

# Construcción y Análisis de Redes Biológicas

Francisco J. Romero Campero  
<http://www.cs.us.es/~fran/>

Dpt. de Ciencias de la Computación e  
Inteligencia Artificial  
Universidad de Sevilla

# Biología Molecular vs Biología de Sistemas

## Reduccionismo vs Sistemas Complejos

### Biología Molecular

- Aproximación reduccionista.
- Estudio de componentes moleculares (genes, proteínas, ...)
- Enfermedades monogénicas.
- Ingeniería genética a un único gen.
- Ingeniería metabólica a una única enzima.

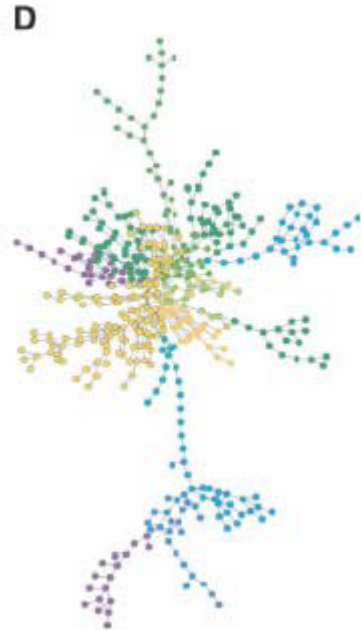
### Biología Molecular de Sistemas

- Aproximación integradora como sistemas complejos.
- Estudio de interacciones entre los componentes moleculares (genes, proteínas, ...)
- Enfermedades complejas.
- Ingeniería genética a sistemas reguladores génicos.
- Ingeniería metabólica a rutas metabólicas completas.

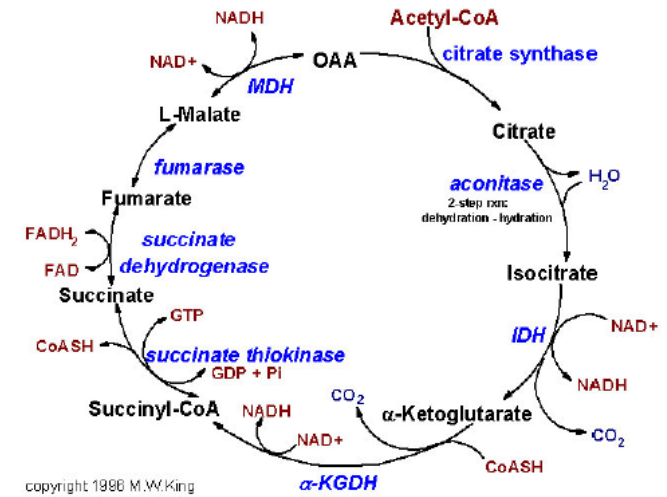
# Redes Biomoleculares



Redes de interacción  
entre proteínas



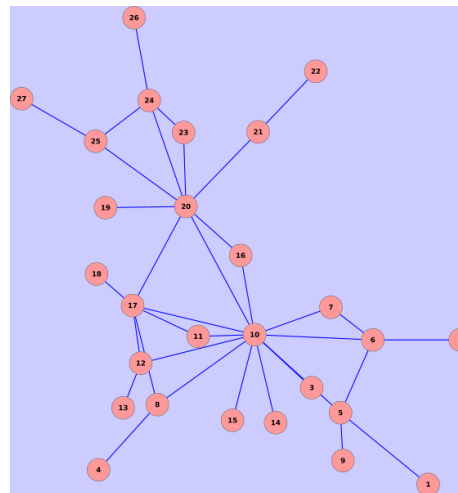
Redes de coexpresión génicas



Redes metabólicas

# Redes Biomoleculares

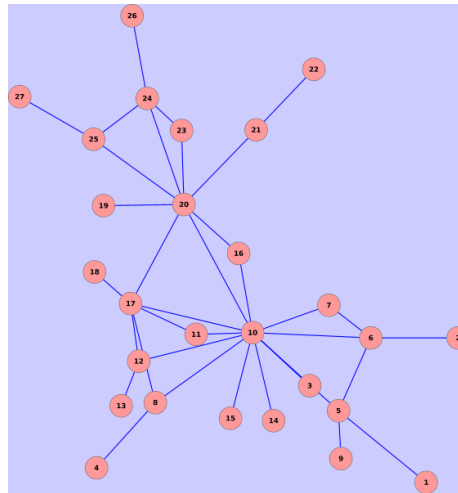
- Una **red** es una representación de las interacciones que tienen lugar entre las entidades que dan lugar a un fenómeno estudiado.
  - Los **nodos** representan las entidades genéricas que constituyen el sistema (genes, proteínas, metabolitos, etc).
  - Las **aristas** entre distintos nodos indican que las correspondientes entidades interactúan de alguna forma.



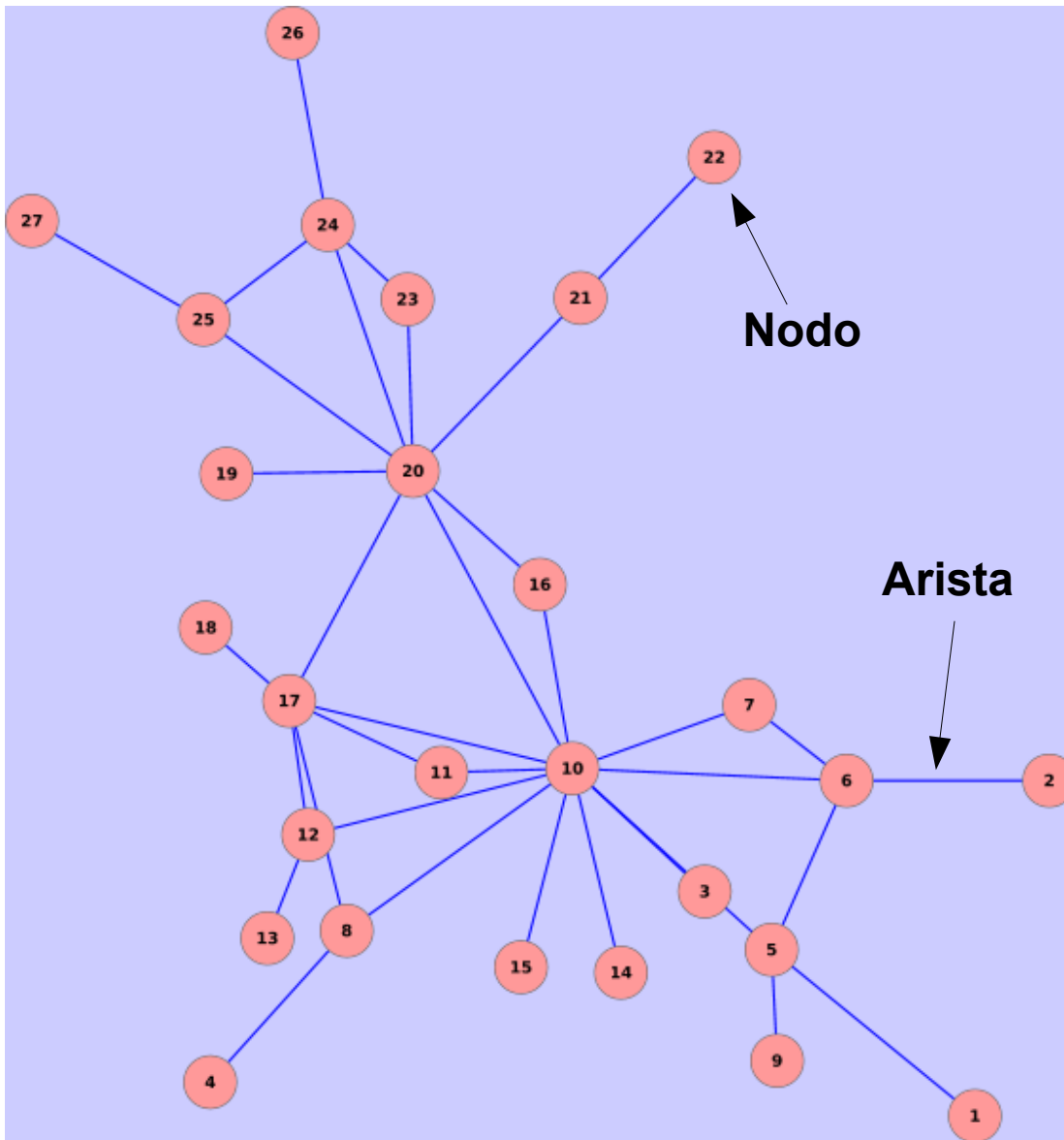
# Definición de Red o Grafo (no dirigido)

Una **red o grafo** **G** es un par de conjuntos (V,E)

- $V=\{v_1,v_2,\dots,v_n\}$  es el conjunto de **vértices** o **nodos**.
- $E=\{(v_i,v_j),(v_i',v_j')\}\dots\dots\}$  es un conjunto de **pares no ordenados** de elementos de V.
- E se denomina **conjunto de aristas** de la red.
- El numero de nodos se denomina **orden** de la red.
- El número de aristas se denomina **tamaño** de la red.



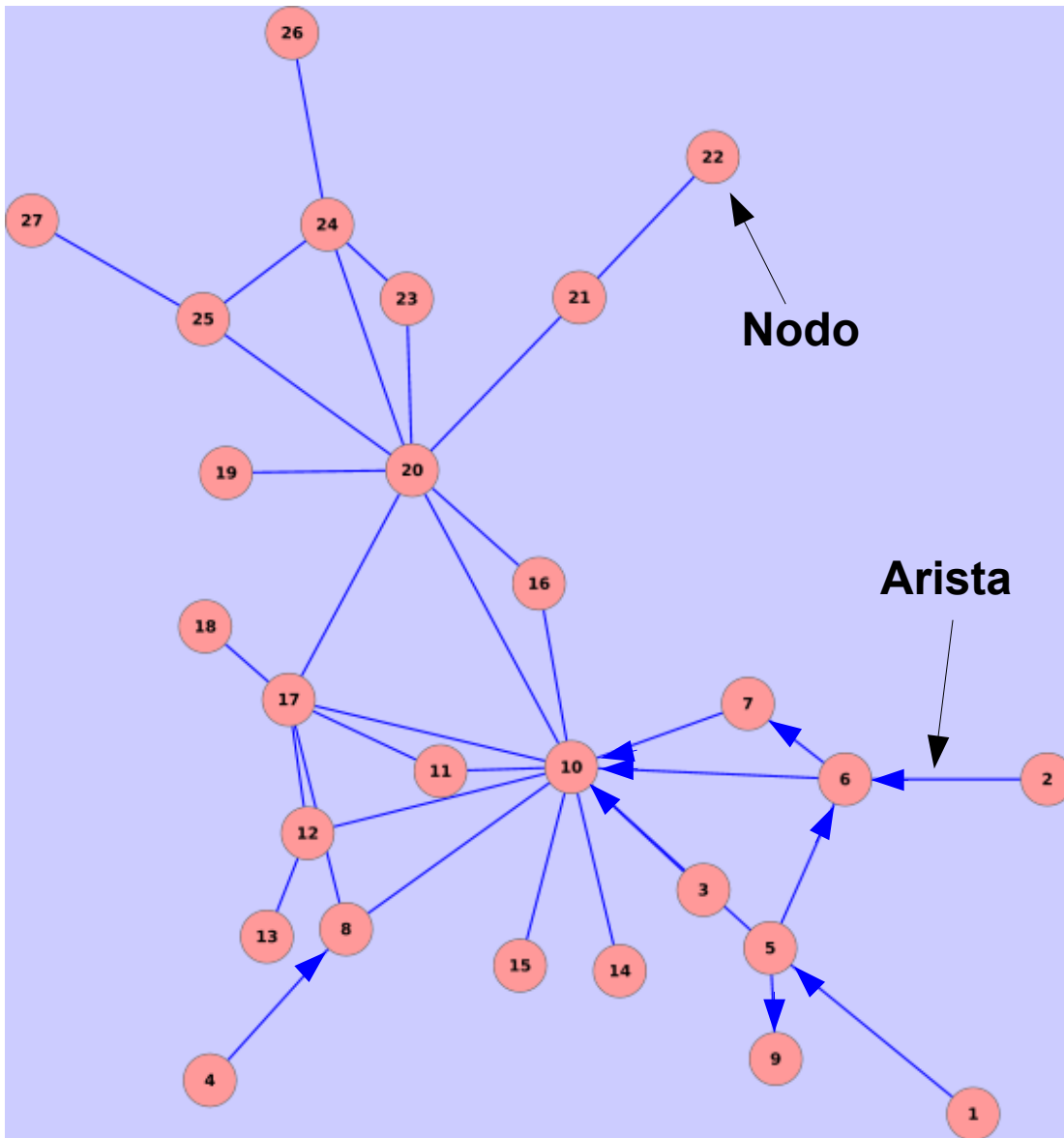
# Definición de Red o Grafo (no dirigido)



$$G = (V, E)$$

- $V = \{1, 2, \dots, 27\}$  es el conjunto de **vértices** o **nodos**.
- $E = \{\{1, 5\}, \{2, 6\}, \{5, 6\}, \{6, 7\}, \{5, 9\}, \{5, 10\}, \{6, 10\}, \{7, 10\}, \{3, 10\}, \{4, 8\}, \{8, 10\}, \{14, 10\}, \{10, 12\}, \{12, 13\}, \{11, 10\}, \{15, 10\}, \{16, 10\}, \{10, 20\}, \{17, 10\}, \{17, 11\}, \{17, 8\}, \{17, 12\}, \{17, 18\}, \{17, 20\}, \{16, 20\}, \{19, 20\}, \{23, 20\}, \{24, 20\}, \{25, 20\}, \{21, 20\}, \{22, 21\}, \{23, 24\}, \{24, 25\}, \{26, 24\}, \{27, 25\}\}$  son las **aristas** de la red.

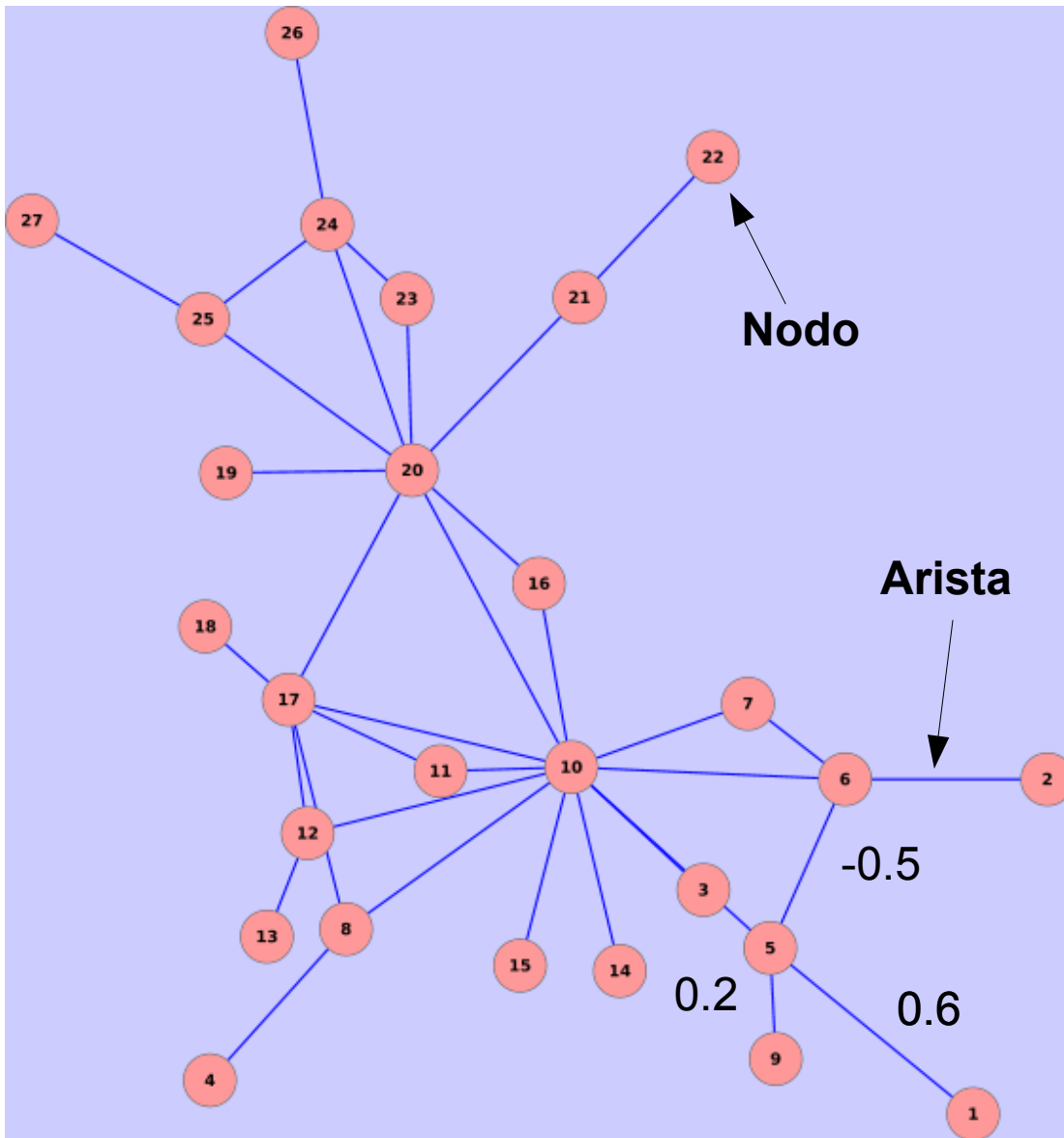
# Definición de Red o Grafo Dirigido



En un **grafo dirigido** las aristas tienen una dirección fija. Las aristas se definen como pares ordenados donde el primer nodo es el origen de la arista y el segundo el destino.

•  $E = \{(1,5), (2,6), (5,6), (6,7), (5,9), (5,10), (6,10), (7,10), (3,10), (4,8), (8,10), (14,10), (10,12), (12,13), (11,10), (15,10), (16,10), (10,20), (17,10), (17,11), (17,8), (17,12), (17,18), (17,20), (16,20), (19,20), (23,20), (24,20), (25,20), (21,20), (22,21), (23,24), (24,25), (26,24), (27,25)\}$  son las **aristas** de la red.

# Definición de Red o Grafo Ponderado



En una **red o grafo ponderado** cada arista tiene asociado un peso o valor numérico que representa una característica de la correspondiente interacción.

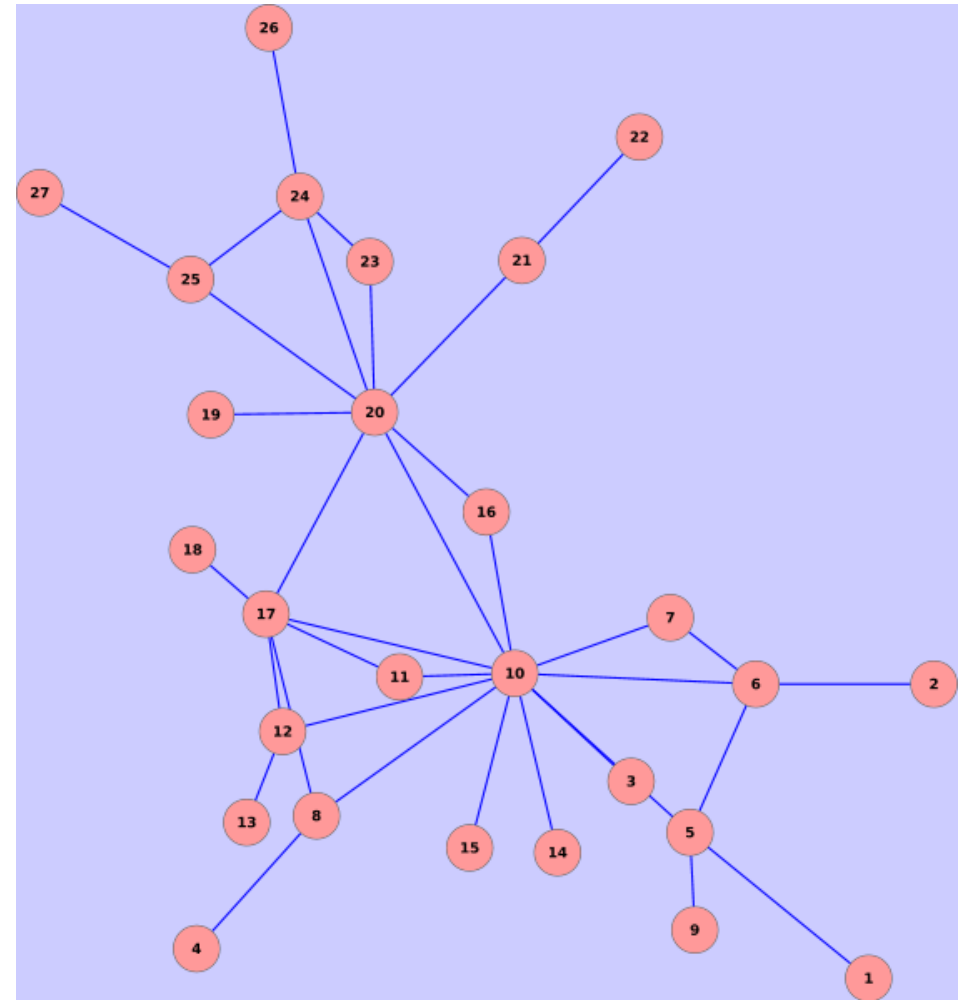


# Especificación de redes o grafos

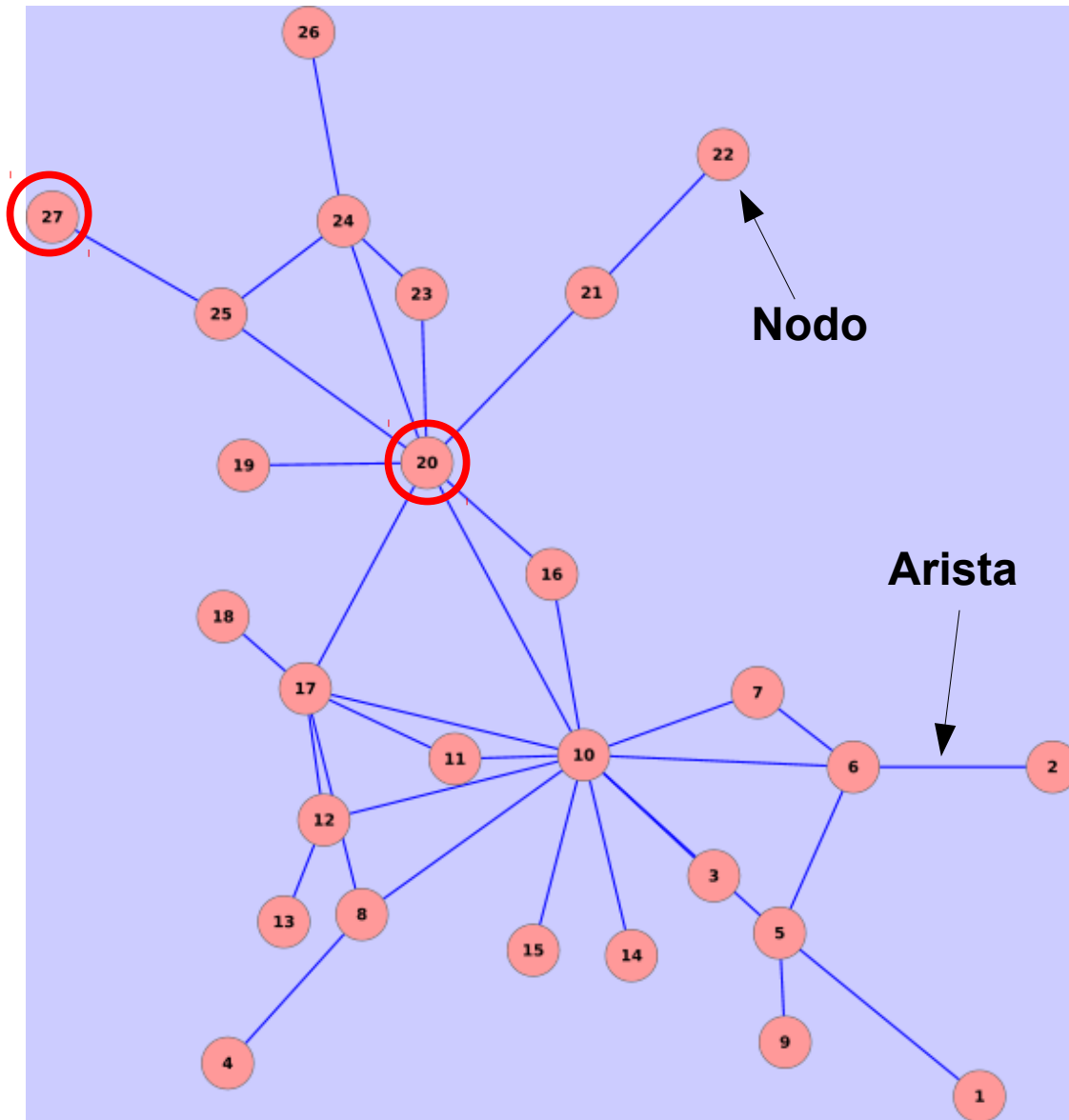
- Dado una red o grafo  $G=(V,E)$  se puede especificar utilizando la matriz de adyacencia  $A = (a_{ij})$  tal que:

$$a_{ij} = \begin{cases} 1 & \text{si y solo si } \{i,j\} \in V \\ 0 & \text{en otro caso} \end{cases}$$

[illegible]



# Grado de un Nodo



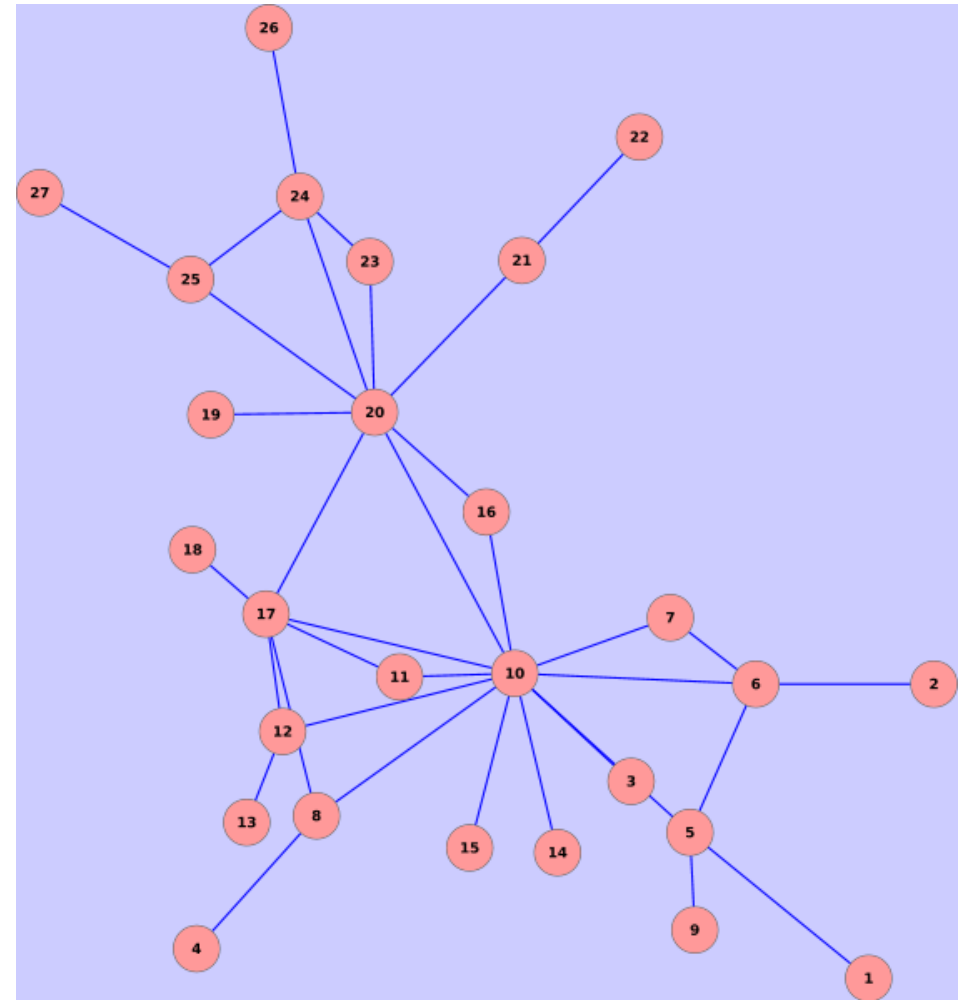
- Dos nodos de un grafo son **vecinos o adyacentes** si existe una arista que los conecta.
- El **grado de un nodo (node degree)** es el número vecinos que tiene dicho nodo.
- En los grafos dirigidos se calcula el **grado de entrada y el grado de salida**.
- Un grafo se dice que es **regular** si todos los nodos tienen el mismo grado.

