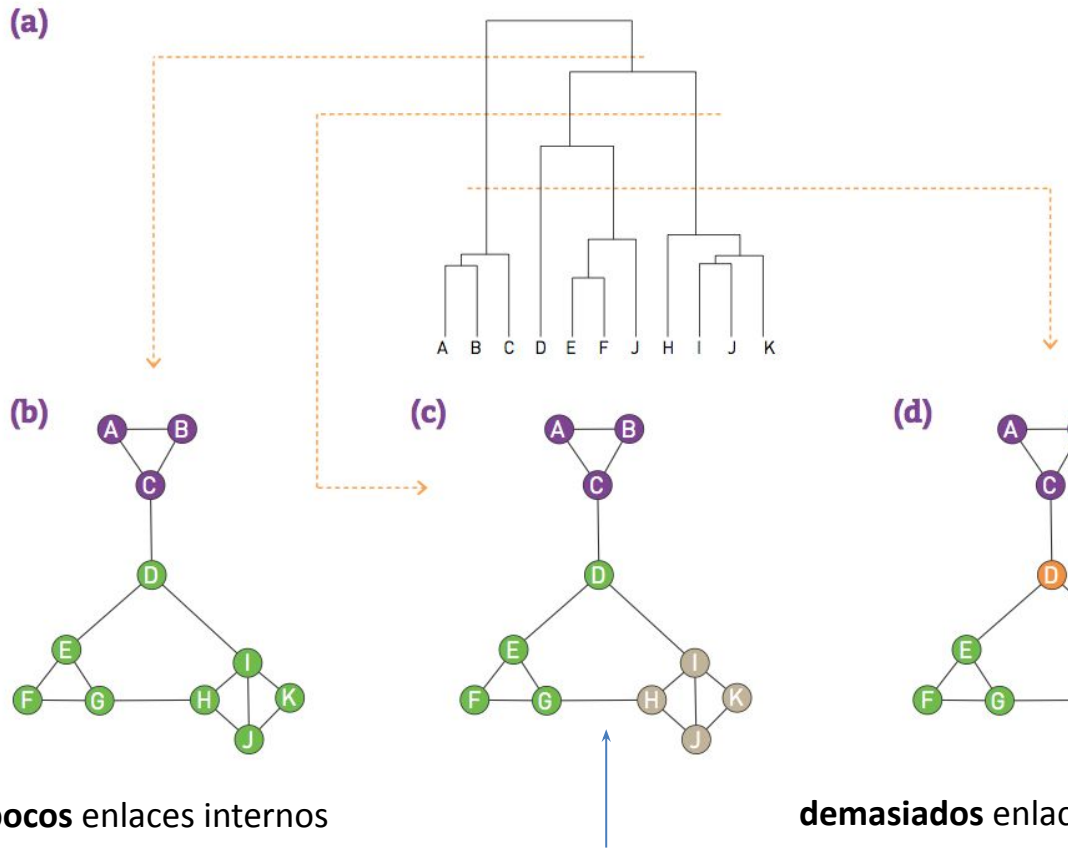
$$Q = \sum_r \left(\frac{L_r}{L} - \left(\frac{k_r}{2L} \right)^2 \right)$$

- Valores altos de modularidad implican mejor valor de asortatividad entre pertenencia a grupos y cableado de la red.
- Una partición de un único cluster tendrá Q=0 (los dos términos son idénticos)
- Si cada nodo pertenece a una comunidad distinta L_r=0 y Q<0



Clusters à la Newman-Girvan

Donde **cortar** el dendrograma para definir los clusters?



pocos enlaces internos

demasiados enlaces externos

Maxima modularidad

Como cuantificar el **acuerdo** entre **cableado** y **partición** en grupos?

