

# Estructura de Comunidades en redes

Sirius Fuenmayor Cardozo  
Grupo de Caos y Sistemas Complejos  
Facultad de Ciencias  
Universidad de los Andes



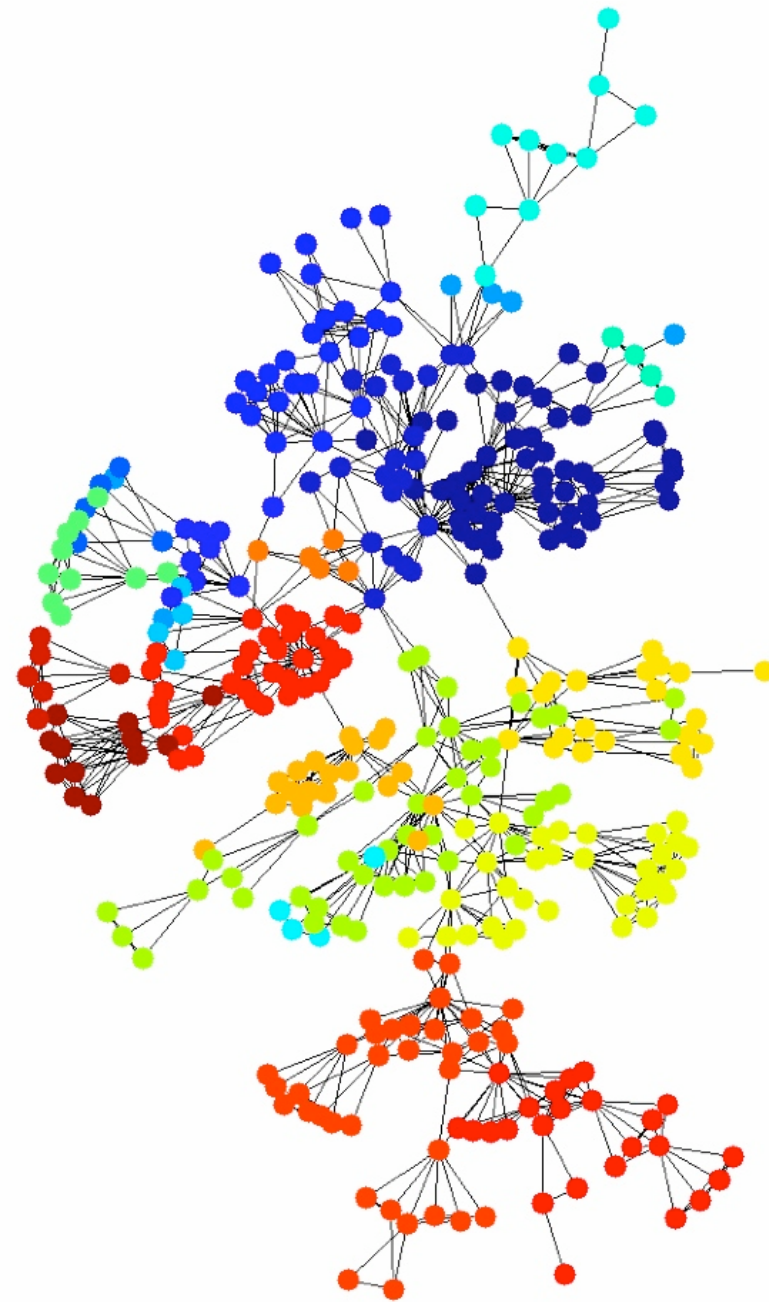
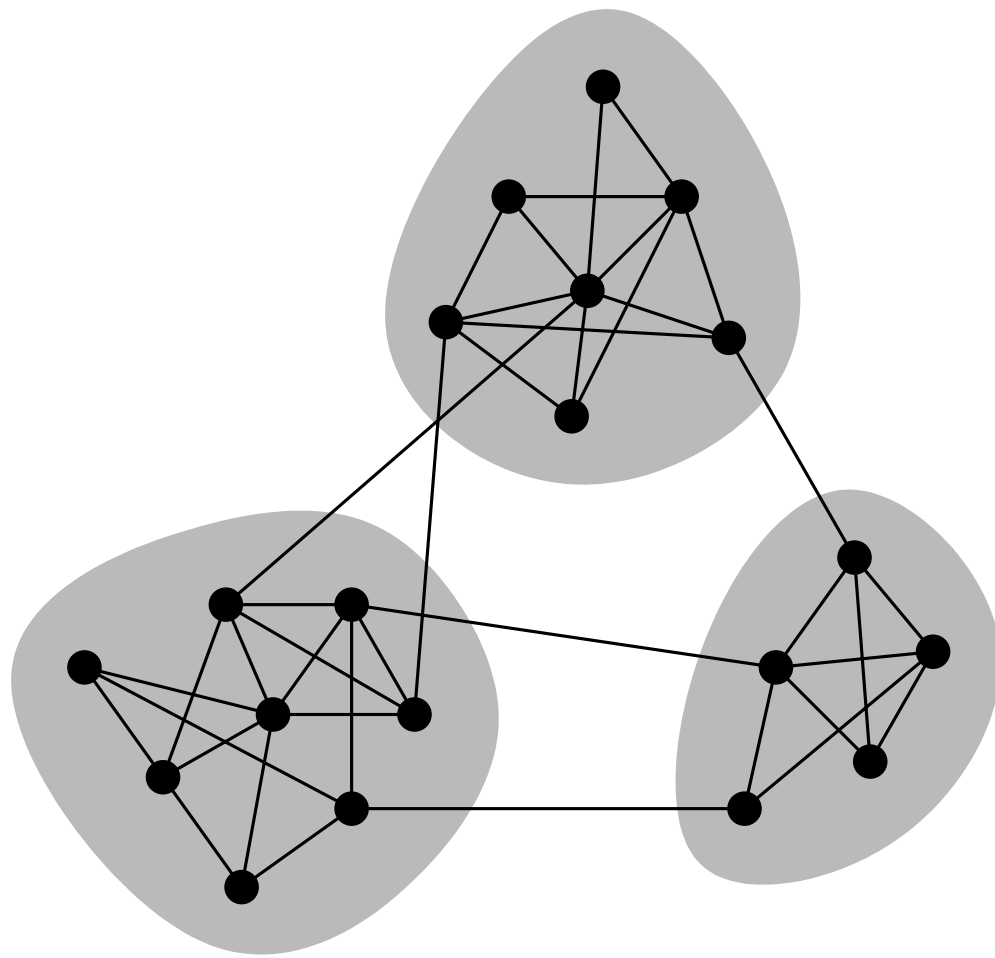
# Resumen

- ¿Que son las comunidades?
- Porque son importantes las comunidades
- Ejemplos de comunidades en algunos tipos de redes reales.
- Métodos de detección de comunidades
- Como se forman las comunidades

# Introducción

- Muchos sistemas se pueden representar como redes:  
nodos = partes, enlaces = interacciones
- Propiedades encontradas en muchas redes:
  - Conectividad de Pequeño Mundo
  - Topología libre de escala:  $P(k) = Ck^{-\gamma}$
  - Estructura de comunidades ✓

# Comunidades - Clusters - Módulos



# Porque estudiar la estructura de comunidades

- La propiedades estructurales de una red afectan su función (transferencia de datos).
- Las comunidades corresponden a divisiones informativas de los nodos de la red:  
*Un modulo en el sentido topológico es también un modulo funcional.*
- Una comunidad tiene propiedades diferentes al resto de la red.