$$\text{Degree}(27) = 1$$

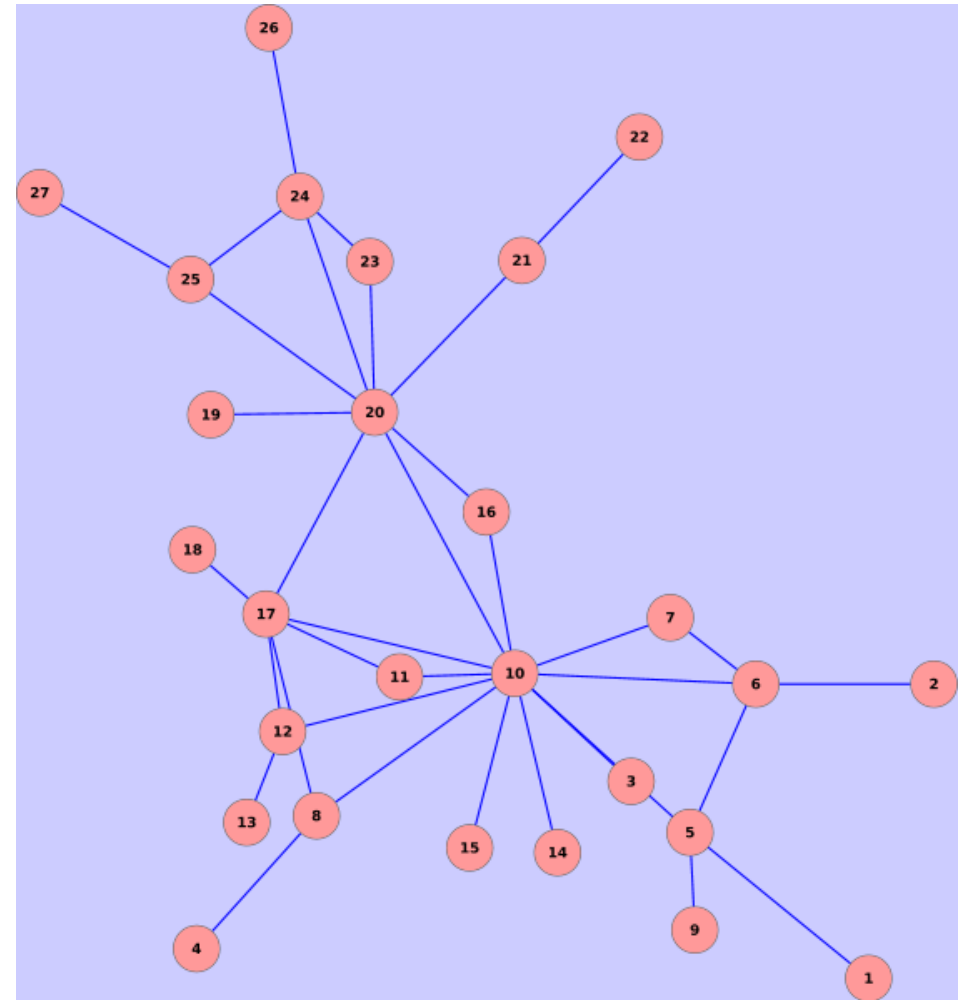
$$\text{Degree}(20) = 8$$

## Cálculo del grado de un nodo

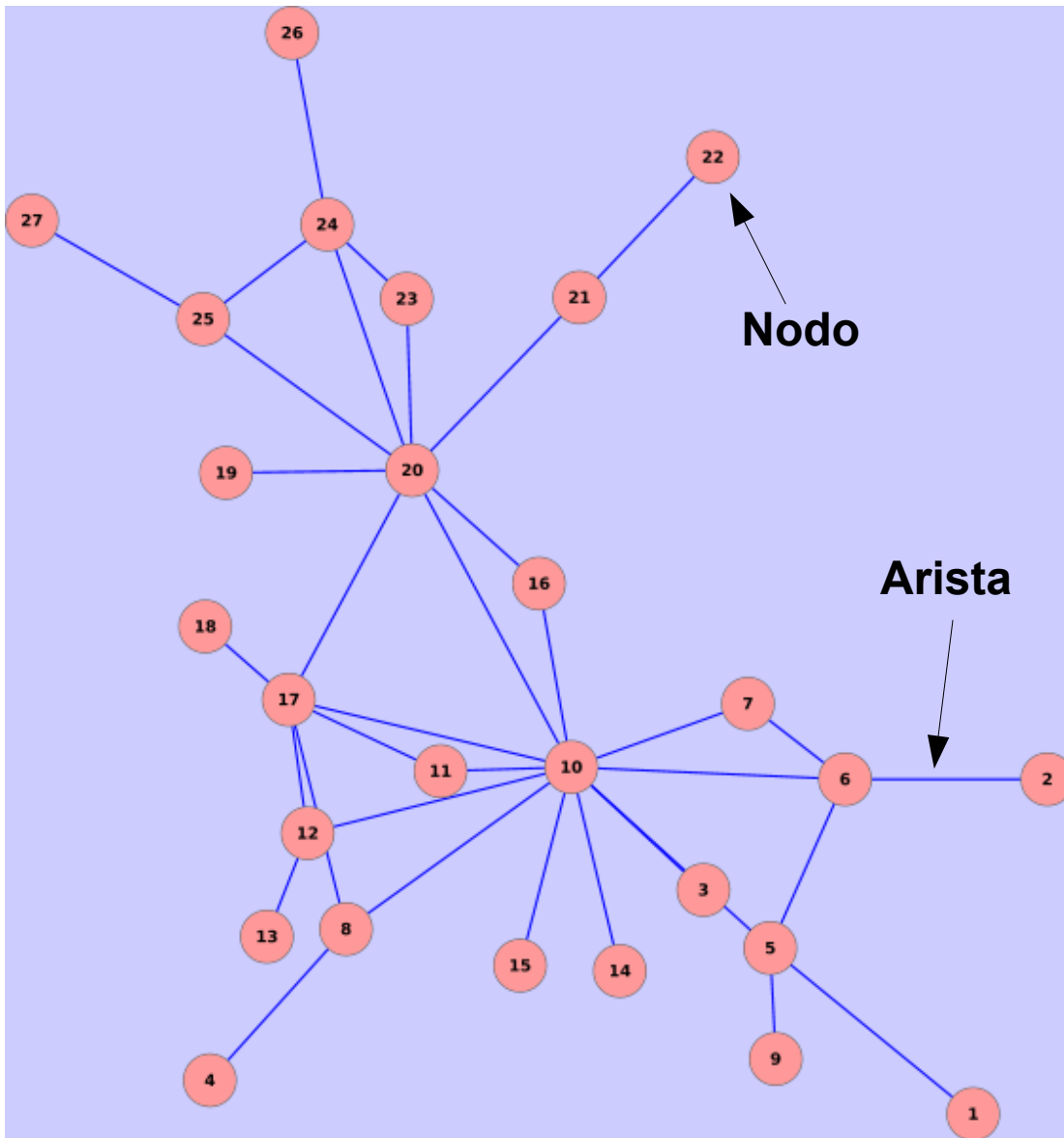
[illegible]



## Cálculo del grado de un nodo

[illegible]

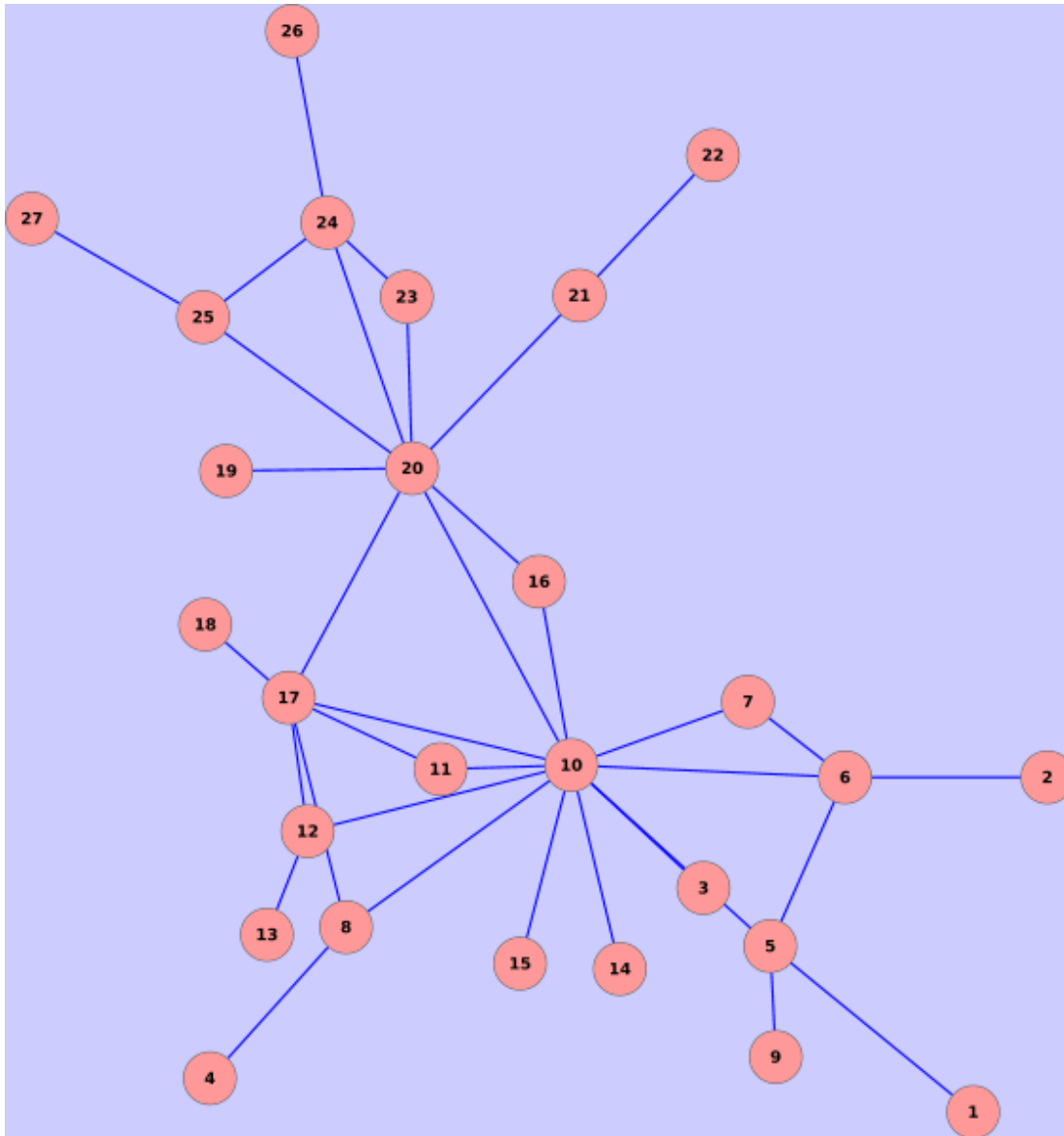
# Coeficiente de agrupamiento



- El **coeficiente de agrupamiento** (clustering coefficient) de un nodo es una medida local que refleja el nivel de agrupamiento que existe entorno a un nodo.
- Se calcula el número de vecinos del nodo correspondiente  $d_v = \text{degree}(v)$ . Entre estos vecinos el número máximo de aristas es  $d_v(d_v - 1) / 2$ . Este valor corresponde al mayor agrupamiento posible.
- Se determina el número real de aristas entre los vecinos de  $v$   $N_v$ .
- Se calcula el coeficiente de agrupamiento como:

$$C_v = \frac{N_v}{\left( \frac{d_v(d_v - 1)}{2} \right)}$$

# Coeficiente de agrupamiento



- **Coeficiente de agrupamiento:**

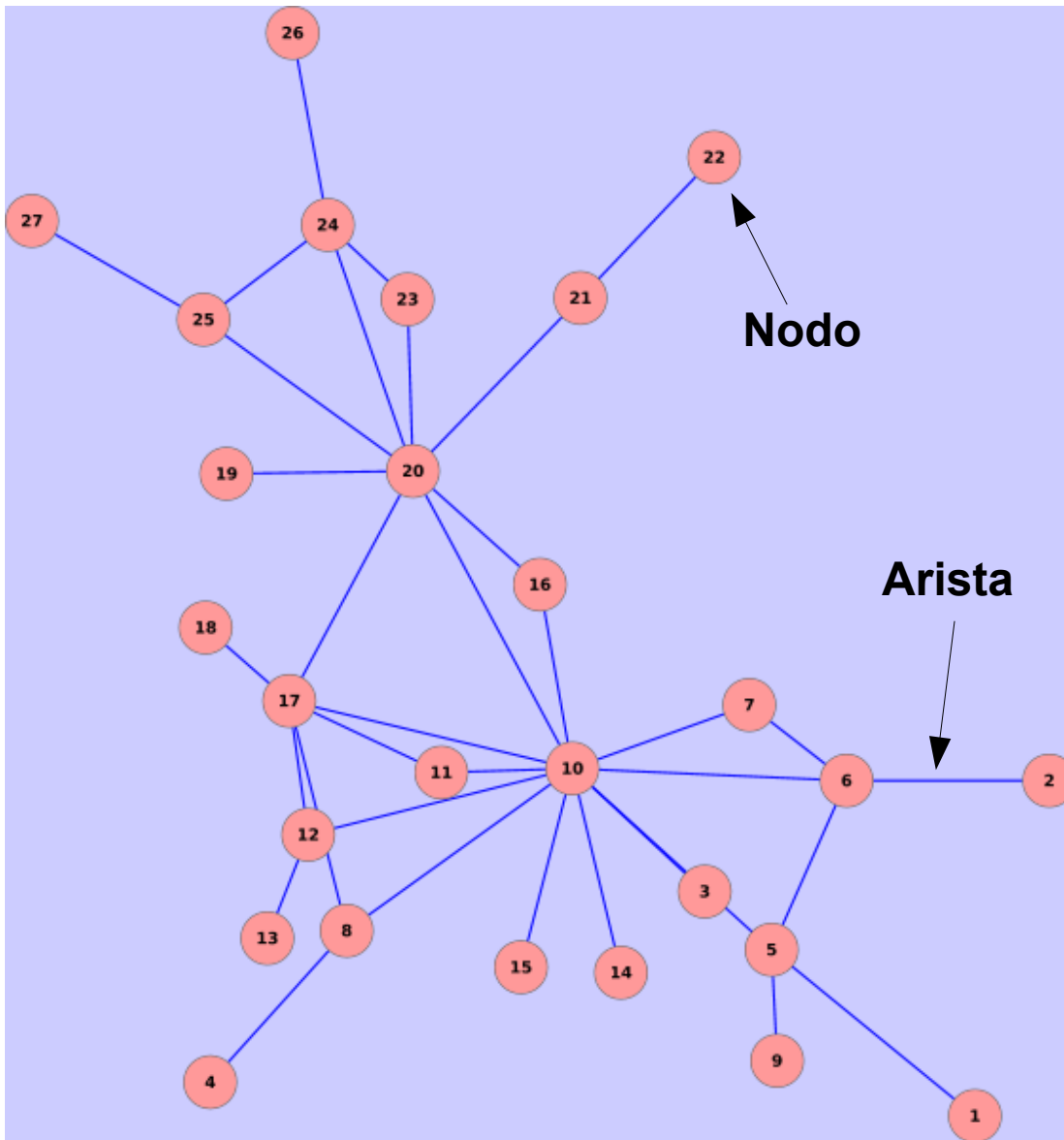
$$C_v = \frac{N_v}{\left(\frac{d_v(d_v-1)}{2}\right)}$$

$$C_{21} = \frac{0}{\left(\frac{2(1)}{2}\right)} = 0$$

$$C_{23} = \frac{1}{\left(\frac{2(1)}{2}\right)} = 1$$

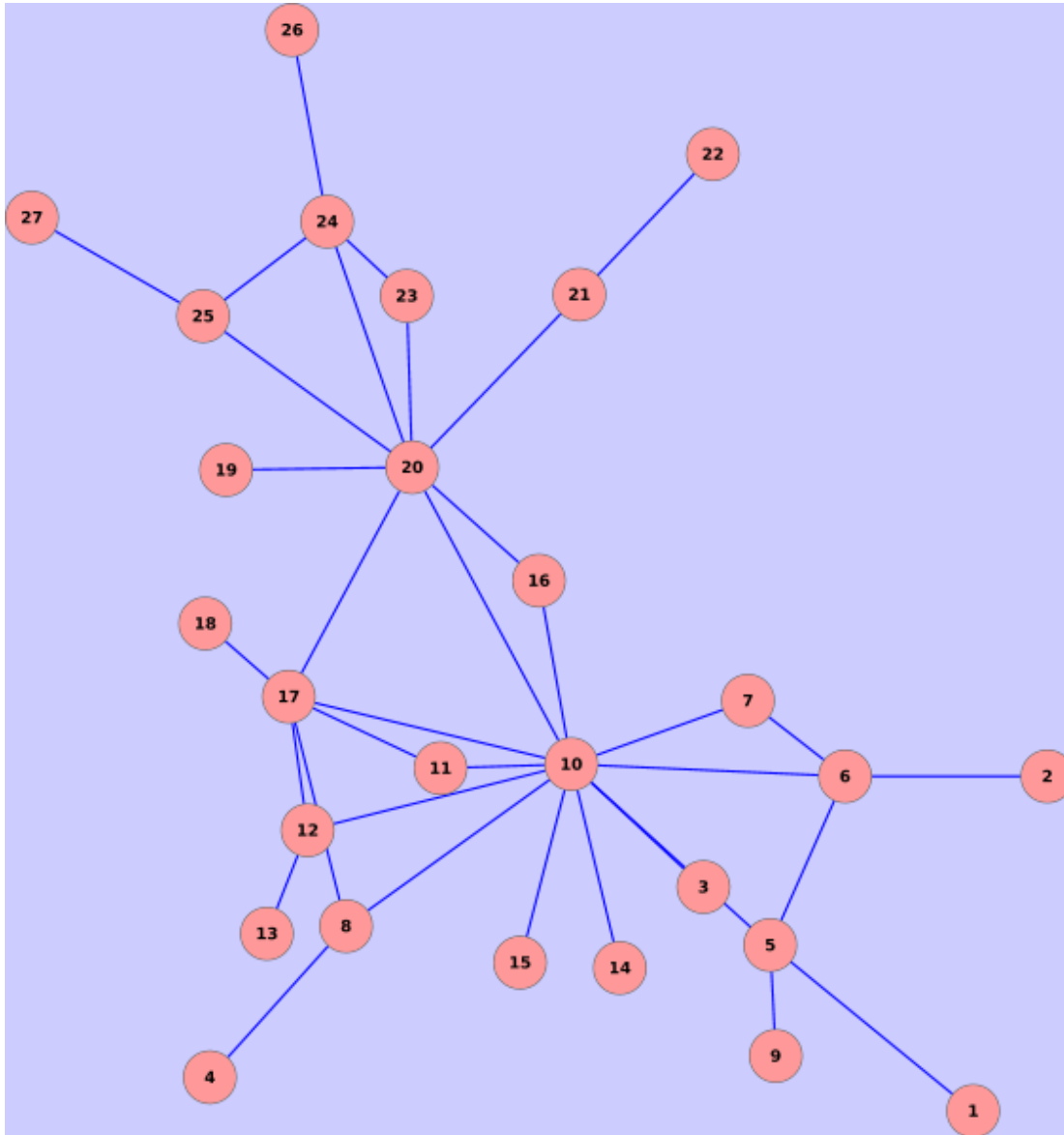
$$C_{20} = \frac{4}{\left(\frac{8(7)}{2}\right)} = 0.14$$

# Definición de Paseo y Camino



- Un **paseo** de un nodo **u** a un nodo **v** es una secuencia de nodos  $\{v_0, v_1, \dots, v_k\}$  con  $v_0 = u$ ,  $v_k = v$  y  $\{v_{i-1}, v_i\}$  rama del grafo.
- El número de aristas del paseo es su **longitud**.
- Un paseo en el cual todos los vertices  $\{v_0, v_1, \dots, v_k\}$  son distintos se denomina **camino**.
- Un **camino mínimo** entre dos nodos es aquel de menor longitud de entre todos los posibles caminos entre ambos nodos.

# Extensión de propiedades de nodos a propiedades globales de redes



- **Distribución del grado de nodos** en un grafo  $G=(V,E)$ :

$$P(k) = m_k / m \text{ donde}$$

$m_k$  es el número de nodos de grado  $k$   
 $m$  es el orden de  $G$

- **Coeficiente de agrupamiento medio** de un grafo  $G=(V,E)$ :

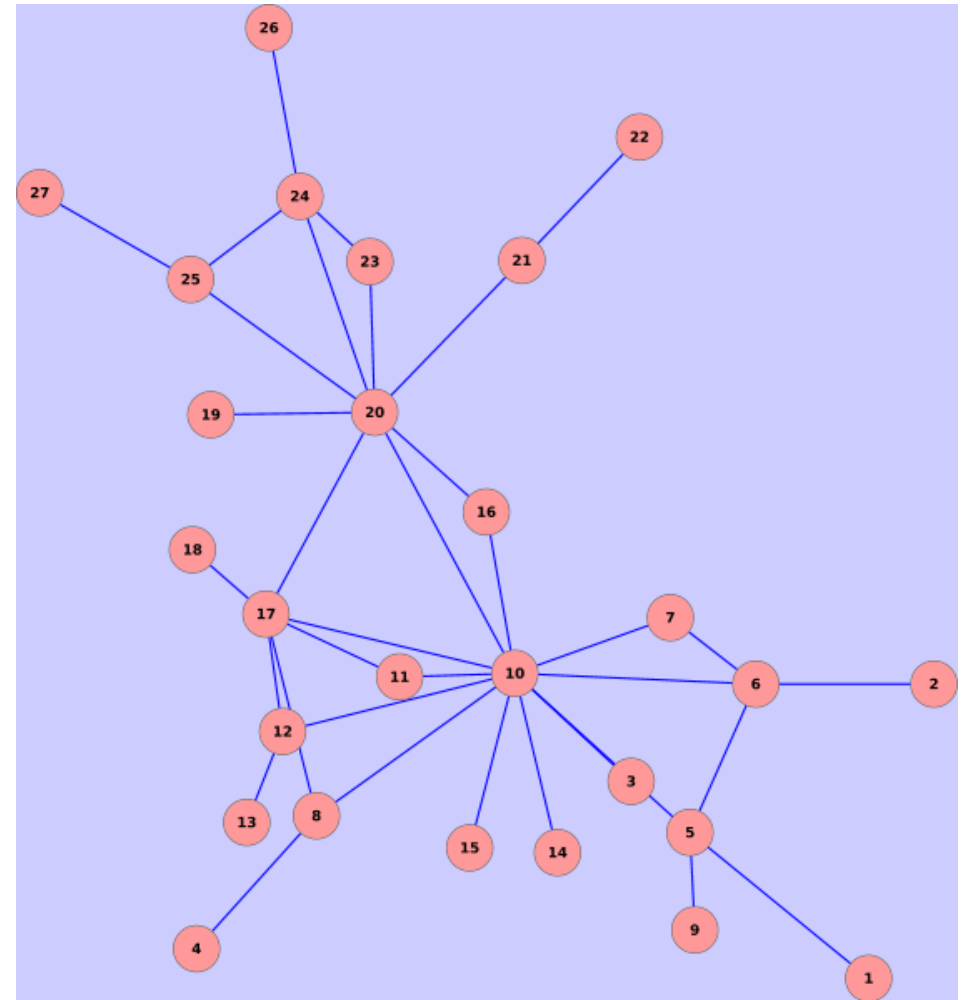
$$C_G = \frac{1}{m} \sum_{i=1}^m C_i$$



# Cálculo de la distribución del grado de nodos

[illegible]

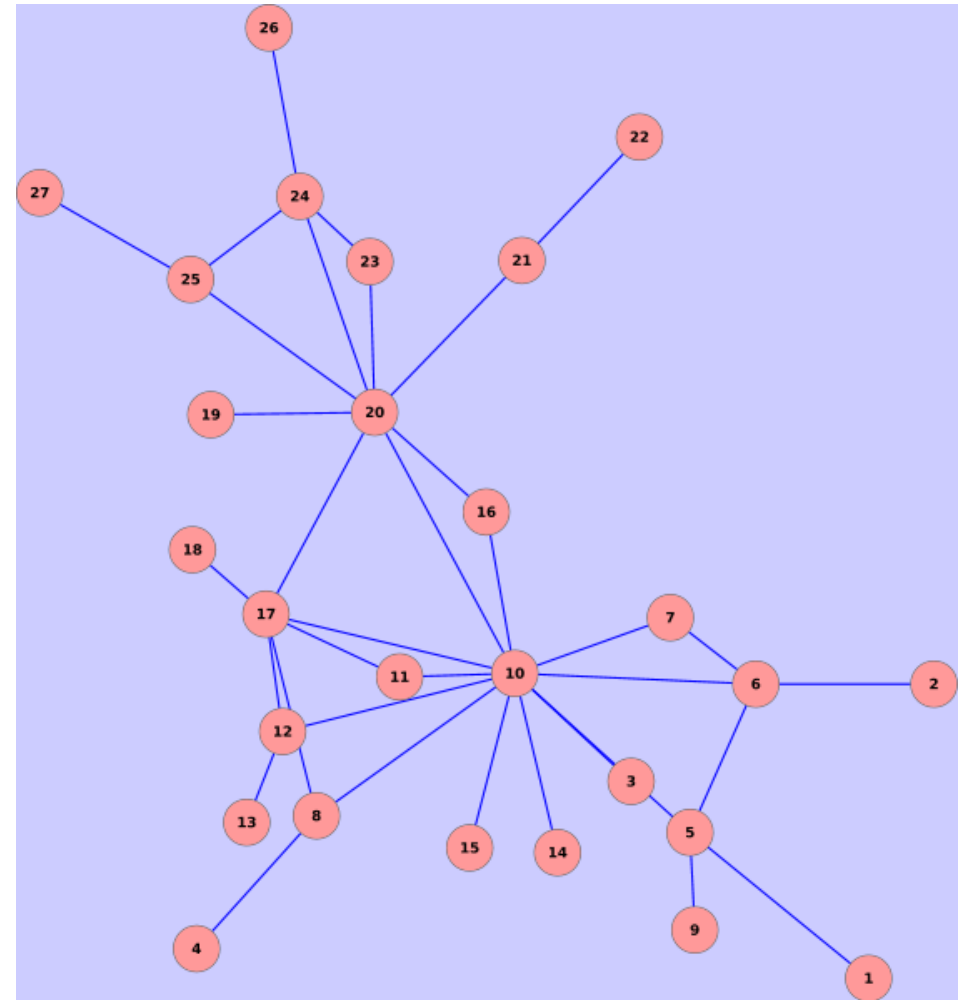
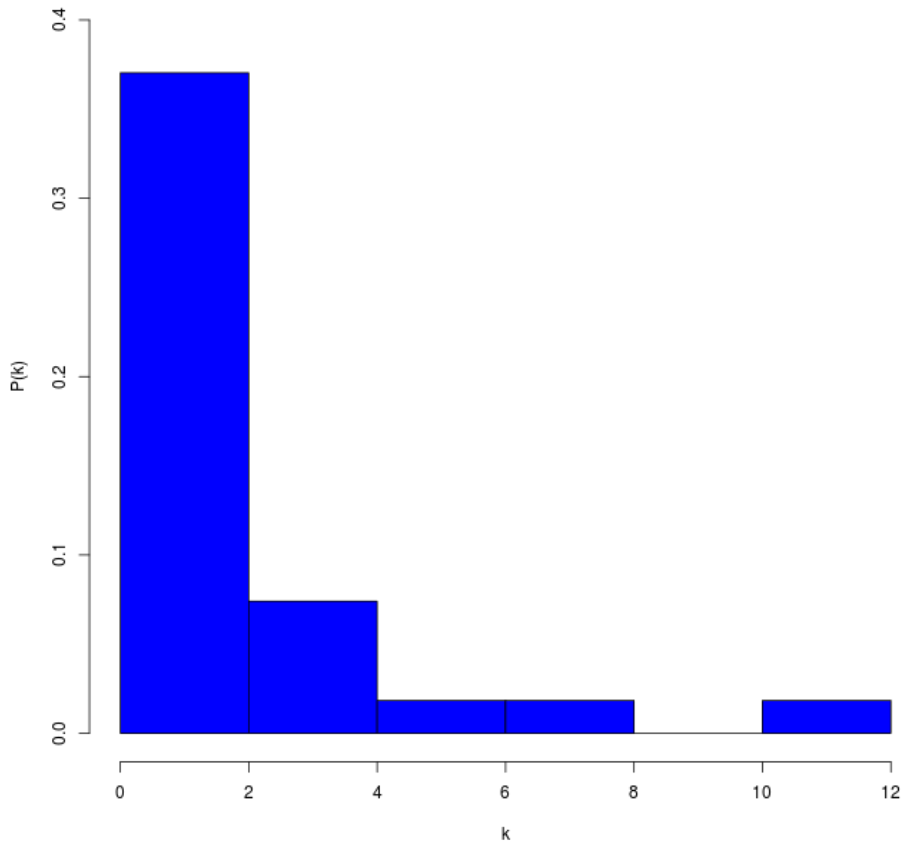
rowSums y hist



