UNIVERSIDAD
DE LOS ANDES
MÉRIDA VENEZUELA

Minería de Grafos

Jose Aguilar

CEMISID, Escuela de Sistemas

Facultad de Ingeniería

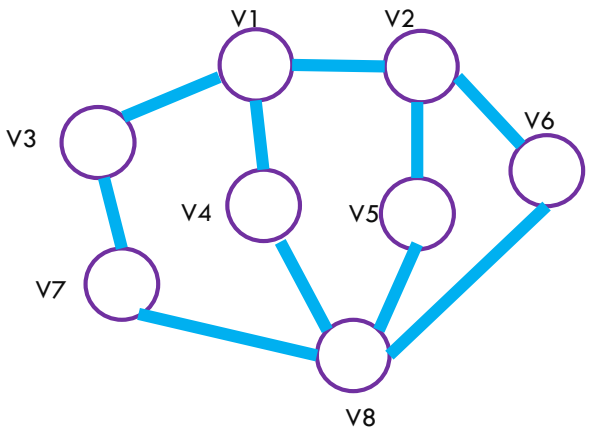
Universidad de Los Andes

Mérida, Venezuela

Grafos

Un grafo **G** es un par ordenado de un conjunto de vértices **V** y un conjunto de aristas **E**

G = (V,E)

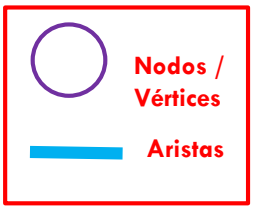


Par ordenado:

(a, b) ≠ (b,a) si a ≠ b

Par No ordenado:

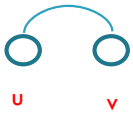
{a,b} = {b,a}



Grafos

Aristas:

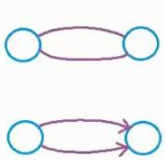
Dirigido
(u, v)
(v, u)

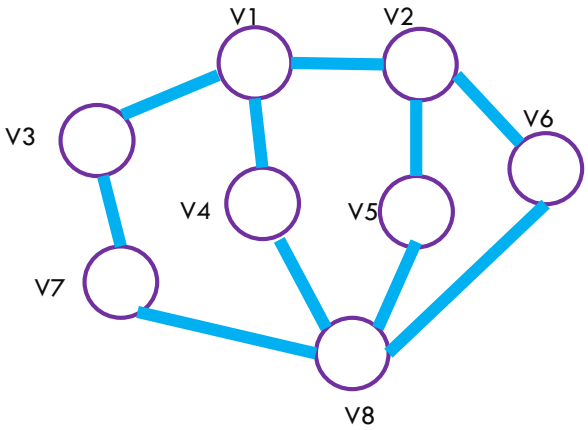


No Dirigido
{u,v}

Bucle

Varias
Aristas





V = {V1. V2. V3, V4, V5, V6, V7, V8}

E = {{v1. v2}, {vi,v3}, {v1,v4}, {v2,v5}, {v2,v6}, {v3, v7}, {v4, v8}. {v7, v8}, {v5.,v8}, {v6, v8}}

Grafos

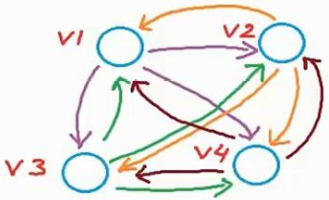
Número de Aristas:

if $|V| = n$

then,

$0 \leq |E| \leq n(n-1)$, if directed

$0 \leq |E| \leq \frac{n(n-1)}{2}$, if undirected



if $|V| = 10$, $|E| \leq 90$

if $|V| = 100$, $|E| \leq 9900$

Modelando Datos con Grafos…

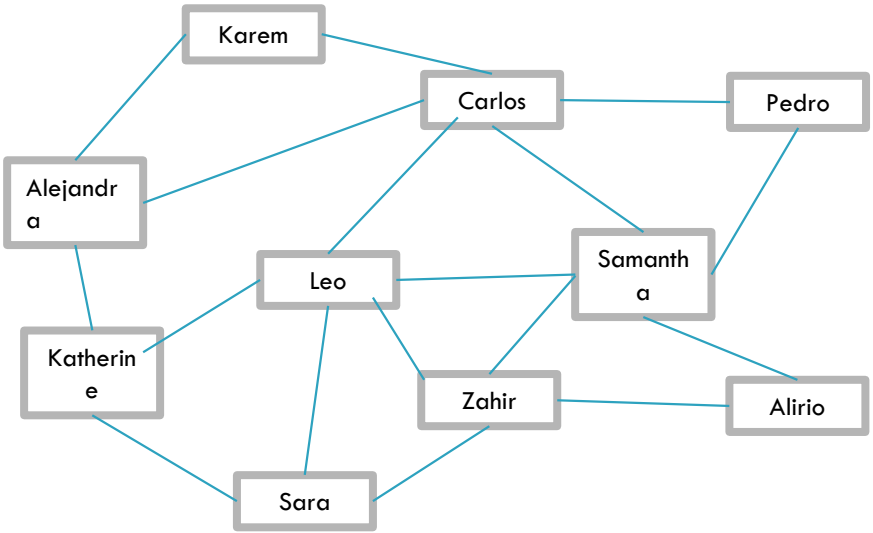
Los grafos son adecuados para la captura de las relaciones arbitrarias entre los diversos elementos.

Instancia		Grafo
Elemento	⟺	Vertice
Atributos Elemento	⟺	Etiquetas Vertices
Relaciones	⟺	Arcos
Tipo de relaciones	⟺	Etiquetas arcos

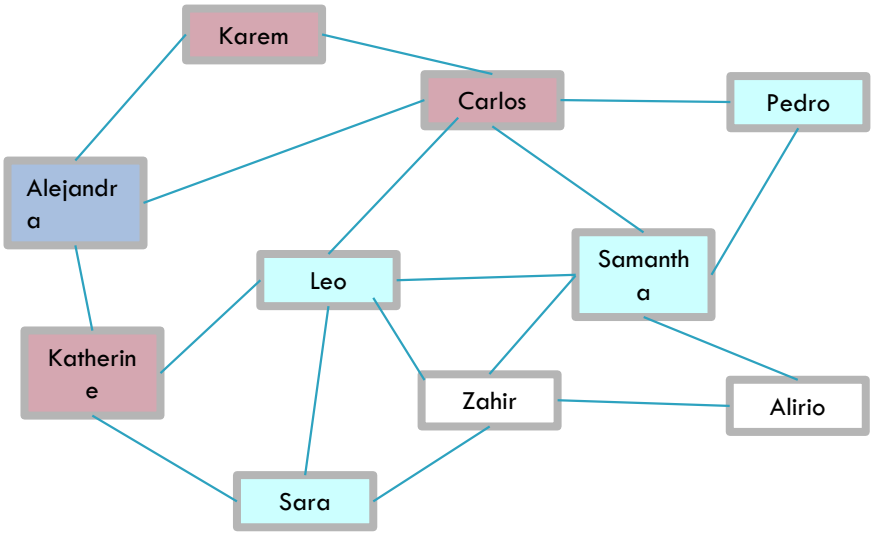
Proporcionan una enorme flexibilidad para el modelado de los datos, ya que permiten al modelador decidir cuáles son el tipo de relaciones a modelar

Grafos

Red Social
FACEBOOK



Grafos



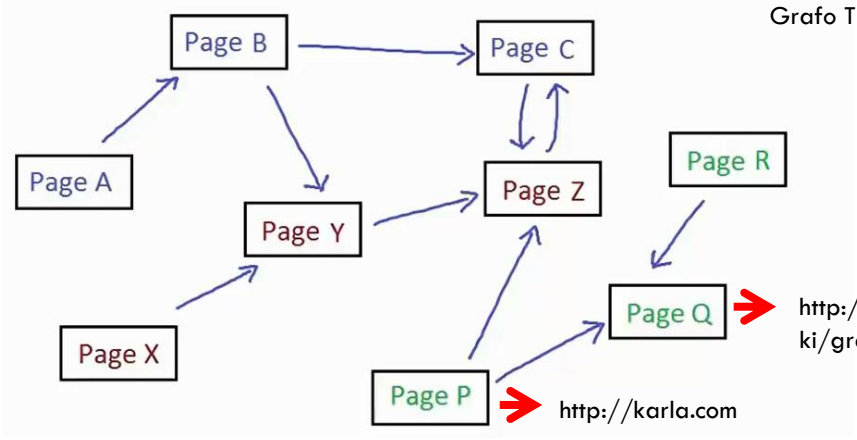
Red Social FACEBOOK

Para Sugerir un amigo a ALEJANDRA hay que encontrar todos los nodos que tengan longitud del camino igual a 2.

World Wide Web

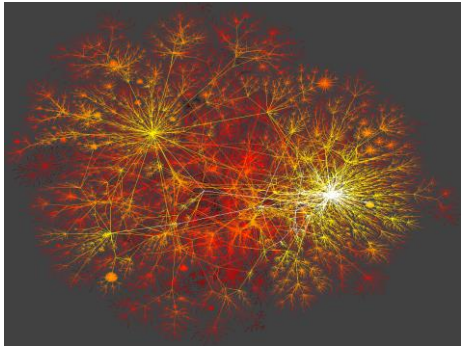
Web – Crawling

Grafo Transversal



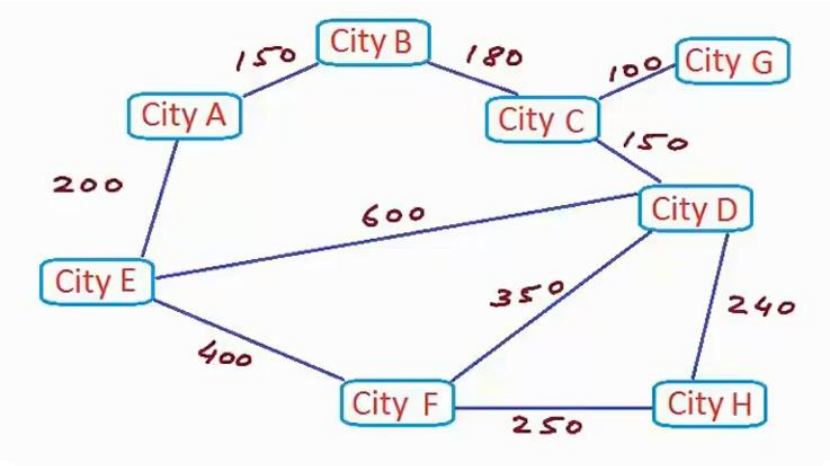
http://wikipedia.org/wiki/grafos

Internet



Grafos

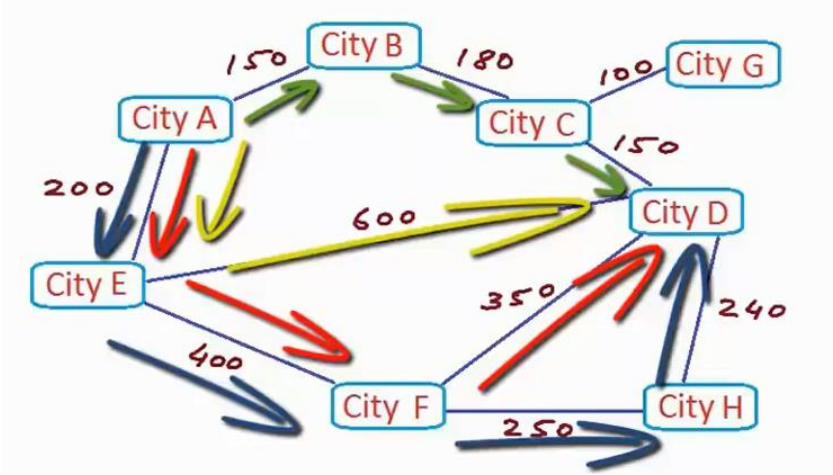
Grafos con Pesos VS Grafos sin Pesos



Red de Carreteras Inter urbanas

Grafos

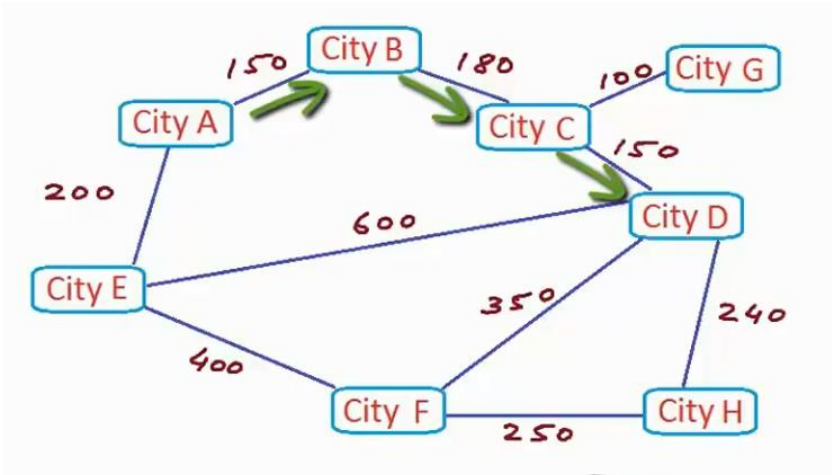
Grafos con Pesos VS Grafos sin Pesos



Red de Carreteras Inter urbanas

Grafos

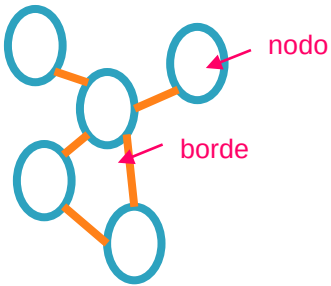
Grafos con Pesos VS Grafos sin Pesos



Red de Carreteras Inter urbanas

Redes

Las redes son colecciones de puntos unidos por líneas.