# Ejemplos de estructura de comunidades en algunas redes

# Computación en Paralelo

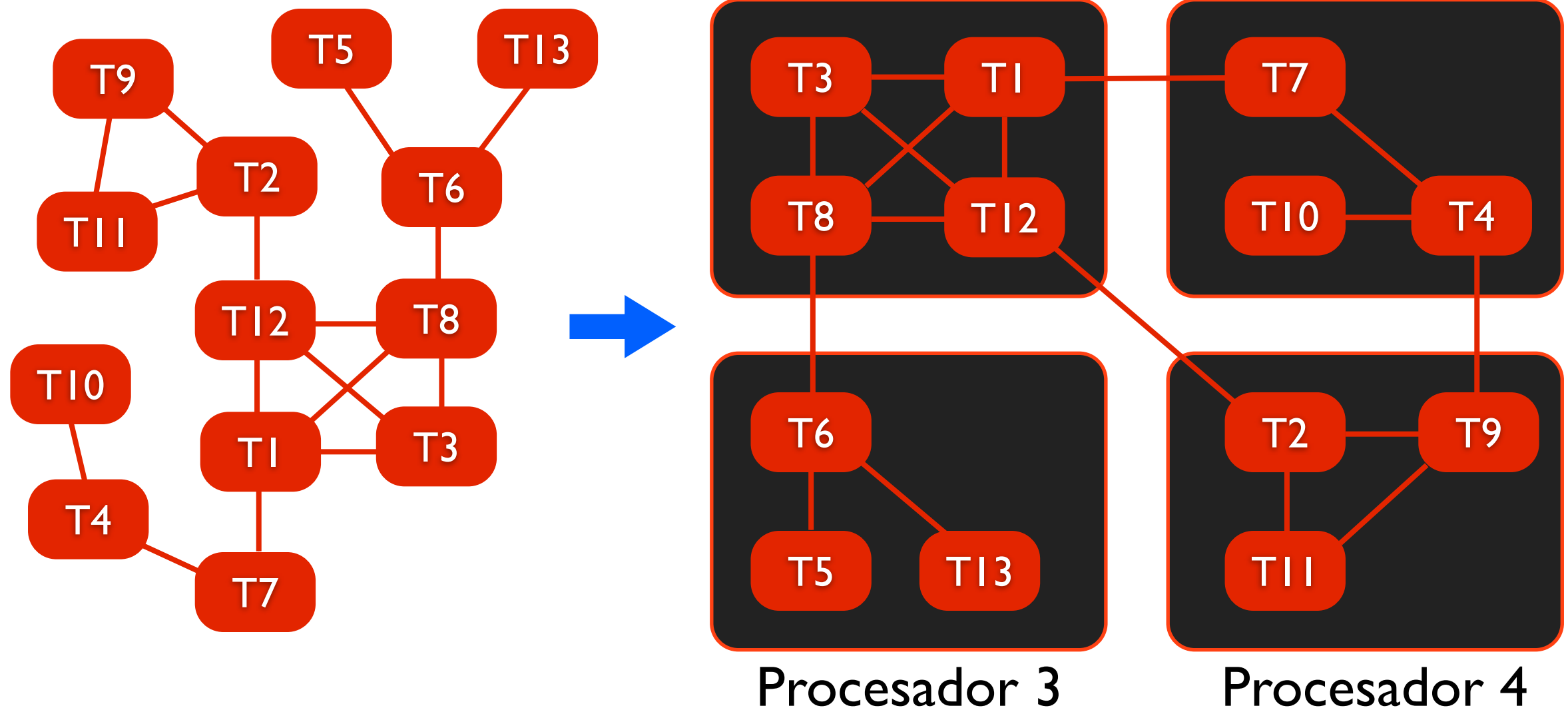
(Particionado de Grafos)

- Como se asignan tareas a un grupo de procesadores tal que:
  - Se balancee la carga entre procesadores
  - Se reduzca la comunicación interprocesador
- El problema fue resuelto por Kernighan y Lin\* en 1970.
- El número de procesadores y la cantidad de tareas que cada procesador puede manejar suele conocerse de antemano.

\* Kernighan, B.W.; Lin, Shen (1970). "An efficient heuristic procedure for partitioning graphs". Bell Systems Technical Journal 49: 291–307.

# Computación en Paralelo

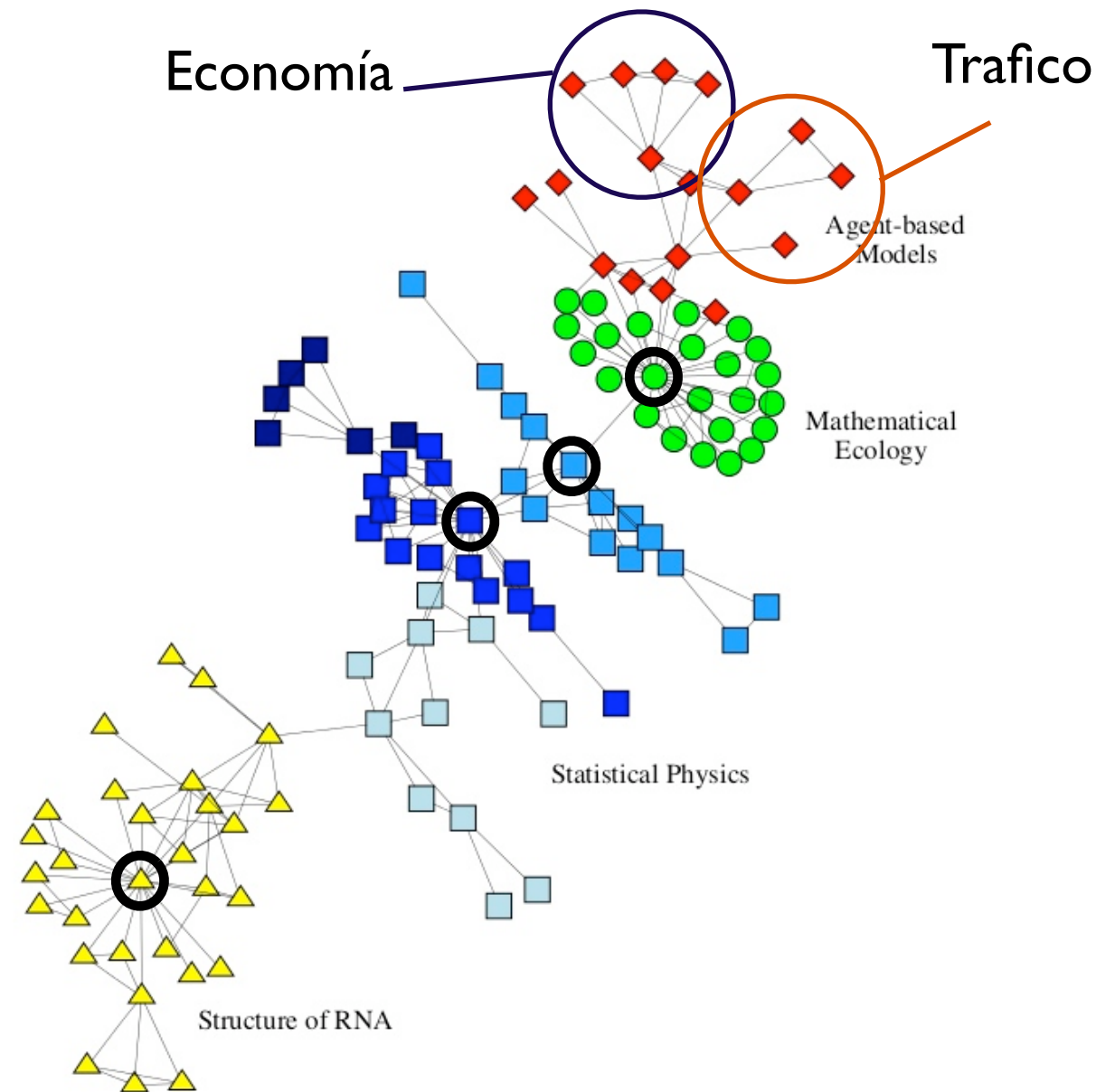
## (Particionado de Grafos)