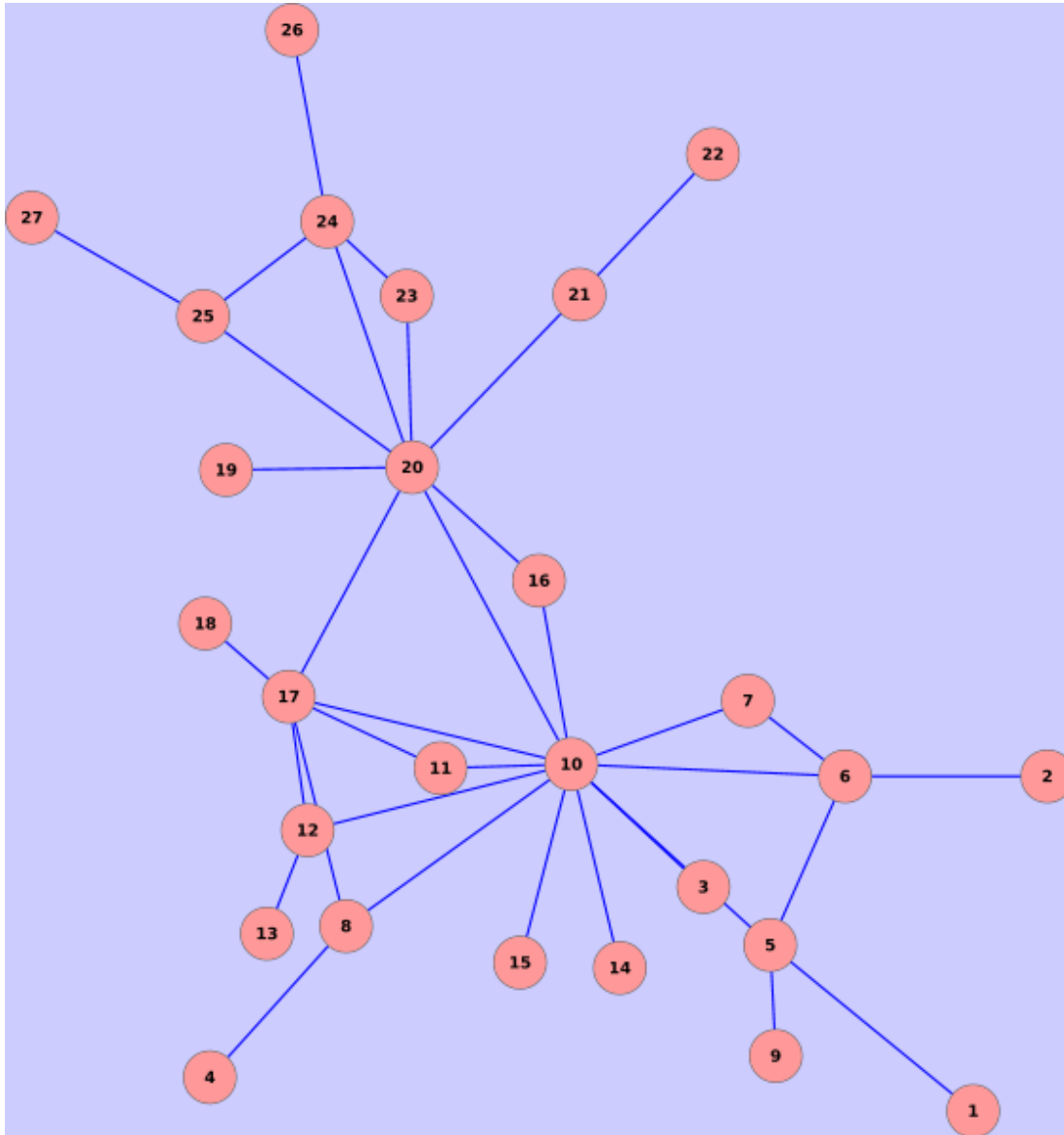
# Cálculo de la distribución del grado de nodos

rowSums y hist

Degree distribution



# Extensión de propiedades de nodos a propiedades globales de redes



- **Distribución del grado de nodos** en un grafo  $G=(V,E)$ :

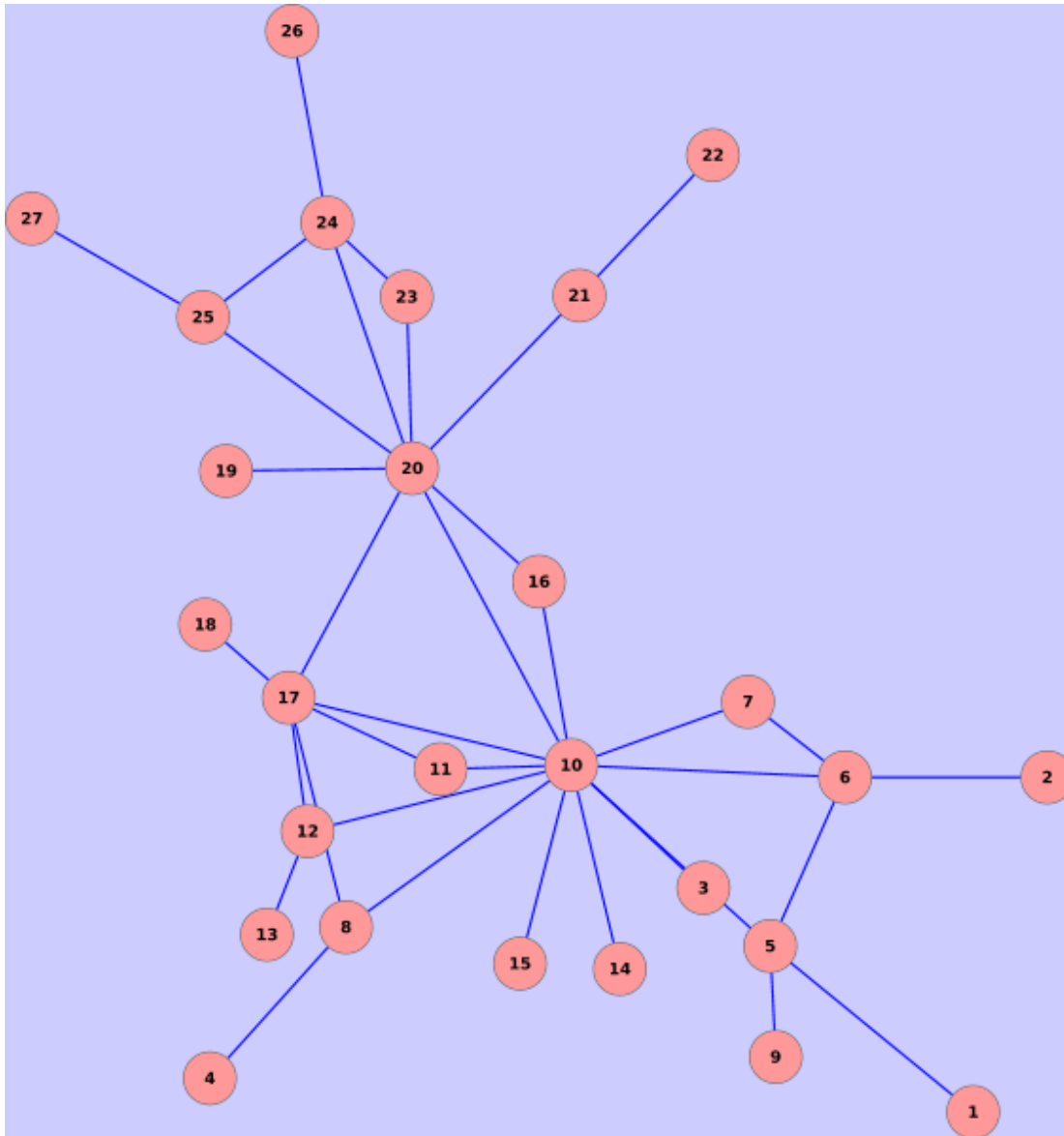
$$P(k) = m_k / m \text{ donde}$$

$m_k$  es el número de nodos de grado  $k$   
 $m$  es el orden de  $G$

- **Coeficiente de agrupamiento medio de un grafo  $G=(V,E)$ :**

$$C_G = \frac{1}{m} \sum_{i=1}^m C_i$$

# Tipos de Redes según su Topología



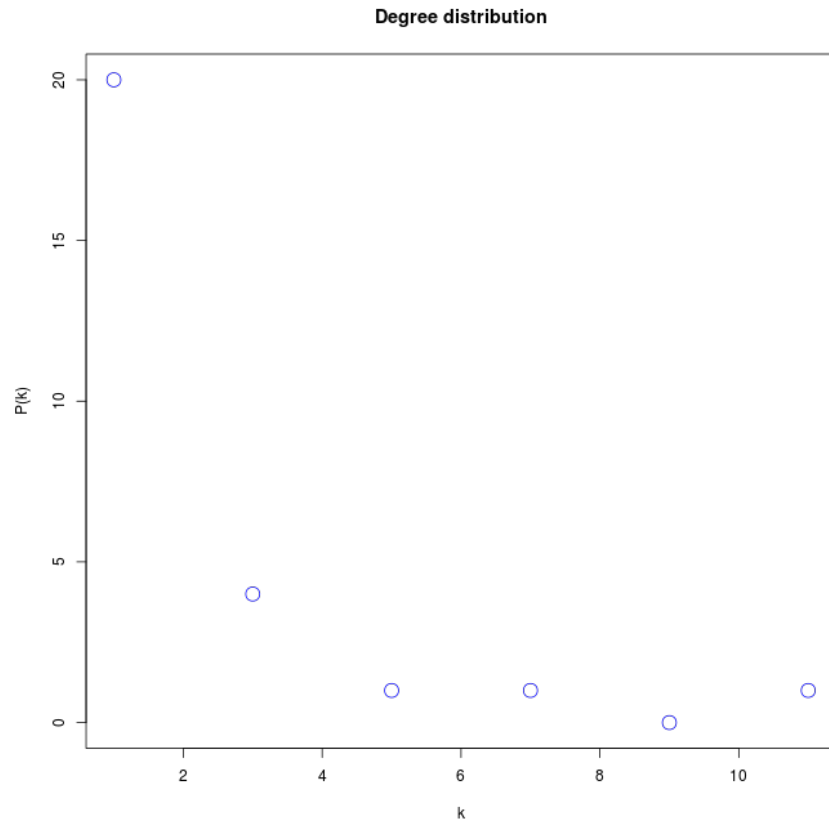
- **Redes libre de escala:** Dada una red  $G=(V,E)$  diremos que es libre de escala si su distribución del grado de nodos sigue una distribución exponencial negativa.

$$P(k) = c * k^{-\gamma}$$

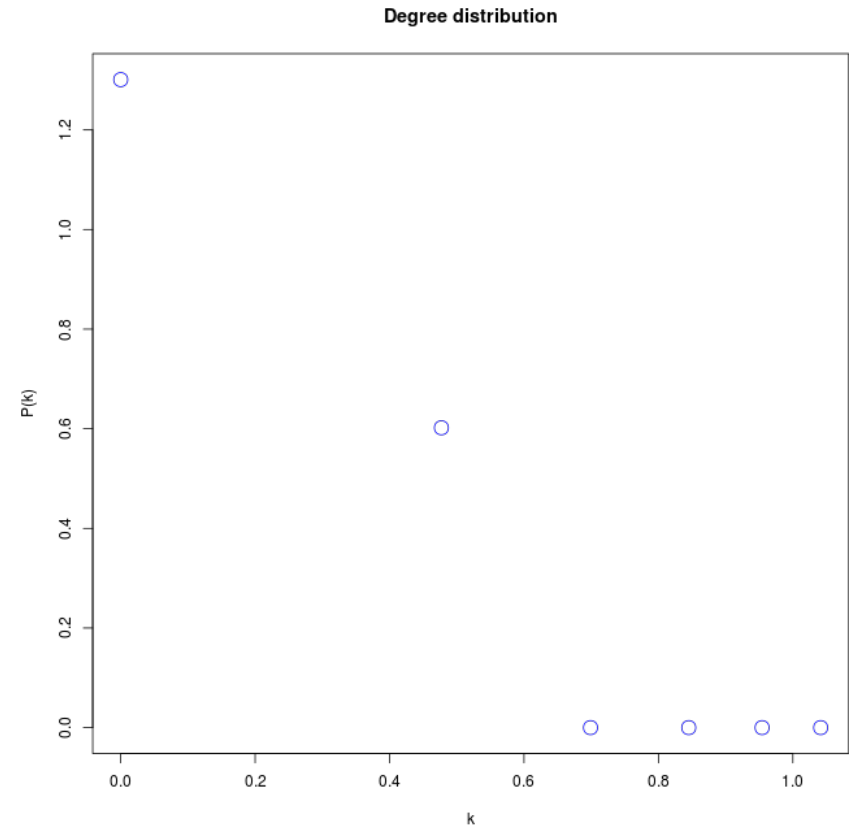
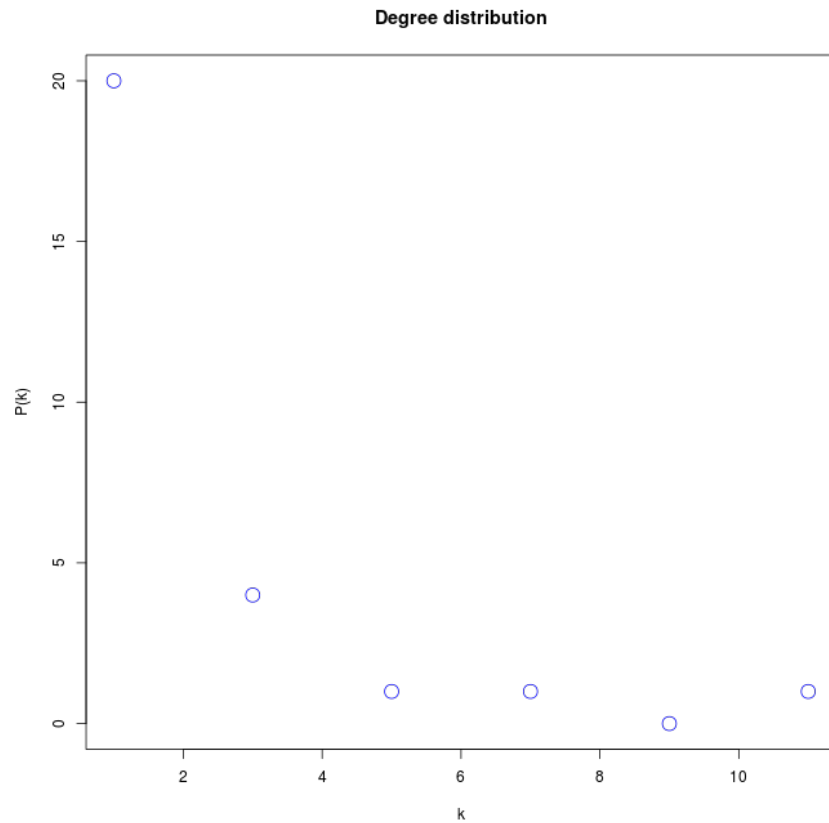
La mayoría de los nodos de este tipo de presentan un número pequeño de vecinos. Sin embargo existen unos pocos nodos destacados que tiene un alto número de veces esto tipo de nodos se denominan **hubs**.

- **Redes de mundo pequeño:** Dada una red  $G=(V,E)$  diremos que es un mundo pequeño si es una red libre de escala que presenta un alto coeficiente medio de agrupamiento. En este tipo de red los caminos entre nodos es pequeño.

# Determinación de Grafos Libres de Escala



# Determinación de Grafos Libres de Escala



# Determinación de Grafos Libres de Escala

Linear regression con lm

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	1.2144	0.1581	7.679	0.00155 **
log10(h[["mids"]])	-1.3402	0.2093	-6.403	0.00306 **

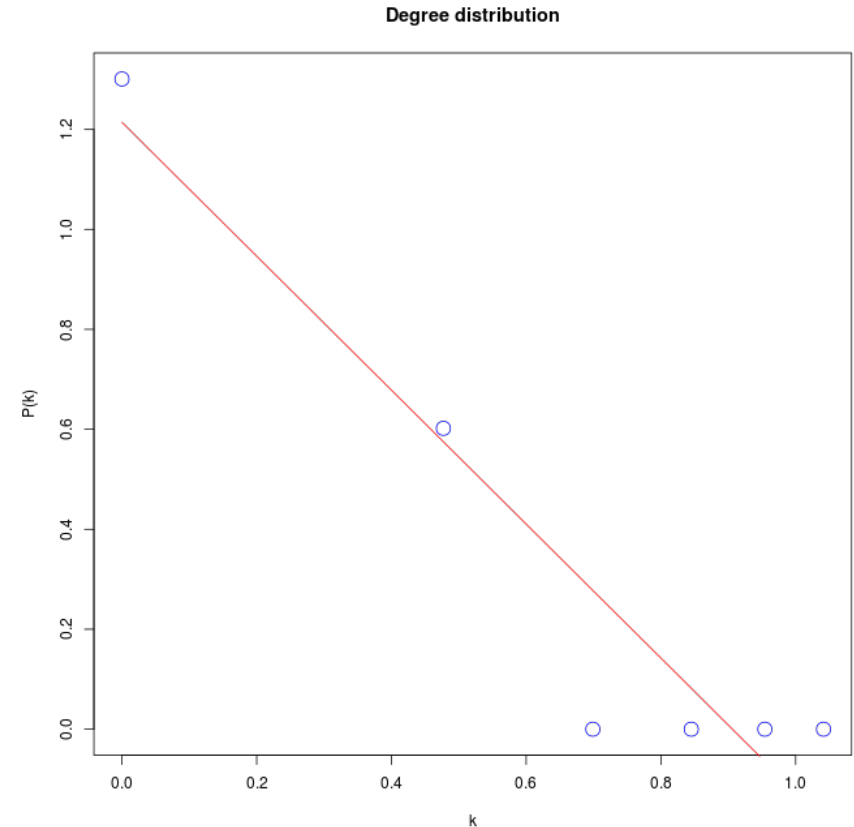
---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1796 on 4 degrees of freedom

Multiple R-squared: 0.9111, Adjusted R-squared: 0.8889

F-statistic: 41 on 1 and 4 DF, p-value: 0.003056



# Determinación de Grafos Libres de Escala

## Linear regression con lm

Coefficients:

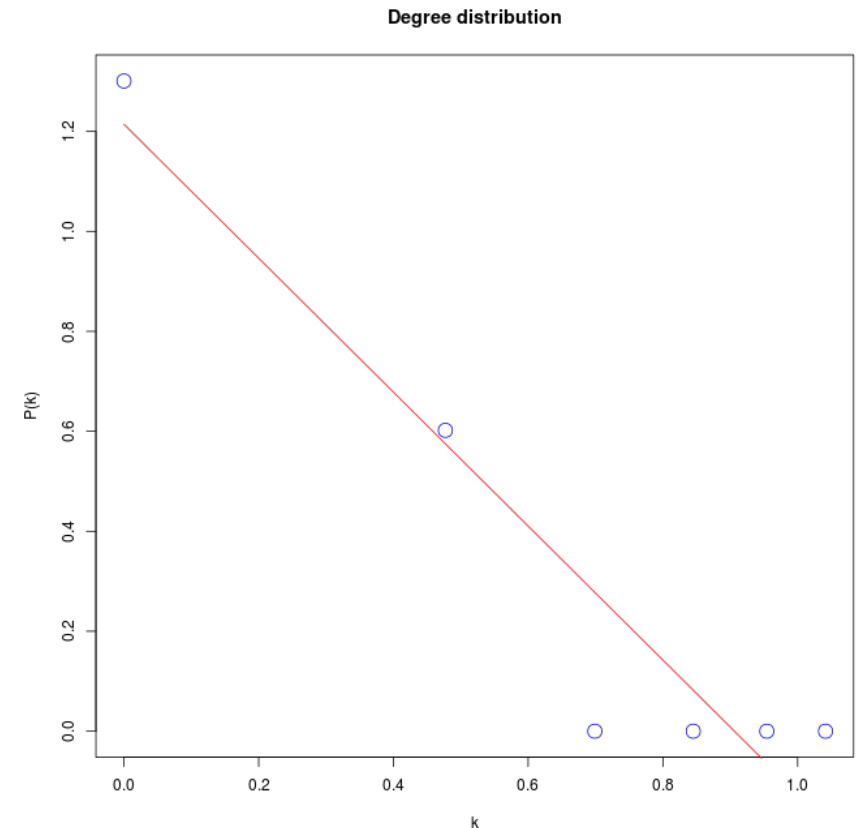
	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	1.2144	0.1581	7.679	0.00155 **
log10(h[["mids"]])	-1.3402	0.2093	-6.403	0.00306 **

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1796 on 4 degrees of freedom  
Multiple R-squared: 0.9111, Adjusted R-squared: 0.8889  
F-statistic: 41 on 1 and 4 DF, p-value: 0.003056

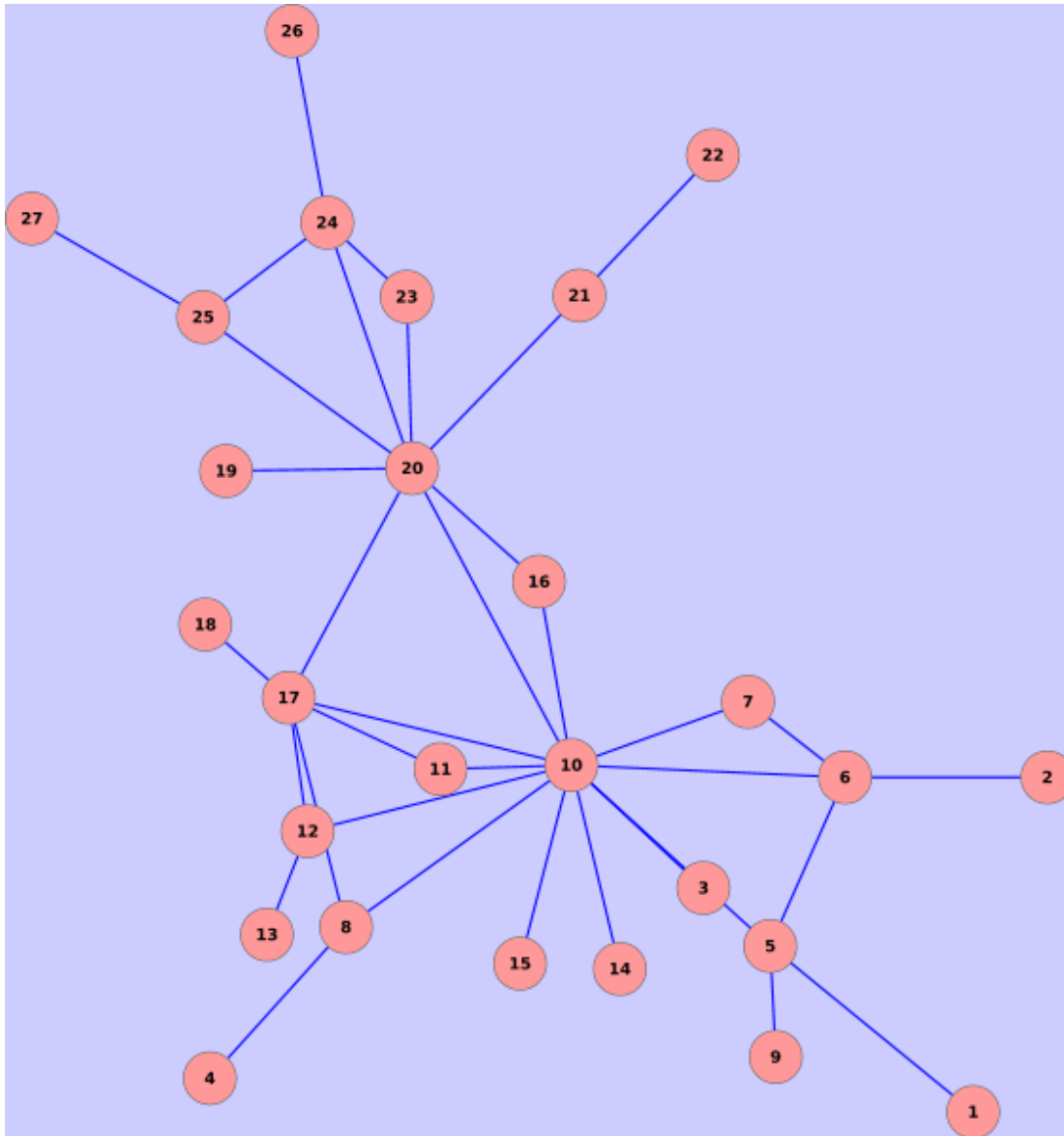
La función de igraph `power.law.fit` que recibe como entrada la distribución del grado de los nodos nos permite realizar un análisis estadístico basado en el test de Kolmogorov-Smirnov sobre el ajuste de la topología de una red a la propiedad libre de escala. Esta función devuelve un objeto donde el valor `KS.p` es el p-valor correspondiente a rechazar la hipótesis nula que en este caso aserta que la red estudiada es libre de escala. Por lo tanto, un valor alto de `KS.p` indica la ausencia de evidencia para afirmar que la red estudiada no es libre de escala



```
> network.degree.distribution <- degree.distribution(example.network)
> fit.scale.free <- power.law.fit(network.degree.distribution)
> fit.scale.free[["KS.p"]]
[1] 0.8990623
```



# Tipos de Redes según su Topología



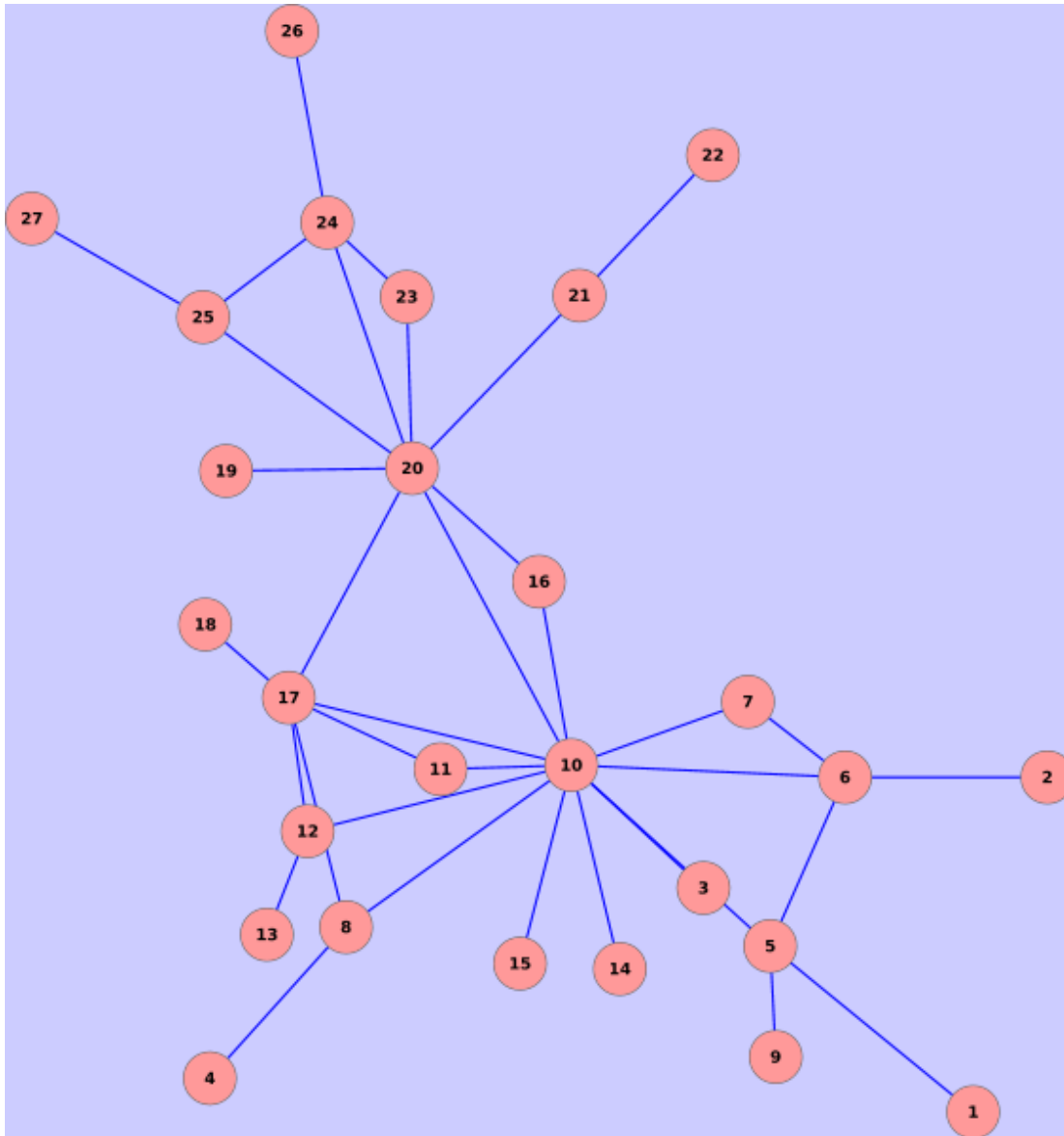
- **Redes libre de escala:** Dada una red  $G=(V,E)$  diremos que es libre de escala si su distribución del grado de nodos sigue una distribución exponencial negativa.

$$P(k) = c * k^{-\gamma}$$

La mayoría de los nodos de este tipo de presentan un número pequeño de vecinos. Sin embargo existen unos pocos nodos destacados que tiene un alto número de veces esto tipo de nodos se denominan **hubs**.

- **Redes de mundo pequeño:** Dada una red  $G=(V,E)$  diremos que es un mundo pequeño si es una red libre de escala que presenta un alto coeficiente medio de agrupamiento. En este tipo de red los caminos entre nodos es pequeño.

# Tipos de Redes según su Topología



- **Redes de mundo pequeño:** Dada una red  $G=(V,E)$  diremos que es un mundo pequeño si es una red libre de escala que presenta un alto coeficiente medio de agrupamiento. En este tipo de red los caminos entre nodos es pequeño.

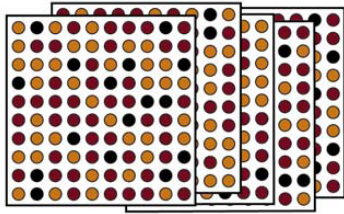
Para comprobar si la longitud media del camino mínimo entre nodos es lo suficientemente pequeña como para considerarla de mundo pequeño es común generar redes libres de escala del mismo orden y tamaño de la estudiada para estimar la probabilidad de que por pura aleatoriedad se obtenga una red similar a la estudiada pero con una longitud media del camino mínimo entre nodos inferior. La función `barabasi.game` permite generar redes libres de escala con el número de nodos proporcionado en el argumento `n`.