fracción enlaces adyacentes a nodos de cluster r

$$Q = \sum_r \left(\frac{L_r}{L} - \left(\frac{k_r}{2L} \right)^2 \right)$$

fracción enlaces entre nodos de cluster r

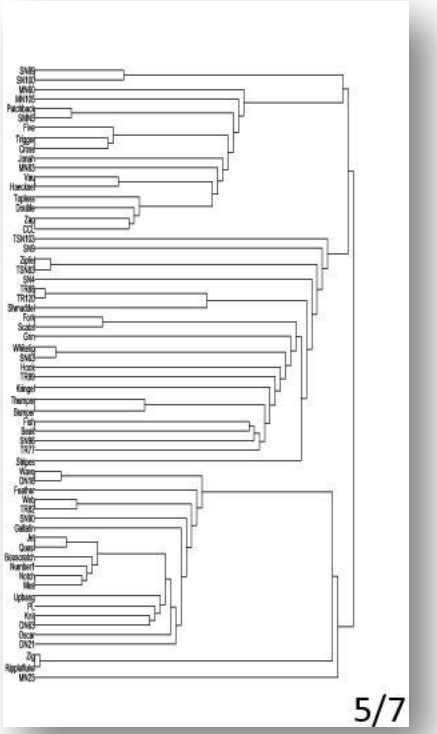
Detección de clusters maximizando Q

PHYSICAL REVIEW E 69, 066133 (2004)

Fast algorithm for detecting community structure in networks

M. E. J. Newman
Department of Physics and Center for the Study of Complex Systems, University of Michigan, Ann Arbor, Michigan 48109-1120, USA
(Received 22 September 2003; revised manuscript received 22 March 2004; published 18 June 2004)

Many networks display community structure—groups of vertices within which connections are dense but between which they are sparser—and sensitive computer algorithms have in recent years been developed for detecting this structure. These algorithms, however, are computationally demanding, which limits their application to small networks. Here we describe an algorithm which gives excellent results when tested on both computer-generated and real-world networks and is much faster, typically thousands of times faster, than previous algorithms. We give several example applications, including one to a collaboration network of more than 50 000 physicists.



Algoritmo codicioso (greedy algorithm) aglomerativo

1. Asignar cada nodo a su propia comunidad
2. Inspeccionar cada par de comunidades conectadas mediante al menos un enlace y estimar ΔQ producida si se combinasen
3. Unir el par de comunidades de mayor ΔQ
4. Guardar partición, P, y valor de modularidad Q respectiva
5. Repetir desde 2 hasta llegar a una única comunidad
6. Elegir la partición que se corresponda a Q máximo

Detección de clusters maximizando Q

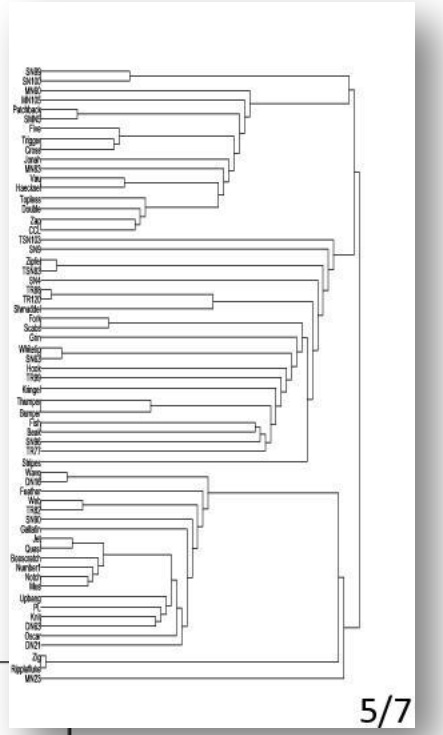
Algoritmo codicioso (greedy algorithm)