Distribución mas eficiente de tareas en un conjunto de procesadores

# Redes de colaboración científica

- ◆ Modelos basados en agentes para estudiar problemas de economía y flujo de tráfico
  - Modelos matemáticos en ecología
  - Física estadística
  - ▲ Estructura del ARN
- ◎ Formación de comunidades en torno a la metodología
  - ◎ El centro de las comunidades corresponde al jefe del grupo de investigación.



Red de colaboración de científicos del Instituto de Santa Fe en Nuevo México

# World Wide Web

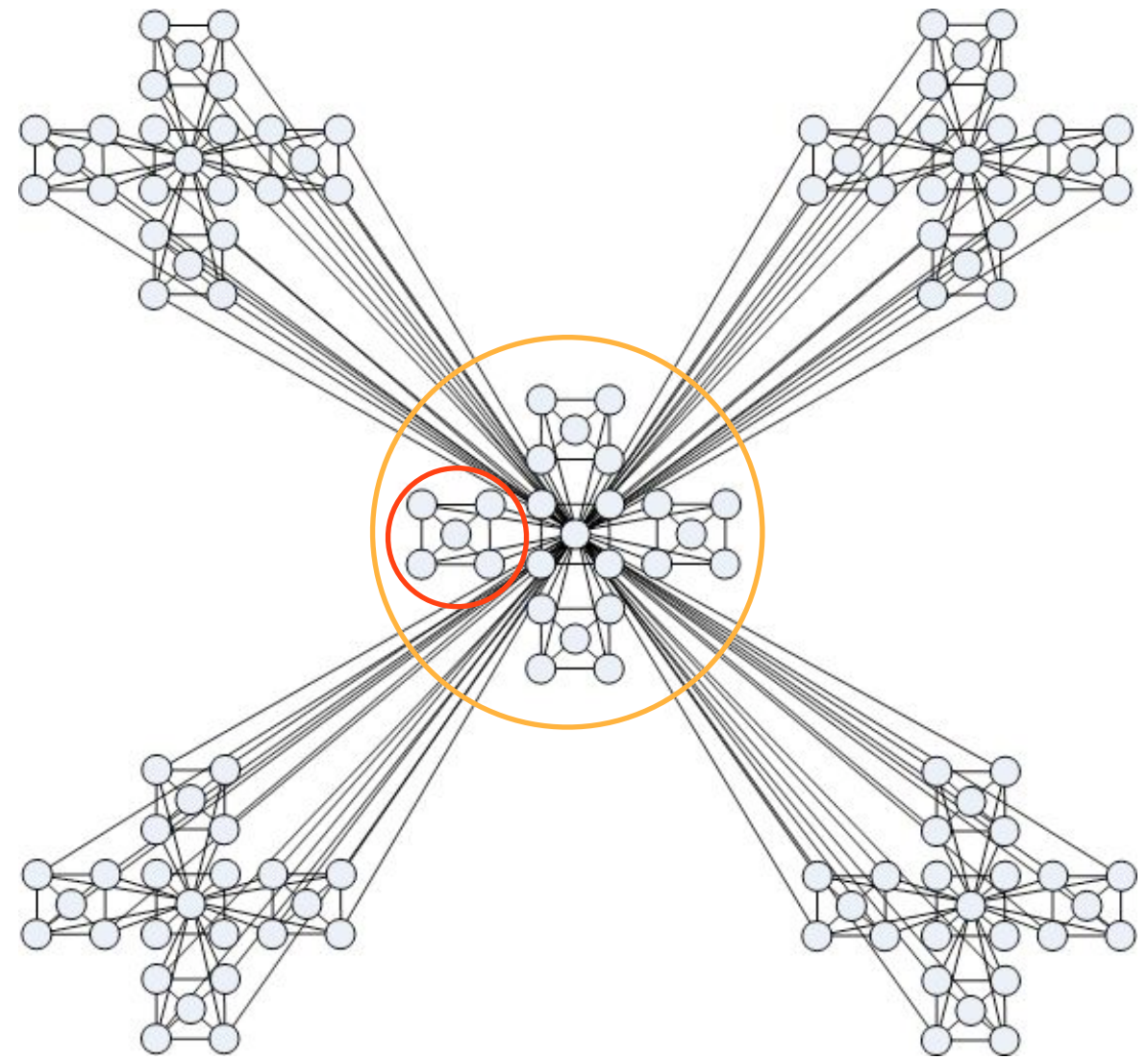
- La WWW se auto-organiza y muestra estructura de comunidades.
- Tal estructura esta basada en enlaces creados explícitamente por los autores y no por coincidencia de contenidos.
- Comunidades formadas por enlaces explícitos también están relacionadas por tópico.