# Redes de Co-expresión Génica

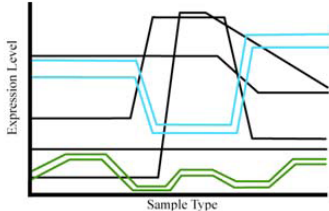
- Las **redes de co-expresión génicas** son un tipo de redes que persiguen integrar información parcial obtenida en diferentes experimentos o análisis de expresión génica. Típicamente se basan en datos transcriptómicos masivos obtenidos utilizando por ejemplo microarrays.
- En una **red de co-expresión génica** los **nodos representan genes** y las **aristas entre nodos representan que los correspondientes nodos se co-expresan** en las distintas muestras de los experimentos analizados.
- La co-expresión entre genes suele medirse utilizando la **correlación entre sus perfiles de expresión**.

# Flujo de Trabajo para la Construcción de Redes de Co-expresión Génica

## Datos Transcriptómicos



## Análisis de la correlación



- **Paso 1:** Análisis de datos transcriptómicos masivos: análisis de expresión diferencial.

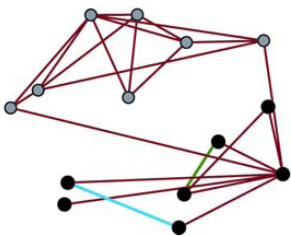
- **Paso 2:** Análisis de la correlación entre los perfiles de expresión.

## Matriz de Adyacencia

	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12	G13	G14
G1	1	0.9	0.9	0.9	0.9	0.8	0.9	0.1	0.9	0.1	0.1	0.8	0.2	0.2
G2	0.9	1	0.9	0.3	0.3	0.7	0.0	0.5	0.3	0.1	0.1	0.2	0.4	0.3
G3	0.9	0.9	1	0.9	0.0	0.2	0.5	0.7	0.6	0.5	0.2	0.6	0.1	0.0
G4	0.9	0.3	0.9	1	0.5	0.3	0.6	0.3	0.0	0.5	0.1	0.2	0.2	0.6
G5	0.9	0.3	0.0	0.5	1	0.1	0.6	0.1	0.3	0.3	0.3	0.5	0.2	0.5
G6	0.8	0.7	0.2	0.3	0.1	1	0.5	0.2	0.1	0.1	0.5	0.3	0.1	0.1
G7	0.9	0.0	0.5	0.6	0.6	0.9	1	0.3	0.1	0.5	0.1	0.3	0.5	0.2
G8	0.1	0.5	0.7	0.3	0.1	0.2	0.3	1	0.9	0.9	0.9	0.8	0.8	0.9
G9	0.9	0.3	0.6	0.0	0.3	0.1	0.1	0.9	1	0.8	0.1	0.3	0.5	0.3
G10	0.1	0.1	0.5	0.5	0.3	0.1	0.5	0.9	0.8	1	0.8	1.0	0.2	0.3
G11	0.1	0.1	0.2	0.1	0.3	0.5	0.1	0.9	0.1	0.8	1	0.5	0.8	0.9
G12	0.8	0.2	0.6	0.2	0.5	0.3	0.3	0.8	0.3	1.0	0.5	1	0.8	0.1
G13	0.2	0.4	0.1	0.2	0.2	0.1	0.5	0.8	0.5	0.2	0.8	0.8	1	0.9
G14	0.2	0.3	0.0	0.6	0.5	0.1	0.2	0.9	0.3	0.3	0.9	0.1	0.9	1

- **Paso 3:** Construcción de la red: determinación de la matriz de adyacencia.

## Visualización de la Red

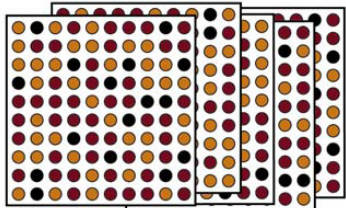


- **Paso 4:** Visualización de la red.

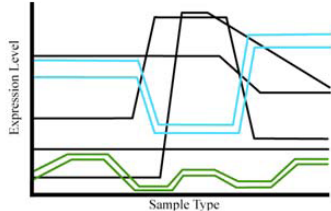
# Paso 1: Análisis de datos transcriptómicos masivos.

## Análisis de la expresión diferencial

### Datos Transcriptómicos



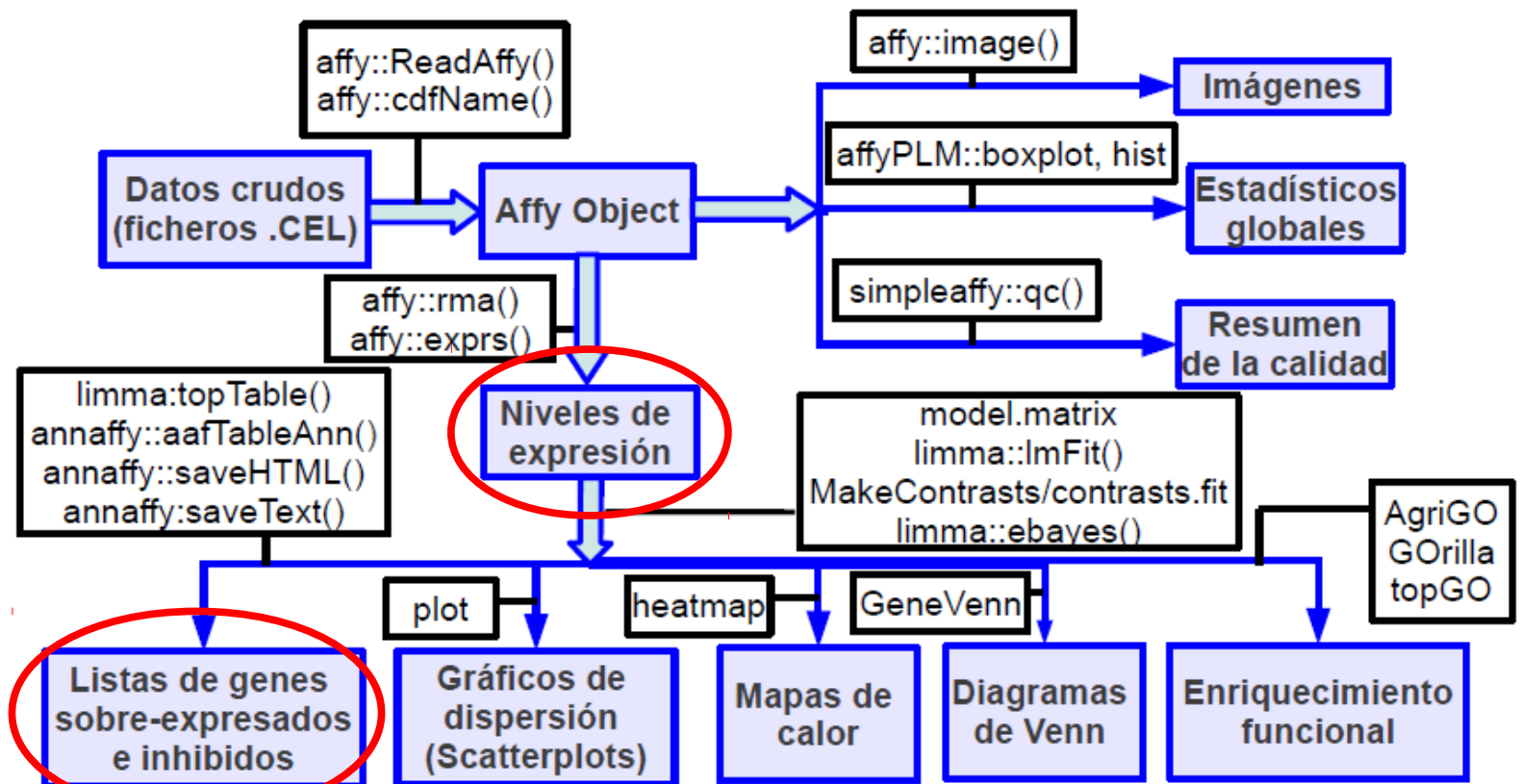
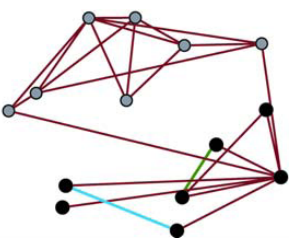
### Análisis de la correlación



### Matriz de Adyacencia

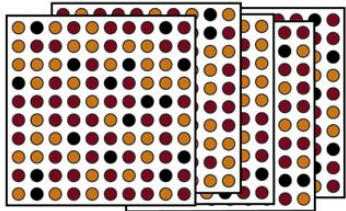
	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12	G13	G14
G1	1	0.9	0.9	0.9	0.9	0.8	0.9	0.1	0.9	0.1	0.1	0.8	0.2	0.2
G2	0.9	1	0.9	0.3	0.3	0.7	0.0	0.5	0.3	0.1	0.1	0.2	0.4	0.3
G3	0.9	0.9	1	0.9	0.0	0.2	0.5	0.7	0.6	0.5	0.2	0.6	0.1	0.0
G4	0.9	0.3	0.9	1	0.5	0.3	0.6	0.3	0.0	0.5	0.1	0.2	0.2	0.6
G5	0.9	0.3	0.0	0.5	1	0.1	0.6	0.1	0.3	0.3	0.3	0.5	0.2	0.5
G6	0.8	0.7	0.2	0.3	0.1	1	0.5	0.2	0.1	0.1	0.5	0.3	0.1	0.1
G7	0.9	0.0	0.5	0.6	0.6	0.9	1	0.3	0.1	0.5	0.1	0.3	0.5	0.2
G8	0.1	0.5	0.7	0.3	0.1	0.2	0.3	1	0.9	0.9	0.8	0.8	0.8	0.9
G9	0.9	0.3	0.6	0.0	0.3	0.1	0.1	0.9	1	0.8	0.1	0.3	0.5	0.3
G10	0.1	0.1	0.5	0.5	0.3	0.1	0.5	0.9	0.8	1	0.8	1.0	0.2	0.3
G11	0.1	0.1	0.2	0.1	0.3	0.5	0.1	0.9	0.1	0.8	1	0.5	0.8	0.9
G12	0.8	0.2	0.6	0.2	0.5	0.3	0.3	0.8	0.3	1.0	0.5	1	0.8	0.1
G13	0.2	0.4	0.1	0.2	0.2	0.1	0.5	0.8	0.5	0.2	0.8	0.8	1	0.9
G14	0.2	0.3	0.0	0.6	0.5	0.1	0.2	0.9	0.3	0.3	0.9	0.1	0.9	1

### Visualización de la Red

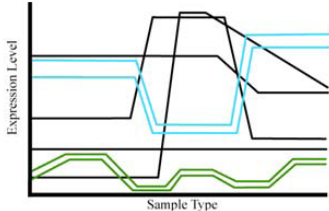


## Paso 2: Análisis de la correlación entre los perfiles de expresión

### Datos Transcriptómicos



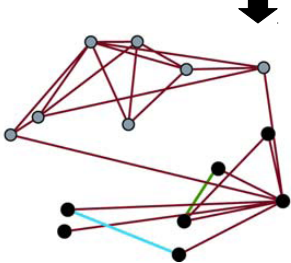
### Análisis de la correlación



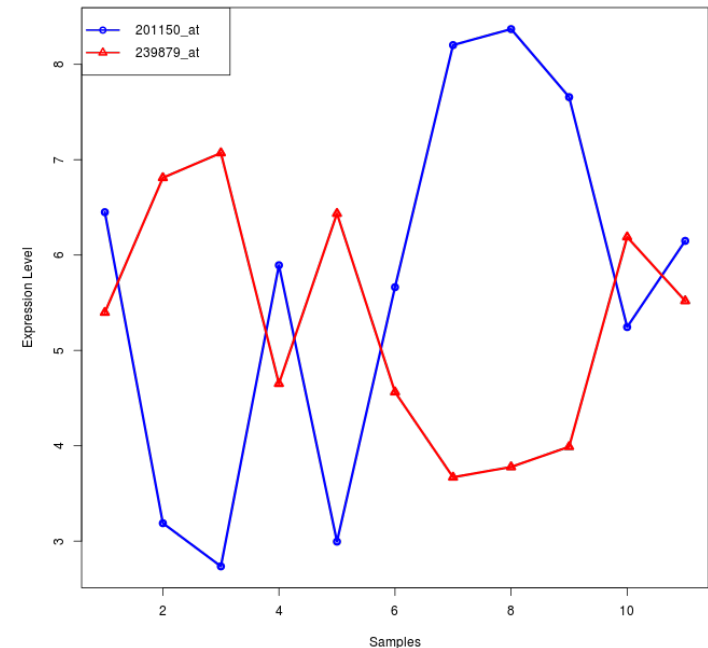
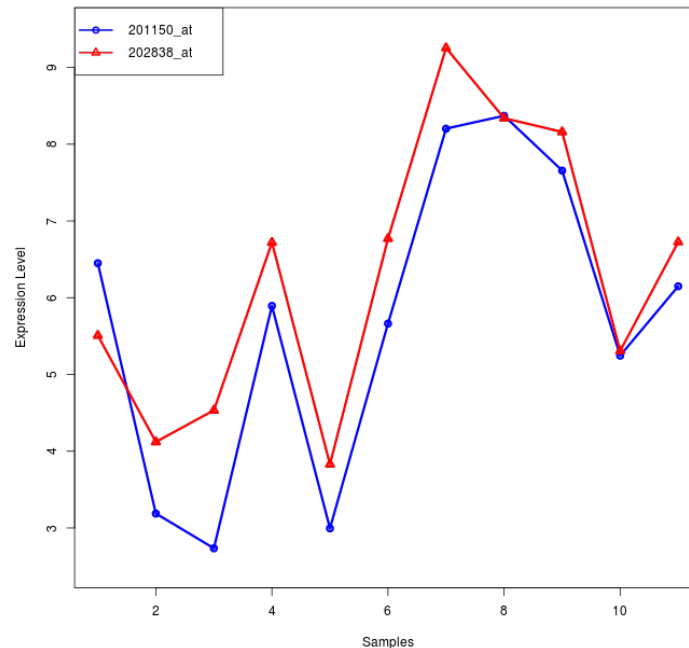
### Matriz de Adyacencia

	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12	G13	G14
G1	1	0.9	0.9	0.9	0.9	0.8	0.9	0.1	0.9	0.1	0.1	0.8	0.2	0.2
G2	0.9	1	0.9	0.3	0.3	0.7	0.0	0.5	0.3	0.1	0.1	0.2	0.4	0.3
G3	0.9	0.9	1	0.9	0.0	0.2	0.5	0.7	0.6	0.5	0.2	0.6	0.1	0.0
G4	0.9	0.3	0.9	1	0.5	0.3	0.6	0.3	0.0	0.5	0.1	0.2	0.2	0.6
G5	0.9	0.3	0.0	0.5	1	0.1	0.6	0.1	0.3	0.3	0.3	0.5	0.2	0.5
G6	0.8	0.7	0.2	0.3	0.1	1	0.5	0.2	0.1	0.1	0.5	0.3	0.1	0.1
G7	0.9	0.0	0.5	0.6	0.6	0.9	1	0.3	0.1	0.5	0.1	0.3	0.5	0.2
G8	0.1	0.5	0.7	0.3	0.1	0.2	0.3	1	0.9	0.9	0.9	0.8	0.8	0.9
G9	0.9	0.3	0.6	0.0	0.3	0.1	0.1	0.9	1	0.8	0.1	0.3	0.5	0.3
G10	0.1	0.1	0.5	0.5	0.3	0.1	0.5	0.9	0.8	1	0.8	1.0	0.2	0.3
G11	0.1	0.1	0.2	0.1	0.3	0.5	0.1	0.9	0.1	0.8	1	0.5	0.8	0.9
G12	0.8	0.2	0.6	0.2	0.5	0.3	0.3	0.8	0.3	1.0	0.5	1	0.8	0.1
G13	0.2	0.4	0.1	0.2	0.2	0.1	0.5	0.8	0.5	0.2	0.8	0.8	1	0.9
G14	0.2	0.3	0.0	0.6	0.5	0.1	0.2	0.9	0.3	0.3	0.9	0.1	0.9	1

### Visualización de la Red



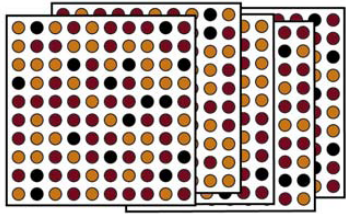
- El criterio seguido para determinar si dos genes se co-expresan en las muestras de los distintos experimentos estudiados se basa usualmente en la **correlación entre sus perfiles de expresión** (niveles de expresión en las distintas muestras).
- Se distingue entre **correlación positiva y negativa**.



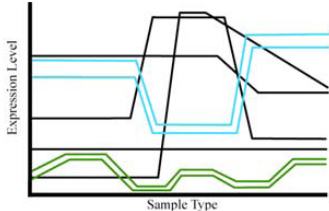


# Paso 2: Análisis de la correlación entre los perfiles de expresión

## Datos Transcriptómicos



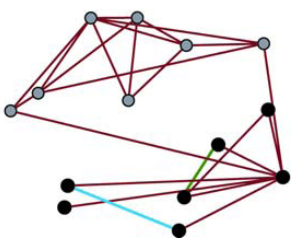
## Análisis de la correlación



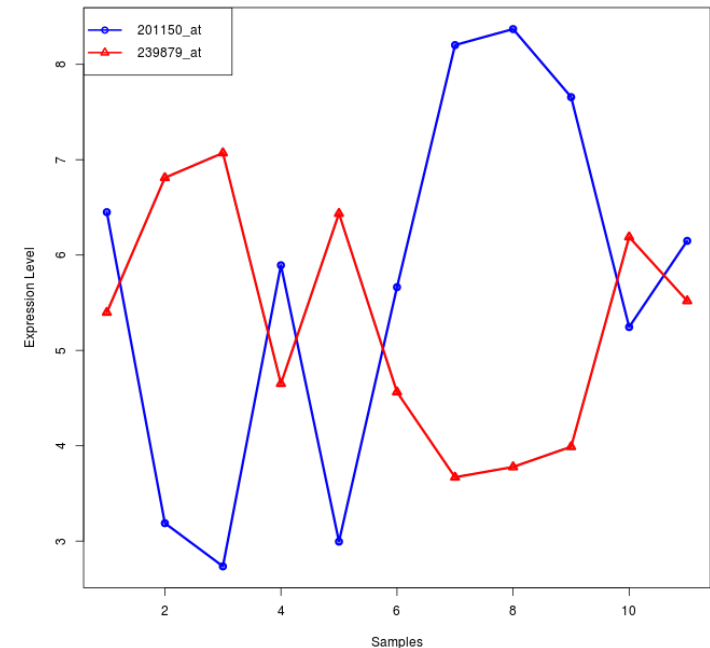
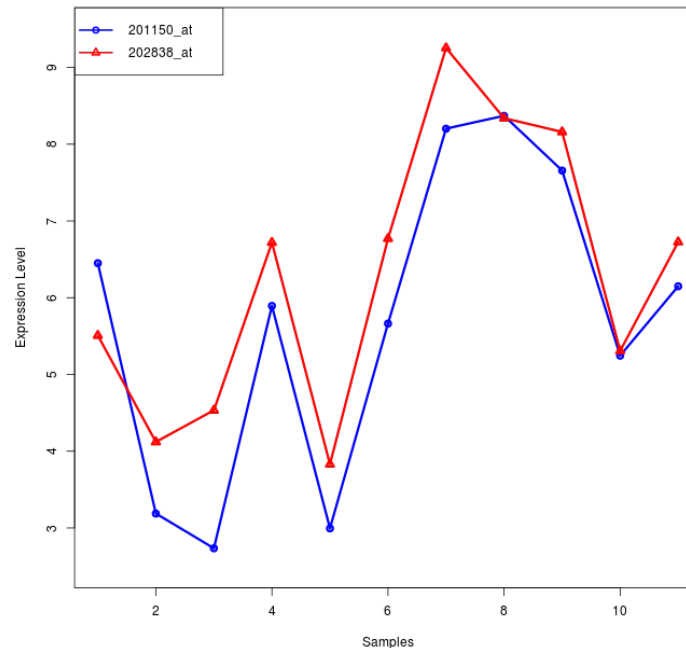
## Matriz de Adyacencia

	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12	G13	G14
G1	1	0.9	0.9	0.9	0.8	0.9	0.1	0.9	0.1	0.1	0.8	0.2	0.2	
G2	0.9	1	0.9	0.3	0.3	0.7	0.0	0.5	0.3	0.1	0.1	0.2	0.4	0.3
G3	0.9	0.9	1	0.9	0.0	0.2	0.5	0.7	0.6	0.5	0.2	0.6	0.1	0.0
G4	0.9	0.3	0.9	1	0.5	0.3	0.6	0.3	0.0	0.5	0.1	0.2	0.2	0.6
G5	0.9	0.3	0.0	0.5	1	0.1	0.6	0.1	0.3	0.3	0.3	0.5	0.2	0.5
G6	0.8	0.7	0.2	0.3	0.1	1	0.5	0.2	0.1	0.1	0.5	0.3	0.1	0.1
G7	0.9	0.0	0.5	0.6	0.6	0.9	1	0.3	0.1	0.5	0.1	0.3	0.5	0.2
G8	0.1	0.5	0.7	0.3	0.1	0.2	0.3	1	0.9	0.9	0.9	0.8	0.8	0.9
G9	0.9	0.3	0.6	0.0	0.3	0.1	0.1	0.9	1	0.8	0.1	0.3	0.5	0.3
G10	0.1	0.1	0.5	0.5	0.3	0.1	0.5	0.9	0.8	1	0.8	1.0	0.2	0.3
G11	0.1	0.1	0.2	0.1	0.3	0.5	0.1	0.9	0.1	0.8	1	0.5	0.8	0.9
G12	0.8	0.2	0.6	0.2	0.5	0.3	0.3	0.8	0.3	1.0	0.5	1	0.8	0.1
G13	0.2	0.4	0.1	0.2	0.2	0.1	0.5	0.8	0.5	0.2	0.8	0.8	1	0.9
G14	0.2	0.3	0.0	0.6	0.5	0.1	0.2	0.9	0.3	0.3	0.9	0.1	0.9	1

## Visualización de la Red

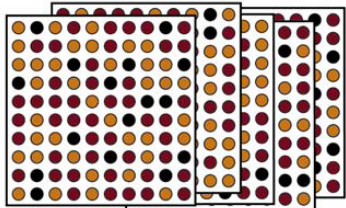


$$coexpr(g_1, g_2) = \begin{cases} cor(per(g_1), per(g_2)) \\ |cor(per(g_1), per(g_2))| \\ |cor(per(g_1), per(g_2))|^\beta \end{cases}$$

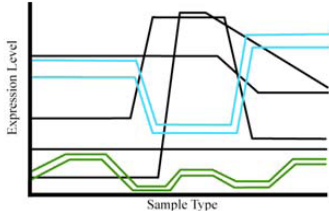


# Paso 3: Construcción de la red: determinación de la matriz de adyacencia.

## Datos Transcriptómicos



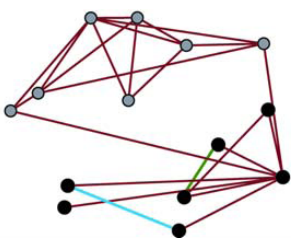
## Análisis de la correlación



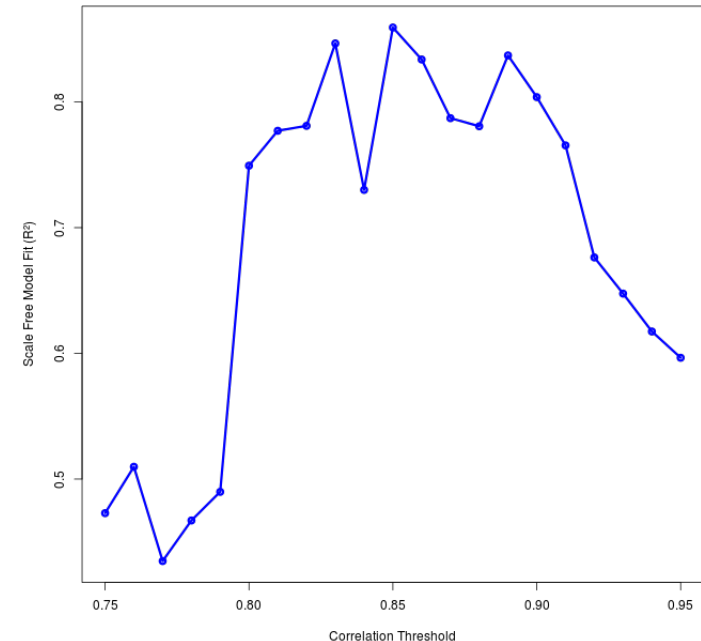
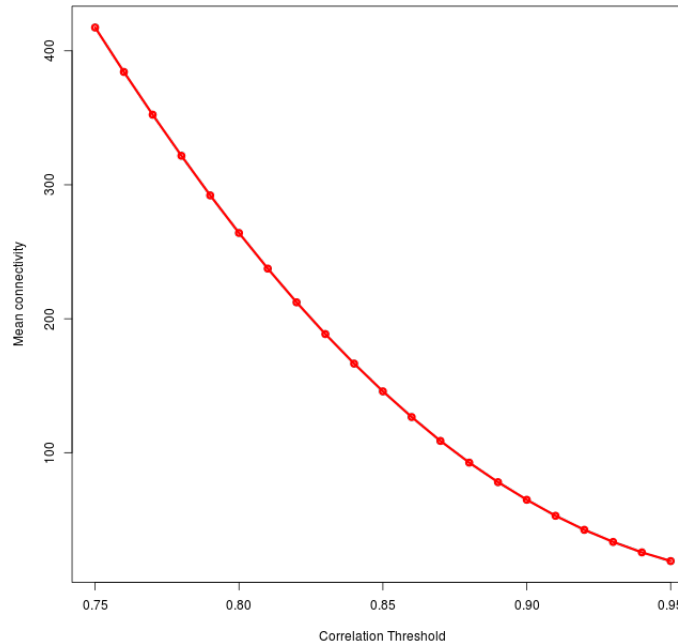
## Matriz de Adyacencia

	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12	G13	G14
G1	1	0.9	0.9	0.9	0.9	0.8	0.9	0.1	0.9	0.1	0.1	0.8	0.2	0.2
G2	0.9	1	0.9	0.3	0.3	0.7	0.0	0.5	0.3	0.1	0.1	0.2	0.4	0.3
G3	0.9	0.9	1	0.9	0.0	0.2	0.5	0.7	0.6	0.5	0.2	0.6	0.1	0.0
G4	0.9	0.3	0.9	1	0.5	0.3	0.6	0.3	0.0	0.5	0.1	0.2	0.2	0.6
G5	0.9	0.3	0.0	0.5	1	0.1	0.6	0.1	0.3	0.3	0.3	0.5	0.2	0.5
G6	0.8	0.7	0.2	0.3	0.1	1	0.5	0.2	0.1	0.1	0.5	0.3	0.1	0.1
G7	0.9	0.0	0.5	0.6	0.6	0.9	1	0.3	0.1	0.5	0.1	0.3	0.5	0.2
G8	0.1	0.5	0.7	0.3	0.1	0.2	0.3	1	0.9	0.9	0.9	0.8	0.8	0.9
G9	0.9	0.3	0.6	0.0	0.3	0.1	0.1	0.9	1	0.8	0.1	0.3	0.5	0.3
G10	0.1	0.1	0.5	0.5	0.3	0.1	0.5	0.9	0.8	1	0.8	1.0	0.2	0.3
G11	0.1	0.1	0.2	0.1	0.3	0.5	0.1	0.9	0.1	0.8	1	0.5	0.8	0.9
G12	0.8	0.2	0.6	0.2	0.5	0.3	0.3	0.8	0.3	1.0	0.5	1	0.8	0.1
G13	0.2	0.4	0.1	0.2	0.2	0.1	0.5	0.8	0.5	0.2	0.8	0.8	1	0.9
G14	0.2	0.3	0.0	0.6	0.5	0.1	0.2	0.9	0.3	0.3	0.9	0.1	0.9	1

## Visualización de la Red



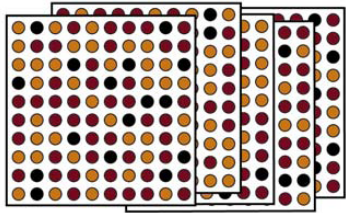
- El paso crítico en la construcción de un red de co-expresión génica consiste en **seleccionar el umbral de corte**, el valor específico de correlación que asumimos es lo suficientemente alto para suponer que ambos genes se coexpresan.
- Usualmente se busca un **compromiso entre lograr un red libre de escala y una alta conectividad**.