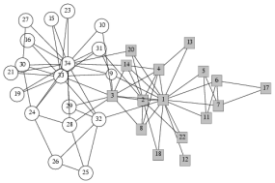
"Red" ≡ "Gráfo"

puntos	líneas	Area de estudio
vértices	bordes, arcos	matemáticas
nodos	enlaces	ciencias de la computación
sitios	lazos	física
actores	vinculo, las relaciones	sociología

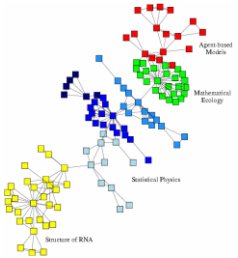
Redes en el mundo real

- **Redes de información:**
  - World Wide Web: hyperlinks
  - Redes de citación
  - Redes de Noticias y Blogs
- **Redes sociales**
  - Organizativas
  - Comunicativas
  - Colaborativas
  - Contactos sexuales
- **Redes tecnológicas:**
  - Energéticas
  - Transporte (aéreo, carreteras, fluviales,…)
  - Telefónicas
  - Internet
  - Sistemas Autónomos

Redes de amistad



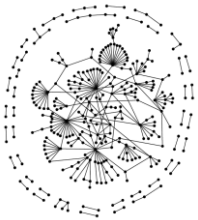
Karate club network



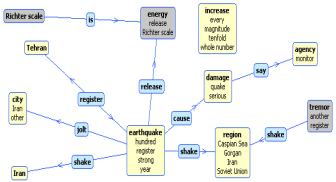
Redes de colaboración

Redes en el mundo real

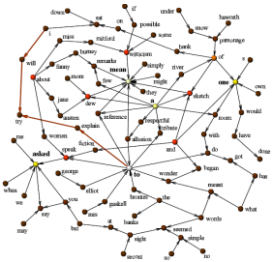
- Redes biológicas
  - Metabólicas
  - Cadenas alimenticias
  - Neuronales
  - Regulación Genética
- Redes de lenguaje
  - Semánticas
  - Lingüísticas
- Redes de software
- …



Interacciones entre las proteínas de la levadura

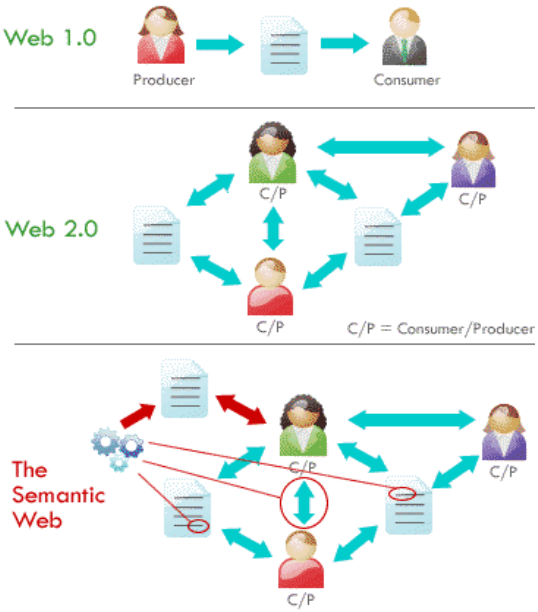


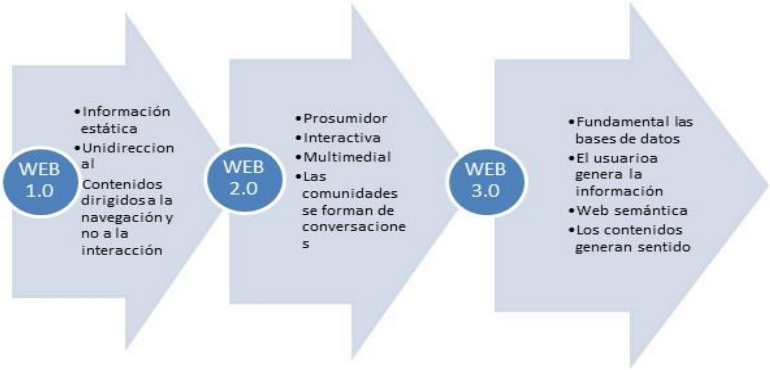
Red semántica



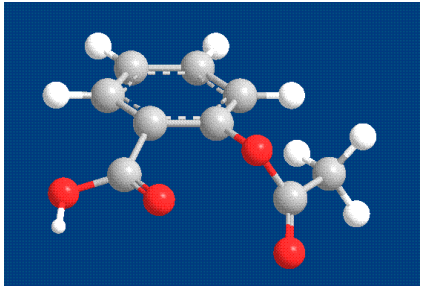
Red Lingüística

LA WEB 2.0 Y LA WEB 3.0

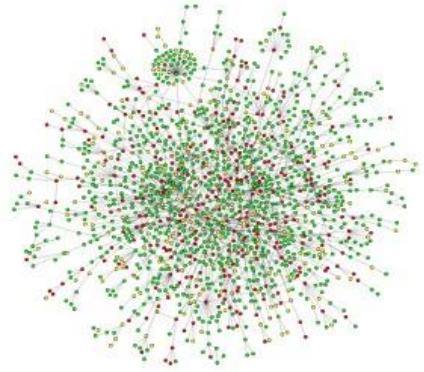




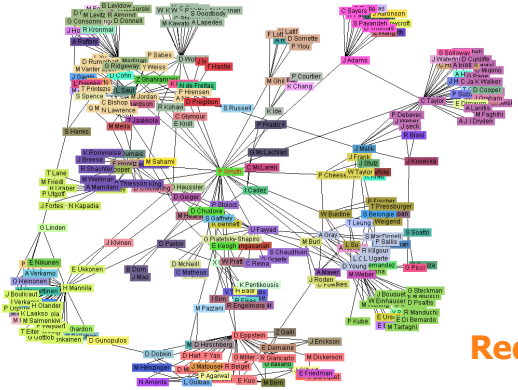
Grafos



Aspirina



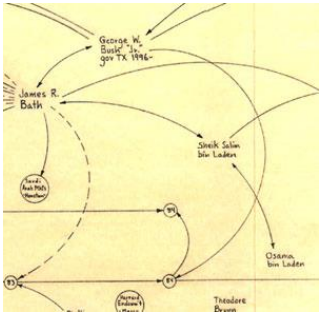
Red de interacción de proteína levadura

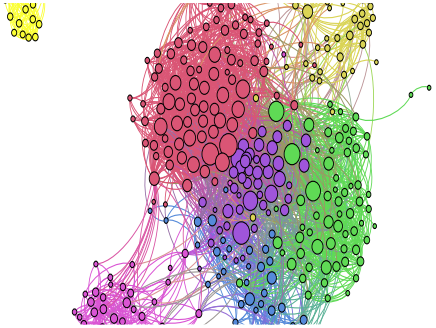


Red Co-autores de libros

Grafos

Mark Lombardi: rastreo y Mapeo fiascos financieros globales en los años 1980 a partir de fuentes públicas, como los artículos de noticias.

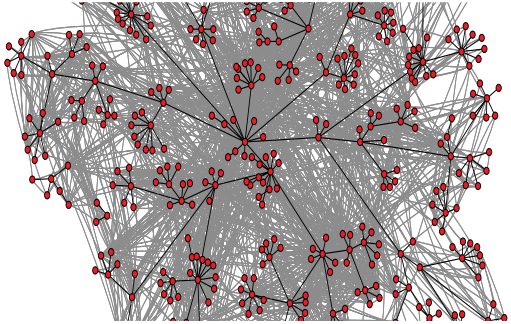




Facebook de alguien

Los colores separan componentes fuertemente conectados de la red.

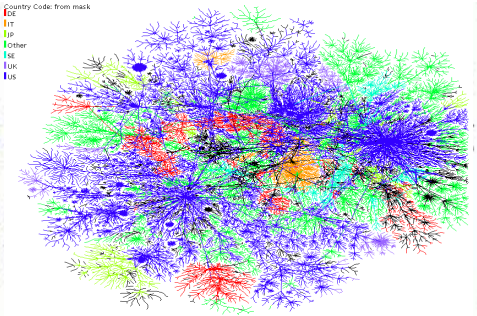
Relación de empleados de una organización



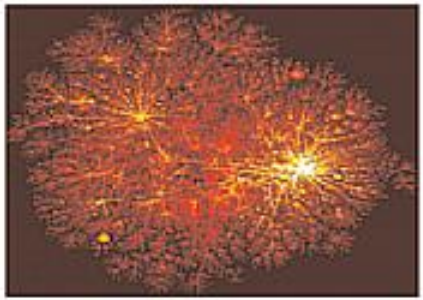
Los arcos negros denotan estructura organizacional y los grises son interacciones por correo electrónico.

Redes Naturales vs. Redes Artificiales

- Las redes naturales evolucionan por adaptación al entorno.
- A diferencia de éstas, el origen de una red artificial, como Internet, está basado en un diseño humano inteligente.
- Sin embargo, presentan características topológicas análogas.



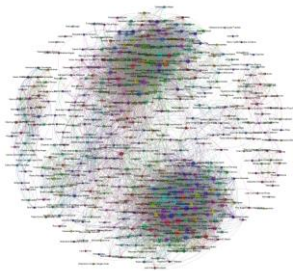
Red Internet



Red de interacciones proteínicas

Red Social

Son las diferentes interacciones que realizamos con nuestros conocidos. Estas pueden ser representadas mediante grafos, donde los nodos son las personas que interactúan y los arcos que los unen son las interacciones entre esas personas.



Red Social

- El **análisis de redes sociales** estudia esta estructura social aplicando la teoría de grafos.

- Se analiza:
  - Si existen estructuras de comunidades ocultas
  - La difusión o las opiniones.
  - La influencia del todo en las partes y viceversa.
  - La difusión de nuevas ideas y prácticas (teoría de difusión de innovaciones).
  - El efecto producido por la acción selectiva de los individuos en la red
  - Grafos de colaboración para ilustrar buenas (amistad, alianza, citas) y malas (odio, ira) relaciones entre los seres humanos.

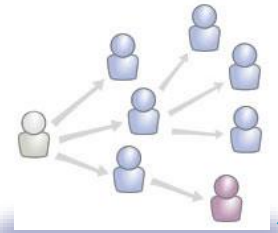
RED SOCIAL VIRTUAL
WEB 2.0 Y 3.0

- **Facebook** es la red social más utilizada del mundo
- **Twitter:** red social de microblogging.
- **LinkedIn** red de usuarios profesionales, y
- **Youtube** red de alojamiento de vídeos.
- **Google+**, apuesta de Google por las redes sociales.
- **Instagram**, red de intercambio de imagenes

RSV

- RED SOCIAL (RS)
- COMUNIDAD VIRTUAL (CV)
- SITIOS DE REDES SOCIALES (SNS: Social Network Services – Servicios de Redes Sociales)

RED SOCIAL VIRTUAL
WEB 2.0 Y 3.0

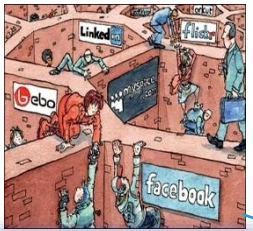


RED SOCIAL (RS)

CONJUNTO DE PERSONAS O ENTIDADES

COMPARTEN INTERESES, VINCULADAS POR CARACTERISTICAS Y OBJETIVOS A FINES

INTERCAMBIAN INFORMACIÓN DE TODA CLASE: FINANCIERA, AMISTAD, OCIO, ACADEMICA, ENTRE OTRA.



COMUNIDAD VIRTUAL (CV)

CONJUNTO DE PERSONAS, ENTIDADES O GRUPOS SOCIALES

CON UN MISMO OBJETIVO O PROPOSITO

SE APOYA EN TECNOLOGIAS, FUNDAMENTALMENTE EN INTERNET

OFRECEN HERRAMIENTAS Y APLICACIONES O RECURSOS INFORMATICOS (SS: SOFTWARE SOCIAL), PARA IMPLEMENTAR LAS CV

SITIOS DE REDES SOCIALES (SNS: Social Network Services – Servicios de Redes Sociales)

FORMAN PARTE DEL SS: SITIOS TIPICOS COMO FACEBOOK, ENTRE OTROS, Y OTROS MAS GENERICOS COMO BLOG, FOROS,.

Herramientas

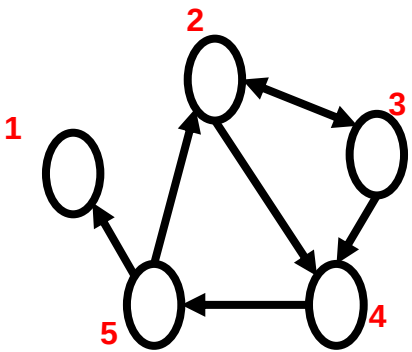
- **Gephi** (visualization and basic network metrics)
- **NetLogo** (modeling network dynamics)
- **Pajek:** amplia funcionalidad basada en menús, incluyendo muchas, muchas métricas de red y manipulaciones
  - pero ... no extensible
- **Guess:** extensibles, herramientas de secuencias de comandos de análisis exploratorio de datos, pero la selección más limitada de métodos incorporados en comparación con Pajek
- **NetLogo:** plataforma general agente basado en la simulación con el apoyo de modelado excelente red
  - muchos de los demos en este curso fueron construidos con NetLogo
- **IGRAPH:** utilizado en la versión de nivel de doctorado. bibliotecas se puede acceder a través de R o Python. Rutinas escalan a millones de nodos. (for programming assignments)

Elementos de un grafo

- Dirigido
  - A -> B
    - A le gusta B, A le dio un regalo a B, A es hijo de B
- No dirigido
  - A <-> B o A – B
    - A y B se gustan, son semejantes
  - Peso (frecuencia de comunicación)
  - ranking (mejor amigo, segundo mejor amigo…)
  - tipo (amigo, pariente, co-trabajador)

Representación de los datos

- Matriz de adyacencia

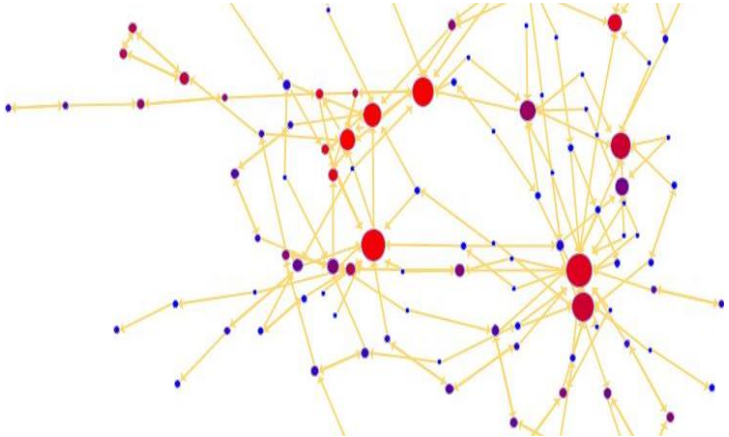


$$A = \begin{matrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 \end{matrix}$$

- Lista de arcos
  - 2, 3
  - 2, 4
  - 3, 2
  - 3, 4
  - 4, 5
  - 5, 2
  - 5, 1

- Lista de adyacencia
  - Todos los vecinos de cada nodo
    - 1:
    - 2: 3 4
    - 3: 2 4
    - 4: 5
    - 5: 1 2
  - Más fácil para redes
    - Grandes
    - Dispersas

Métricas



¿Cuál es el nodo con más arcos?