| Francis Crick Community |                                                                           | Stephen Hawking Community |                                                                            | Ronald Rivest Community |                                                                     |
|-------------------------|---------------------------------------------------------------------------|---------------------------|----------------------------------------------------------------------------|-------------------------|---------------------------------------------------------------------|
| Score                   | Site title or description                                                 | Score                     | Site title or description                                                  | Score                   | Site title or description                                           |
| 80                      | Biography of Francis Harry Compton Crick (Nobel Foundation)               | 85                        | Professor Stephen W. Hawking's Web pages                                   | 86                      | Ronald L. Rivest home page                                          |
| 79                      | Biography of James Dewey Watson (Nobel Foundation)                        | 46                        | <i>Stephen Hawking's Universe</i> (PBS)                                    | 29                      | "Chaffing and Winnowing: Confidentiality without Encryption"        |
| 51                      | The Nobel Prize in Physiology or Medicine 1962 (Nobel Foundation)         | 17                        | The Stephen Hawking pages                                                  | 20                      | Thomas H. Cormen's home page at Dartmouth                           |
| 50                      | "Biographical Sketch of James Dewey Watson" (Cold Spring Harbor Lab.)     | 15                        | "Stephen Hawking Builds Robotic Exoskeleton" (parody in <i>The Onion</i> ) | 9                       | "The Mathematical Guts of RSA Encryption"                           |
| 41                      | "A Structure for Deoxyribose Nucleic Acid" ( <i>Nature</i> , 2 Apr. 1953) | 14                        | Stephen Hawking and Intel                                                  | 8                       | German news story on Cryptography                                   |
| ...                     |                                                                           | ...                       |                                                                            | ...                     |                                                                     |
| 1                       | <i>Felix D'Herelle and the Origins of Molecular Biology</i> (Amazon.com)  | 1                         | "Did the Cosmos Arise from Nothing?" (MSNBC)                               | 1                       | Phil Zimmermann's PGP Web page                                      |
| 1                       | Biography of Gregor Mendel                                                | 1                         | Spanish page for <i>Stephen Hawking's Universe</i>                         | 1                       | "A Very Brief History of Computer Science"                          |
| 1                       | Magazine: <i>HMS Beagle Home</i>                                          | 1                         | Relativity Group at DAMTP, Cambridge                                       | 1                       | Cormen/Leiserson/Rivest: Introduction to Algorithms                 |
| 1                       | The Alfred Russel Wallace Page                                            | 1                         | Millennium Mathematics Project                                             | 1                       | Security and encryption links                                       |
| 1                       | US Human Genome Project 5 Year Plan                                       | 1                         | Particle physics education and information sites                           | 1                       | HotBot Directory: Computers & Internet, Computer Science, People: R |

Cinco primeras y cinco últimas páginas de las comunidades asociadas a las paginas web de tres científicos. La puntuación indica el número total de enlaces entrantes y salientes que una página Web tiene con otras páginas de su comunidad

# Redes Metabólicas

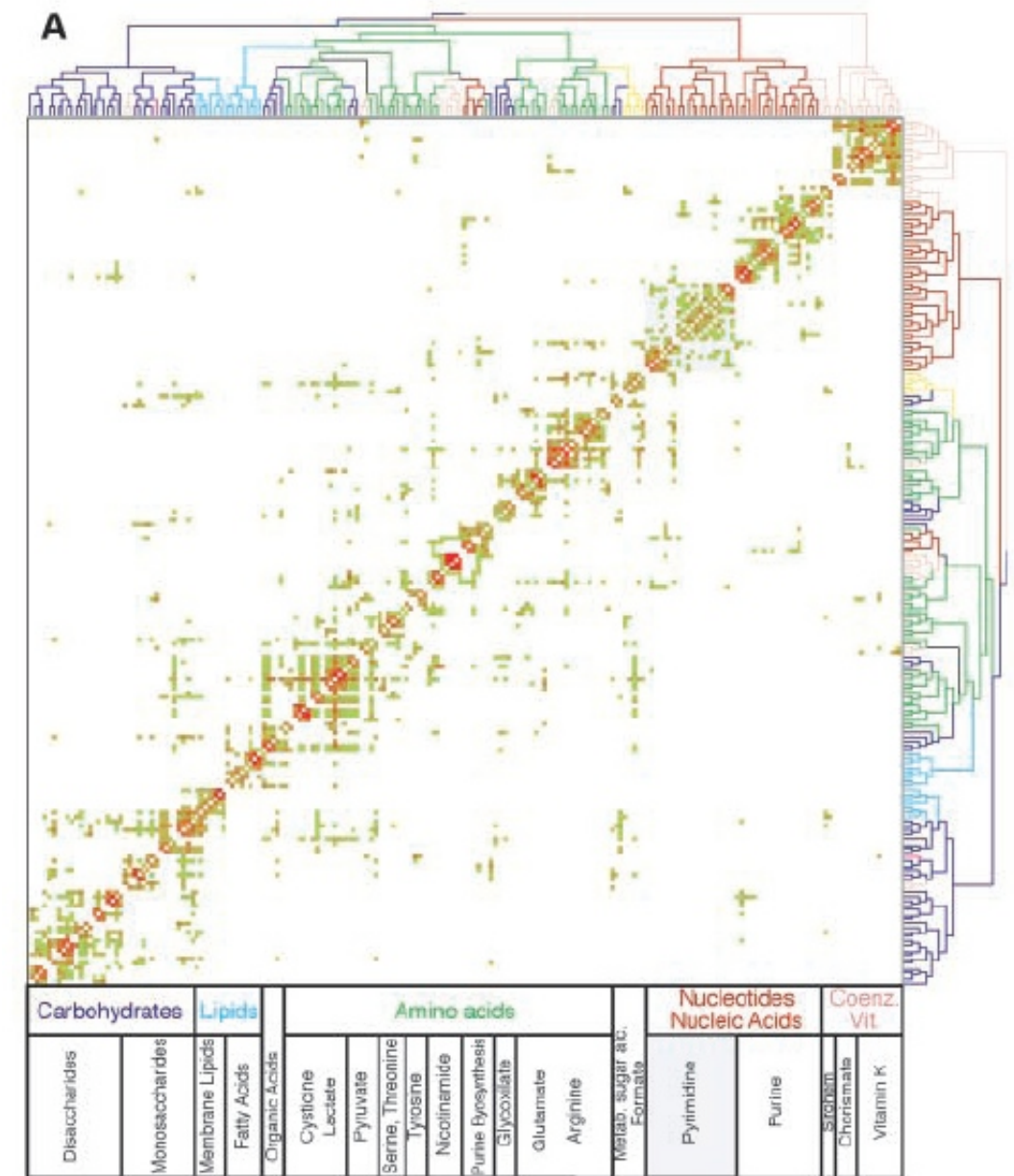
- La modularidad en las redes metabólicas posee una topología especial llamada “modularidad jerárquica”
- Pequeños módulos con alto clustering se unen para formar módulos mayores.



Red con modularidad jerárquica

# Redes Metabólicas

- ¿La modularidad topologica esta relacionada con la modularidad funcional en las redes metabólicas?
- “En E. coli, la modularidad jerárquica descubierta coincide estrechamente con las funciones metabólicas conocidas.”  
Esto permite inferir la funcionalidad de un grupo de metabolitos en base a el modulo al que pertenecen.



Matriz de solapamiento topológico de la red metabólica de E. Colí

# Algoritmos de detección

# Algoritmos de detección

- Edge Betweenness Method, M. Girvan and M. E. Newman (GN)
- Fast greedy modularity optimization, A. Clauset, M. E. Newman, and C. Moore (Clauset et al.)
- Exhaustive modularity optimization via simulated annealing, R. Guimerá, M. Sales-Pardo (Sim ann.)
- Multi-Level Aggregation Method based on modularity, V. D. Blondel, J.-L. Guillaume, R. Lambiotte (Blondel et al.)
- Divisive algorithm based on the edge-clustering coefficient, F. Radicchi, C. Castellano, F. Cecconi (Radicchi et al.)
- Clique Percolation Method for finding communities, G. Palla, I. Derenyi, I. Farkas (Cfinder)
- Graph clustering by flow simulation, S. van Dongen (MCL)

# Algoritmos de detección

- Optimal compression of network topology information, M. Rosvall and C.T. Bergstrom (Infomod)
- Compression of the information of random walks on the network, M. Rosvall and C.T. Bergstrom (Infomap)
- Detection based on the spectral properties of the graph, L. Donetti and M.A. Muñoz (DM)
- Node classification based on similar patterns of connection, M. E. J. Newman and E.A. Leicht (EM)
- Minimization of the Hamiltonian of a Potts-like spin model representing node membership, P. Ronhovde and Z. Nussinov (RN)

El resultado de los algoritmos es una ‘partición’ de la red.