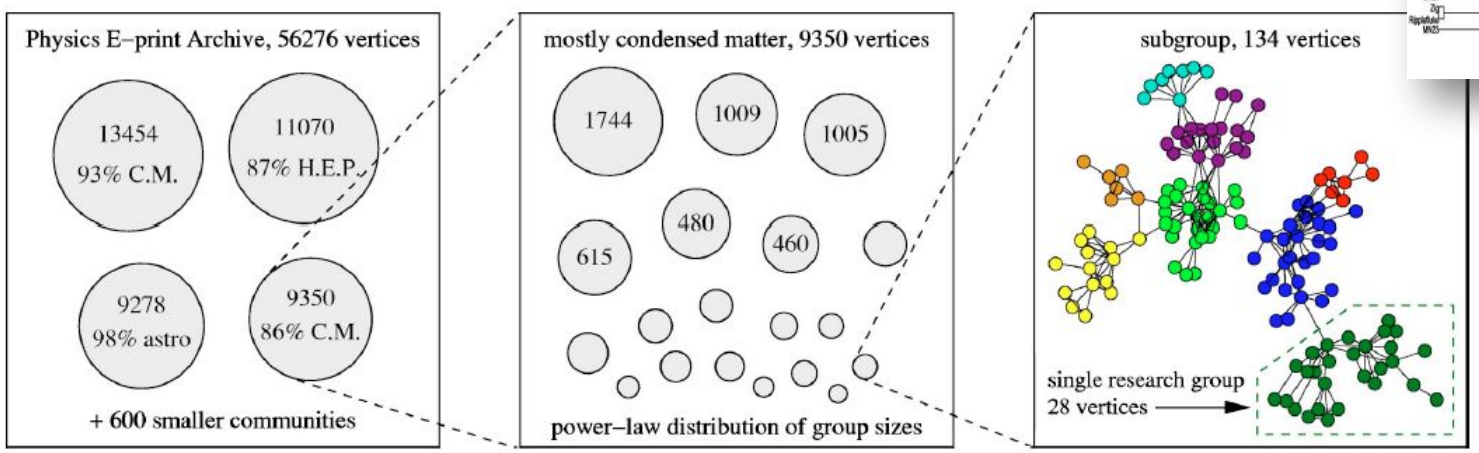
PHYSICAL REVIEW E 69, 066133 (2004)

Fast algorithm for detecting community structure in networks

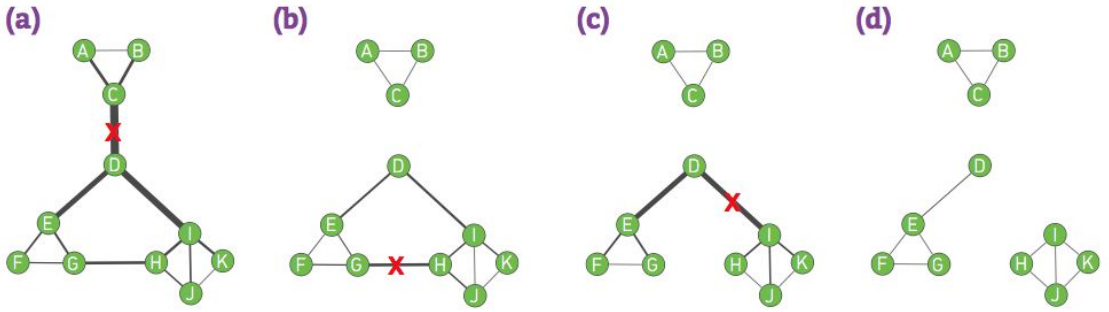
M. E. J. Newman
Department of Physics and Center for the Study of Complex Systems, University of Michigan, Ann Arbor, Michigan 48109-1120, USA
(Received 22 September 2003; revised manuscript received 22 March 2004; published 18 June 2004)

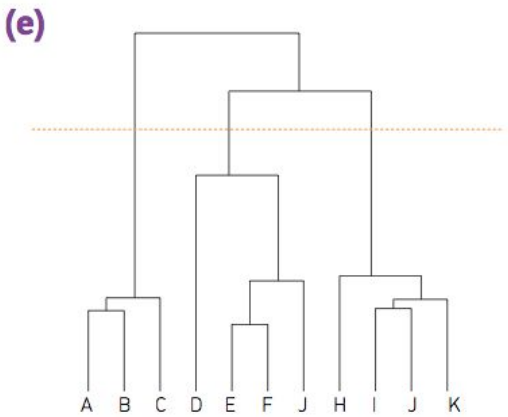
Many networks display community structure—groups of vertices within which connections are dense but between which they are sparser—and sensitive computer algorithms have in recent years been developed for detecting this structure. These algorithms, however, are computationally demanding, which limits their application to small networks. Here we describe an algorithm which gives excellent results when tested on both computer-generated and real-world networks and is much faster, typically thousands of times faster, than previous algorithms. We give several example applications, including one to a collaboration network of more than 50 000 physicists.