Métricas de redes

- **Cada métrica de red da respuesta a las siguientes preguntas:**

- pregunta: **¿Quién es más central?**

  1) METRICA DE RED: centralidad

  a)Centralidad de grado (degree centrality).
  1) Indegree o grado de entrada
  2) Outdegree o grado de salida
  b) Centralidad de cercanía (closeness centrality).
  c) Centralidad de intermediación (Betweenness centrality).

- pregunta: **¿Todo está conectado?**

  2) METRICA DE RED: los componentes conectados

  - Componentes fuertemente conectados:
  -Componentes Débilmente conectados:

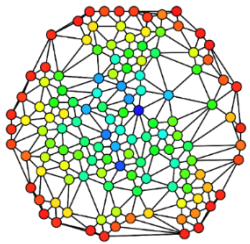
  3) METRICA DE RED: tamaño de componente gigante(giant component)

- pregunta: **¿A qué distancia están las cosas?**

  4) METRICA DE RED: rutas más cortas

- pregunta: **¿Cómo densa son?**

  5) METRICA DE RED: densidad grafo



Métricas: Propiedades de los nodos de la Red

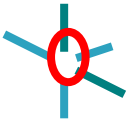
- Conexiones
  - indegree
    cuantos arcos estan dirigidos al nodo
  - outdegree
    arcos que salen del nodo
  - degree (in or out)
    todos los arcos del nodo, entrada y salida

indegree=3

$$\sum_{i=1}^{n} A_{ij}$$

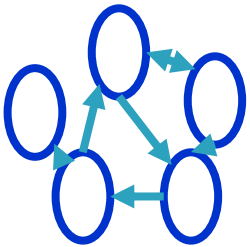
outdegree=2

$$\sum_{j=1}^{n} A_{ij}$$



degree=5

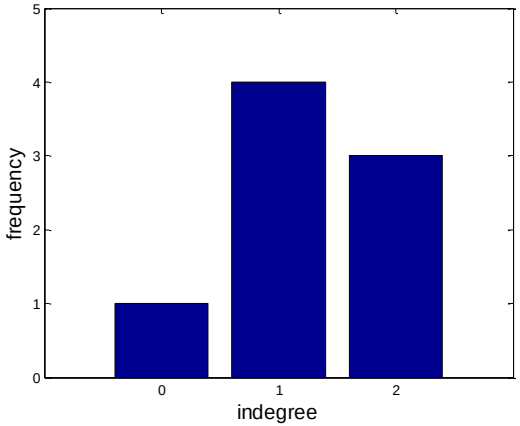
- Degree sequence: Lista ordenada de los grados de cada nodo
  - In-degree sequence:
    - [2, 2, 2, 1, 1, 1, 1, 0]
  - Out-degree sequence:
    - [2, 2, 2, 2, 1, 1, 1, 0]
  - (undirected) degree sequence:
    - [3, 3, 3, 2, 2, 1, 1, 1]



Métricas: Propiedades de los nodos de la Red

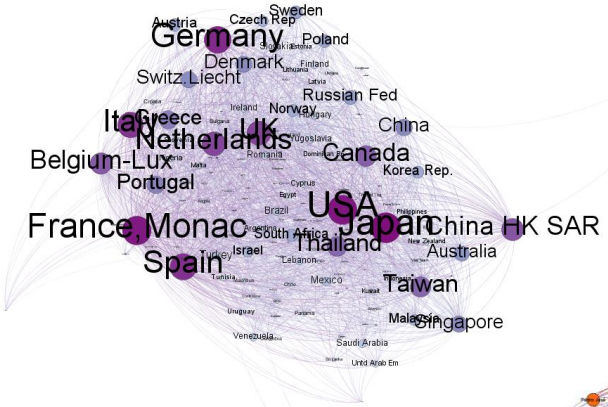
- Degree distribution: La frecuencia con la que ocurre cada grado

  - In-degree distribution:
    - [(2,3) (1,4) (0,1)]
  - Out-degree distribution:
    - [(2,4) (1,3) (0,1)]
  - (undirected) distribution:
    - [(3,3) (2,2) (1,3)]

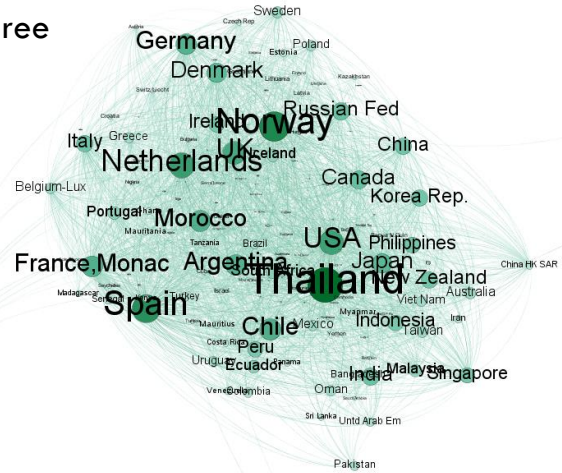


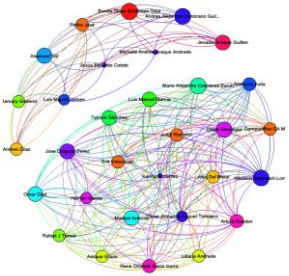
Métricas

InDegree



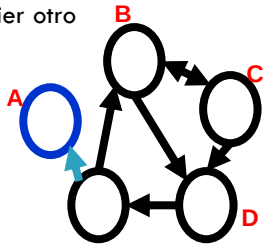
OutDegree

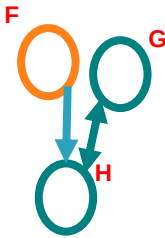




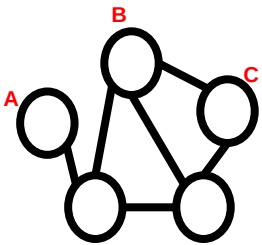
Métricas: Componentes conectados

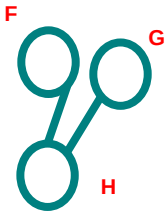
- **Componentes fuertemente conectados:**
  - Cada nodo dentro del componente se puede llegar desde cualquier otro nodo en el componente siguiendo los enlaces dirigidos
    - Componentes fuertemente conectados
      - B C D E
      - La
      - G H
      - F





- **Componentes Débilmente conectados:** cada nodo se puede llegar desde cualquier otro nodo siguientes enlaces en cualquier dirección
  - Componentes débilmente conectados
    - A B C D E
    - G H F

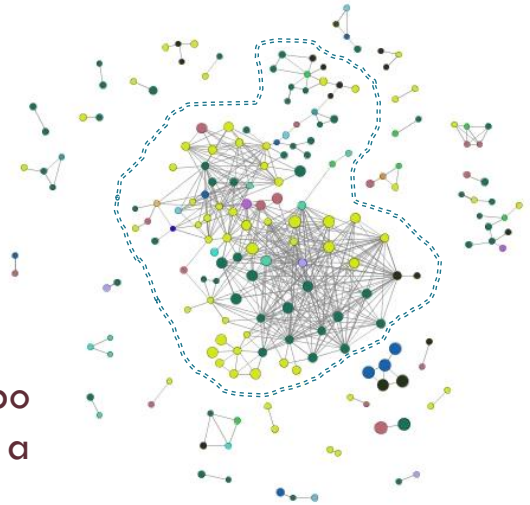




- En las **redes no dirigidos** se habla simplemente de "**componentes conectados** '

Métricas: Componentes conectados

- Si el componente más grande ocupa una región significativa de la red o grafo, es llamado **giant component**

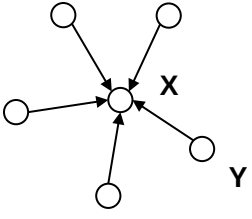


El componente gigante, consiste en un grupo de nodos enlazados entre si, y que agrupan a la mayoría de los nodos de la red.

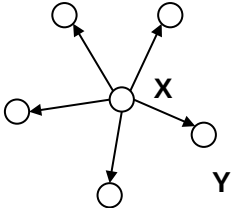
El componente gigante aparece también en casi todas las redes sociales.

Métricas: Centralidad

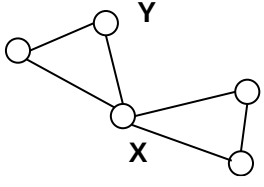
medida posible de un vértice en un grafo, que determina su importancia relativa dentro de éste



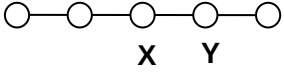
indegree



outdegree



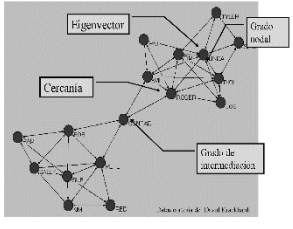
Betweenness (intermediación)



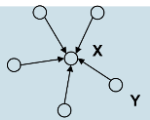
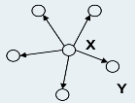
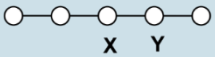
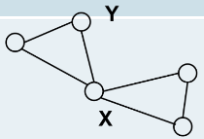
Closeness (cercanía)

La centralidad de vector propio («eigenvector centrality»).

Cuatro Aspectos de la Centralidad



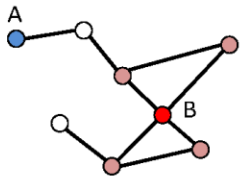
Métricas principales en la determinación de la centralidad de una unidad

Métricas de centralidad		
a) **Centralidad de grado (*degree centrality*)**: Es una métrica de qué tantas conexiones directas tiene una unidad con otras unidades. Una unidad con alta centralidad de grado sirve como "conector" o "hub" de la red.	**Indegree** o grado de entrada	
	outdegree o grado de salida	
b) **Centralidad de cercanía (*closeness centrality*)**: Esta métrica indica que tan "cerca" se encuentra una unidad de la red de las otras, considerando tanto conexiones directas como indirectas. Dado que una unidad con alta centralidad de cercanía puede interactuar fácilmente con otras unidades, tiene la visibilidad del comportamiento de la red en su conjunto – y puede influir en ella.	**closeness** o cercanía	
c) **Centralidad de intermediacion (Betweenness centrality)**: Esta métrica es un índice de en qué tantas rutas más cortas entre 2 unidades cualesquiera de la red se encuentra una unidad dada. Estas unidades tienen el control del flujo de información dentro de la red.	**betweenness** o intermediación	

Métricas: Centralidad

MEDIDAS LOCALES DE CENTRALIDAD | **Centralidad de grado**

No dirigida

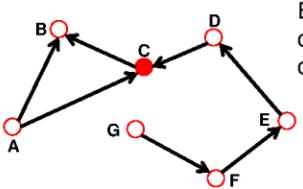


Centralidad de grado de un actor (C_D): número de enlaces que lo conectan con otros

$$C_D(\mathrm{A}) = k_{\mathrm{A}} = 1 \qquad C_D(\mathrm{B}) = k_{\mathrm{B}} = 4$$

C_D(i) se define en {0,*g*-1}, siendo *g* el número de nodos de la componente conexa

Dirigida



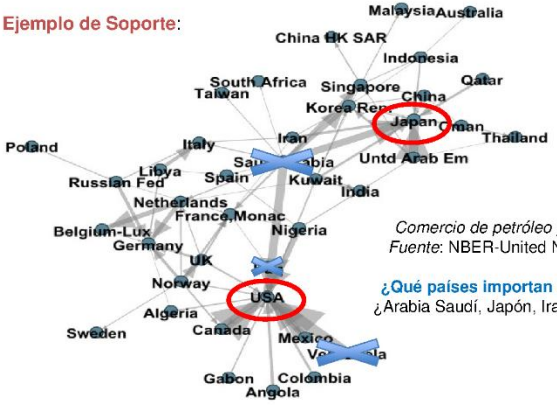
En redes dirigidas, se define el **Prestigio** de entrada (*in-degree*), denominado **Soporte**, y el **Prestigio** de salida (*out-degree*), denominado **Influencia**:

$$P_D^{in}(\mathrm{C}) = k_C^{in} = 2 \qquad P_D^{out}(\mathrm{C}) = k_C^{out} = 1$$

Ambos se definen en {0,*g*-1}

MEDIDAS LOCALES DE CENTRALIDAD | Soporte

Ejemplo de Soporte:

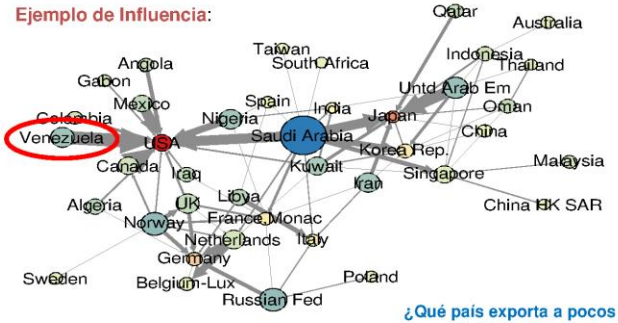


Comercio de petróleo y derivados, 1998
Fuente: NBER-United Nations Trade Data

¿Qué países importan de muchos otros?
¿Arabia Saudí, Japón, Iraq, USA, Venezuela?