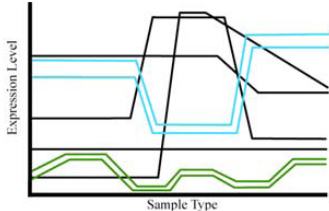


# Paso 4: Visualización de la Red

## Datos Transcriptómicos



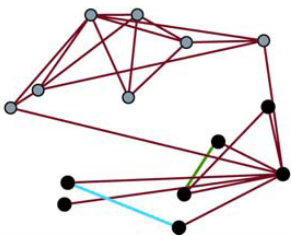
## Análisis de la correlación



## Matriz de Adyacencia

	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12	G13	G14
G1	1	0.9	0.9	0.9	0.9	0.8	0.9	0.1	0.9	0.1	0.1	0.8	0.2	0.2
G2	0.9	1	0.9	0.3	0.3	0.7	0.0	0.5	0.3	0.1	0.1	0.2	0.4	0.3
G3	0.9	0.9	1	0.9	0.0	0.2	0.5	0.7	0.6	0.5	0.2	0.6	0.1	0.0
G4	0.9	0.3	0.9	1	0.5	0.3	0.6	0.3	0.0	0.5	0.1	0.2	0.2	0.6
G5	0.9	0.3	0.0	0.5	1	0.1	0.6	0.1	0.3	0.3	0.5	0.2	0.5	
G6	0.8	0.7	0.2	0.3	0.1	1	0.5	0.2	0.1	0.1	0.5	0.3	0.1	0.1
G7	0.9	0.0	0.5	0.6	0.6	0.9	1	0.3	0.1	0.5	0.1	0.3	0.5	0.2
G8	0.1	0.5	0.7	0.3	0.1	0.2	0.3	1	0.9	0.9	0.9	0.8	0.8	0.9
G9	0.9	0.3	0.6	0.0	0.3	0.1	0.1	0.9	1	0.8	0.1	0.3	0.5	0.3
G10	0.1	0.1	0.5	0.5	0.3	0.1	0.5	0.9	0.8	1	0.8	1.0	0.2	0.3
G11	0.1	0.1	0.2	0.1	0.3	0.5	0.1	0.9	0.1	0.8	1	0.5	0.8	0.9
G12	0.8	0.2	0.6	0.2	0.5	0.3	0.3	0.8	0.3	1.0	0.5	1	0.8	0.1
G13	0.2	0.4	0.1	0.2	0.2	0.1	0.5	0.8	0.5	0.2	0.8	0.8	1	0.9
G14	0.2	0.3	0.0	0.6	0.5	0.1	0.2	0.9	0.3	0.3	0.9	0.1	0.9	1

## Visualización de la Red



- Existen diferentes herramientas para la visualización de redes. En esta asignatura utilizaremos **Cytoscape**.
- El formato estándar más simple de especificación de una red que admite **cytoscape** consiste en el **formato gml**.

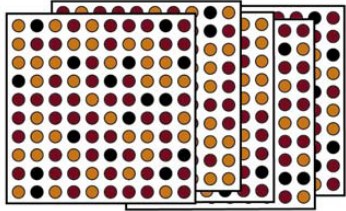
**File → Import → Network → File → fichero.gml**

- Existen diferentes algoritmos para la organización visual de redes, por ejemplo, **organic**, **spring**, **spring-weighted** etc.

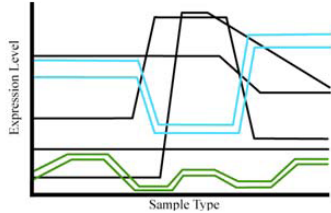
**Layout → yFiles Layouts → Organic**  
**VizMapper → Current Visual Style → Solid**

# Paso 4: Visualización de la Red

## Datos Transcriptómicos



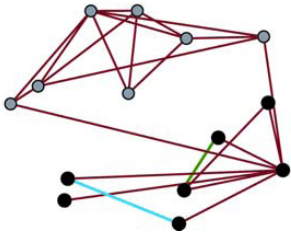
## Análisis de la correlación



## Matriz de Adyacencia

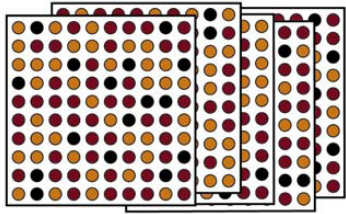
	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12	G13	G14
G1	1	0.9	0.9	0.9	0.9	0.8	0.9	0.1	0.9	0.1	0.1	0.8	0.2	0.2
G2	0.9	1	0.9	0.3	0.3	0.7	0.0	0.5	0.3	0.1	0.1	0.2	0.4	0.3
G3	0.9	0.9	1	0.9	0.0	0.2	0.5	0.7	0.6	0.5	0.2	0.6	0.1	0.0
G4	0.9	0.3	0.9	1	0.5	0.3	0.6	0.3	0.0	0.5	0.1	0.2	0.2	0.6
G5	0.9	0.3	0.0	0.5	1	0.1	0.6	0.1	0.3	0.3	0.3	0.5	0.2	0.5
G6	0.8	0.7	0.2	0.3	0.1	1	0.5	0.2	0.1	0.1	0.5	0.3	0.1	0.1
G7	0.9	0.0	0.5	0.6	0.6	0.9	1	0.3	0.1	0.5	0.1	0.3	0.5	0.2
G8	0.1	0.5	0.7	0.3	0.1	0.2	0.3	1	0.9	0.9	0.9	0.8	0.8	0.9
G9	0.9	0.3	0.6	0.0	0.3	0.1	0.1	0.9	1	0.8	0.1	0.3	0.5	0.3
G10	0.1	0.1	0.5	0.5	0.3	0.1	0.5	0.9	0.8	1	0.8	1.0	0.2	0.3
G11	0.1	0.1	0.2	0.1	0.3	0.5	0.1	0.9	0.1	0.8	1	0.5	0.8	0.9
G12	0.8	0.2	0.6	0.2	0.5	0.3	0.3	0.8	0.3	1.0	0.5	1	0.8	0.1
G13	0.2	0.4	0.1	0.2	0.2	0.1	0.5	0.8	0.5	0.2	0.8	0.8	1	0.9
G14	0.2	0.3	0.0	0.6	0.5	0.1	0.2	0.9	0.3	0.3	0.9	0.1	0.9	1

## Visualización de la Red

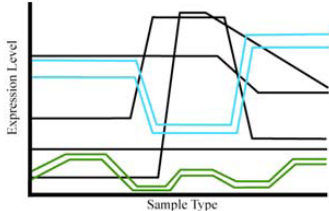


# Análisis de Redes de Co-expresión Génica

## Datos Transcriptómicos



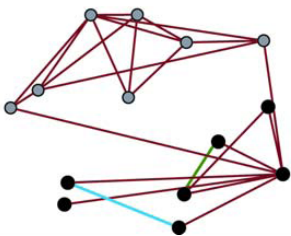
## Análisis de la correlación



## Matriz de Adyacencia

	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12	G13	G14
G1	1	0.9	0.9	0.9	0.9	0.8	0.9	0.1	0.9	0.1	0.1	0.8	0.2	0.2
G2	0.9	1	0.9	0.3	0.3	0.7	0.0	0.5	0.3	0.1	0.1	0.2	0.4	0.3
G3	0.9	0.9	1	0.9	0.0	0.2	0.5	0.7	0.6	0.5	0.2	0.6	0.1	0.0
G4	0.9	0.3	0.9	1	0.5	0.3	0.6	0.3	0.0	0.5	0.1	0.2	0.2	0.6
G5	0.9	0.3	0.0	0.5	1	0.1	0.6	0.1	0.3	0.3	0.3	0.5	0.2	0.5
G6	0.8	0.7	0.2	0.3	0.1	1	0.5	0.2	0.1	0.1	0.5	0.3	0.1	0.1
G7	0.9	0.0	0.5	0.6	0.6	0.9	1	0.3	0.1	0.5	0.1	0.3	0.5	0.2
G8	0.1	0.5	0.7	0.3	0.1	0.2	0.3	1	0.9	0.9	0.9	0.8	0.8	0.9
G9	0.9	0.3	0.6	0.0	0.3	0.1	0.1	0.9	1	0.8	0.1	0.3	0.5	0.3
G10	0.1	0.1	0.5	0.5	0.3	0.1	0.5	0.9	0.8	1	0.8	1.0	0.2	0.3
G11	0.1	0.1	0.2	0.1	0.3	0.5	0.1	0.9	0.1	0.8	1	0.5	0.8	0.9
G12	0.8	0.2	0.6	0.2	0.5	0.3	0.3	0.8	0.3	1.0	0.5	1	0.8	0.1
G13	0.2	0.4	0.1	0.2	0.2	0.1	0.5	0.8	0.5	0.2	0.8	0.8	1	0.9
G14	0.2	0.3	0.0	0.6	0.5	0.1	0.2	0.9	0.3	0.3	0.9	0.1	0.9	1

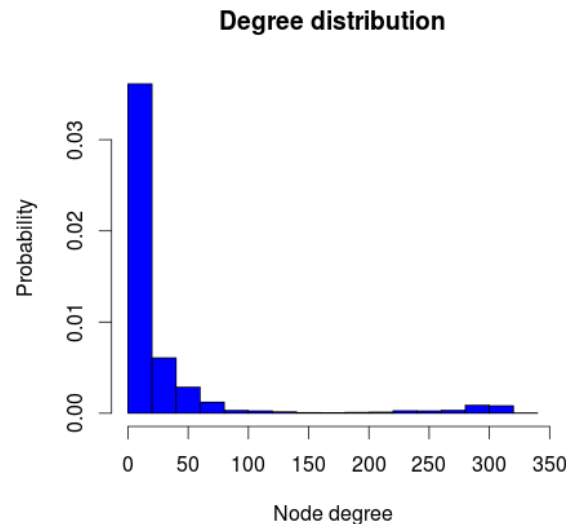
## Visualización de la Red



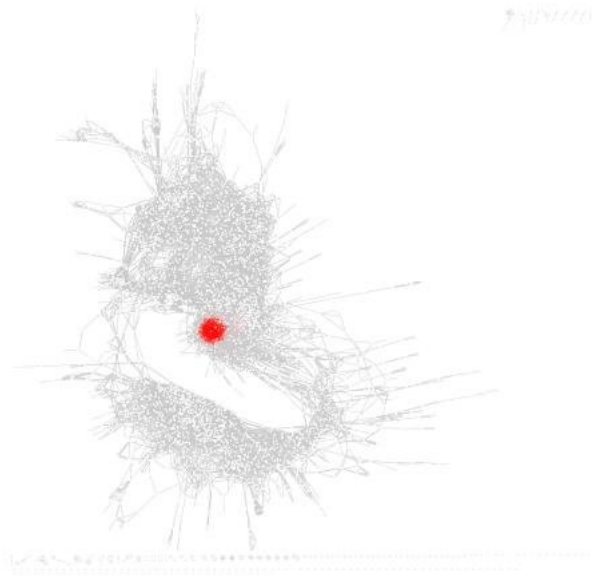
- Existen multitud de técnicas para el análisis de redes de co-expresión.
- Como introducción en esta asignatura nos centraremos en:
  - **Análisis de la topología** (estructura de conectividad de la red).
  - Búsqueda de patrones globales mediante **técnicas de clustering**.
  - Enriquecimiento de **términos de ontología de genes**.

# Análisis de la Topología de la Red

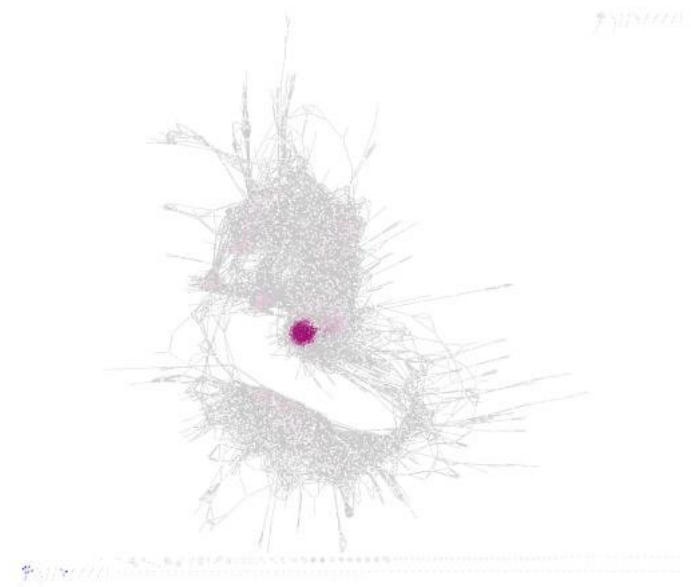
- Este análisis se centra en el estudio de **propiedades topológicas locales** tales como el *grado de un nodo* y el *coeficiente de agrupamiento* así como de las correspondientes extensiones a **propiedades globales**, *distribución del grado de nodos* o *coeficiente de agrupamiento medio*.
- En este apartado también se determina si la red construída es **libre de escala**, de **mundo pequeño** y se analizan los **hubs** de la red.



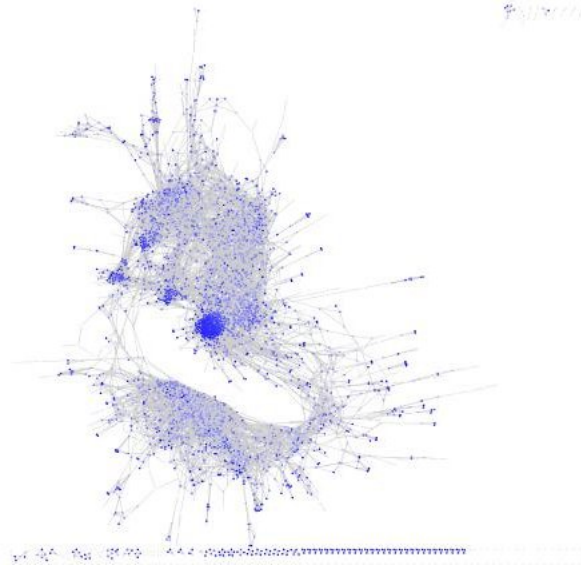
# Análisis de la Topología de la Red



**Hub scores**



**Node degree**



**Clustering coefficient**

# Identificación de Patrones Globales: Clustering

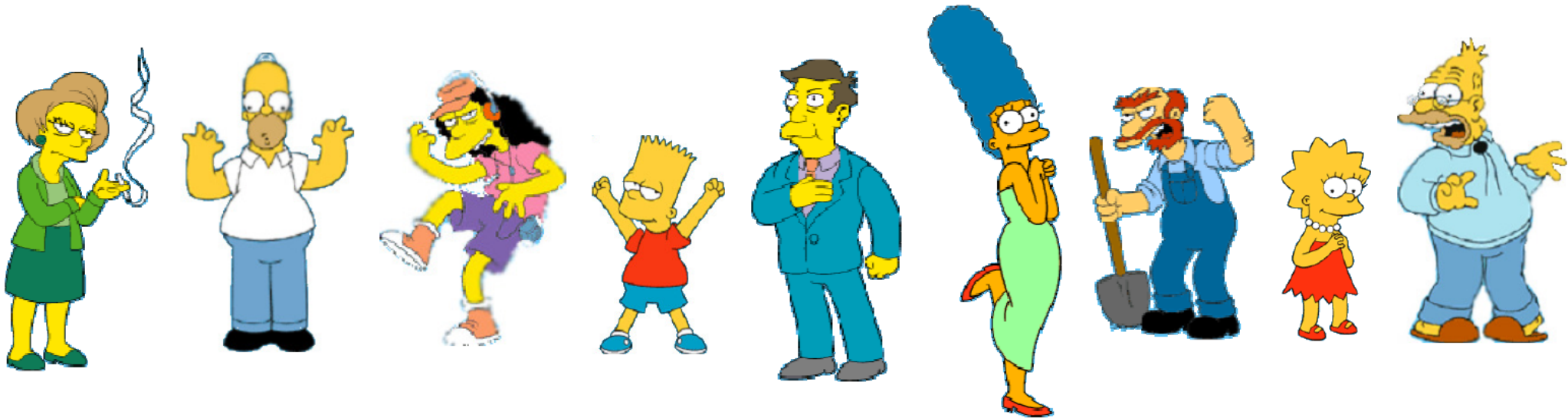
- **Clustering** es una técnica de **minería de datos (data mining)** dentro de la disciplina de *Inteligencia Artificial* que identifica de forma *automática* agrupaciones o clústeres de elementos de acuerdo a una medida de similitud entre ellos.
- El **objetivo fundamental** de las técnicas de clustering consiste en identificar grupos o clústeres de elementos tal que:
  - La similitud media entre elementos del mismo clúster sea alta. **Similitud intra-clúster alta.**
  - La similitud media entre elementos de distintos clústeres sea baja. **Similitud inter-clúster baja.**

# Identificación de Patrones Globales: Clustering

- Las distintas técnicas de clustering tienen una gran diversidad de aplicaciones:
  - Revelación la estructura interna de los datos analizados según sus características.
  - Procesamiento de datos previo a técnicas de análisis más complejas tales como la identificación de marcadores génicos.
  - Asignación de funciones a genes desconocidos.
  - Estudios de enfermedades complejas.
  - Estudios evolutivos
  - Etc.



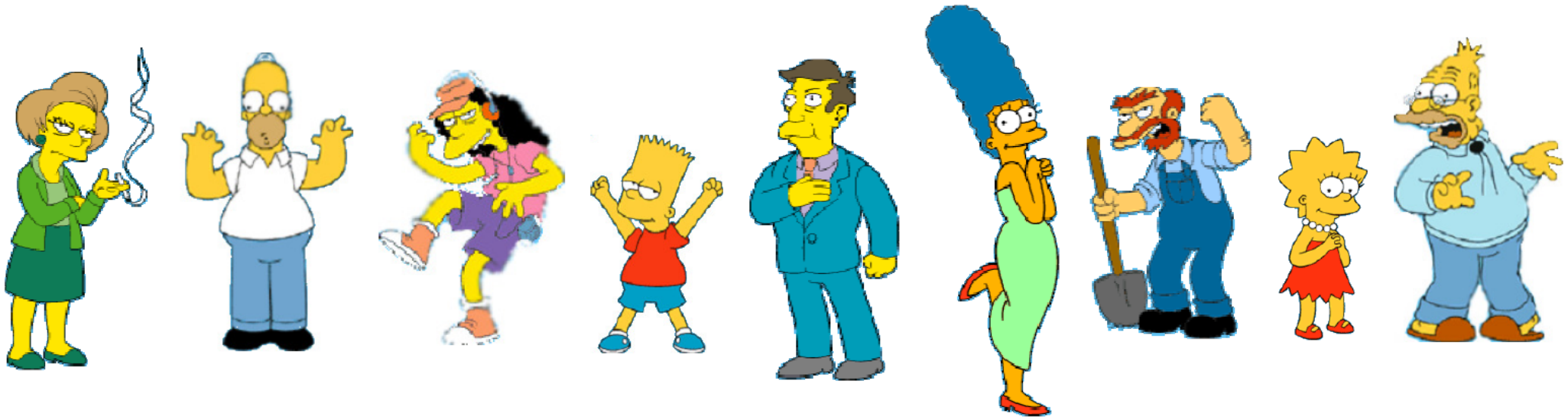
# Elección de una Medida de Similitud



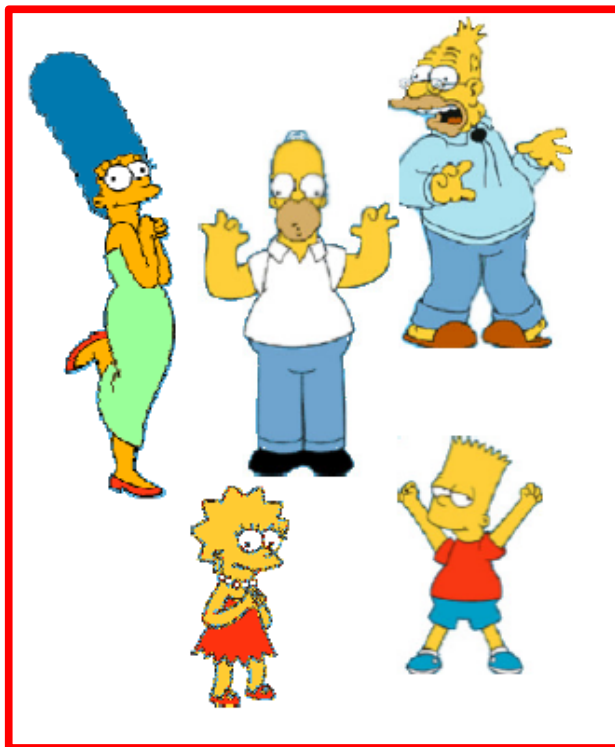
- La identificación de clústeres o grupos de elementos se basa en una medida de similitud. Diferentes medidas de similitud dan lugar a diferentes clústeres.



# Elección de una Medida de Similitud



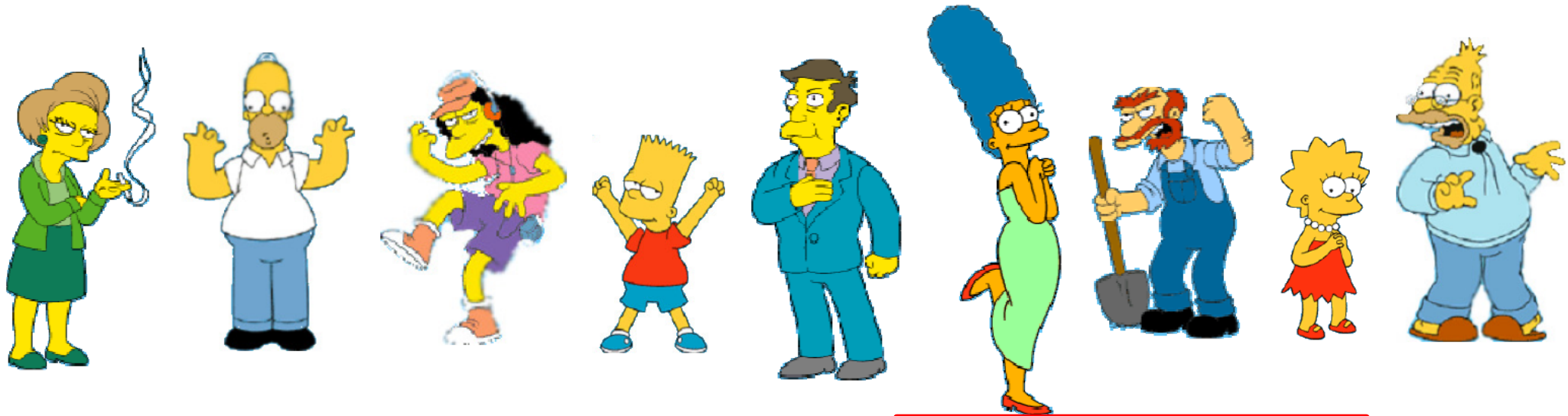
Familia  
Simpson



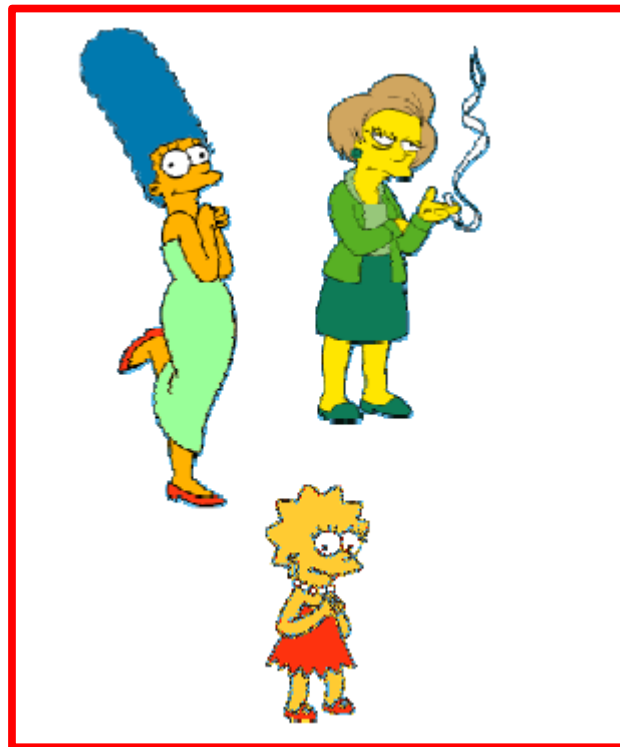
Empleados  
del  
colegio



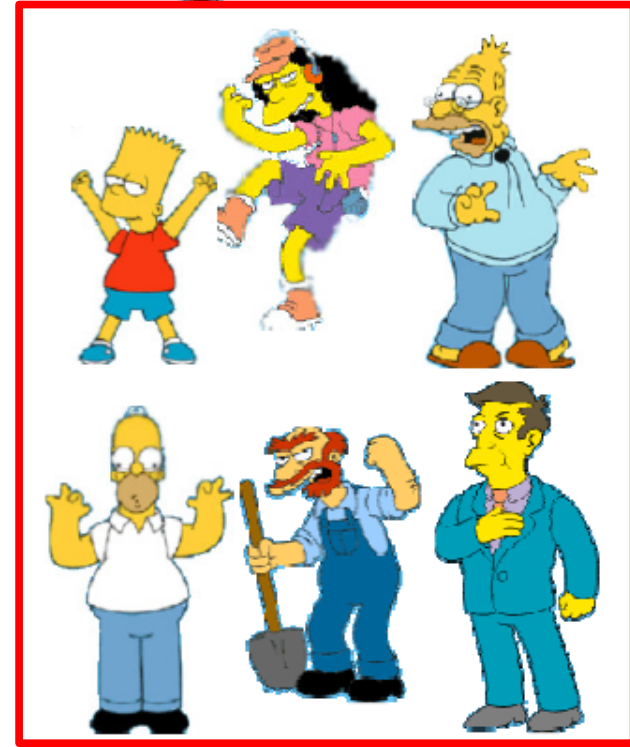
# Elección de una Medida de Similitud



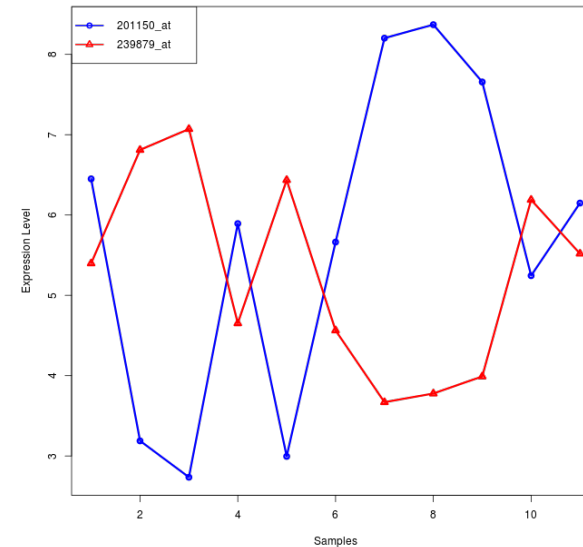
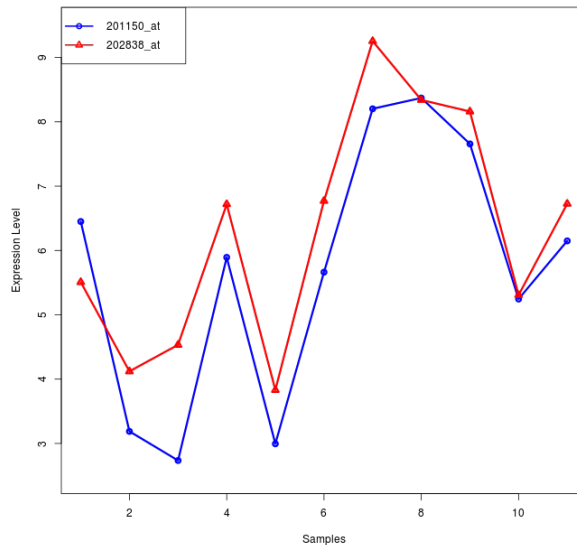
Mujeres



Hombres



# Elección de una Medida de Similitud



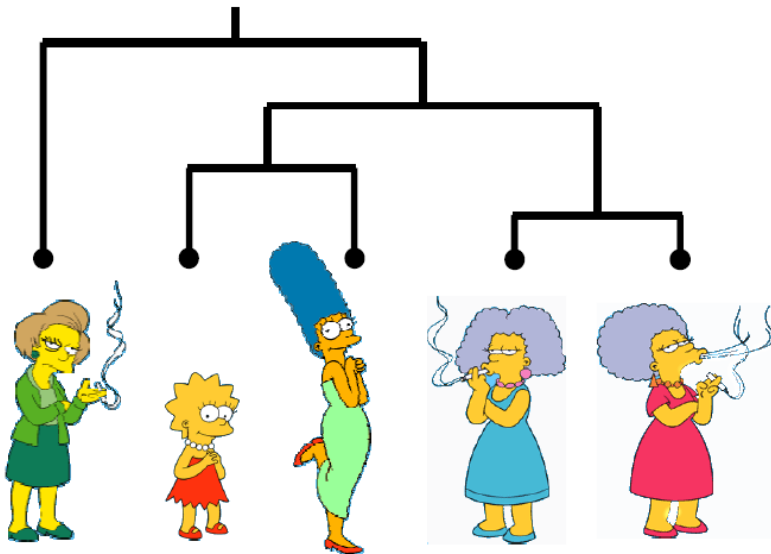
- En redes de co-expresión génica una de las posibles medidas de similitud que se utilizan con mayor frecuencia está basada en la **correlación de Pearson**:

$$D(g_1, g_2) = 1 - \text{cor}(g_1, g_2)$$

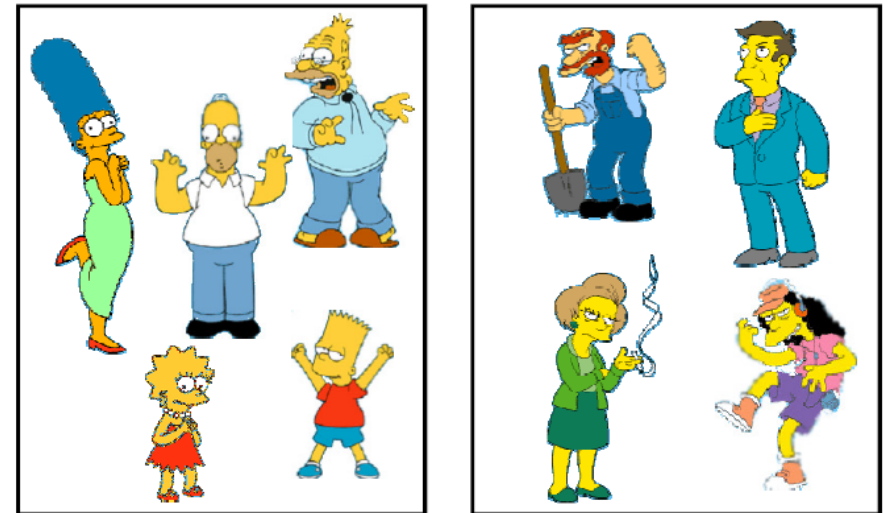
# Elección de una Técnica de Clustering

- Existen principalmente **dos tipos diferentes** de técnicas de clustering:

## Clustering Jerárquico



## Clustering de Partición



# Clustering Jerárquico










- La técnica de clustering jerárquico construye un **dendograma o árbol** que representa las relaciones de similitud entre los distintos elementos.
- Explorar todos los posibles árboles es computacionalmente intratable. Por lo tanto, suelen seguirse **algoritmos aproximados** guiados por determinadas heurísticas.
- Existen dos aproximaciones diferentes al clustering jerárquico:
  - **Clustering jerárquico aglomerativo:** se comienza con tantos clústeres como individuos y consiste en ir formando (aglomerando) grupos según su similitud.
  - **Clustering jerárquico de división:** se comienza con un único clúster y consiste en ir dividiendo clústeres según la disimilitud entre sus componentes.