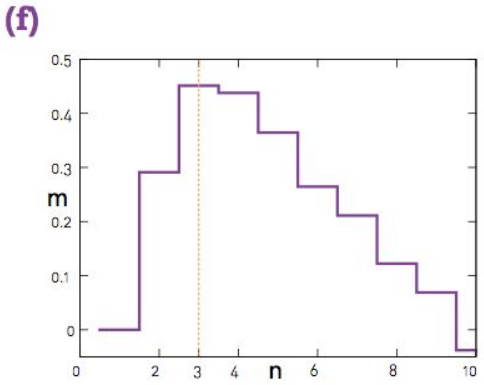




Entonces…cual partición?







$$Q = \sum_r \left(\frac{L_r}{L} - \left(\frac{k_r}{2L} \right)^2 \right)$$

Limitaciones de la Modularidad

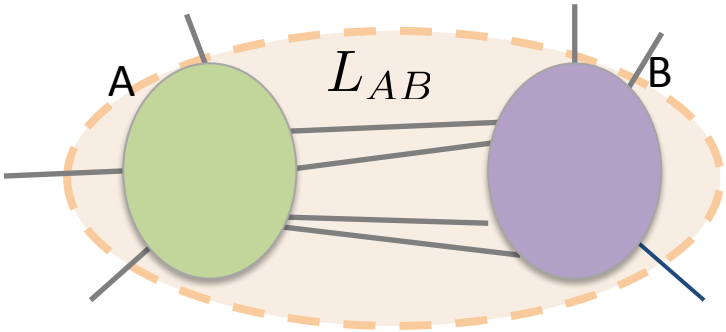
- Límite de resolución
- Máximo global, no muy diferente de máximos locales (*spin-glass landscape*)

Limitaciones de la Modularidad

- Límite de resolución
- Máximo global, no muy diferente de máximos locales (*spin-glass landscape*)

Supongamos que tenemos dos grupos y analizamos juntarlos

$$Q = \sum_r \left(\frac{L_r}{L} - \left(\frac{k_r}{2L} \right)^2 \right)$$



k_x : grado total de grupo X
L_A : nro enlaces en A
L_B : nro enlaces en B
L_{AB}: nro enlaces entre A y B

$$\Delta Q = \left[\frac{L_{AB} + L_A + L_B}{L} - \left(\frac{k_{AB}}{2L} \right)^2 \right] - \left[\frac{L_A}{L} - \left(\frac{k_A}{2L} \right)^2 + \frac{L_B}{L} - \left(\frac{k_B}{2L} \right)^2 \right]$$

$$k_{AB} = k_A + k_B$$

Q despues de mergear

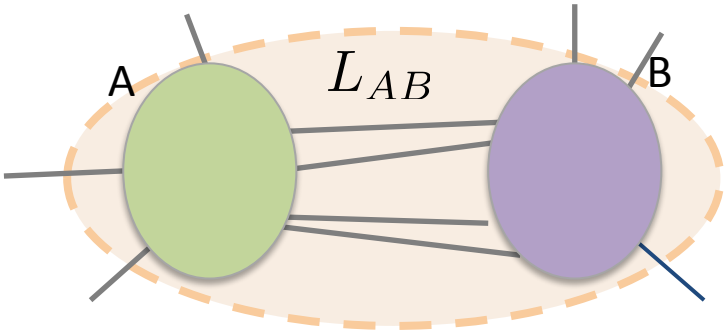
Q antes de mergear

Limitaciones de la Modularidad

- Límite de resolución
- Máximo global, no muy diferente de máximos locales (*spin-glass landscape*)

Supongamos que tenemos dos grupos y analizamos juntarlos

$$Q = \sum_r \left(\frac{L_r}{L} - \left(\frac{k_r}{2L} \right)^2 \right)$$



k_x : grado total de grupo X
L_A : nro enlaces en A
L_B : nro enlaces en B
L_{AB}: nro enlaces entre A y B

$$\Delta Q = \left[\frac{L_{AB} + L_A + L_B}{L} - \left(\frac{k_{AB}}{2L} \right)^2 \right] - \left[\frac{L_A}{L} - \left(\frac{k_A}{2L} \right)^2 + \frac{L_B}{L} - \left(\frac{k_B}{2L} \right)^2 \right] \qquad k_{AB} = k_A + k_B$$