MEDIDAS LOCALES DE CENTRALIDAD | Influencia

Ejemplo de Influencia:



¿Qué país exporta a pocos países (*out-degree* bajo) **pero lo hace en gran cantidad** (grosor de los arcos = volumen exportado)**?**
¿Arabia Saudí, Japón, Irak, USA, Venezuela?

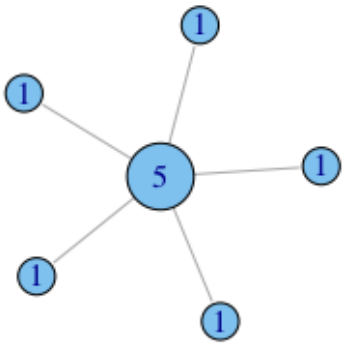
Métricas: Centralidad

$$C'_D(p_k) = \frac{\sum_{i=1}^{n} a(p_i, p_k)}{n-1}$$

$$C_D = \frac{\sum_{i=1}^{g} \left[C_D(n^*) - C_D(i) \right]}{[(N-1)(N-2)]}$$

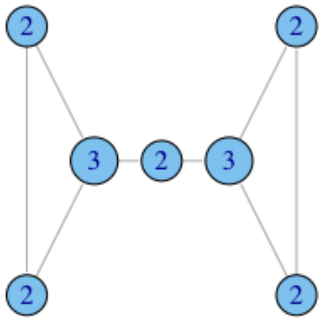
Máximo valor de conexiones posibles en la red

Formula de centralidad general de Freeman's



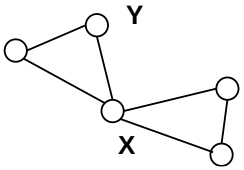
$C_D = 1.0$

$C_D = 0.167$



$C_D = 0.167$

Métricas: betweenness

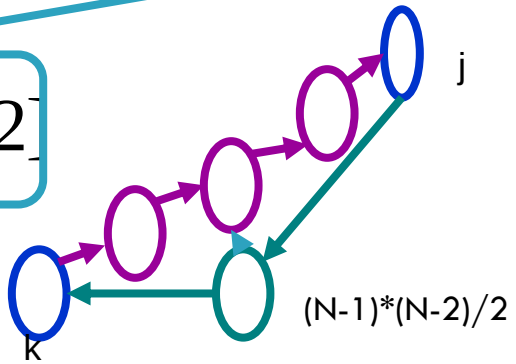


Donde g_{jk} = numero de caminos cortos que conectan *jk*
$g_{jk}(i)$ = numero de caminos cortos en los que el nodo i se encuentra.

$$C_B(i) = \sum_{j<k} g_{jk}(i)/g_{jk}$$

Pares de vértices posibles excluyendo el del mismo nodo

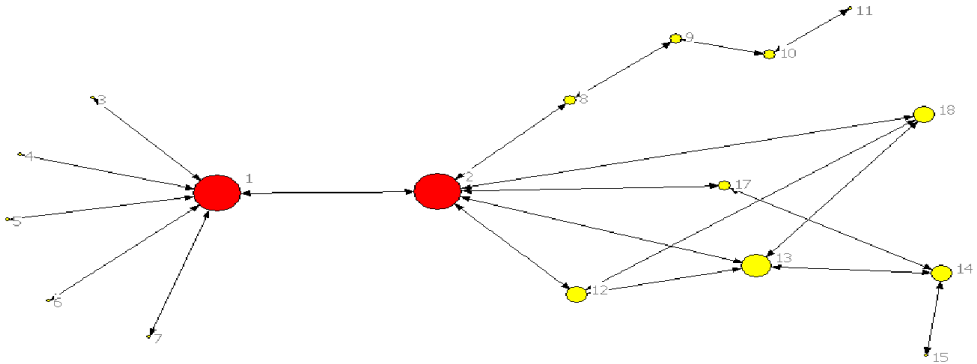
$$C_B'(i) = C_B(i)/[(n-1)(n-2)/2]$$



La centralidad del grado de intermediación ve al nodo con una posición favorable en la medida que el nodo está situado entre los caminos geodésicos entre otros pares de actores en la red.

MEDIDAS LOCALES DE CENTRALIDAD

Grado vs. Intermediación (1)

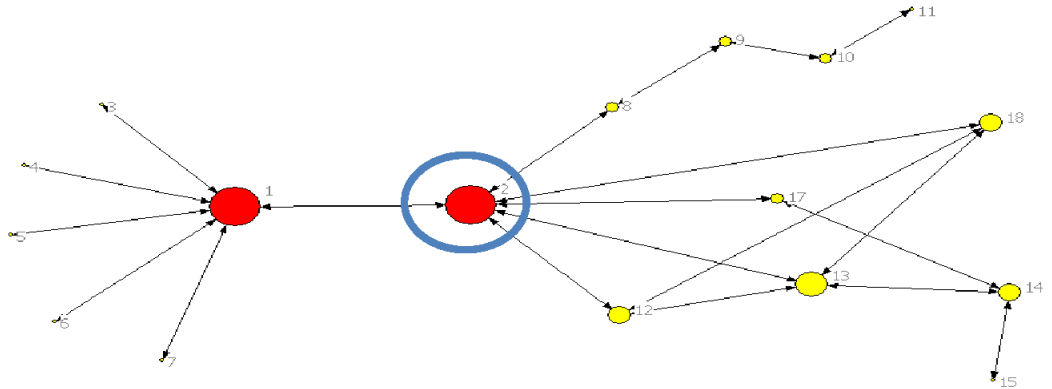


El grado de los nodos en rojo (1 y 2) es el mismo (6) pero, evidentemente, no son igual de importantes

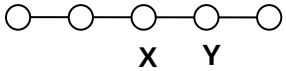
¿Por qué? Por la debilidad de esta medida, que solo toma en cuenta los vínculos inmediatos (a un nivel local) dejando de lado los vínculos indirectos (a nivel global)

MEDIDAS LOCALES DE CENTRALIDAD

Grado vs. Intermediación (2)



Nombre	Grado (C_D) Max = 17	Intermediación (C_B) Max = 17·16/2 = 136
1	6	70.00
2	6	96.50

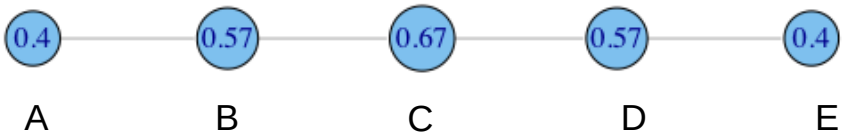


Métricas: closeness

Distancia promedio del camino mas corto entre un nodo a todos los nodos.

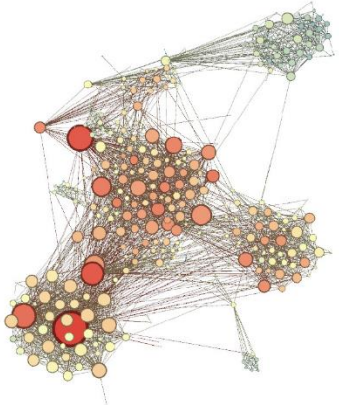
Closeness Centrality: $$C_c(i) = \left[\sum_{j=1}^{N} d(i,j) \right]^{-1}$$

Normalized Closeness Centrality $$C'_C(i) = (C_C(i))/(N-1)$$

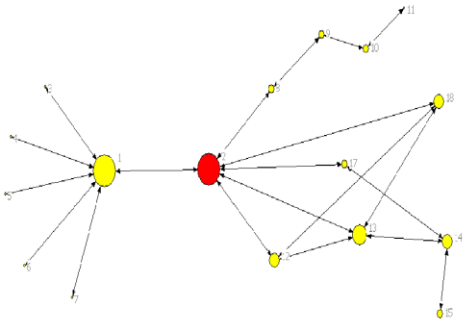


$$C'_c(A) = \frac{\left[\sum_{j=1}^{N} d(A,j) \right]^{-1}}{N-1} = \left[\frac{1+2+3+4}{4} \right]^{-1} = \left[\frac{10}{4} \right]^{-1} = 0.4$$

MEDIDAS LOCALES DE CENTRALIDAD | Grado vs. Cercanía



Red Personal de Contactos de Facebook de Oscar Cordón: el tamaño de los nodos indica el **grado** y el color la **cercanía** (más azul, menor valor; más rojo, mayor valor)

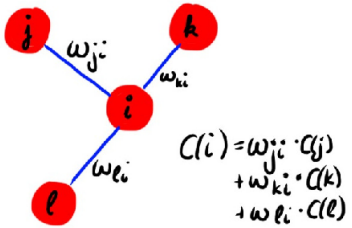


Actor	Lejanía	CercaníaN
2	34.000	50.000
1	40.000	42.500
13	42.000	40.476
17	44.000	38.636
8	44.000	38.636
12	45.000	37.778
18	45.000	37.778
14	52.000	32.692
6	56.000	30.357
5	56.000	30.357
7	56.000	30.357
3	56.000	30.357
4	56.000	30.357
9	56.000	30.357
15	66.000	25.758
10	70.000	24.286
16	82.000	20.732
11	86.000	19.767

MEDIDAS LOCALES DE CENTRALIDAD | Centralidad de vector propio (1)

La **Centralidad de vector propio** se basa en que la centralidad de un nodo concreto depende de cómo de centrales sean sus vecinos (**prominencia**)

La idea básica es que el poder y el status de un actor (***ego***) se define recursivamente a partir del poder y el status de sus vecinos (***alters***)



w_{ij} (a_{ij}) corresponde a la entrada de la matriz de adyacencia. Puede ser binaria {0,1} o un peso numérico

La medida es válida para redes dirigidas (**Prestigio de rango**) y no dirigidas

Es una versión más elaborada de la Centralidad de grado al asumir que no todas las conexiones tienen la misma importancia. No se tiene en cuenta la cantidad sino la calidad de las mismas

Centralidad de vector propio (2)

La medida de Centralidad de vector propio, C_{VP}, se define como una combinación lineal (o una **suma**, si los enlaces ponderados no están ponderados) de los valores de todos los actores que apunten a *i*:

$$C_{VP}(i) = a_{1i} \cdot C_{VP}(1) + a_{2i} \cdot C_{VP}(2) + ... + a_{ni} \cdot C_{VP}(n)$$

Para calcular los valores de C_{VP} para los *n* actores se construye un sistema de *n* ecuaciones con *n* incógnitas que se representa de forma matricial

Si $\boldsymbol{C}=(C_{VP}(1), \ldots, C_{VP}(n))^T$ es el vector transpuesto que almacena los *n* valores de C_{VP} ($\boldsymbol{C}$ es un vector columna) y $\boldsymbol{A}$ es la matriz de adyacencia, entonces:

$$\boldsymbol{C} = \boldsymbol{A}^T \cdot \boldsymbol{C}$$

Esta ecuación coincide con la **ecuación característica para encontrar los vectores y valores propios de la matriz $\boldsymbol{A}^T$**. $\boldsymbol{C}$ es un vector propio de $\boldsymbol{A}^T$

Métricas: Eigenvector centrality Bonacich

$$c_i(b) = \sum_j (a + bc_j)A_{ji}$$

$$c(b) = a(I - bA)^{-1}A1$$

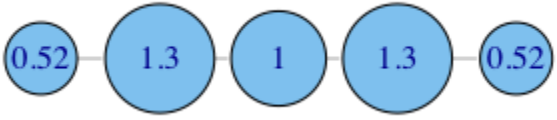
- α is a normalization constant
- β determines how important the centrality of your neighbors is
- **A** is the adjacency matrix (can be weighted)
- **I** is the identity matrix (1s down the diagonal, 0 off-diagonal)
- **1** is a matrix of all ones.