# ¿Cual es el mejor algoritmo?

Para encontrar el mejor algoritmo se deben realizar pruebas sobre una red modelo.

Como se construye la red modelo:

- Distribución de grado dada por  $p(k) = k^{-\gamma_1}$
- Se asigna a cada nodo un grupo de nodos.
- La relación entre la conectividad de un nodo con su comunidad y el resto de la red se controla con un parámetro  $\mu$ , donde

$$\mu = \frac{\text{total de conexiones de } i}{\text{conexiones de } i \text{ con su comunidad}}$$

# ¿Cual es el mejor algoritmo?

## Como se construye la red modelo:

Se puede demostrar que aparecen comunidades cuando:

$$\mu = \frac{N - n_c^{max}}{N}$$

Donde  $N$  es el tamaño de la red y  $n_c^{max}$  el de la comunidad mas grande.

$\mu$  controla la definición de las comunidades y permite conocer la estructura de comunidades a priori.

# ¿Cual es el mejor algoritmo?

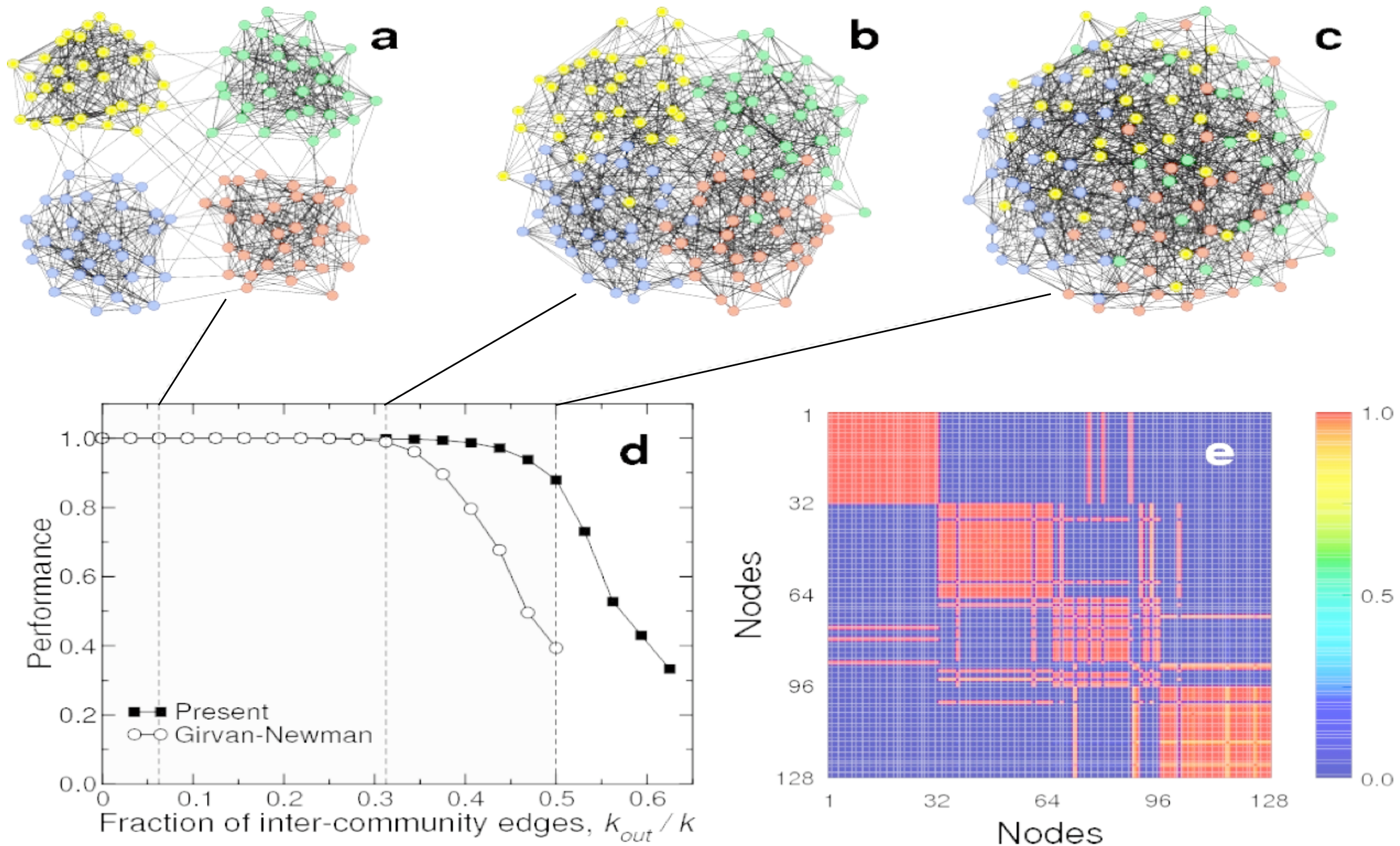
El algoritmo se evalúa comparando la partición detectada con la conocida a priori.

Si  $X$  es la partición detectada y  $Y$  la partición conocida a priori la comparación se hace por medio la *información mutua normalizada* de  $X$  y  $Y$ :

$$I_{norm} = \frac{2 (H(X) + H(X|Y))}{H(X) + H(Y)}$$

Donde  $H(X) = - \sum_x P(X) \log P(X)$   
 $I_{norm}(X, Y)$  es igual a 1 si las particiones son idénticas  
e igual a 0 si son independientes.