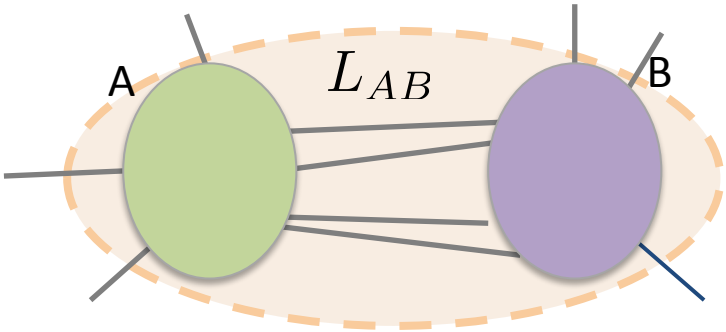
$$\Delta Q = \left[\frac{L_{AB}}{L} - \frac{k_A k_B}{2L^2} \right]$$

Si ΔQ>0 se deberia promover la fusion

Limitaciones de la Modularidad

- Límite de resolución
- Máximo global, no muy diferente de máximos locales (*spin-glass landscape*)

Supongamos que tenemos dos grupos y analizamos juntarlos $Q = \sum_r \left(\frac{L_r}{L} - \left(\frac{k_r}{2L} \right)^2 \right)$



k_x : grado total de grupo X
L_A : nro enlaces en A
L_B : nro enlaces en B
L_{AB}: nro enlaces entre A y B

(9.57) Barabasi $\Delta Q = \frac{L_{AB}}{L} - \frac{k_A k_B}{2L^2} = \frac{1}{L}\left(L_{AB} - \frac{k_A k_B}{2L} \right)$

Si $L > \frac{k_A k_B}{2} \Rightarrow \quad \Delta Q > 0$

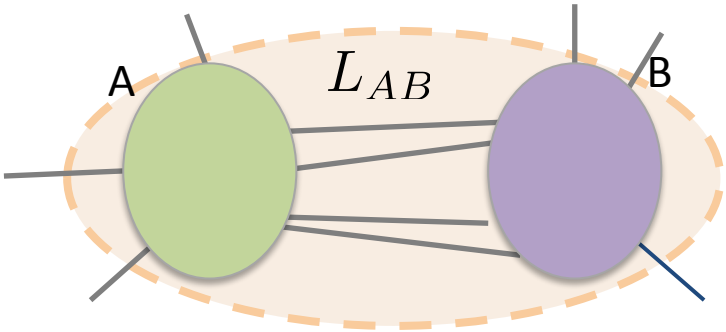
No sólo depende del cableado de A y B, sino del número de enlaces, L, de la red completa !

y maximizar Q impicará unir ambos clusters **SIEMPRE** !!! (aunque L_{AB}=1)

Limitaciones de la Modularidad

- Límite de resolución
- Máximo global, no muy diferente de máximos locales (*spin-glass landscape*)

Supongamos que tenemos dos grupos y analizamos juntarlos $Q = \sum_r \left(\frac{L_r}{L} - \left(\frac{k_r}{2L} \right)^2 \right)$



k_x : grado total de grupo X
L_A : nro enlaces en A
L_B : nro enlaces en B
L_{AB}: nro enlaces entre A y B