# Clustering Jerárquico Aglomerativo

Esta técnica comienza con una matriz de similitud que contiene las distancias entre los distintos elementos a agrupar. En nuestro caso esta matriz se calcula a partir de la matriz de correlaciones.

$$D(\text{Marge Simpson}, \text{Lisa Simpson}) = 8$$

$$D(\text{Marge Simpson}, \text{Marge Simpson}) = 1$$










				
0	8	8	7	7
	0	2	4	4
		0	3	3
			0	1
				0

# Clustering Jerárquico Aglomerativo

Esta técnica comienza con una matriz de similitud que contiene las distancias entre los distintos elementos a agrupar. En nuestro caso esta matriz se calcula a partir de la matriz de correlaciones.

$$D(\text{Marge Simpson}, \text{Lisa Simpson}) = 8$$

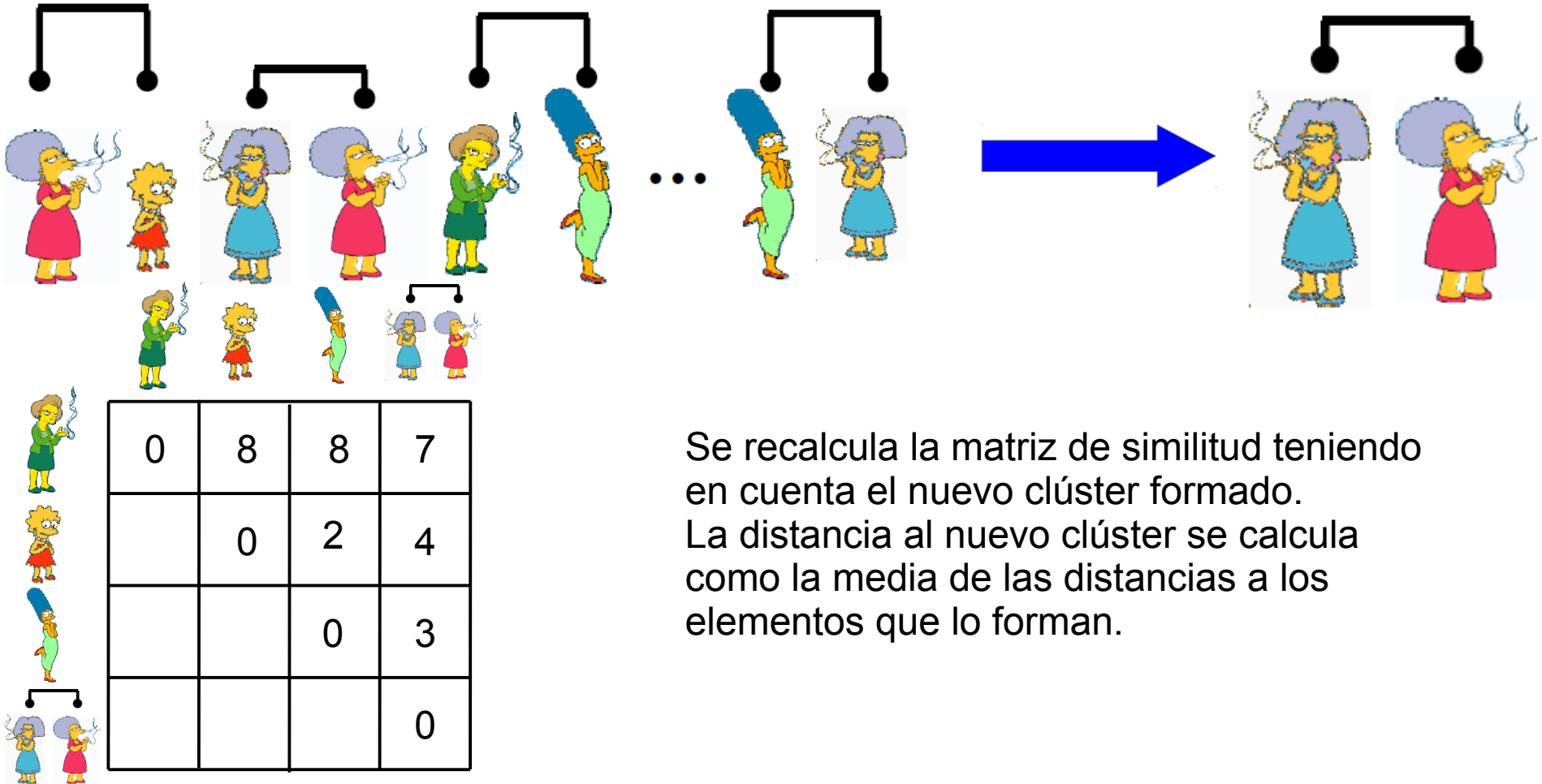
$$D(\text{Marge Simpson}, \text{Marge Simpson}) = 1$$

				
0	8	8	7	7
	0	2	4	4
		0	3	3
			0	1
				0



# Clustering Jerárquico Aglomerativo

Consideramos todas las agrupaciones posibles y elegimos la mejor según la matriz de similitud.



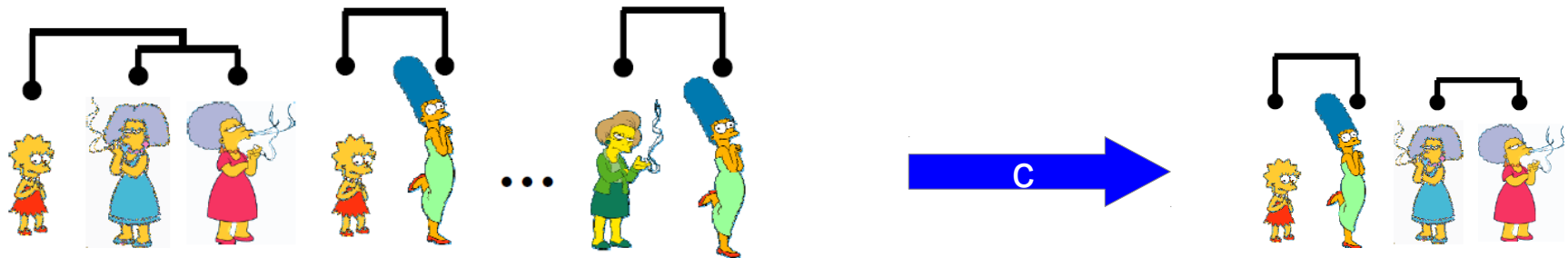
Se recalcula la matriz de similitud teniendo en cuenta el nuevo clúster formado. La distancia al nuevo clúster se calcula como la media de las distancias a los elementos que lo forman.



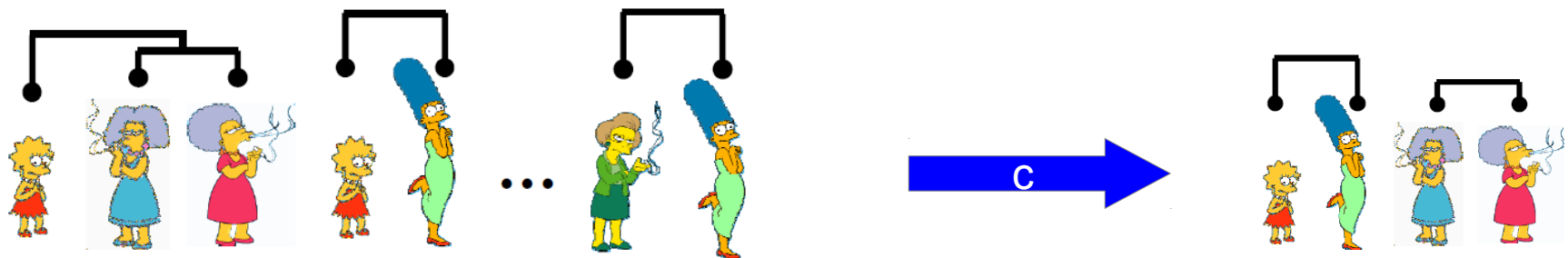
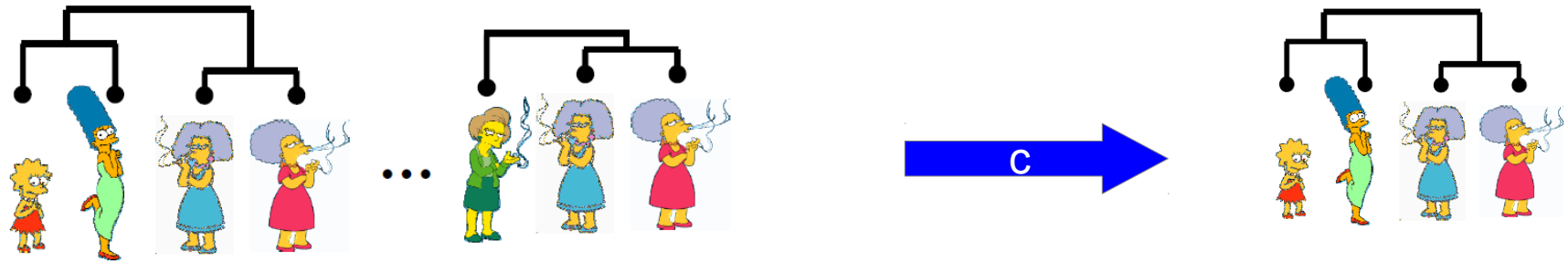
# Clustering Jerárquico Aglomerativo



# Clustering Jerárquico Aglomerativo



# Clustering Jerárquico Aglomerativo



# Ventajas / Desventajas del Clustering Jerárquico

- En el clustering jerárquico no es necesario especificar en número de clústeres a priori. Es posible seleccionarlo a posteriori según un umbral de corte.
- La estructura jerárquica es cercana a la intuición humana.
- La principal desventaja consiste en la acumulación de errores. Errores que se comenten en un paso de agrupamiento se propagan durante el resto de la construcción del dendograma sin ser posible su reajuste.

# Clustering Jerárquico en R

- Utilizaremos como matriz de similitudes o distancias:

$$D(g_1, g_2) = 1 - \text{cor}(g_1, g_2)$$

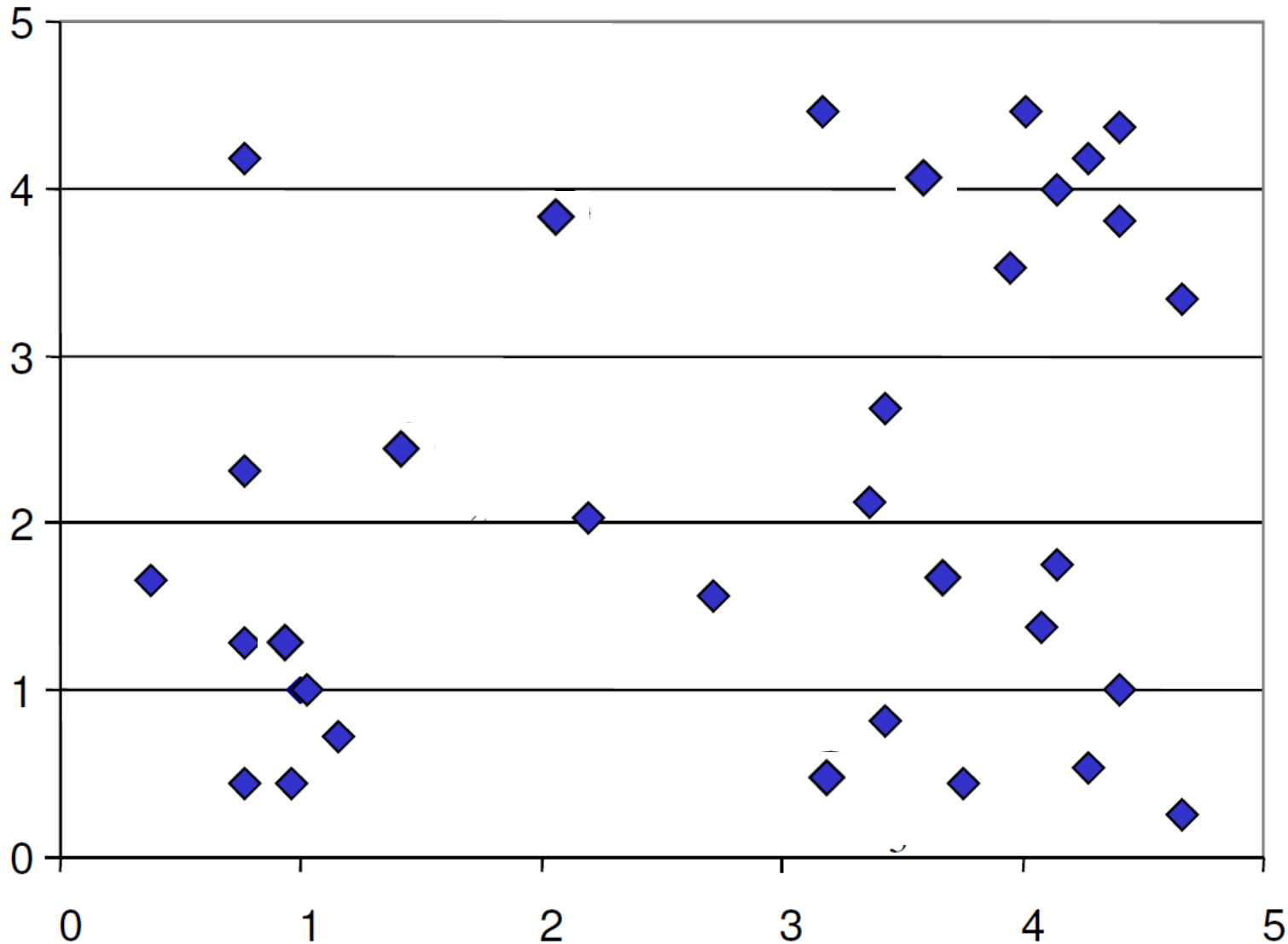
- Los paquetes R a utilizar son **impute** y **WGCNA**.
- La función que realiza el clustering jerárquico se llama **hclust**. Recibe como entrada la matriz de similitudes a usar como distancia (as.dist) y el método para recalcular la matriz de distancias tras cada agrupamiento.
- Para determinar los clústeres formados a un cierto umbral de corte se utiliza la función **cutree** que recibe como entrada el clustering jerárquico, y el número de clústeres a formar.

# Clustering de Partición en torno a Centroides

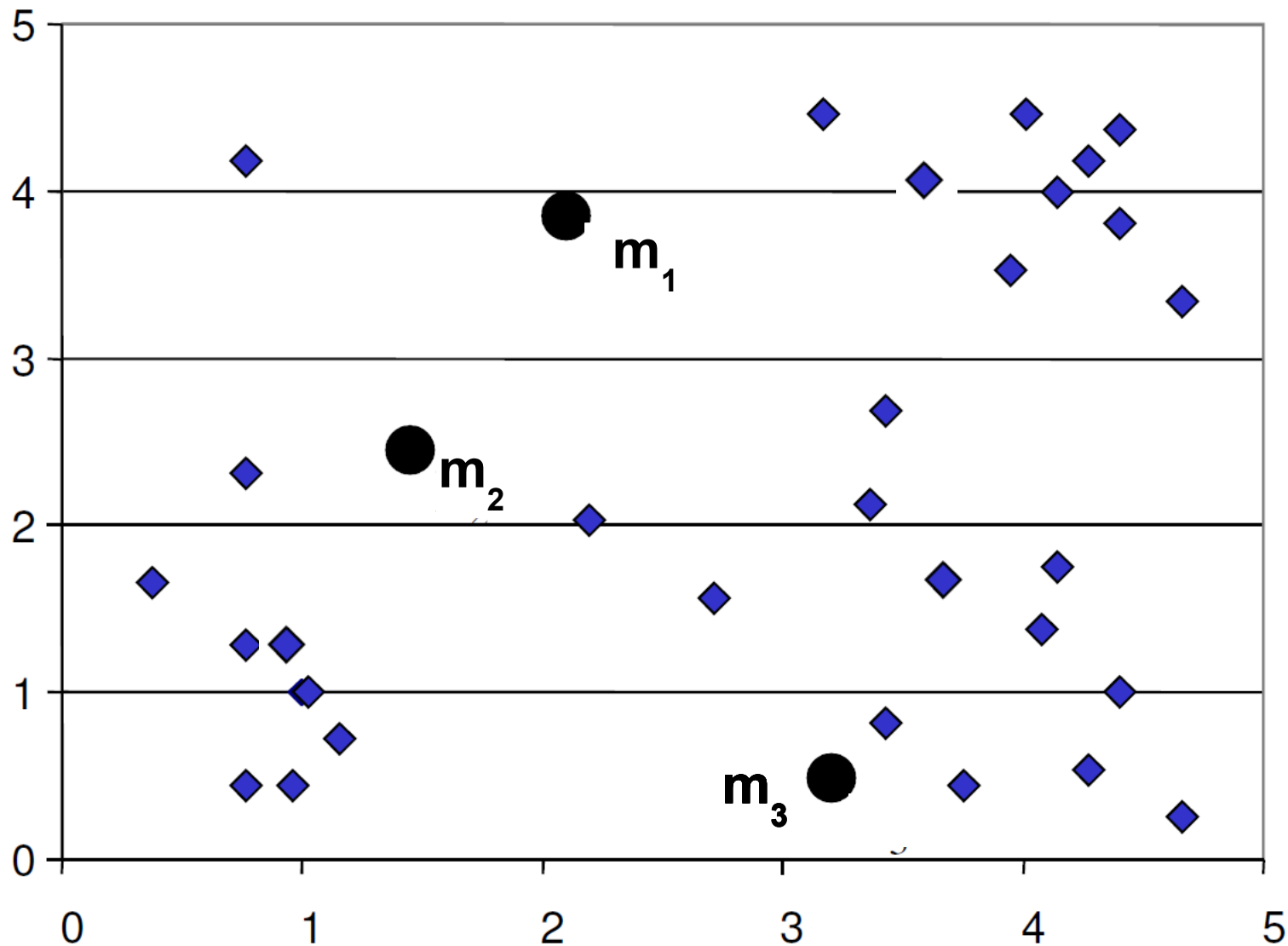
- La técnica de clustering de **partición en torno a centroides (PAM)** realiza una distribución de los elementos entre un número prefijado de clústeres o grupos. Esta técnica recibe como **dato de entrada el número de clústers a formar** además de los elementos a clasificar y la matriz de similitudes.
- Explorar todas las posibles particiones es computacionalmente intratable. Por lo tanto, suelen seguirse **algoritmos aproximados** guiados por determinadas heurísticas.
- En lugar de construir un árbol el objetivo en PAM consiste en agrupar los elementos entorno a **elementos centrales llamados centroides** a cada clúster.
- Definimos el **centroide** de un clúster como aquel elemento que minimiza la suma de las similitudes al resto de los elementos del clúster.

$$m_C = \operatorname{argmin}_{m \in C} \sum_{m_j \in C} \operatorname{dist}(m, m_j)$$

# Paso 1: Seleccionar k centroides aleatoriamente

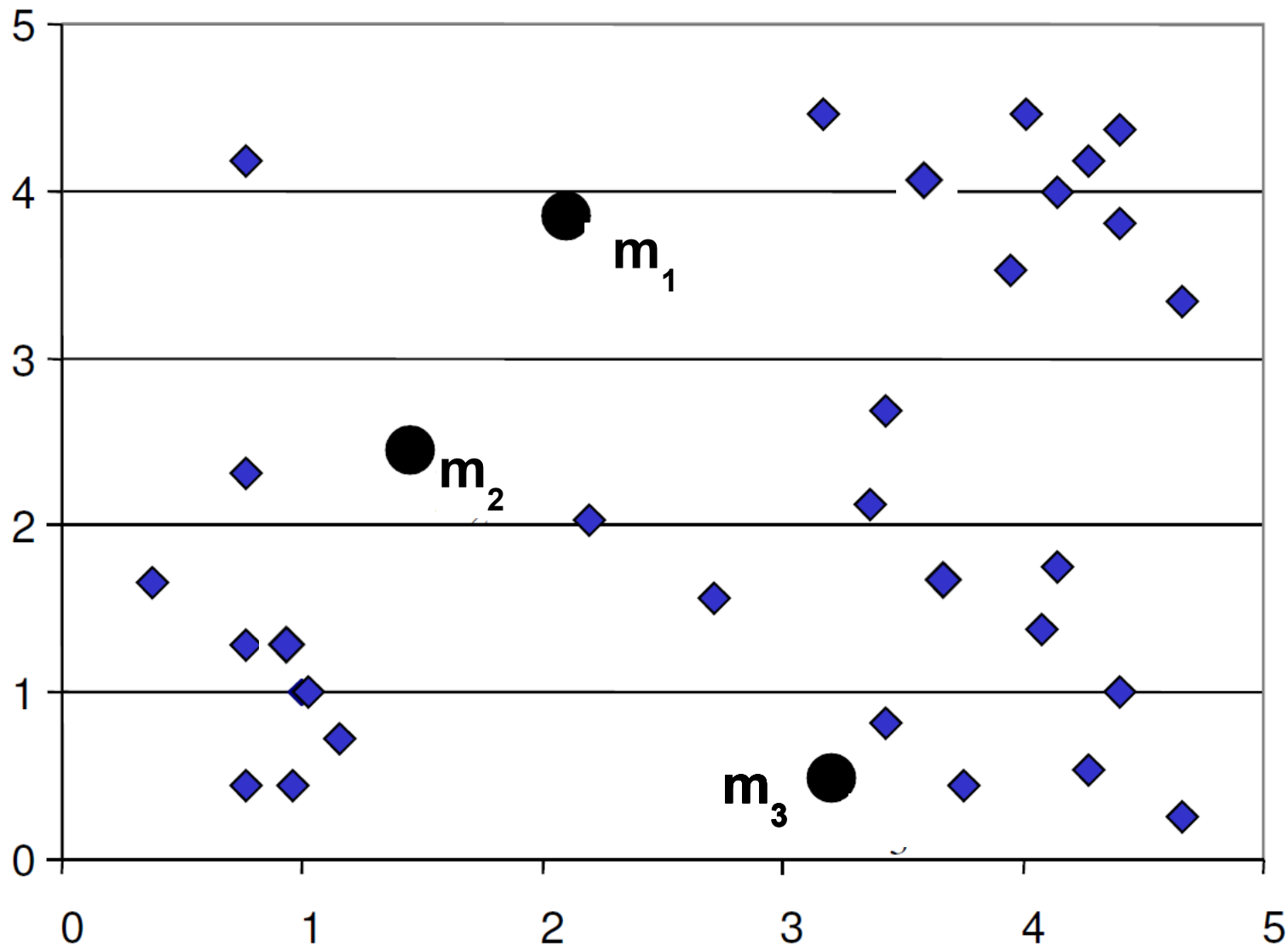


# Paso 1: Seleccionar k centroides aleatoriamente

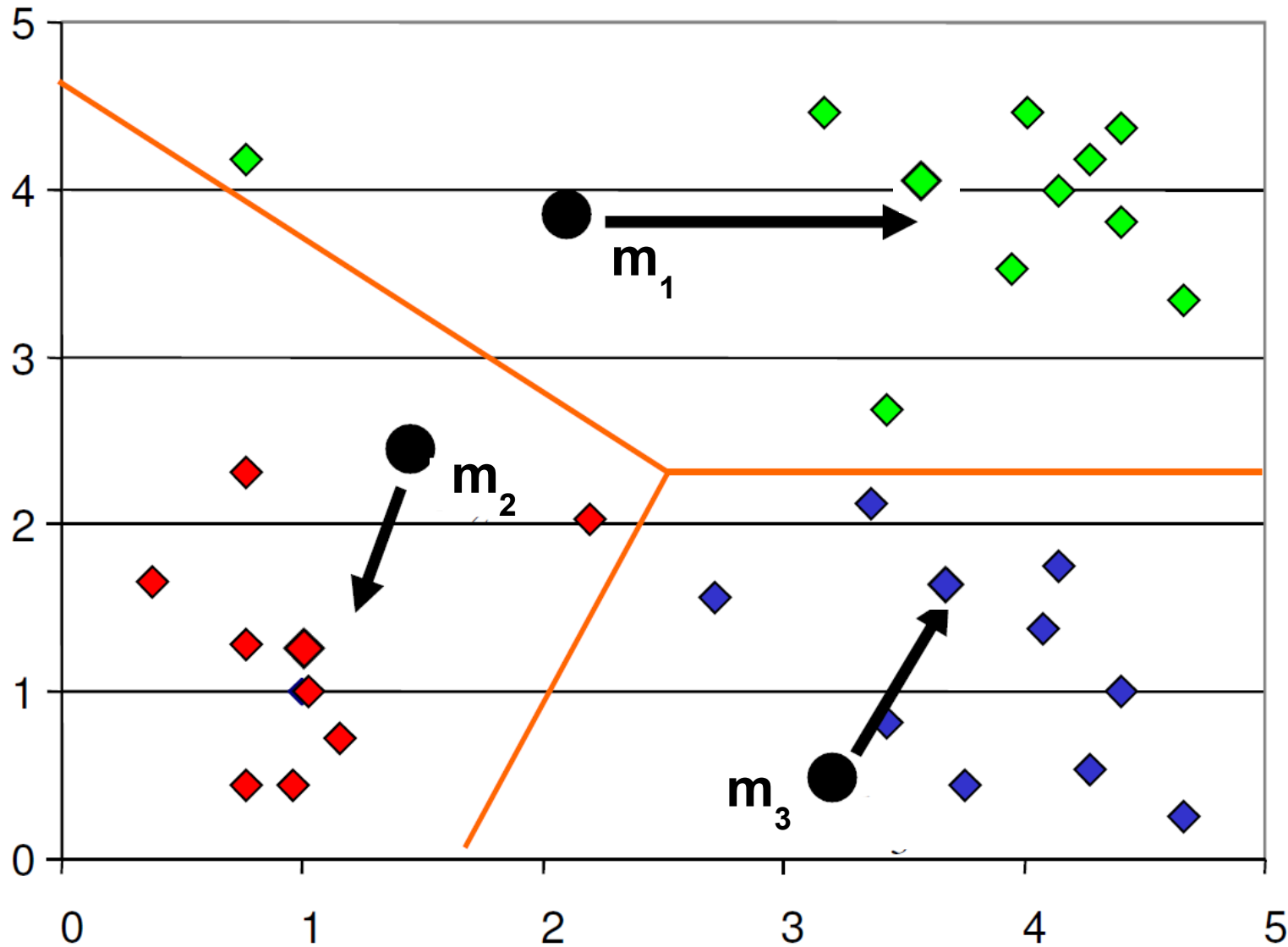




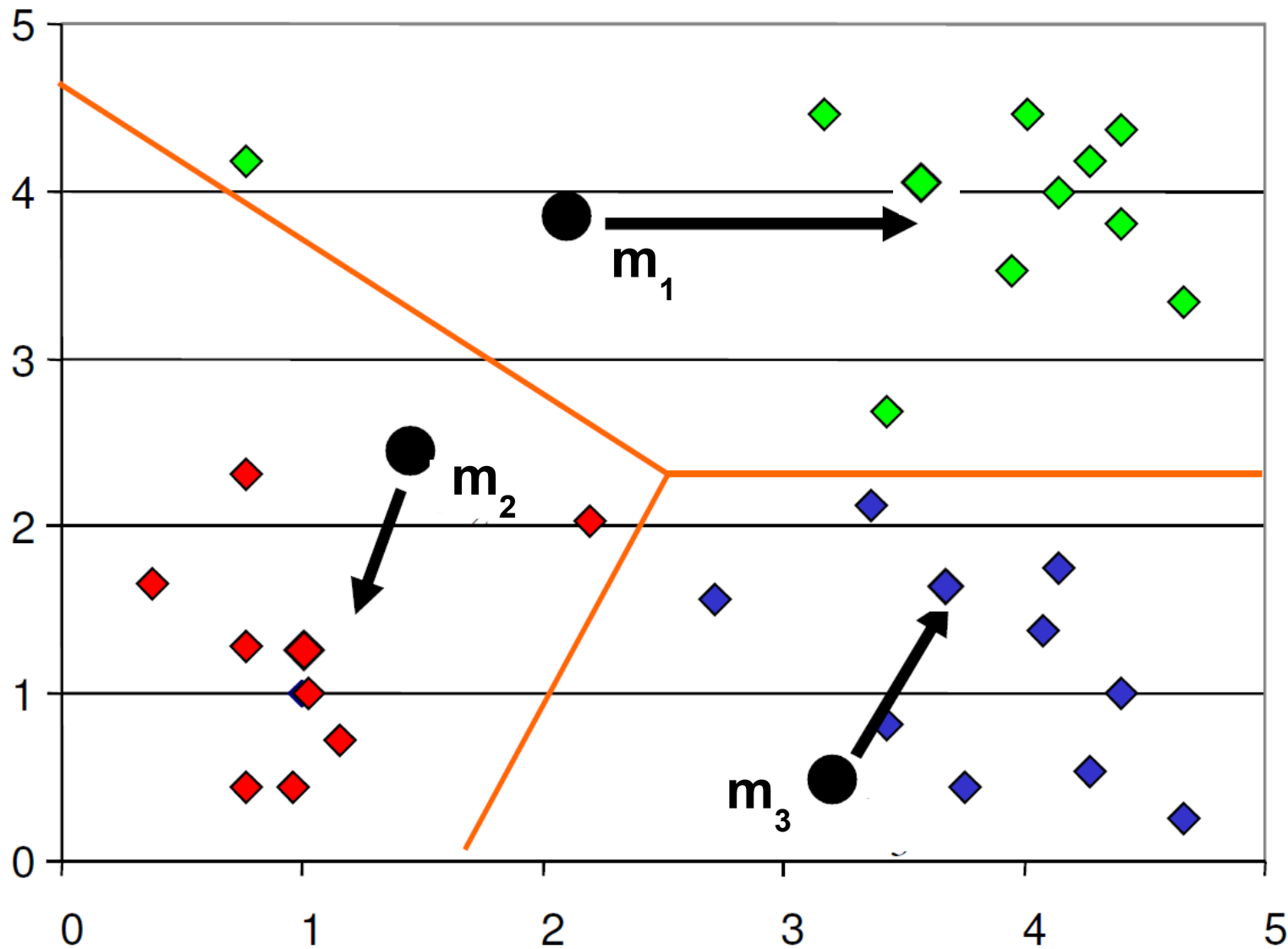
## Paso 2: Clear k clústeres asignando cada elemento al centroide más cercano