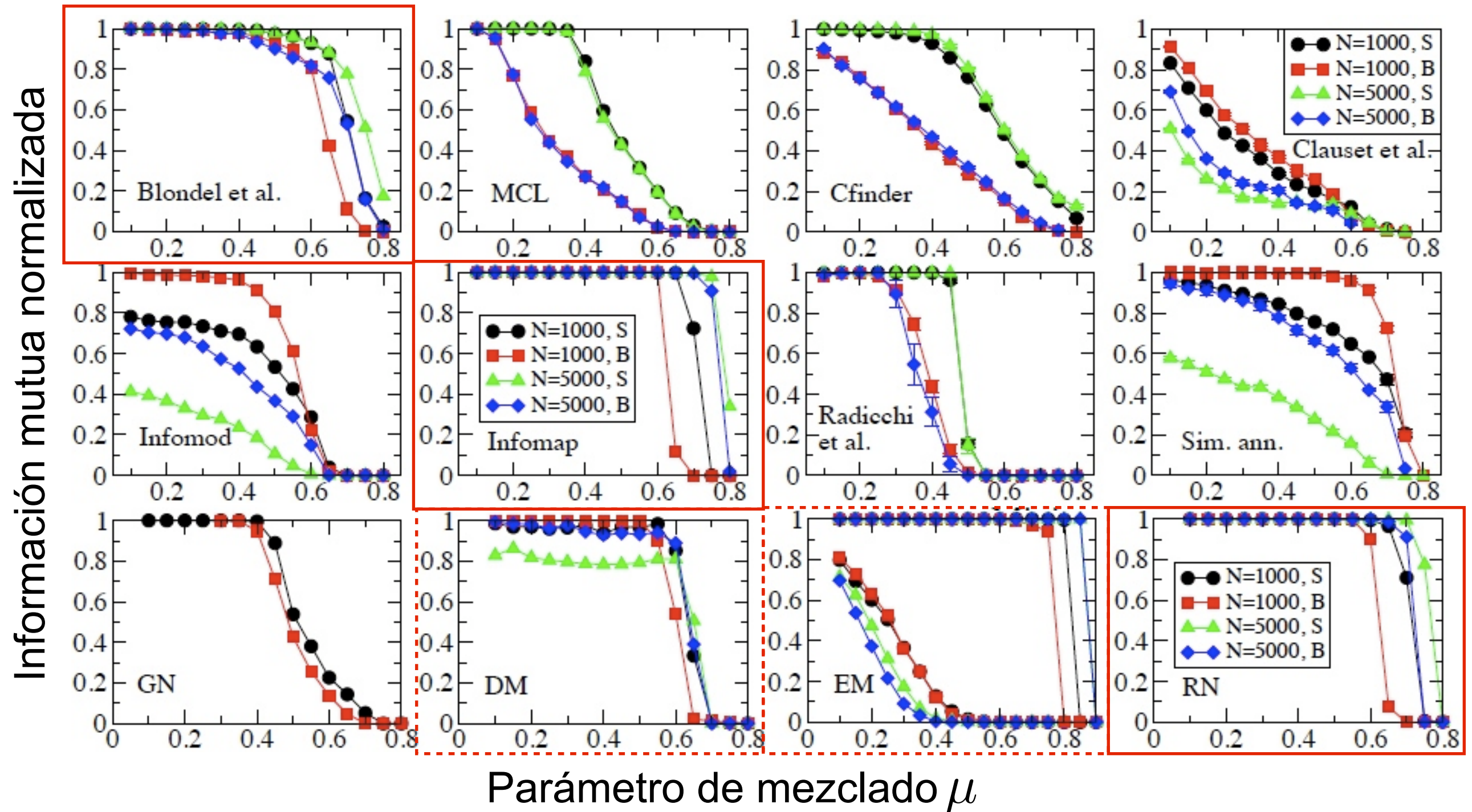
# ¿Cual es el mejor algoritmo?



S, Fortunato. "Community structure in graphs." Institute for Scientific Interchange Foundation. September 2008.

# ¿Cual es el mejor algoritmo?

Resultados de calcular  $I_{norm}$  vs  $\mu$  para los diferentes algoritmos:



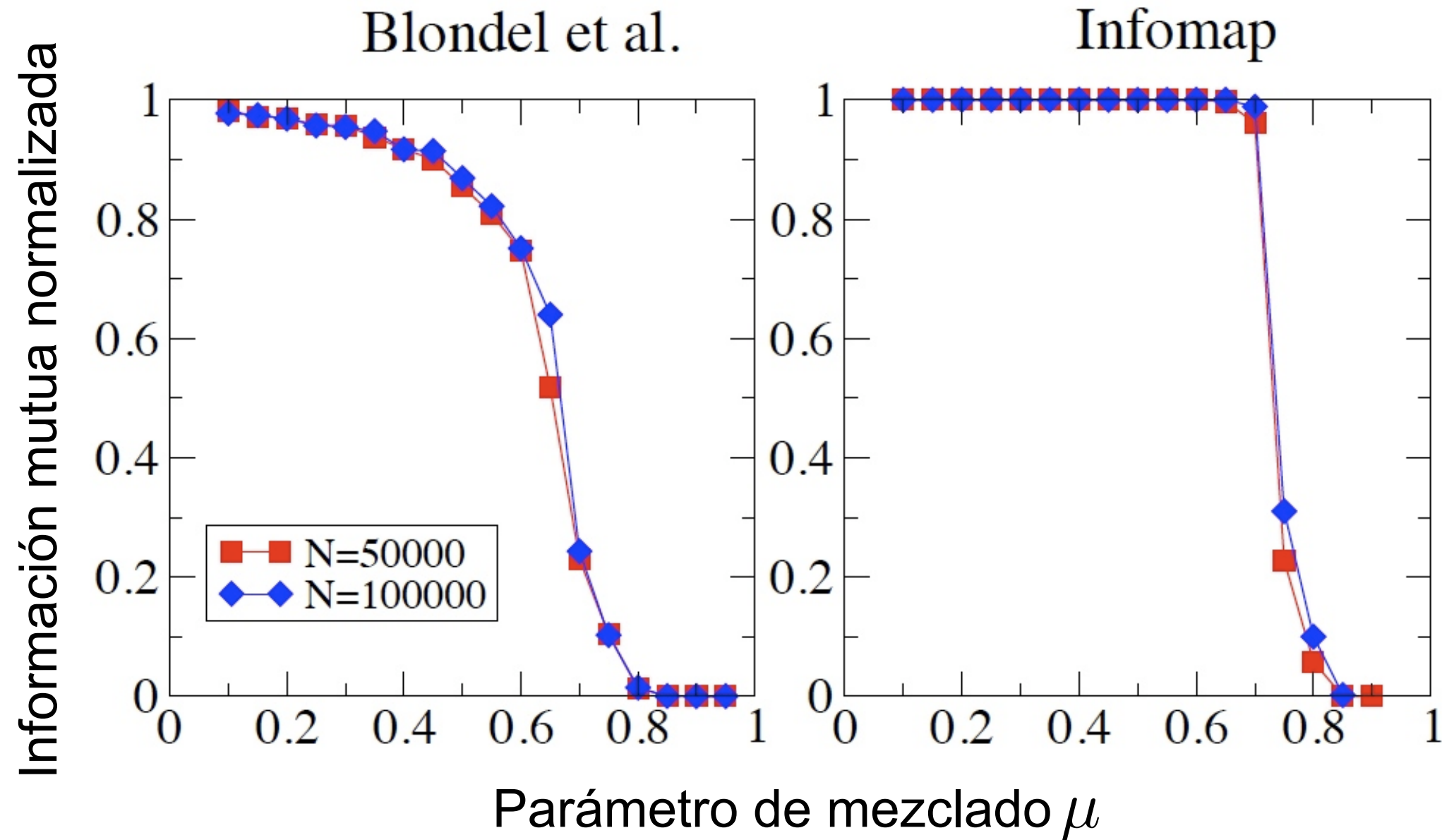
# ¿Cual es el mejor algoritmo?

Conclusiones (para redes no dirigidas ni pesadas):

- A medida que  $\mu$  crece las comunidades se hacen menos definidas y por tanto es posible apreciar la sensibilidad del algoritmo de detección.
- Métodos con mejor desempeño:
  - ➔ ***Infomap***
  - ➔ ***RN***
  - ➔ ***Blondel et al.***
- *Infomap* y *Blondel et al.* son rápidos: su tiempo de ejecución es lineal con el tamaño de la red.