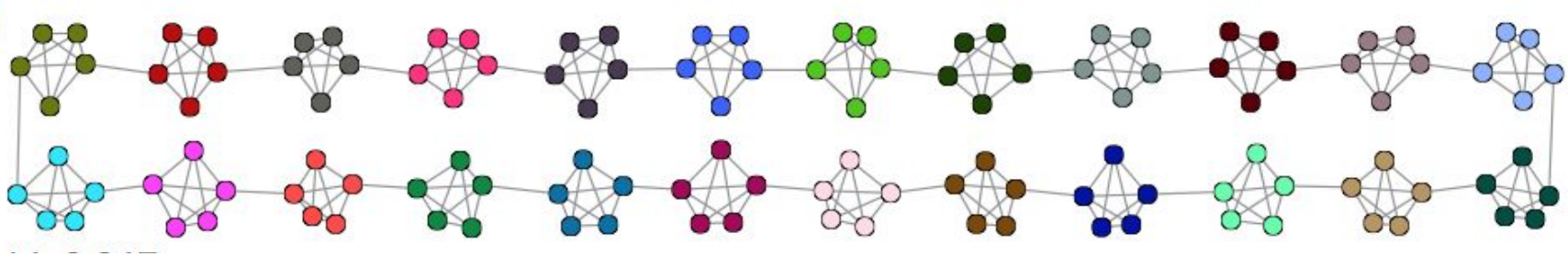
(9.57) Barabasi $\Delta Q = \frac{L_{AB}}{L} - \frac{k_A k_B}{2L^2} = \frac{1}{L}\left(L_{AB} - \frac{k_A k_B}{2L} \right)$ Si $L > \frac{k_A k_B}{2} \Rightarrow \Delta Q > 0$

Si asumimos $k_A \sim k_B \equiv \kappa \Rightarrow \kappa < \sqrt{2L}$

Algoritmos que maximicen Q **no podrán identificar** comunidades de ***tamaño*** menor a κ (!)
Se suele hacer zoom…

Limite de resolucion…ejemplo

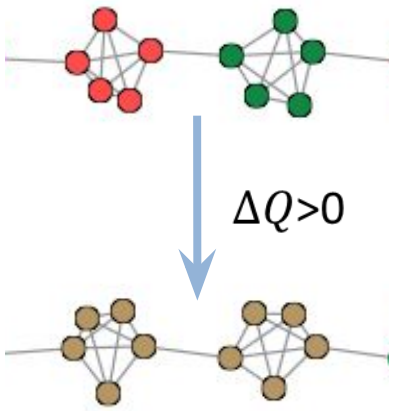
Partición ***natural*** para una red conformada por 24 5-cliques



Pero…en esta red, unir dos cliques es negocio!

$$\Delta Q = \frac{L_{AB}}{L} - \frac{k_A k_B}{2L^2}$$

$$= \frac{L_{AB}}{L} - \frac{1}{L}\frac{k_A k_B}{2L} = \frac{1}{L}\left(L_{AB} - \frac{k_A k_B}{2L}\right) > \frac{1}{L}\left(L_{AB} - 0.\widehat{75}\right)$$



$$L = 10 * 24 + 24 = 264$$

$$k_A = k_B = 20$$

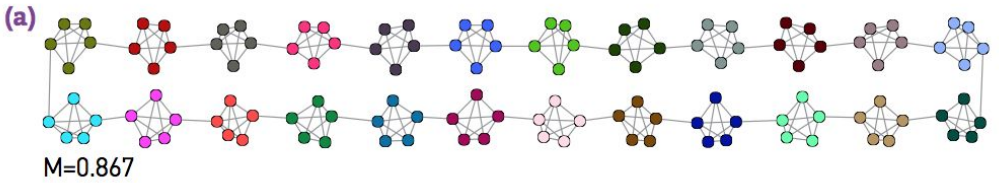
$$\frac{k_A k_B}{2L} = \frac{400}{528} < 1$$

$$\Delta Q > 0 \quad \text{si } L_{AB} \geq 1$$

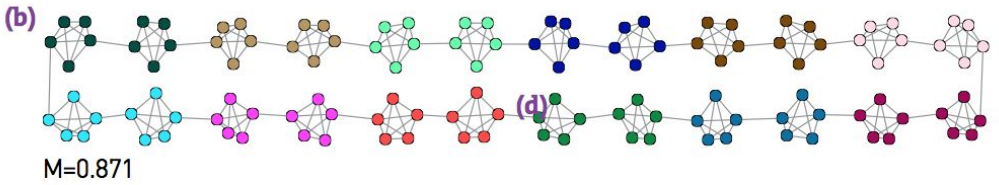
Limitaciones de la Modularidad

- Límite de resolución
- Máximo global, no muy diferente de máximos locales (*spin-glass landscape*)