Si $\beta > 0$, los nodos tienen mayor centralidad cuando tienen enlaces con nodos con alta centralidad

β=.25

Si $\beta < 0$, los nodos tienen mayor centralidad cuando tienen enlaces a nodos no tan centrales

β=-.25



Otra medida local de centralidad basada en distancias es la **Centralidad de excentricidad (C_E)**. Se define como la inversa de la **excentricidad** (la máxima distancia geodésica) entre un actor y cualquier otro actor de la red:

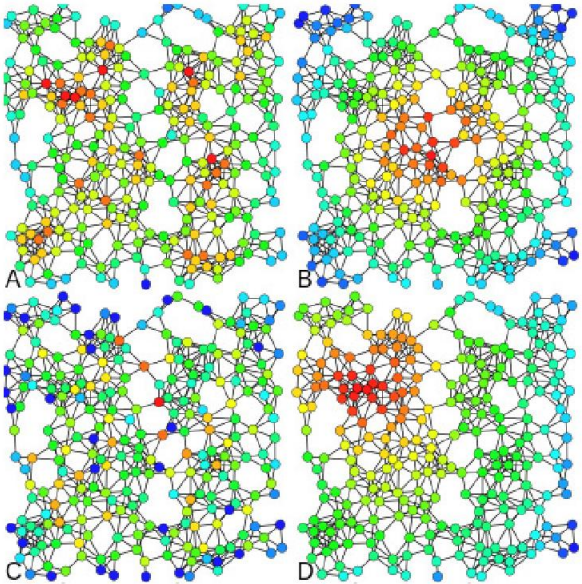
$$C_E(i) = \frac{1}{\max_{j \in V(G)/i} d(i, j)} \qquad C'_E(i) = C_E(i) / g - 1$$

Los actores con un mayor valor de excentricidad se denominan **actores periféricos**, los de menor valor forman el **centro de la red**

MEDIDAS LOCALES DE CENTRALIDAD | Comparativa

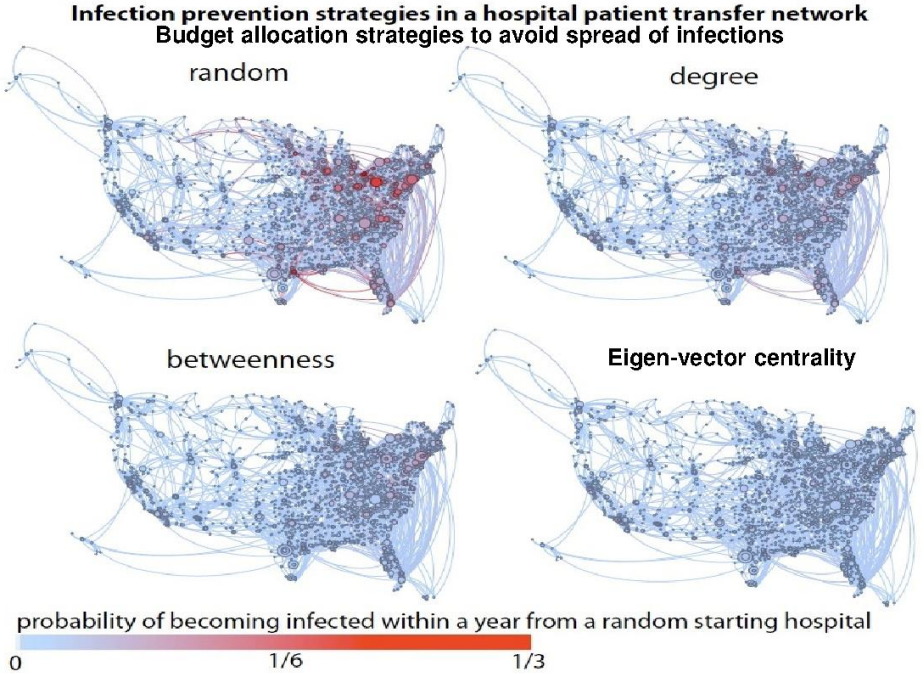
A) centralidad de grado; B) cercanía; C) intermediación; D) centralidad de vector propio

azul = menor valor



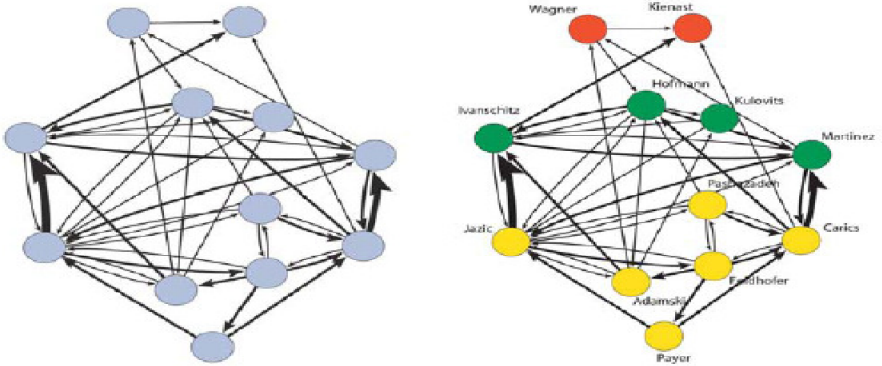
rojo = mayor valor

Infecciones en hospitales EEUU por transferencia de pacientes (4)



APLICACIONES — Análisis de juego en equipos de fútbol (1)

Let's look at an example on the soccer field.
The Rapid Vienna network consists of 12 persons (11 players on the field and one substitute), and we observed who passed the ball to whom during the course of a match. The resulting graph consists of a quantity of players (nodes) and a quantity of passes (arcs). To the left is the graph that depicts Rapid's passing game during the last 15 minutes of a soccer match between Rapid Vienna and Sturm Graz on December 7, 2003. As soon as we add additional information such as the players' names and their positions (red = attack, green = midfield, yellow = defense), we have produced a network. Networks are graphs with additional information about nodes and / or arcs.

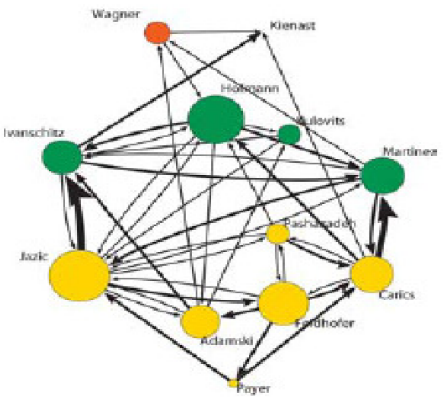


Rapid's passing game during the last 15 minutes of a soccer match between Rapid Vienna and Sturm Graz on December 7, 2003 (data by Harald Katzmair and Helmut Neundlinger). Left: graph, Right: network

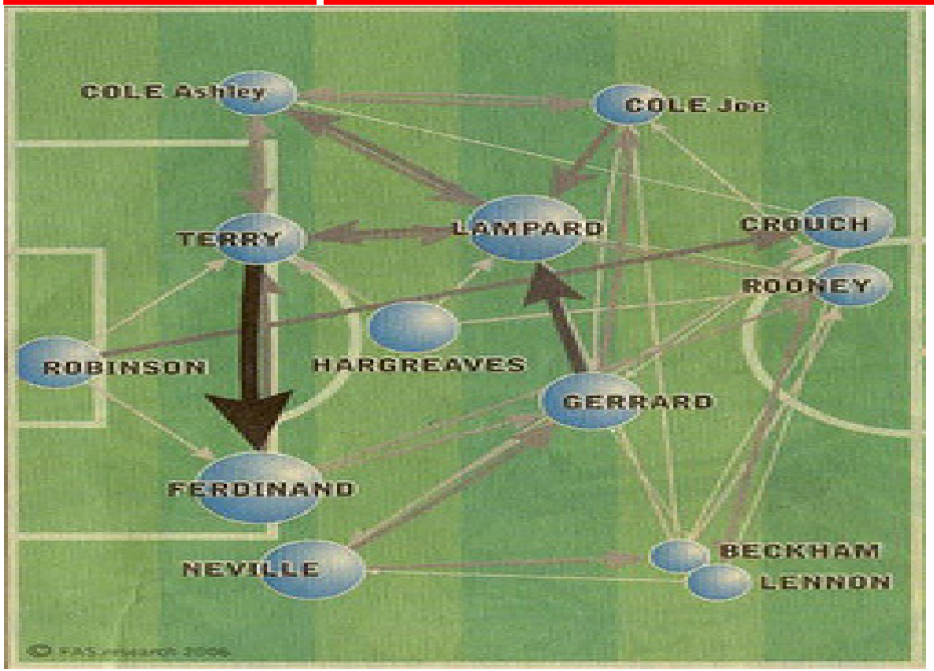
APLICACIONES | Análisis de juego en equipos de fútbol (2)

Posibles factores de análisis:

- ¿Qué jugador ha iniciado más pases (grado ponderado de salida)? **Jazic**
- ¿Qué jugador ha recibido más pases (grado ponderado de entrada)? **Jazic**
- ¿Quién ha controlado el juego del Rapid (centralidad)? **Jazic** y **Hoffman**
- ¿Qué jugadores han estado implicados en jugadas con el mayor número de pases (caminos)? **Jazic**, **Hofmann**, **Feldhofer**, **Martinez** y **Carics**
- ¿Quién ha jugado con quién y quién no (análisis de los enlaces)? **Ni un solo pase de Ivanschitz a Wagner**
- ¿Qué grupos de jugadores han compuesto la columna vertebral del equipo (análisis de triadas)? **Por ejemplo, Feldhofer-Carics-Pashazadeh**
- ¿Qué jugadores han tenido un rol similar (análisis de enlaces)? **Por ejemplo, Ivanschitz / Martinez**



APLICACIONES | Análisis de juego en equipos de fútbol (3)



Chelsea FC:

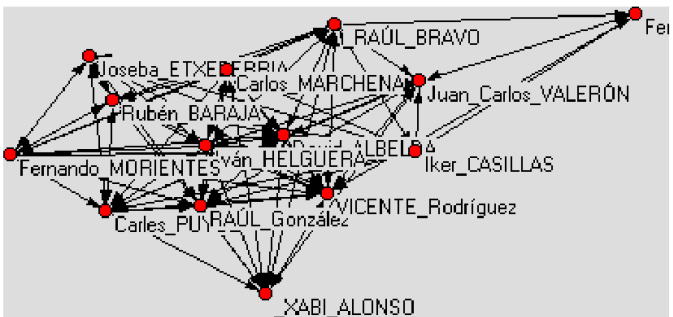
Robinson	1
Cole Ashley	2
Terry	3
Ferdinand	4
Neville	5
Cole Joe	6
Lampard	7
Heargraves	8
Gerrard	9
Beckham	10
Lennon	11
Crouch	12
Rooney	13

APLICACIONES — Análisis de juego en equipos de fútbol (4)

J.J. Merelo. Redes contra redes: el fútbol es así. http://atalaya.blogalia.com/historias/19642

En la lista Redes hemos propuesto analizar las redes que ser forman en los partidos de la Eurocopa de Portugal 2004. Con un programilla me bajo las páginas de estadísticas y les paso un programilla en Perl que extrae las estadísticas de pases

Los nodos son futbolistas y los enlaces, pases. Vamos a ver cómo fue la red del España 1 – Rusia 0

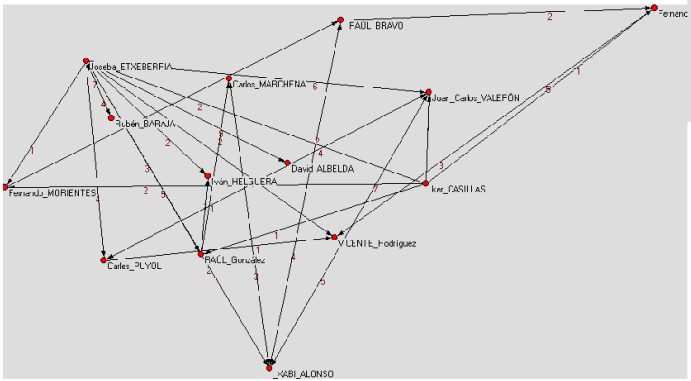


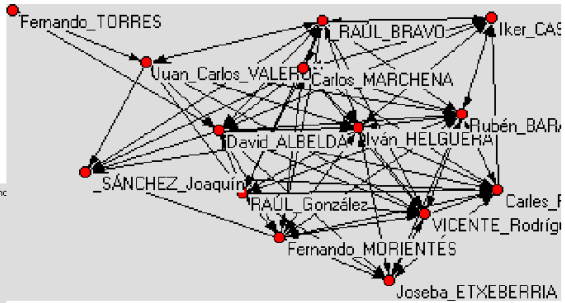
Una red donde, curiosamente, el jugador que tiene más centralidad es Iker Casillas, cuando debería ser un medio como Baraja

APLICACIONES — Análisis de juego en equipos de fútbol (5)

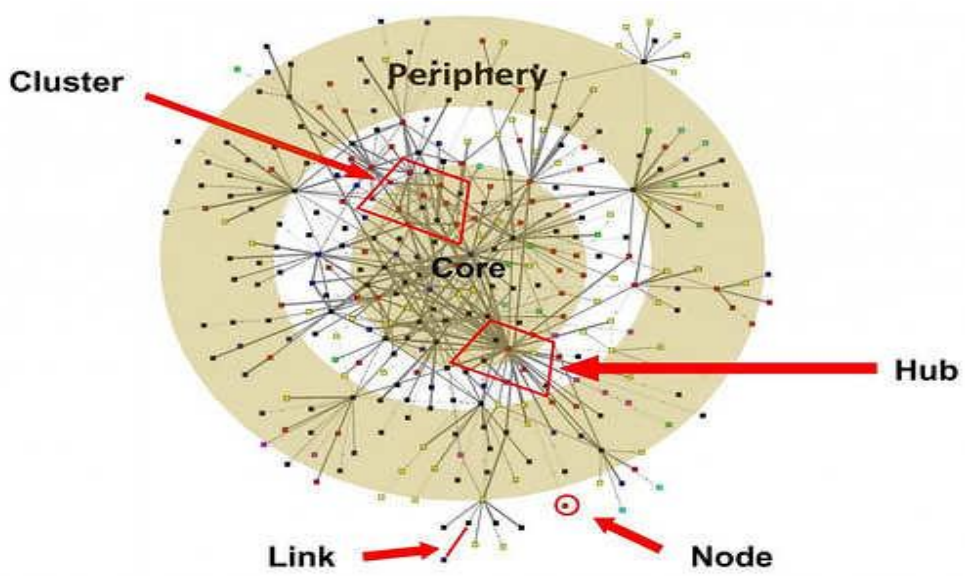
La situación no cambió mucho en el segundo partido (Grecia 1 – España 1) salvo que, en este caso, Albelda, Baraja y Helguera organizaron un poco más el juego. Y Fernando Torres a su bola, claro

Casi el 90% de los pases fueron los mismos. Esta es la diferencia de las dos redes (sin Joaquín porque es un nodo nuevo)





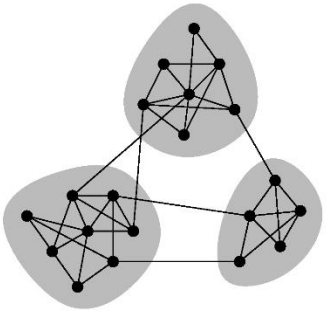
Es curioso ver también que la "autoridad" de la red es Vicente, un extremo. Lo lógico sería que las autoridades fueran los delanteros, pero Morientes y Raúl se hallan ahí perdidos, en la maraña de la red