# ¿Cual es el mejor algoritmo?

## *Blondel et al. vs Infomap*



# ¿Cual es el mejor algoritmo?

Conclusiones (para redes en general sin overlap):

- El metodo con el mejor desempeño para redes en general es el *Infomap*:

*"We conclude that the Infomap method by Rosvall and Bergstrom is the best performing on the set of benchmarks we have examined here."* <sup>†</sup>

- Los algoritmos por *Blondel et al.* y por *Ronhovde y Nussinov (RN)* son también eficaces y deberían utilizarse junto con infomap para obtener información independiente del algoritmo.

<sup>†</sup> A. Lancichinetti, S. Fortunato. Community detection algorithms: a comparative analysis. Physical Review E 80, 056117 (2009)

# ¿Cual es el mejor algoritmo?

## Observaciones:

- El único de los métodos que puede estudiar comunidades con *overlap* es *Cfinder*.
- Contras del marco LFR:
  - Coeficiente de agrupamiento bajo: por lo que no se han probado los algoritmos cuando el clustering es alto, propiedad que se observa en las redes reales.
- Tampoco se ha comprobado el desempeño de los métodos en redes con *modularidad jerarquica*.

† A. Lancichinetti, S. Fortunato. Community detection algorithms: a comparative analysis. Physical Review E 80, 056117 (2009)

# Detección basada en la Modularidad

- Si una red existen grupos de nodos con una densidad de enlaces pero esta densidad es la que se espera por azar no se puede concluir que exista una estructura de comunidades.
- Esta idea, de que la estructura de comunidades en una red corresponde a una disposición estadísticamente fuera de lugar de enlaces, se puede cuantificar mediante la *Modularidad*.
- Algunos métodos se aprovechan de esta medida para detectar las comunidades.

# Detección basada en la Modularidad

Como se define la modularidad:

La suma se hace para todos los pares de nodos  $i$  y  $j$

Este termino es positivo solo si los nodos  $i$  y  $j$  están conectados

La contribución de los nodos  $i$  y  $j$  sera diferente de 0 si pertenecen a la misma comunidad.

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

Donde:  $m$  es el numero de enlaces,  $c_i$  es la comunidad a la que pertenece el nodo  $i$ ,  $A_{ij}$  es el valor de la matriz de adyacencia y  $\delta(c_i, c_j)$  es la delta de Kronecker.

# Detección basada en la Modularidad

La modularidad también se puede escribir como:

$$Q = \sum_i^c (e_{ii} - a_i^2)$$

$$\text{Con } e_{ii} = \sum_{ij} \frac{A_{ij}}{2m} \delta(c_i, c_j) \text{ y } a_i = \frac{k_i}{2m}$$

$e_{ii}$  representa el numero de enlaces dentro de la comunidad y  $a_i$  el numero de enlaces que parte fuera de  $c$ .  $Q$  puede tomar valores en en el rango  $[-1, 1]$ . Valores positivos indican la presencia de comunidades.

# Detección basada en la Modularidad

- Supongamos que tenemos una red con estructura de comunidades pero no sabemos cuantas comunidades hay ni cuales nodos pertenecen a cada comunidad.
- Calculando la modularidad para todas las particiones posibles podemos encontrar la estructura real de comunidades en la red; es aquella de la partición con la mayor modularidad.

# Detección basada en la Modularidad

El algoritmo para la detección basado en esta idea seria:

- Para un numero de módulos desde 1 hasta el numero de nodos.
  - Tomar todas las posibles divisiones de los nodos en esos módulos.
  - Calcular la modularidad en cada caso.
- Finalmente seleccionar la modularidad máxima y la división de nodos asociada.

# Detección basada en la Modularidad

Problemas con los métodos basados en la optimizar la modularidad:

- Calcular la modularidad para todas las posibles particiones de una red resulta imposible:

$$\# \text{ de particiones} = \sum_{k=1}^n (n! / k!(n-k)!) = 2^n$$

- Para una red con 100 nodos:  $2^{100} = 1.27 \times 10^{30}$  particiones posibles!
- Debemos hacer alguna suposición para la búsqueda inteligente (heurística) de las particiones con modularidad máxima.

# Detección basada en la Modularidad

Heurísticas utilizadas para la optimización de la modularidad:

- Recocido Simulado (Guimera and Amaral)
- Optimización extrema (J. Duch and A. Arenas)
- Algoritmos voraces (Clauset et al.)
- Reformulación de la modularidad en términos de las propiedades espectrales de la red. (Newman)

Las dos ultimas heurísticas han resultado efectivas sin embargo poseen un problema inherente al concepto de modularidad llamado *limite de resolución*.

Modularity and community structure in networks, M. E. J. Newman, Proc. Natl. Acad. Sci. USA 103, 8577–8582 (2006).

Wikipedia, 2011. Modularity (networks). [en línea] Disponible en: <[http://en.wikipedia.org/wiki/Modularity\\_\(networks\)](http://en.wikipedia.org/wiki/Modularity_(networks))> [Consultado el 17 Noviembre 2011].