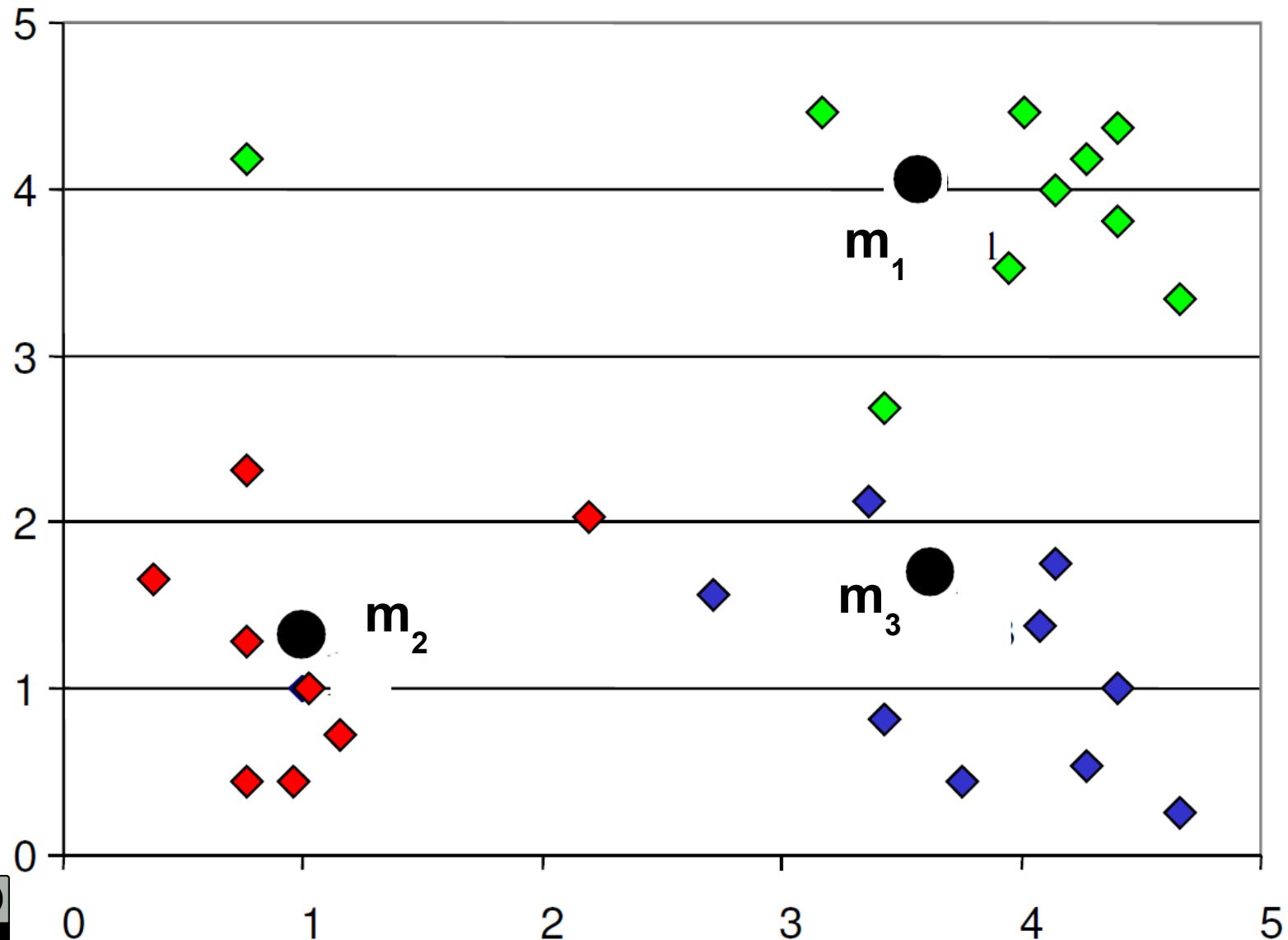
## Paso 2: Clear k clústeres asignando cada elemento al centroide más cercano



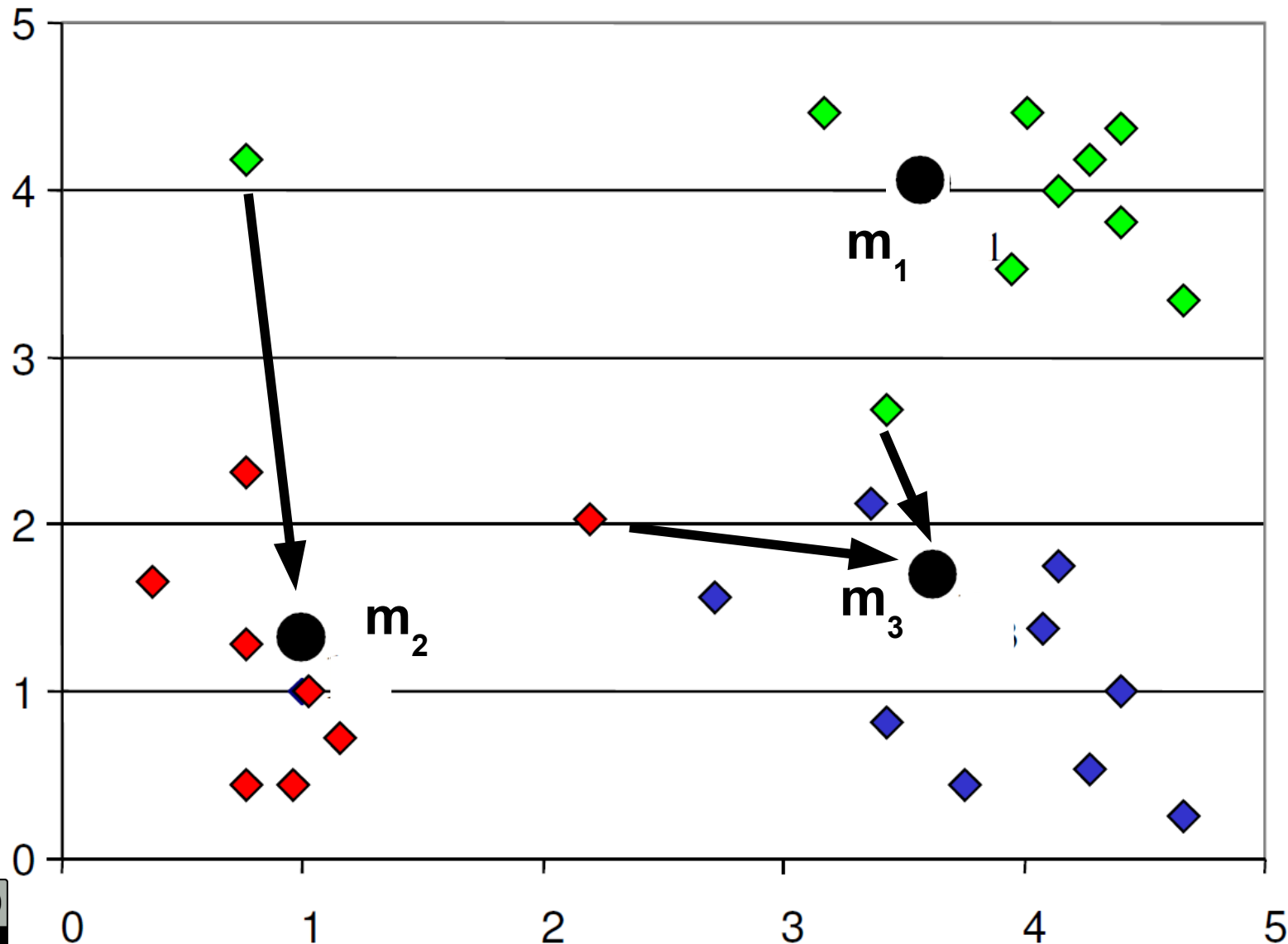
# Paso 3: Calcular nuevos centroides como aquellos elementos que minimizan la suma de las distancias al resto de elementos del clúster



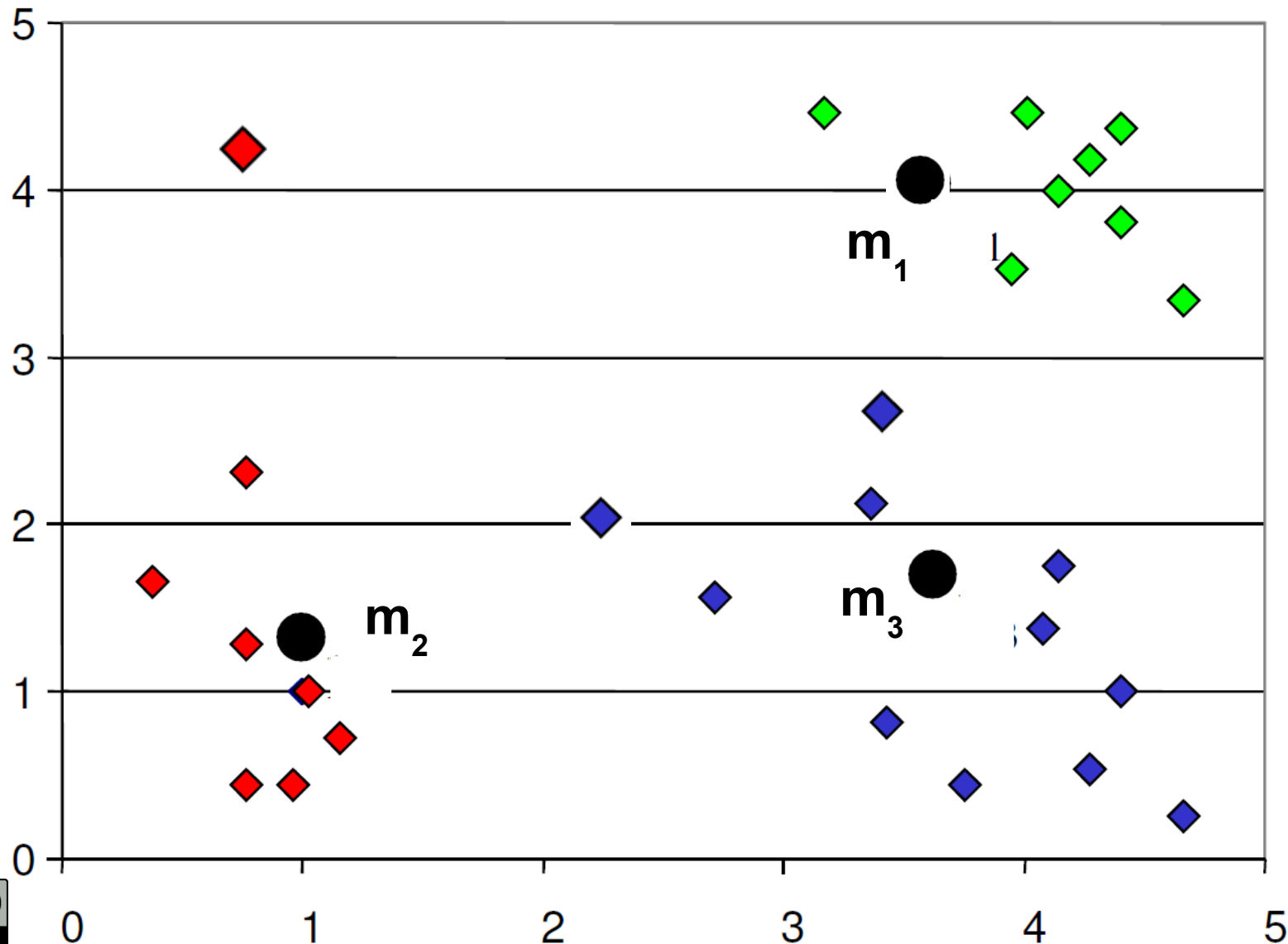
# Paso 3: Calcular nuevos centroides como aquellos elementos que minimizan la suma de las distancias al resto de elementos del clúster



**Paso 4: Volver al paso 2 mientras haya cambio en los clústeres o se alcance un número máximo de iteraciones.**



**Paso 4: Volver al paso 2 mientras haya cambio en los clústeres o se alcance un número máximo de iteraciones.**



# Ventajas / Desventajas de la Partición entorno a Centroides

- En cada iteración de PAM se realiza un reajuste y mejora de los clústeres contruidos de esta forma se evita la propagación de errores.
- Además de formar clústeres este algoritmo devuelve el elemento más central en cada clúster.
- La principal desventaja que presenta PAM consiste en la necesidad de fijar de antemano un número de clústeres a formar.

# PAM Clustering en R

- Utilizaremos como matriz de similitudes o distancias:

$$D(g_1, g_2) = 1 - \text{cor}(g_1, g_2)$$

- Los paquetes R a utilizar son **impute** y **WGCNA**.
- La función que realiza el clustering de partición entorno a centroides se llama **pam**. Recibe como entrada la matriz de similitudes a usar como distancia (as.dist) y el número de clústeres a generar.
- De igual forma que para el clustering jerárquico para la visualización en cytoscape del PAM clustering es necesario generar el fichero de atributos de genes correspondiente, cargarlo y utilizar vizmapper para seleccionar los colores apropiados.

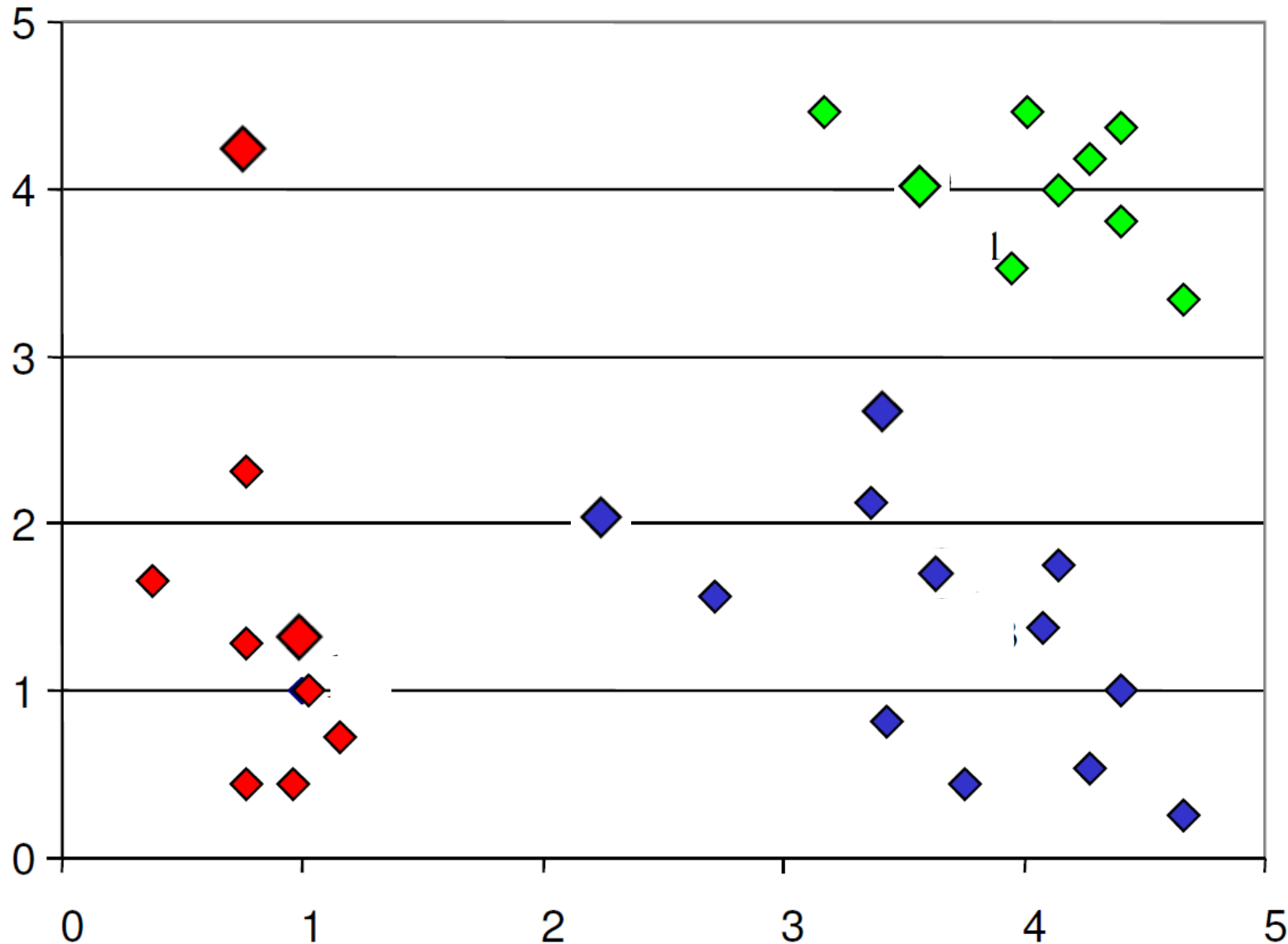


# Medida de Calidad de un proceso de Clustering, su Silueta

- Durante el flujo de trabajo de clustering existen tres puntos claves donde se toman decisiones que determinan la identificación final de grupos o clústeres de genes:
  - Elección de la **medida de similitud** o distancia.
  - Elección del **algoritmo de clustering**.
  - Elección del **número de clústers** a identificar.
- Para determinar la mejor elección posible es necesario fijar un criterio para mediar la calidad del resultado proporcionado por un flujo de trabajo de clustering.
- El objetivo general perseguido por las técnicas de clustering consiste en identificar grupos o clústeres compactos. Es decir, clusteres con una **similitud intra-clúster alta** y una **similitud inter-clúster baja**. Esta idea intuitiva se formaliza en el concepto de **silueta** de un cluster.

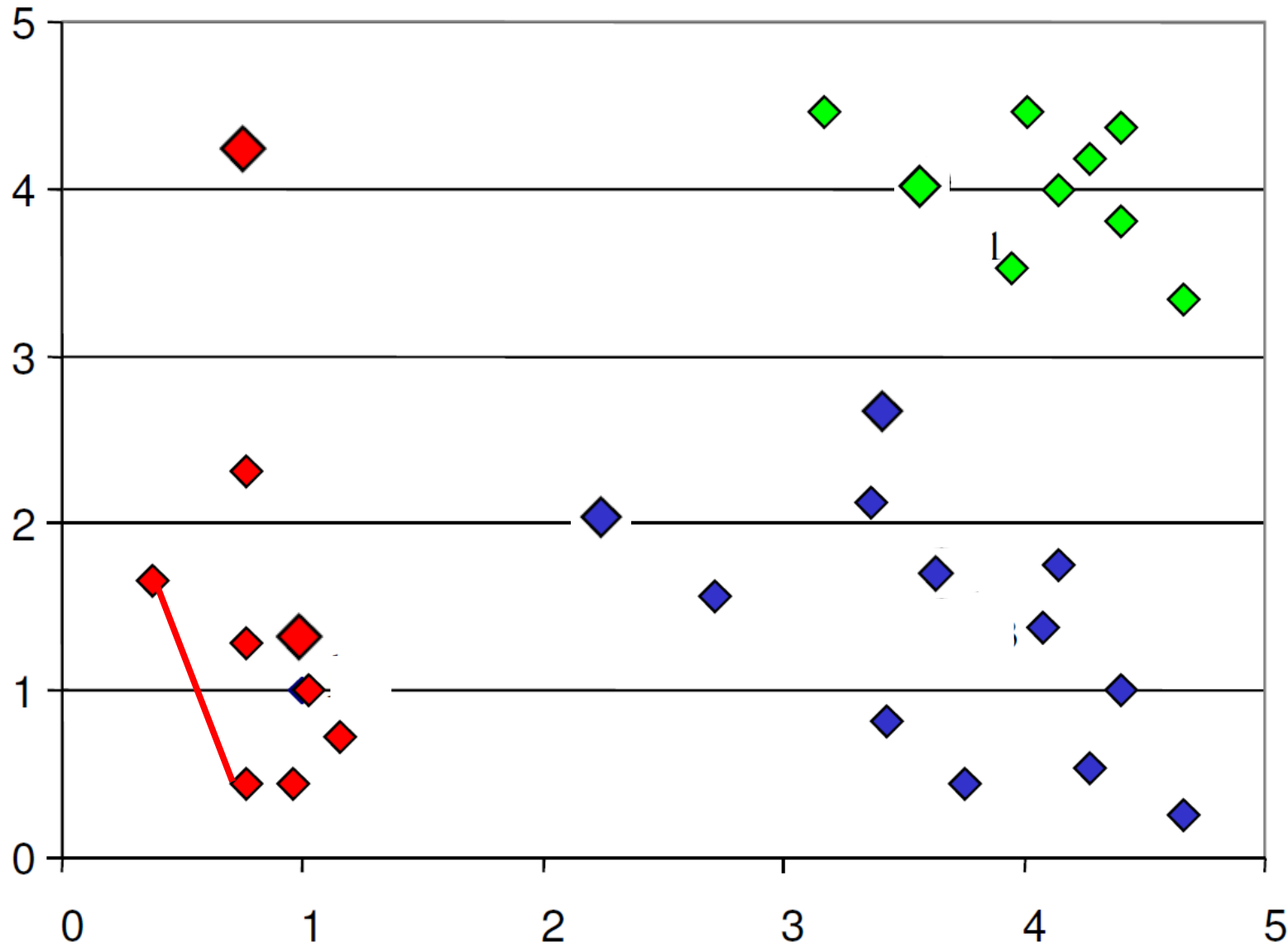
# Medida de Calidad de un proceso de Clustering, su Silueta

Para calcular la silueta de un cluster  $C$  para cada elemento  $s_i$  en  $C$  calculamos primero  $a(s_i)$  la media de las distancias entre si y todos los  $s_j$  en  $C$ .



# Medida de Calidad de un proceso de Clustering, su Silueta

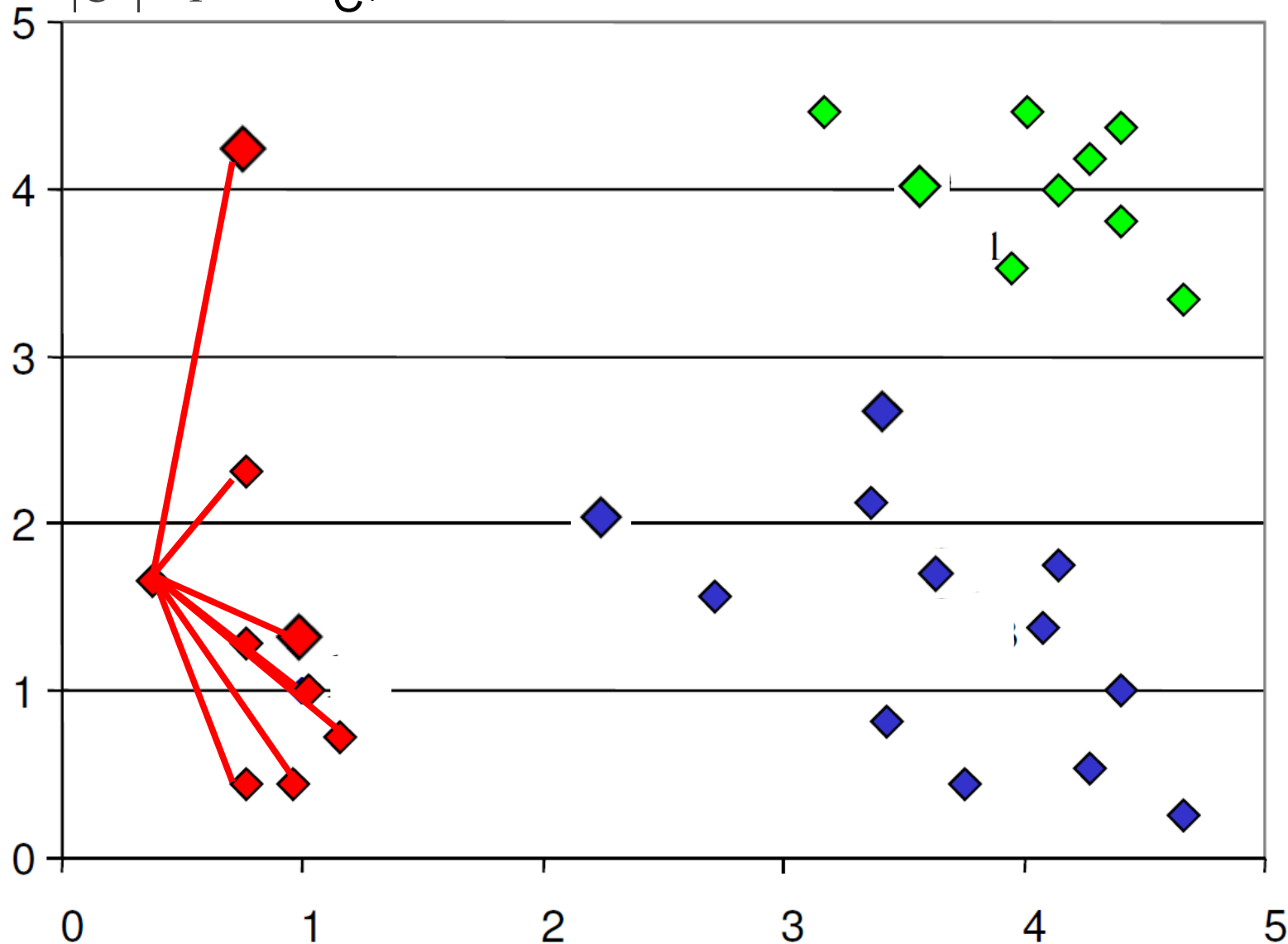
Para calcular la silueta del cluster  $C^1$  para cada elemento  $s_i$  en  $C^1$  calculamos primero  $a(s_i)$  la media de las distancias entre  $s_i$  y todos los  $s_j$  en  $C^1$ .