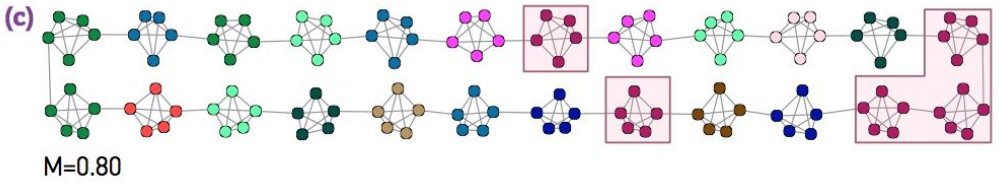
Partición ***natural***



Partición ***optima***

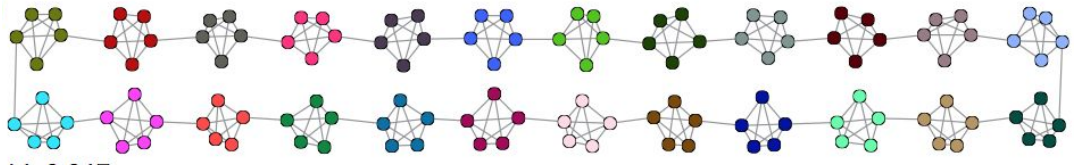


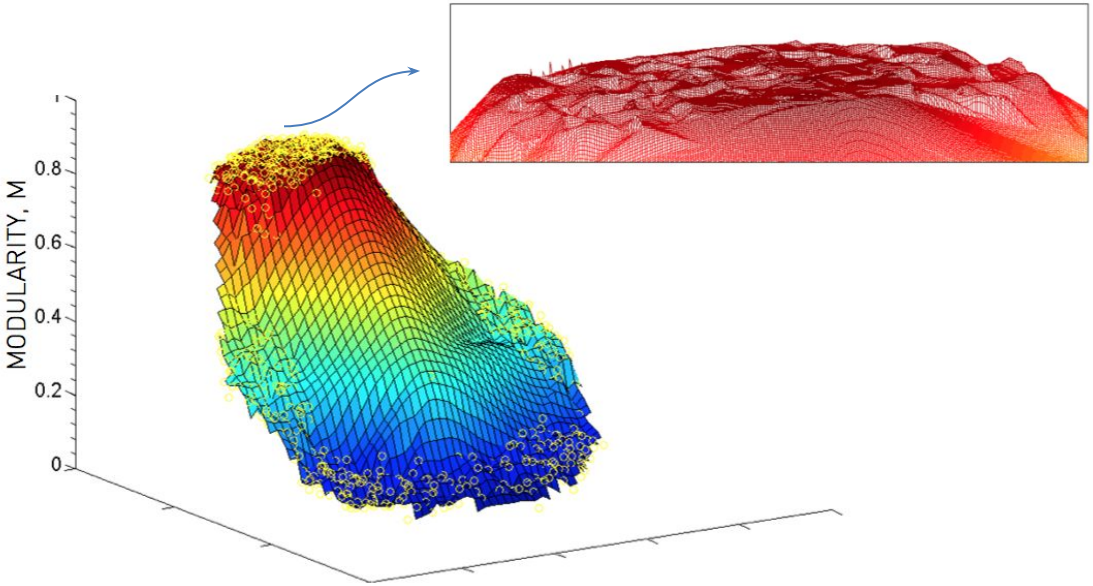
Partición ***aleatoria***



Limitaciones de la Modularidad

- Límite de resolución
- Máximo global, no muy diferente de máximos locales (*spin-glass landscape*)





Modularidad estimada para 997 particiones. No se observa un máximo claro para Q.
Particiones muy diferentes en su naturaleza, podrian presentar valores comparables de Q

Good 2010 PRE Performance of modularoty maximization

4 hipótesis para que funcione

- **Suponemos** que existe una estructura en comunidades embebida en la red, i.e. en A_{ij}

- Criterio de **Conectividad** y **densidad**. Una comunidad se corresponde con un subgrafo conexo localmente denso

- Redes aleatorias carecen de una estructura de comunidades

- La modularidad de una partición sobre una red permite identificar particiones óptimas en el sentido de mixing asortativo entre pertenencia a comunidades y conexionado (entender las limitaciones de esto!)