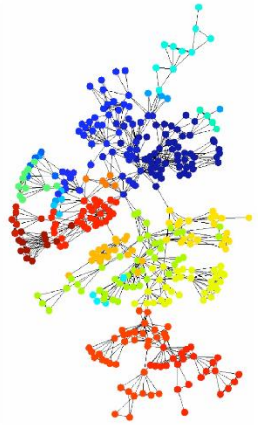

Cluster
Periphery
Core
Hub
Link
Node

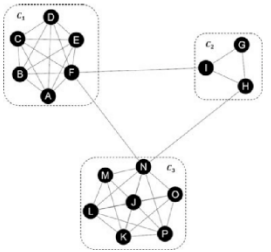
Métricas: Comunidades

- Mutualidad
  - Cada miembro conoce a todos los miembros
- Frecuencia
  - Cada miembro conoce al menos k miembros del grupo
- Cercanía
  - Los miembros están separados por máximo de n saltos

Comunidades - Clusters - Módulos



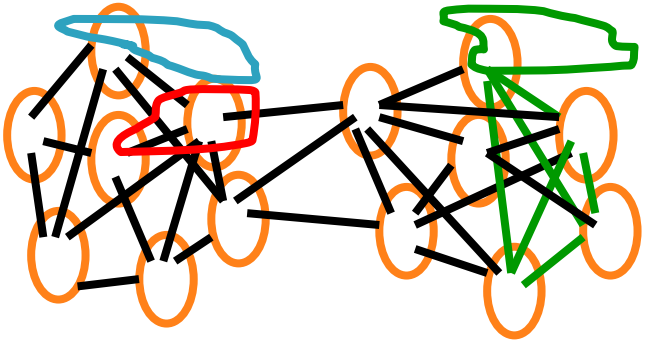




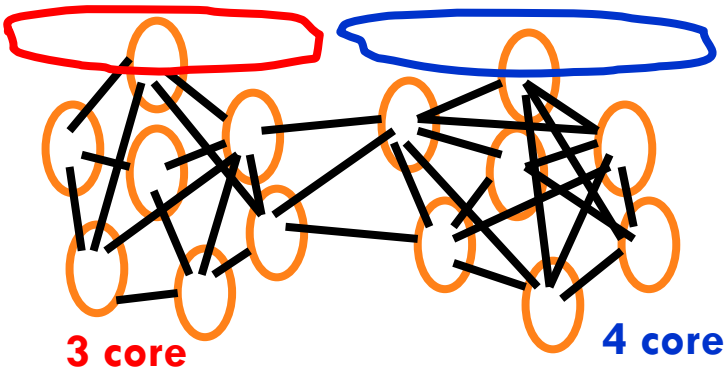
- En esta red, hay **tres comunidades**: C_1, C_2 y C_3
- Cada comunidad está formada por un grafo completo (un **clique**) de tamaño variable ($C_1 = K_6$, $C_2 = K_3$ y $C_3 = K_7$)
- La densidad de enlaces entre las comunidades es muy baja. Los pocos enlaces que existen son **puentes**

Cliques y K-core

Cada miembro del grupo posee un link a todos los miembros del grupo



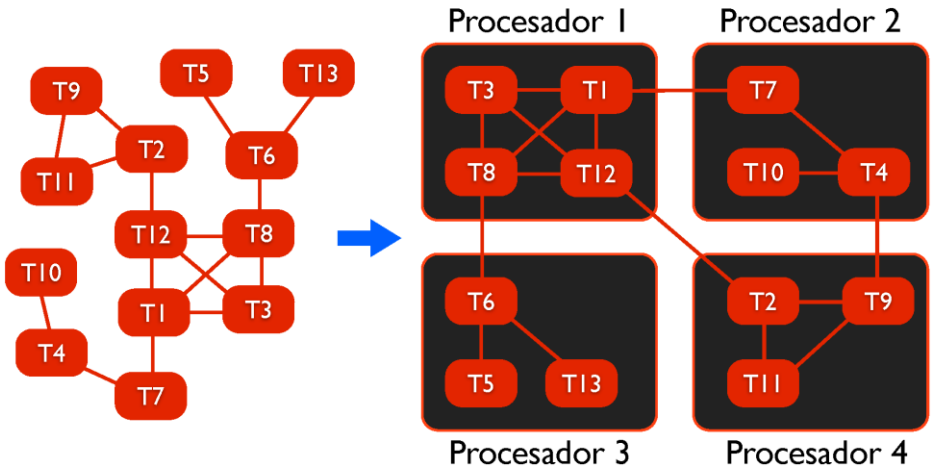
- Si se pierde un enlace, deja de ser un clique
- No es interesante que todos estén conectados con todos
- No hay medidas de centralidad dentro de un clique



Cada miembro del grupo está conectado con k otros miembros del grupo

Computación en Paralelo

(Particionado de Grafos)



Distribución mas eficiente de tareas en un conjunto de procesadores

Redes de colaboración científica

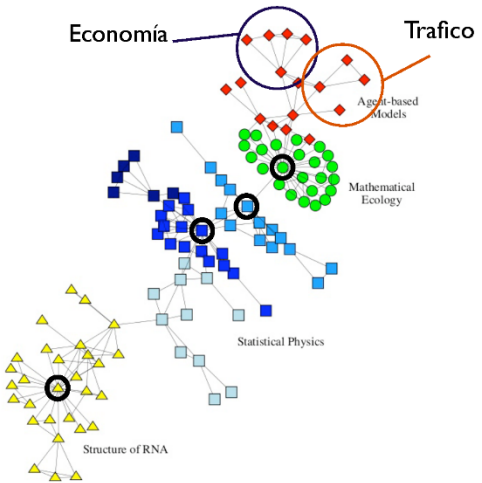
◆ Modelos basados en agentes para estudiar problemas de economía y flujo de tráfico

● 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

M. Girvan and M. E. J. Newman (2002). "Community structure in social and biological networks". Proc. Natl. Acad. Sci. USA 99 (12): 7821–7826. doi:10.1073/pnas.122653799. PMC 122977. PMID 12060727.

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	*Stephen Hawking's Universe* (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 *The Onion*)	9	"The Mathematical Guts of RSA Encryption"
41	"A Structure for Deoxyribose Nucleic Acid" (*Nature*, 2 Apr. 1953)	14	Stephen Hawking and Intel	8	German news story on Cryptography
...		...		...	
1	*Felix D'Herelle and the Origins of Molecular Biology* (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 *Stephen Hawking's Universe*	1	"A Very Brief History of Computer Science"
1	Magazine: *HMS Beagle Home*	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

G. Flake, S. Lawrence, C. L. Giles, and F. Coetzee. Self-organization and identification of web communities. IEEE Computer, 35:3, March 2002.

Métricas: Comunidades

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)

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

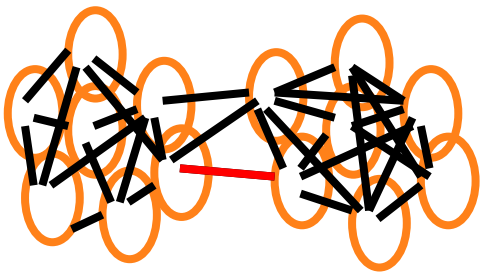
Betweenness clustering

Entrada: Grafo inicial G(E,V)

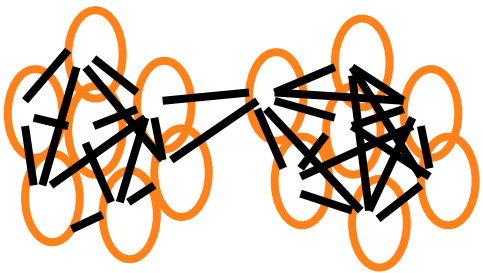
Inicio:

1. Calcular la intermediación para todas las aristas en la red.
2. Generar un nuevo grafo al remover las aristas con un valor alto de intermediación.
3. Recalcular la intermediación para todas las aristas del nuevo grafo.
4. Repetir el paso 2 hasta que todas las aristas tengan un nivel bajo de intermediación.

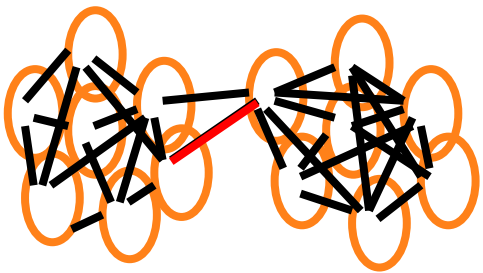
Betweenness clustering



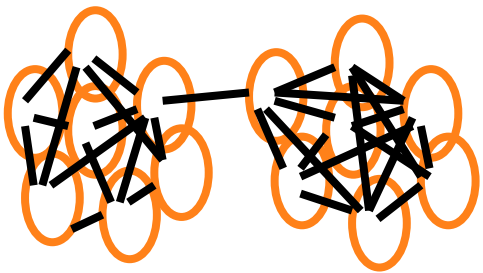
Betweenness clustering



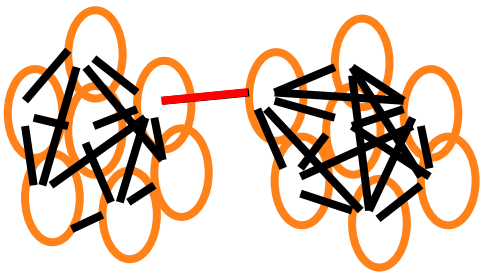
Betweenness clustering



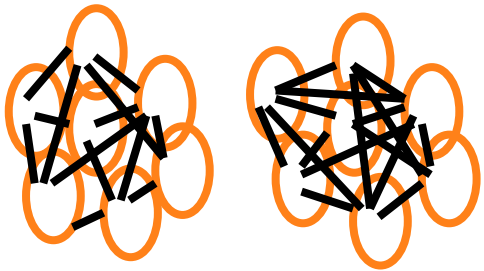
Betweenness clustering

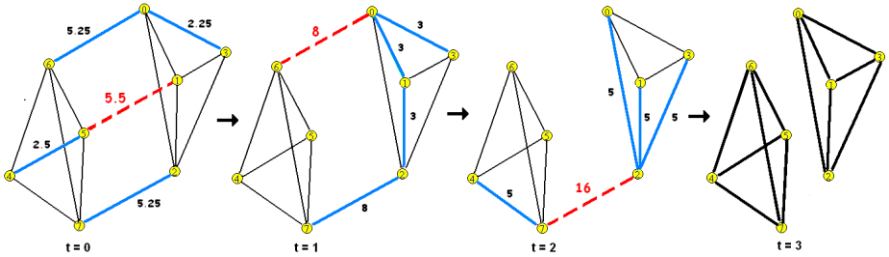


Betweenness clustering



Betweenness clustering

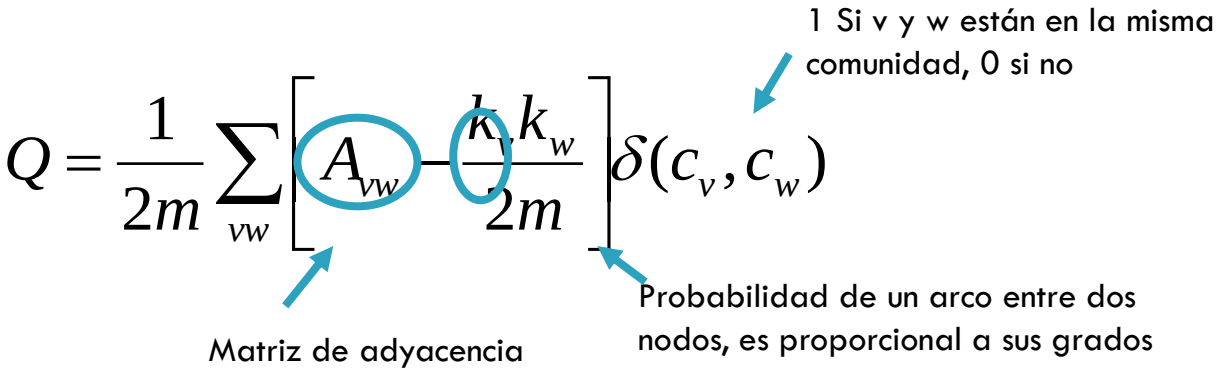




Modularidad

- Métrica diseñada para dividir la red en módulos, clusters, comunidades.
- Es usada para maximizar métodos para hallar comunidades
- Una red con alta modularidad significa que:
  - Es muy densa entre nodos de una misma comunidad
  - Pero con conexiones dispersas entre nodos de comunidades distintas

Modularidad



Detección basada en la Modularidad

La modularidad también se puede escribir como:

$$Q = \sum_{i}^{c} \left(e_{ii} - a_i^2 \right)$$

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

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

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.

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

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

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)

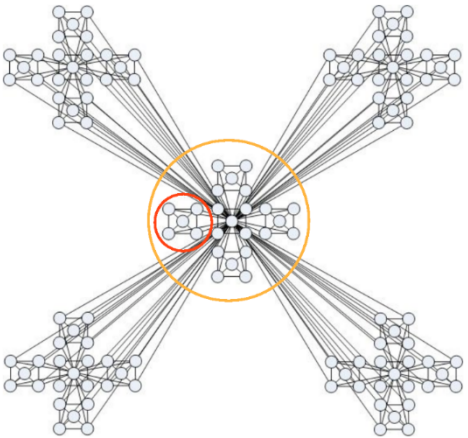
Las dos ultimas heurísticas han resultado efectivas sin embargo poseen un problema inherente al concepto de modularidad llamado *limite de resolución*.