# Medida de Calidad de un proceso de Clustering, su Silueta

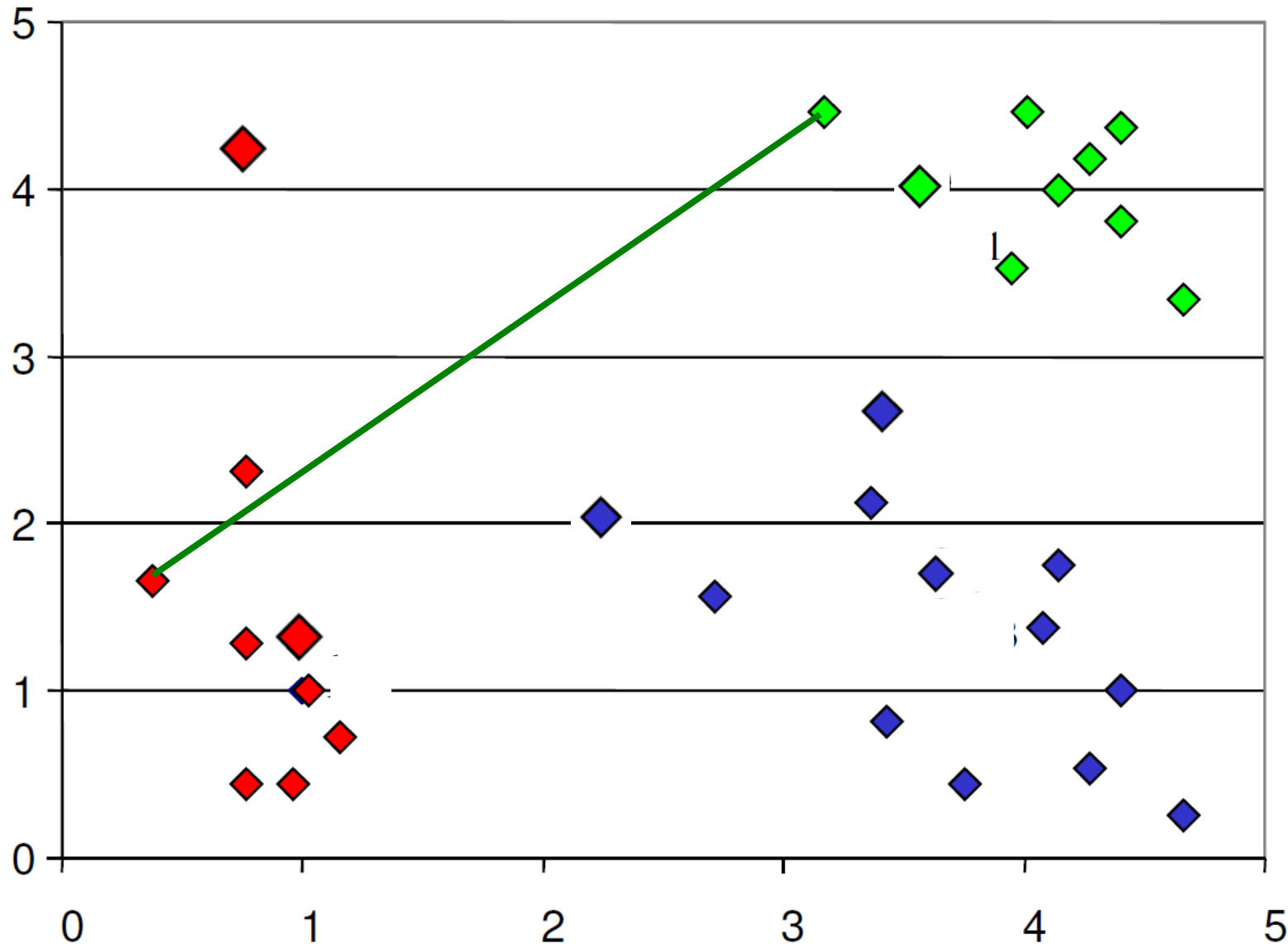
$$a(s_i) = \frac{\sum_{s_j \in C^1} d(s_i, s_j)}{|C^1| - 1}$$

$a(s_i)$  constituye una medida de la distancia intracluster en  $C^1$



# Medida de Calidad de un proceso de Clustering, su Silueta

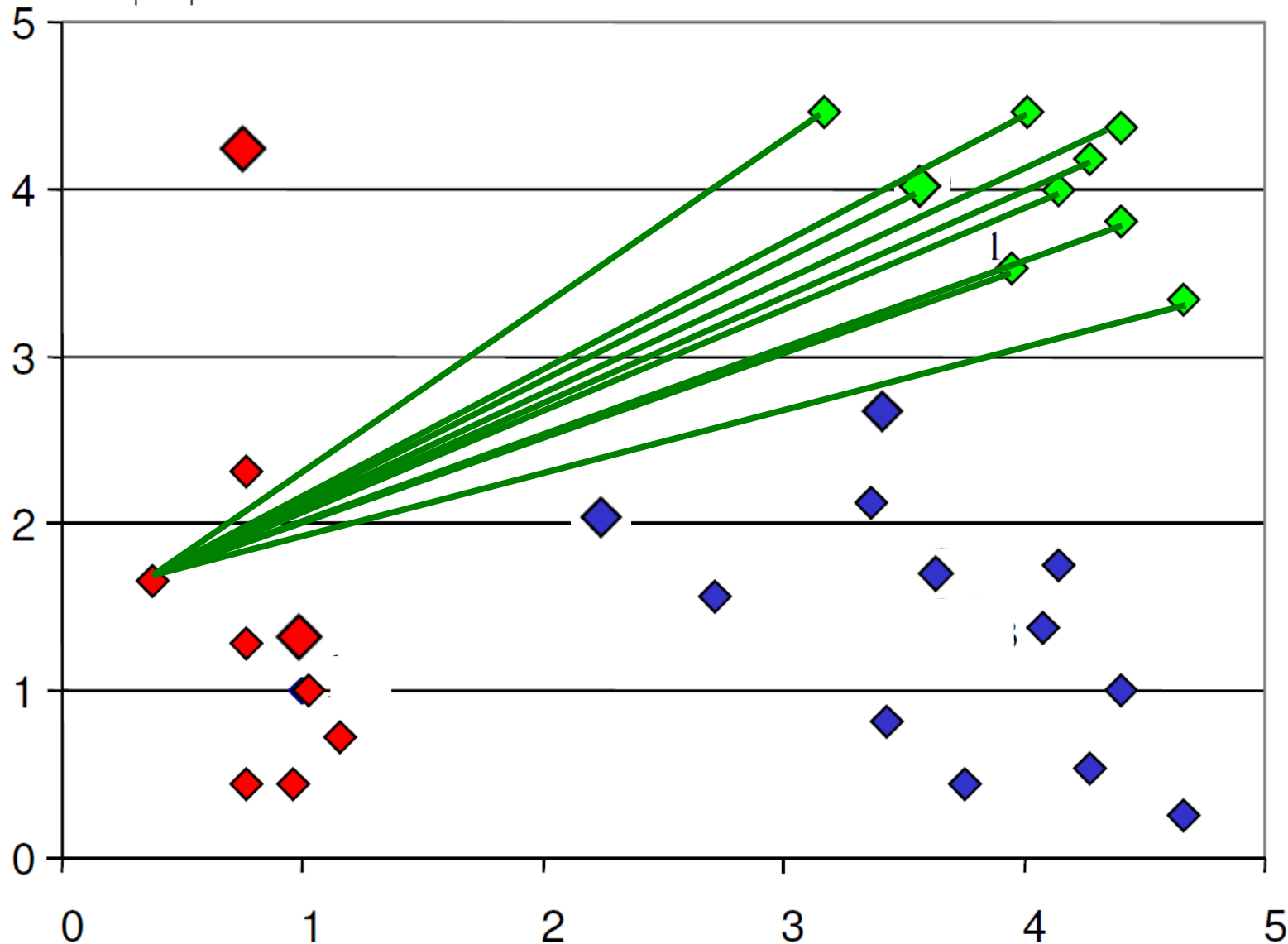
Para calcular una medida de la distancia intercluster entre el cluster  $C^1$  y el resto para cada elemento  $s_i$  en  $C^1$  calculamos  $d(s_i, C^k)$  la media de las distancias entre  $s_i$  y todos los  $s_j$  en  $C^k$ .



# Medida de Calidad de un proceso de Clustering, su Silueta

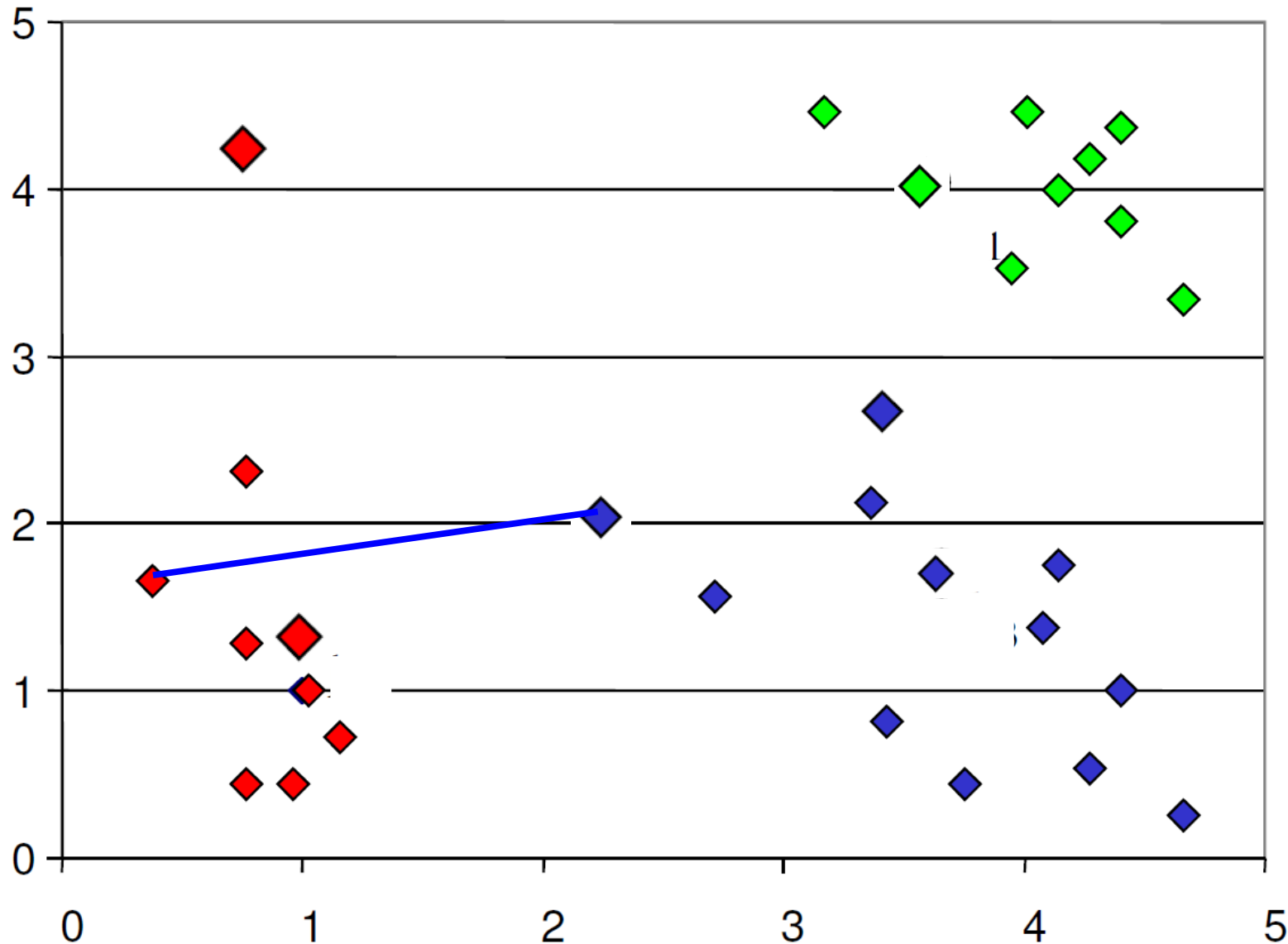
$$d(s_i, C^k) = \frac{\sum_{s_j \in C^k} d(s_i, s_j)}{|C^k|}$$

$d(s_i, C^k)$  constituye una medida de la distancia intercluster



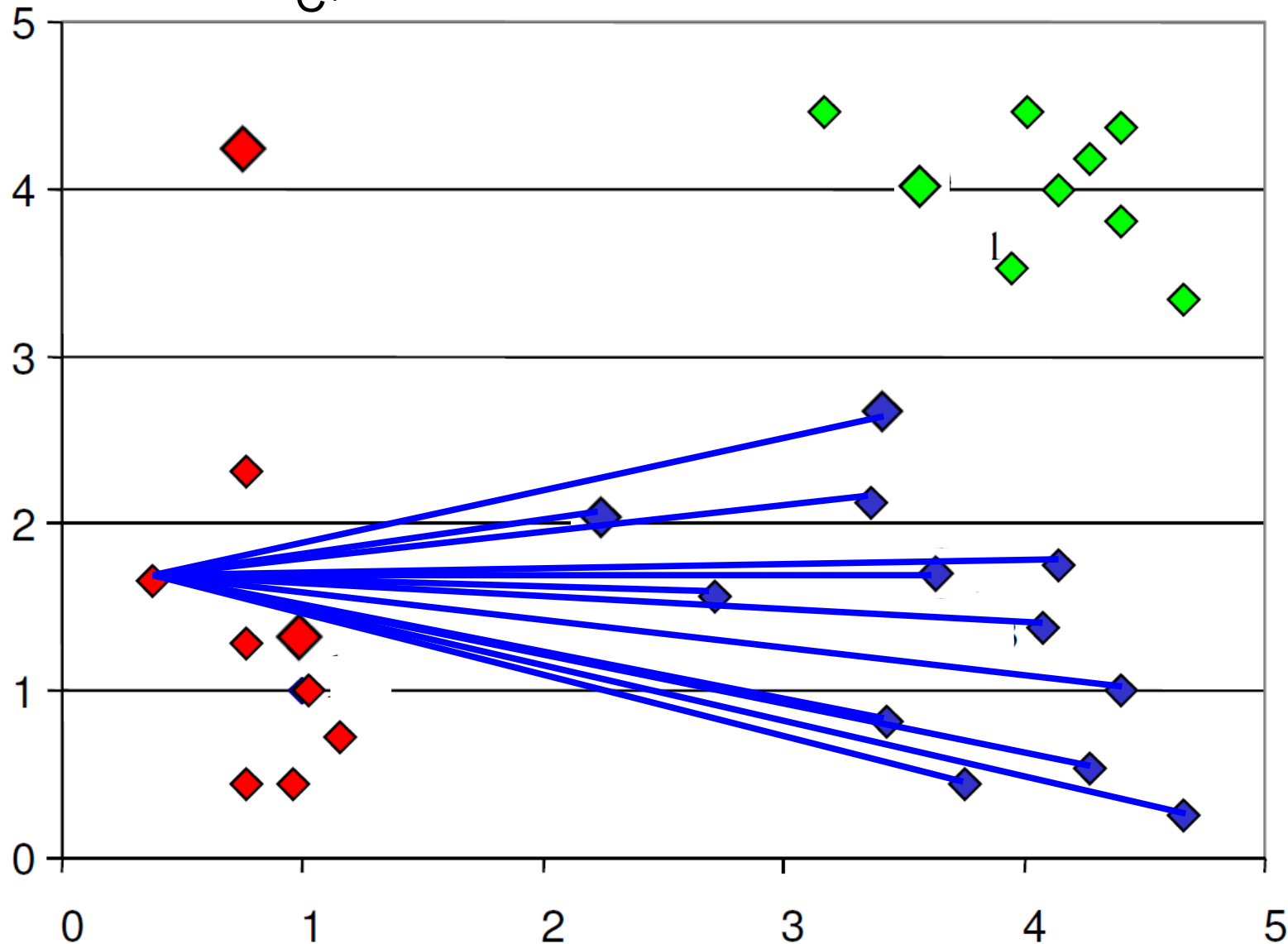
# Medida de Calidad de un proceso de Clustering, su Silueta

Para calcular una medida de la distancia intercluster entre el cluster  $C^1$  y el resto para cada elemento  $s_i$  en  $C^1$  calculamos  $b(s_i) = \min_k d(s_i, C^k)$ .



# Medida de Calidad de un proceso de Clustering, su Silueta

$b(s_i) = \min_k d(s_i, C^k)$   $\mathbf{b}(\mathbf{s}_i)$  constituye una medida de la distancia intercluster en  $C^1$





# Medida de Calidad de un proceso de Clustering, su Silueta

$$a(s_i) = \frac{\sum_{s_j \in C^1} d(s_i, s_j)}{|C^1| - 1} \quad \mathbf{a(s_i)}$$
 constituye una medida de la distancia intracluster en  $C^1$

$$b(s_i) = \min_k d(s_i, C^k) \quad \mathbf{b(s_i)}$$
 constituye una medida de la distancia intercluster en  $C^1$

Se define la silueta  $\mathbf{s(s_i)}$  como:

$$s(s_i) = \frac{b(s_i) - a(s_i)}{\max(a(s_i), b(s_i))}$$

Se define la silueta de un cluster  $C$ ,  $s(C)$  como:

$$s(C) = \frac{\sum_{s_i \in C} s(s_i)}{|C|}$$

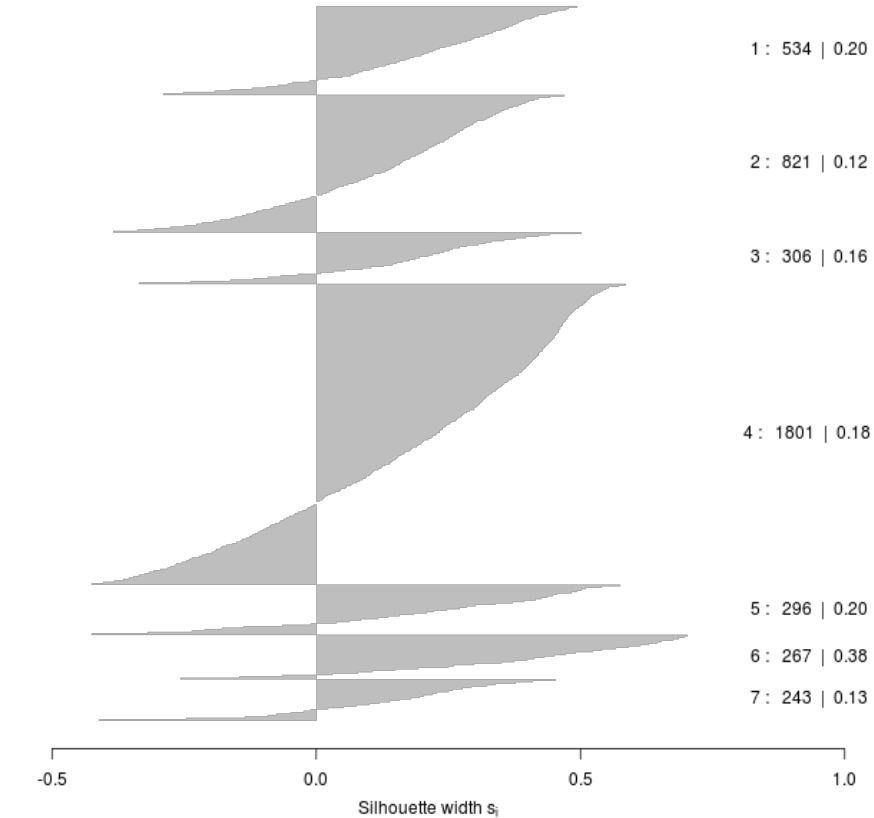
Se define la silueta del resultado de un proceso de clustering  $C_1, \dots, C_n$  como:

$$s(C_1, \dots, C_n) = \frac{\sum_{i=1}^n s(C_i)}{n}$$

# Medida de Calidad de un proceso de Clustering, su Silueta

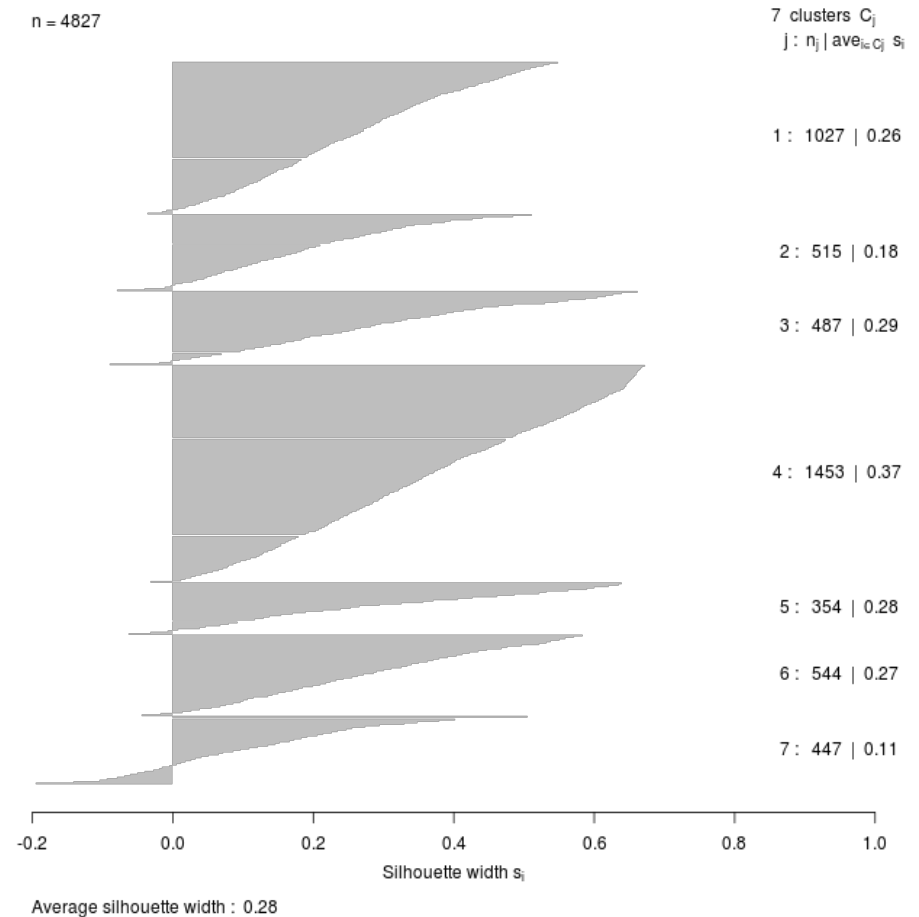
## Hierarchical 7 clusters

n = 4268

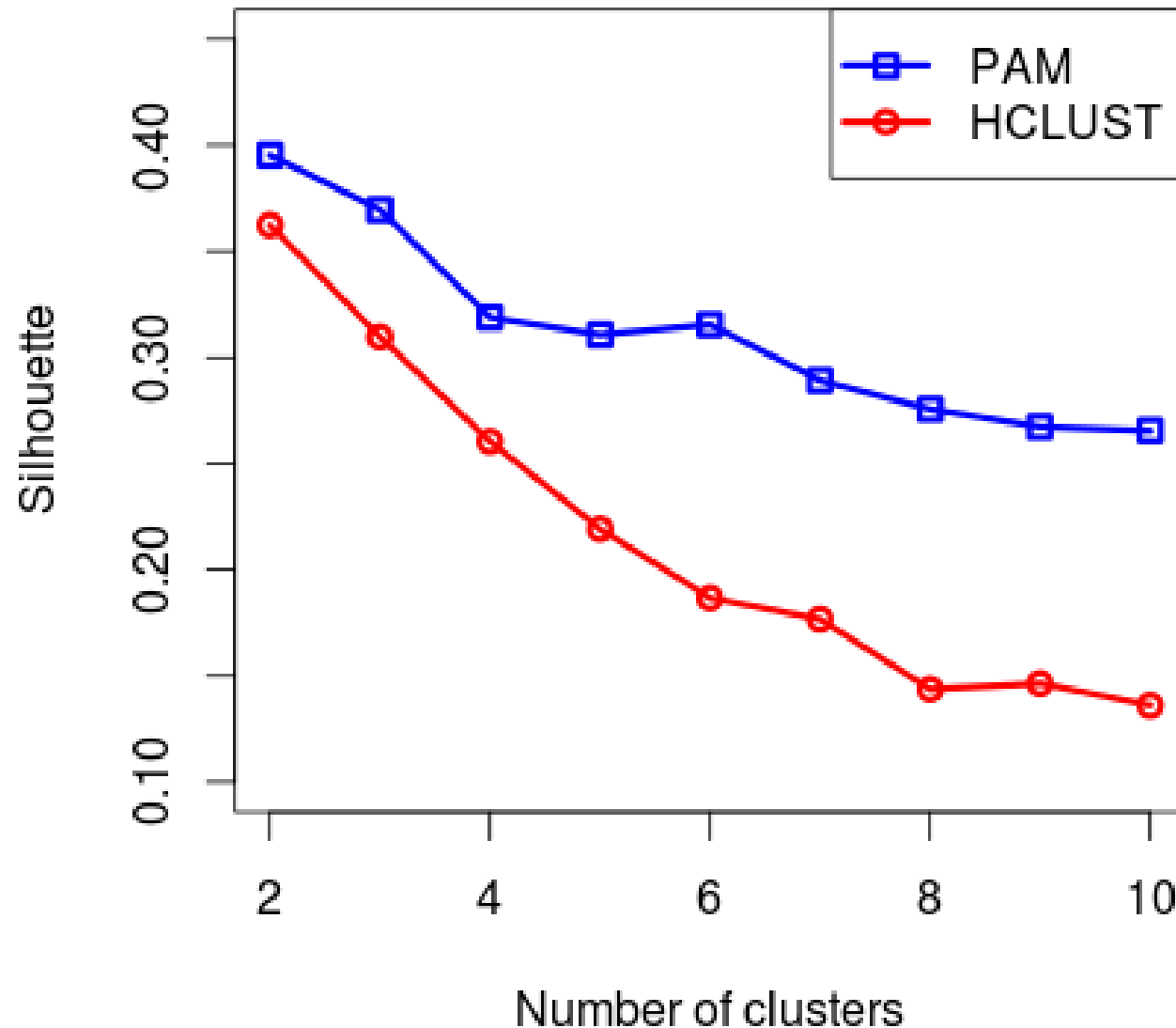


## PAM 7 clusters

n = 4827



# Medida de Calidad de un proceso de Clustering, su Silueta



# Visualización de Clustering en Cytoscape

- Es necesario generar un fichero de texto con dos columnas. La primera columna debe contener los nombres de los genes o nodos de la red y la segunda debe contener los atributos a importar, por ejemplo el número del cluster al que pertenece cada gen.
- Para cargar atributos en Cytoscape:

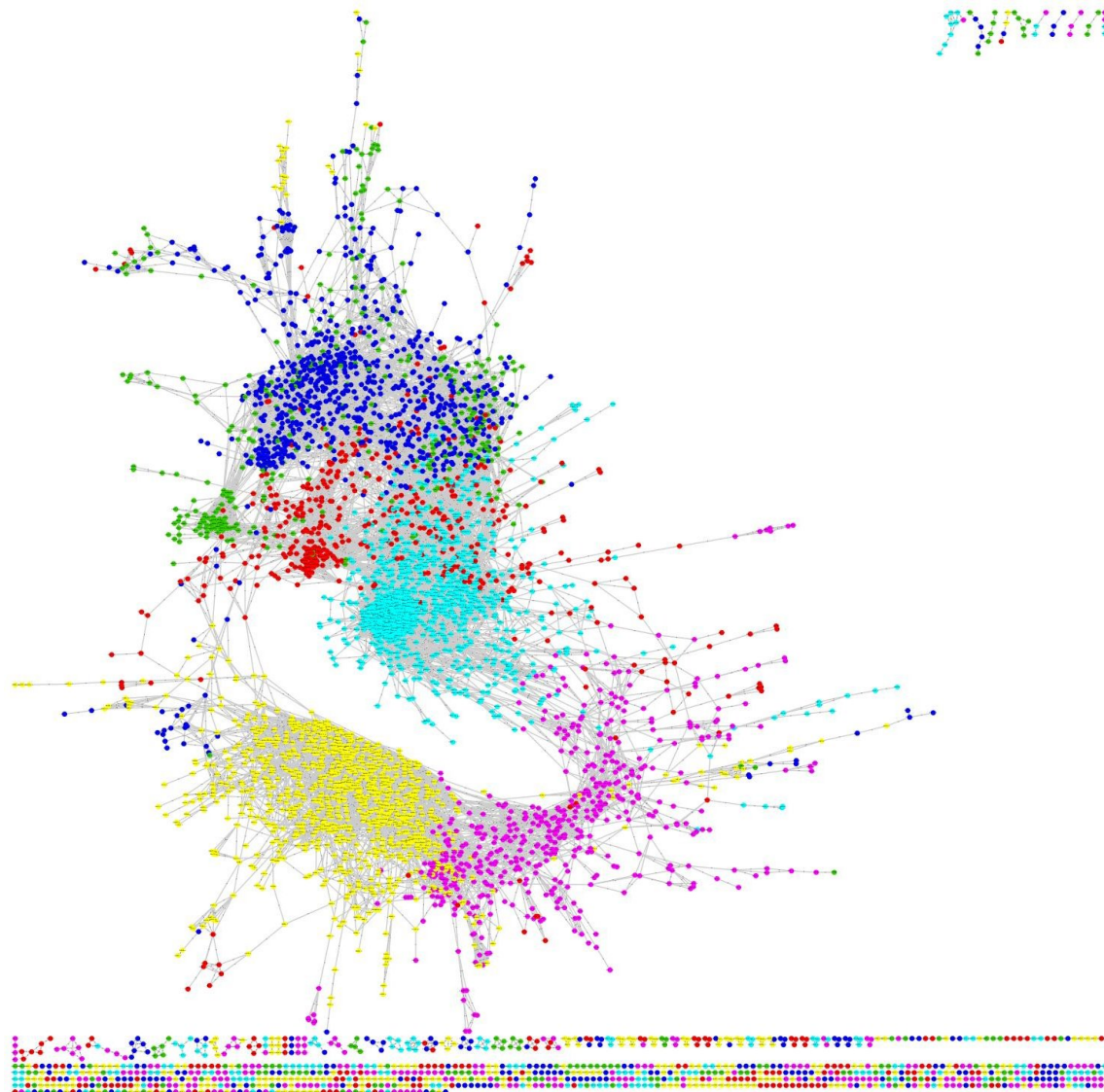
**File → Import → Table → File**

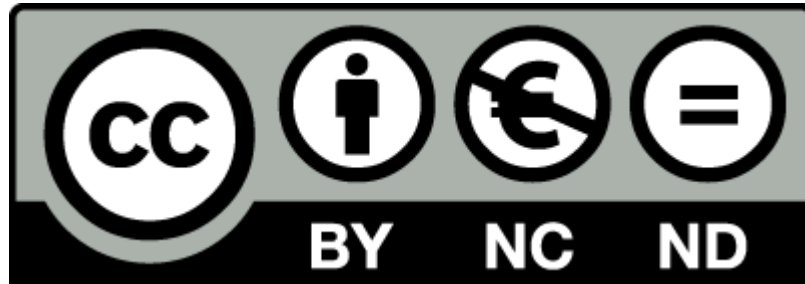
**Show Text File Import Options → Delimiter (space) → Transfer first line...**

- Para cambiar de color a los genes según su modulo:

**Vizmapper → Node Fill color → Cluster → Mapping Type = Discrete mapping → Selección de colores**

# Visualización de Clustering en Cytoscape





This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-nd/3.0/>.

---