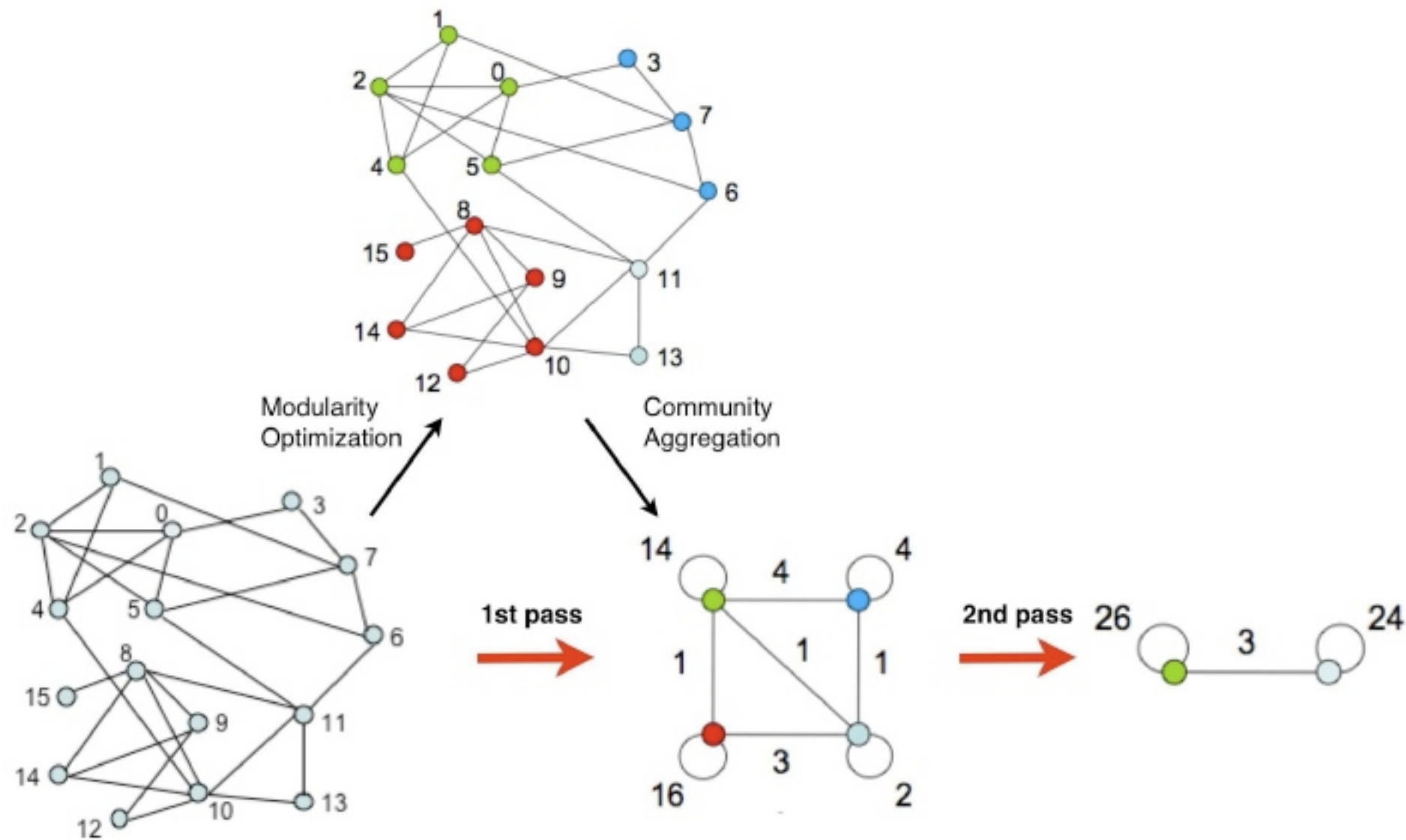
# Detección basada en la Modularidad

El método de Agregación multi-nivel de Blondel et al:

- En el inicio cada nodo de la red es una comunidad.
- Luego los nodos se fusionan en comunidades mas grandes con el fin de maximizar la modularidad, estas nuevas comunidades se convierten en súper nodos y el proceso anterior se repite.
- Este proceso se aplica repetidamente y de forma secuencial para todos los nodos hasta que no se pueda obtener una mejora.

# Detección basada en la Modularidad

El método de Agregación de niveles múltiples de Blondel et al:



Fast unfolding of community hierarchies in large networks, V.D. Blondel, J.-L. Guillaume, R. Lambiotte and E. Lefebvre, J. Stat. Mech., (2008) P10008

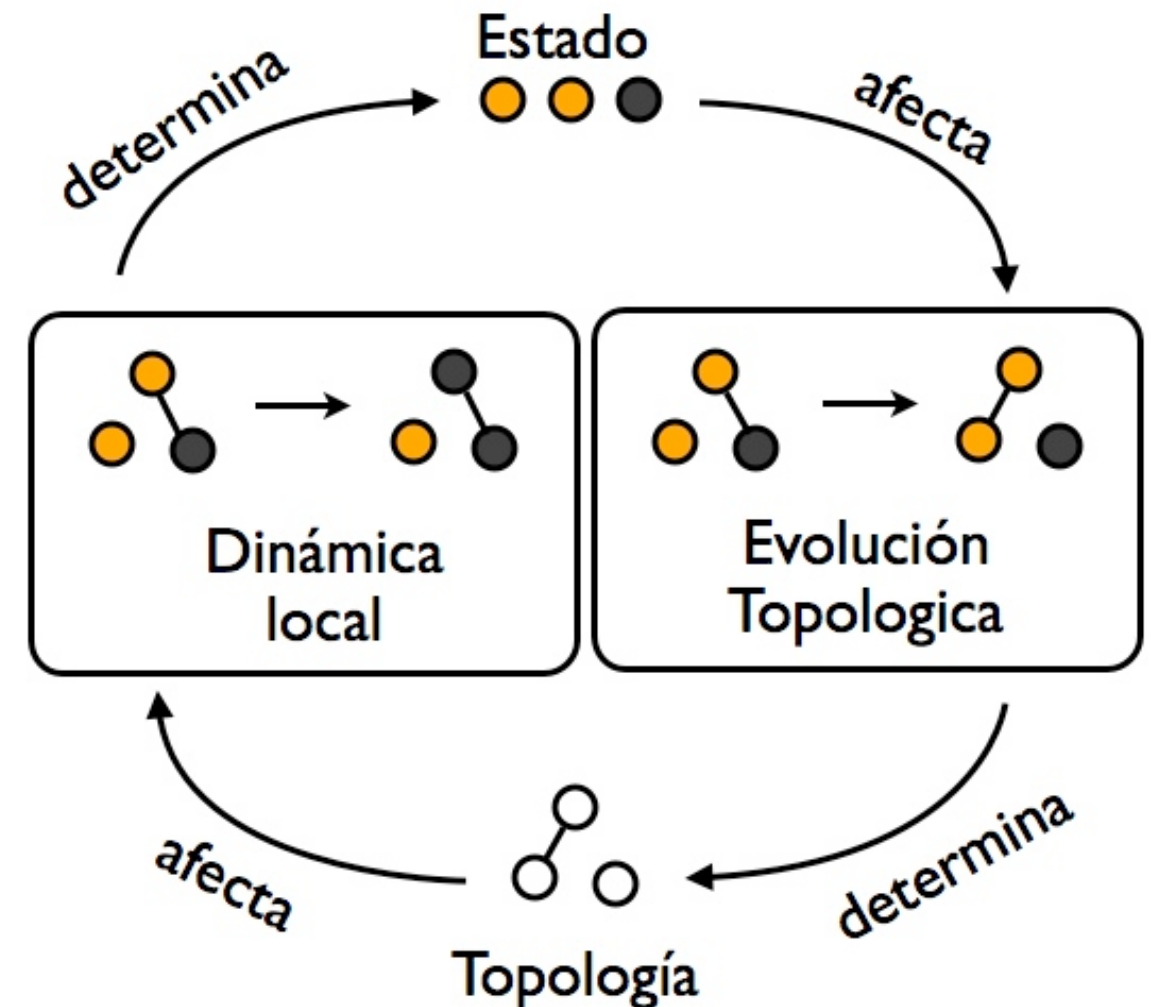
# Detección basada en la Modularidad

El método de Agregación de niveles múltiples de Blondel et al:

- Se ofrece una estructura jerárquica de toda la red.
- En cada nivel (resolución), hay un valor máximo de la modularidad para cierto número de grupos.
- El límite de resolución se evita ya que en la primera fase del método solo consiste en el desplazamiento de individual de nodos y es poco probable que se superpongan comunidades de este modo.  
La superposición puede ocurrir en las fases posteriores pero en este caso es visible.

# Redes Coevolutivas

En común encontrar redes donde la evolución de la topología esta relacionada con el estado de la red y viceversa. En estas redes se forma un circuito de retroalimentación entre el estado y la topología de la red y se conocen como *Redes Coevolutivas*.



# Formación de comunidades en una red coevolutiva

- Propuesta de una dinámica que puede conducir a la formación de comunidades en una red coevolutiva.
- Implementación de la dinámica en ISyS con el fin de generar redes y detectar la estructura de comunidades en estas.
- Estudio de la relación entre los parámetros de la dinámica y las características de la red así como el número y características de las comunidades.