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

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

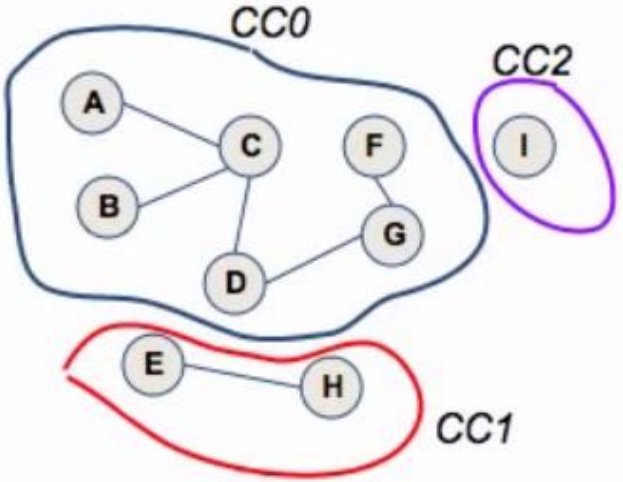
Redes Metabólicas

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

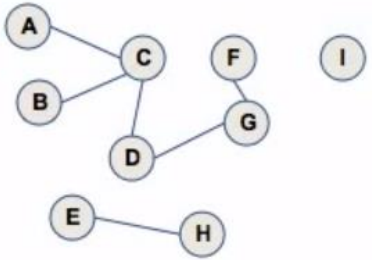


Red con modularidad jerárquica

Ravasz, E., Somera, A. L., Mongru, D.A., Oltvai, Z. N. and Barabasi, A. L. (2002). Hierarchical organization of modularity in metabolic networks. Science 297, 1551-5



¿Cuáles son los subgrafos conexos de el grafo?

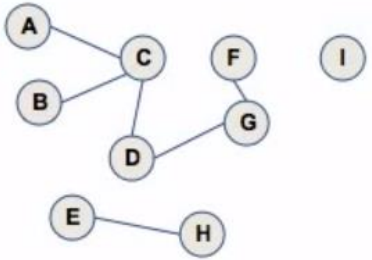


Estrategia:

n = 0

Hasta haber recorrido todos los nodos:

- Comenzando desde un nodo que no haya sido visitado, recorrer, marcando con n.
- n++



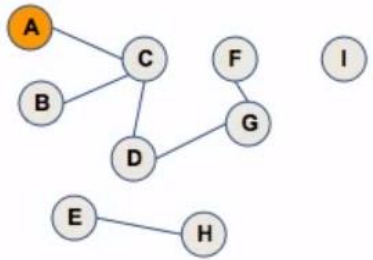
Estrategia:

n = 0

Hasta haber recorrido todos los nodos:

- Comenzando desde un nodo que no haya sido visitado, recorrer, marcando con n.
- n++

(CC0)



n = 0

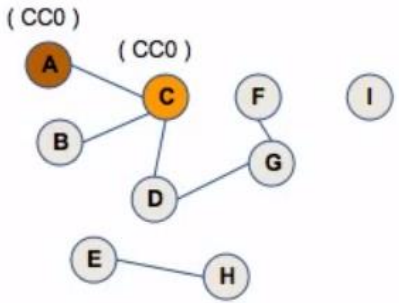
BreadthFirstTraversal(A)

Estrategia:

n = 0

Hasta haber recorrido todos los nodos:

- Comenzando desde un nodo que no haya sido visitado, recorrer, marcando con n.
- n++



n = 0

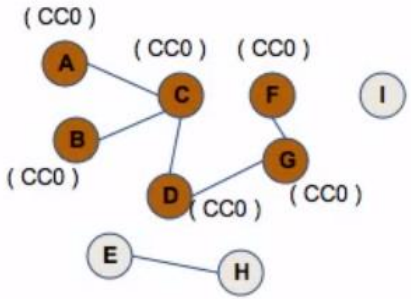
BreadthFirstTraversal(A)

Estrategia:

n = 0

Hasta haber recorrido todos los nodos:

- Comenzando desde un nodo que no haya sido visitado, recorrer, marcando con n.
- n++



n = 0

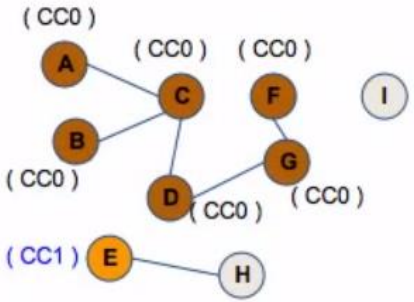
BreadthFirstTraversal(A)

Estrategia:

n = 0

Hasta haber recorrido todos los nodos:

- Comenzando desde un nodo que no haya sido visitado, recorrer, marcando con n.
- n++



n = 1

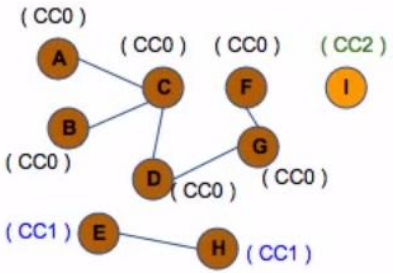
BreadthFirstTraversal(E)

Estrategia:

n = 0

Hasta haber recorrido todos los nodos:

- Comenzando desde un nodo que no haya sido visitado, recorrer, marcando con n.
- n++



n = 2

BreadthFirstTraversal(I)

Estrategia:

n = 0

Hasta haber recorrido todos los nodos:

- Comenzando desde un nodo que no haya sido visitado, recorrer, marcando con n.
- n++

MEDIDAS GLOBALES

Existen varias medidas globales en SNA. La mayoría son las mismas empleadas para analizar cualquier otro tipo de red, que ya hemos estudiado:

1. **Diámetro y Radio**
2. **Distancia media**
3. **Grado Medio**
4. **Densidad**
5. **Coeficiente de Clustering Global**
6. **Reciprocidad**

Otras, denominadas **medidas de Centralización**, son **agregaciones a nivel de grupo de actores** de las medidas locales de Centralidad que permiten comparar redes

MEDIDAS GLOBALES — Diámetro y Radio

Diámetro (d_{max}): longitud del camino mínimo más largo de la red

En redes grandes, se puede determinar con el algoritmo de búsqueda primero en anchura

Equivale al valor máximo de excentricidad para todos los nodos de la red:

$$E(i) = \max_{j \in V(G)/i} d(i,j) \qquad d_{\max} = \max \{E(i) : i \in V(G)\}$$

En el contexto del SNA, esta métrica da una idea de la proximidad entre pares de actores en la red, indicando cómo de lejos están en el peor de los casos

Las redes más dispersas suelen tener un mayor diámetro que las más densas al existir menos caminos entre cada par de nodos

Radio (**r**): Valor mínimo de excentricidad para toda la red:

$$r = \min \{E(i) : i \in V(G)\}$$

- Densidad: Actividad global de la red
  - En redes no dirigidas:
    $\Delta = L/[g\,(g-1)/2]$
  - En redes dirigidas:
    $\Delta = L / g\,(g-1)$

L, número de enlaces
g, número de nodos

MEDIDAS GLOBALES — Distancia media

Distancia media (<d>) para un grafo dirigido:

$$\langle d \rangle \equiv \frac{1}{2L_{\max}} \sum_{i,j \neq i} d_{ij}$$

d_{ij} es la distancia geodésica entre los nodos *i* y *j*

En un grafo no dirigido $d_{ij}=d_{ji}$. De este modo, sólo es necesario contar la longitud de los caminos una vez:

$$\langle d \rangle \equiv \frac{1}{L_{\max}} \sum_{i,j>i} d_{ij}$$

La medida da una idea de cómo de lejos están los distintos actores en promedio. En SNA representa la eficiencia del flujo de información en la red

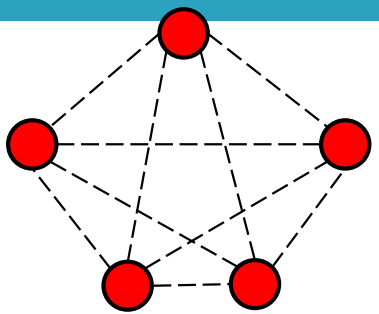
Medidad de red : Densidad de una red no dirigida

En el siguiente grafo:

Las conexiones posibles son:

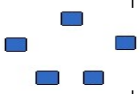
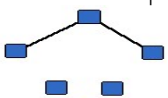
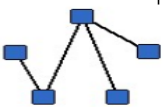
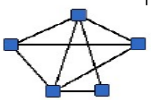
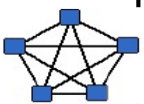
10 conexiones posibles en el grafo.

Pos.= posibles
Eff . = efectivas



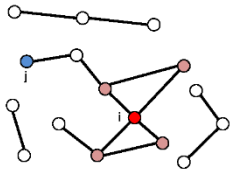
Medidas de red
Densidad de una red no dirigida

$$d = \frac{n^{\circ}\ effective\ edges}{n^{\circ}\ possibleedges}$$

Eff.=0 Pos.=10	Eff.=2 Pos.=10	Eff.=4 Pos.=10	Eff.=8 Pos.=10	Eff.=10 Pos.=10
d=0	d=0.2	d=0.4	d=0.8	d=1
				

MEDIDAS GLOBALES — Grado medio

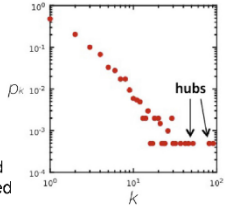
No dirigida



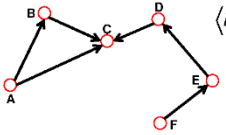
$$\langle k \rangle \equiv \frac{1}{N}\sum_{i=1}^{N} k_i$$

$$\langle k \rangle \equiv \frac{2L}{N}$$

N – número de nodos de la red
L – número de enlaces de la red



Dirigida



$$\langle k^{in} \rangle \equiv \frac{1}{N}\sum_{i=1}^{N} k_i^{in}, \quad \langle k^{out} \rangle \equiv \frac{1}{N}\sum_{i=1}^{N} k_i^{out}, \quad \langle k^{in} \rangle = \langle k^{out} \rangle$$

$$\langle k \rangle = \frac{L}{N}$$

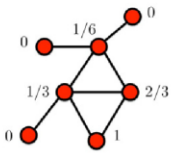
MEDIDAS GLOBALES — Densidad y Coeficiente de Clustering

La **densidad D** = L/L_{max} mide el grado de conectividad de la red social a nivel global

El coeficiente de clustering C_i mide la densidad local, el ratio de vecinos de un nodo que están conectados entre sí

El **coeficiente de clustering medio (<*C*>)** es una medida global que indica la probabilidad de que dos vecinos de un nodo de la red escogido aleatoriamente estén conectados entre sí:

$$C_i = \frac{2L_i}{k_i(k_i-1)} \qquad \langle C \rangle = \frac{1}{N}\sum_{i=1}^{N} C_i$$



$$\langle C \rangle = \frac{13}{42} \approx 0.310$$

Las redes sociales reales son redes de mundos pequeños (*small worlds*) y suelen tener valores de <*C*> altos. Eso implica que la transitividad entre nodos aparece más y con más fuerza, incrementando la probabilidad de que se formen cliques

MEDIDAS GLOBALES | Reciprocidad

La **reciprocidad (R)** en un grafo dirigido mide la tendencia de pares de actores a tener conexiones mutuas entre ellos

La forma más habitual de calcularla es como un ratio entre el número de conexiones (*díadas*) mutuas (*#mut*) y el número de conexiones totales en la red, las mutuas y las asimétricas (*#asim*):

$$R = \frac{\#mut}{\#mut + \#asim}, \quad R \in [0,1]$$

Indica la probabilidad de que dos actores de la red apunten el uno al otro

Una *díada asimétrica* es un par de actores que presentan un arco en una u otra dirección pero no en ambas. Una díada mutua incluye los arcos en las dos direcciones

APLICACIONES — Aprendizaje Colaborativo por Ordenador (1)

Aprendizaje Colaborativo por Ordenador: énfasis en las relaciones entre los actores en un curso on-line en BSCW

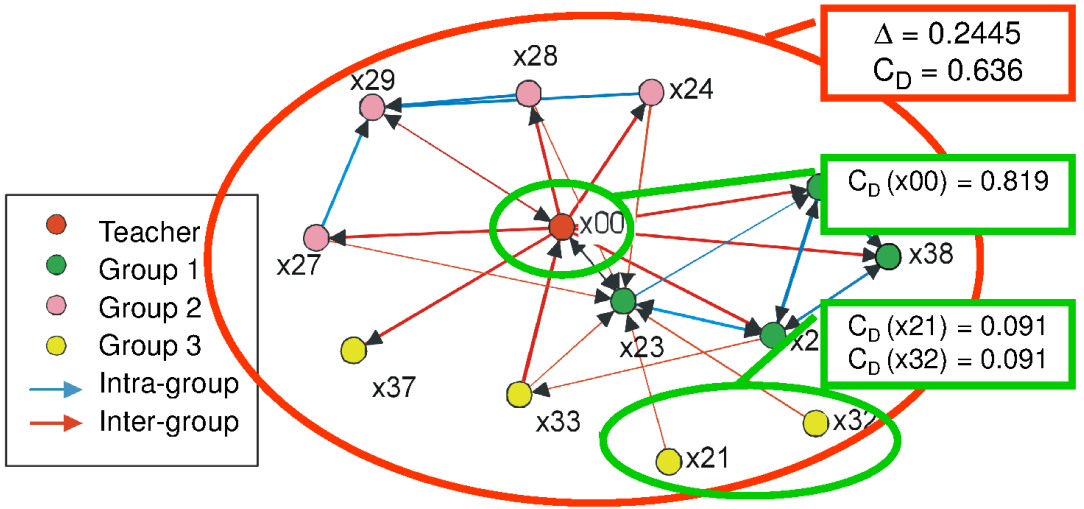
Relación entre el profesor y los alumnos en un curso, así como entre los propios alumnos de distintos grupos

Red Social con distintas variantes: trimodal egocéntrica (tres grupos de alumnos-profesor) y trimodal, unimodal completa (todos los alumnos con todos) y trimodal completa (miembros de un grupo con los de otros)

Pregunta global: ¿Cómo ayudar a los profesores a monitorizar aspectos colaborativos de aprendizaje mediante la tecnología?

Análisis con dos medidas globales (Densidad de la red Δ y Centralización de grado C_D) y dos medidas locales (Centralidades de grado y de cercanía)

APLICACIONES | **Aprendizaje Colaborativo por Ordenador (2)**



APLICACIONES | Aprendizaje Colaborativo por Ordenador (3)

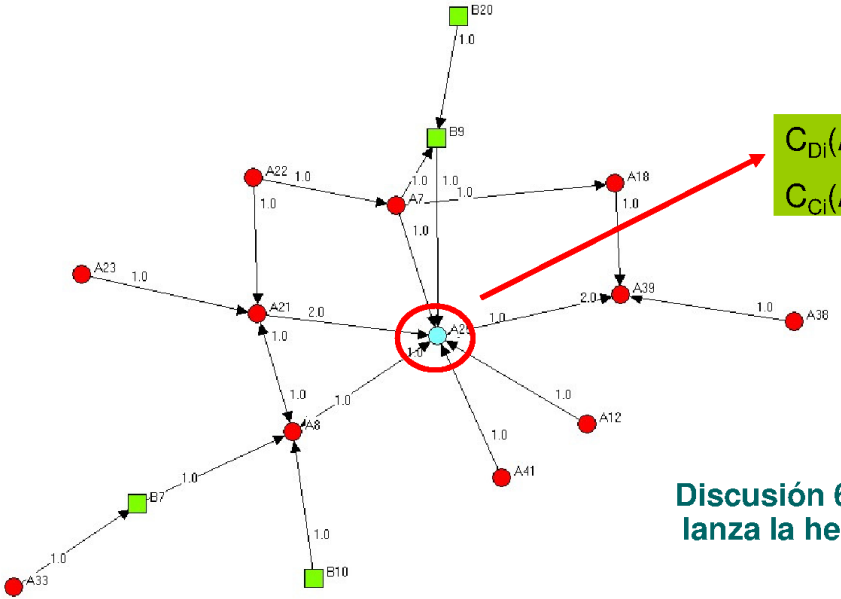
Análisis de influencia de los actores en las discusiones del curso

Workshop on interaction analysis approaches (CSCL 2009)

ID	Title	Thread_ID	Parent_ID	date	Student Name	School
21	Solutions of it is Energy Conservation?	3	0	2007-04-30	A41	A
22	I think it is too late for us to solve the global warming	3	21	2007-05-19	A5	A

```
<SESSION id="Thread_3" date=30.04.2007>
<ACTION>
  <ACT.TIMESTAMP>19.05.2007 00:00:00</ACT.TIMESTAMP>
    <ACT.SOURCE ref="A5" />
      <ACT.DESC>
        <ACT.DIR type="Debate">
          <ACT.DIR.DEST ref="A41" />
        </ACT.DIR>
      </ACT.DESC>
</ACTION>
</SESSION>
```

APLICACIONES — Aprendizaje Colaborativo por Ordenador (4)

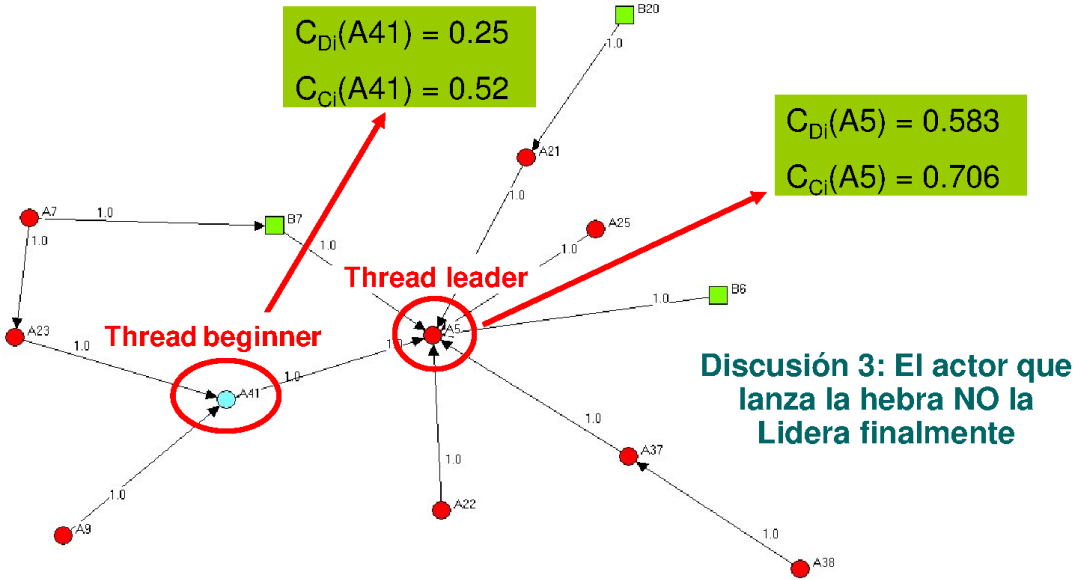


$C_{Di}(A25) = 0.467$

$C_{Ci}(A25) = 0.625$

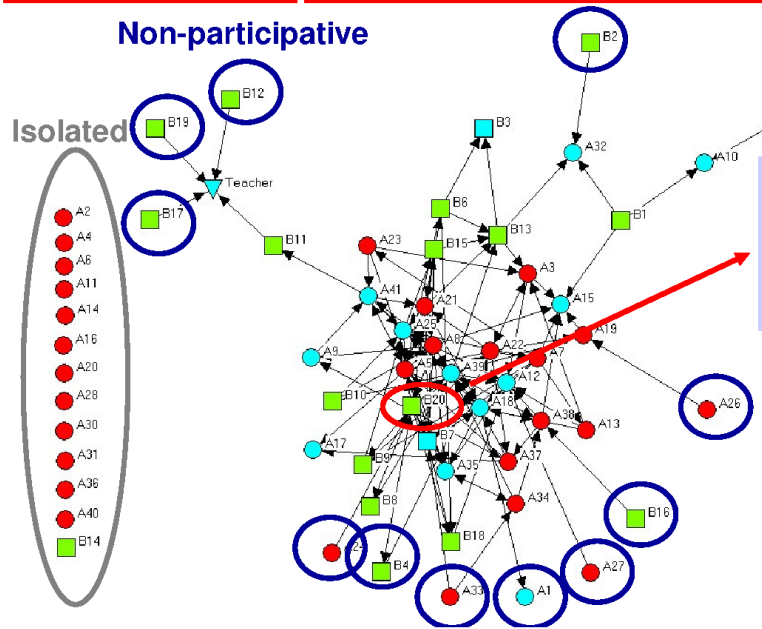
Discusión 6: El actor que lanza la hebra es el líder

APLICACIONES | Aprendizaje Colaborativo por Ordenador (5)



Discusión 3: El actor que lanza la hebra NO la Lidera finalmente

APLICACIONES | Aprendizaje Colaborativo por Ordenador (6)



Dynamizer

$C_{Do}(B20) = 16$

$C_{Do\text{-}sessions}(B20) = 0.308$

$C_{Di}(B20) = 4$ (17^{th} value)

Detección de roles

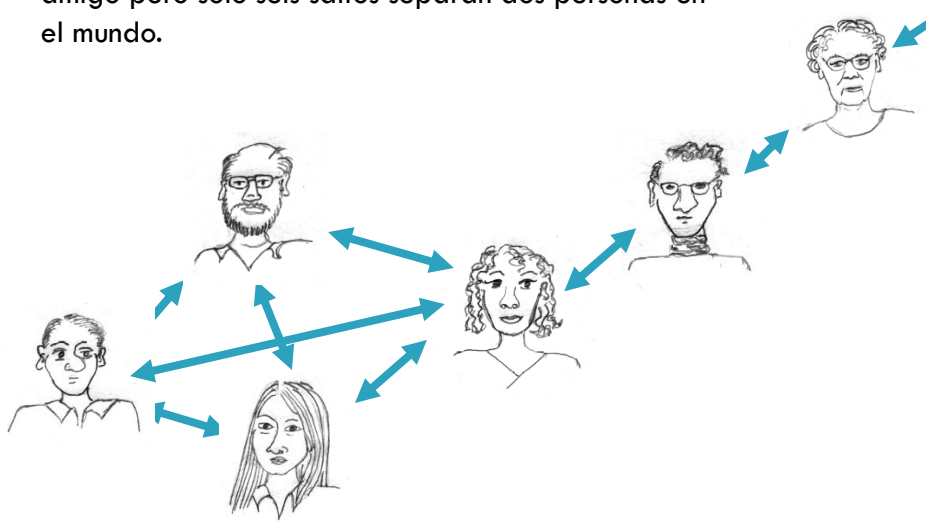
Pequeños mundos

- Fenómeno donde:
  - Todos los nodos pueden ser alcanzado dentro de un número pequeño de saltos
  - La distancia L de la red (puede ser medida con el avg shorted path) crece de manera logarítmica a medida que se agregan nodos a la red.
  - Redes que presentan este fenómeno:
    - Redes sociales, wikis, el Internet, redes genéticas, de proteínas, mapas de camino, cadenas alimenticias, grafos de llamadas.

Pequeños mundos

Pequeños mundos:

un amigo de un amigo es también con frecuencia un amigo pero sólo seis saltos separan dos personas en el mundo.

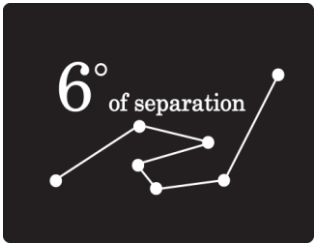


redes de pequeños mundos

El mundo es un pañuelo

La teoría de los Seis Grados de Separación

Según esta teoría sólo seis niveles nos separan de cualquier persona del planeta. Sólo Seis pasos. Seis grados.



Facebook acorta los "seis grados de separación"

¿De qué me sirve?

- Propiedades de small world network:
  - Poseen cliques definidos, en los cuales puedo definir comunidades
  - Sin importar el tamaño en nodos de la red, un nodo puede llegar a otro en un camino mínimo pequeño
  - Presentan gran cantidad de hubs, nodos con muchas conexiones

Modelos de construcción de Redes

- Dos procedimientos principales:
  - Modelos Estadísticos
    - Modelo de Grafos aleatorios
    - Modelo de Wattz-Strogatz
  - Modelos Dinámicos
    - Enlace Preferencial
    - Duplicación

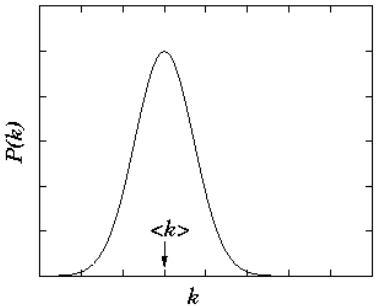
Modelos de construcción de Redes

- Modelo de Grafos aleatorios

  En Matemáticas se denomina **grafo aleatorio** a un **grafo** que es **generado** por algún **tipo de proceso aleatorio.**

- Nodos conectados al azar
- Número de aristas incidentes en cada nodo se distribuye Poisson

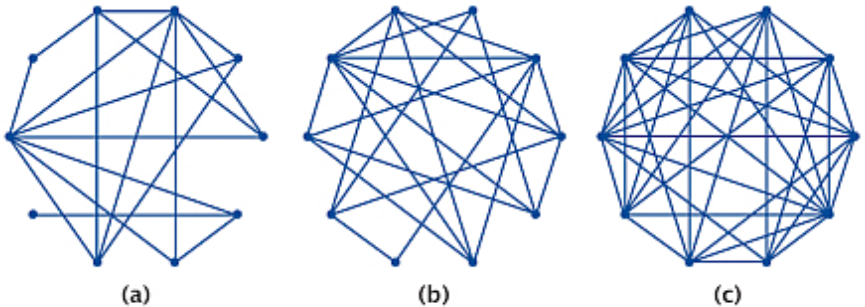
Distribución de Poisson



Modelo de Grafos Aleatorios

Iniciador de la Teoría de Redes Complejas: Erdos-Renyi (50's).

- Construye la red enlazando nodos elegidos al azar según determinada probabilidad



Grafos Aleatorios con distinta probabilidad de conexión entre pares de nodos (A) p=0.3, (B) p=0.5 y (C) p=0.8

En este **modelo Erdös–Rényi** se tiene que **un nuevo nodo se enlaza con igual probabilidad con el resto de la red,**

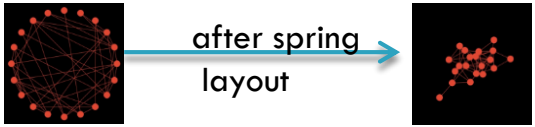
Cómo detectarlos

- Grafos aleatorios:
  - El modelo de Erdos-Renyi presenta un avg shorted path relativamente pequeño (al igual que el fenómeno de pequeños mundos)
  - Pero se diferencian en que el modelo de grafo aleatorio de Erdos-Renyi posee un clustering coef pequeño y las redes de pequeños mundo lo poseen alto

Erdös and Rényi Random Graph

Suposiciones:

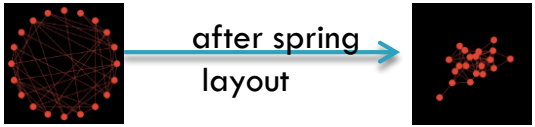
- Nodos se conectan de manera aleatoria
- Grafo no dirigido
- Probabilidad P de dos nodos se conecten, M numero de enlaces.



Erdös and Rényi Random Graph

Caracteristicas:

- El max degree obtenido de los nodos no se diferencia mucho al avg degree
- A medida que N aumenta el avg degree aumenta
- No se observan grandes hubs (nodos puentes).



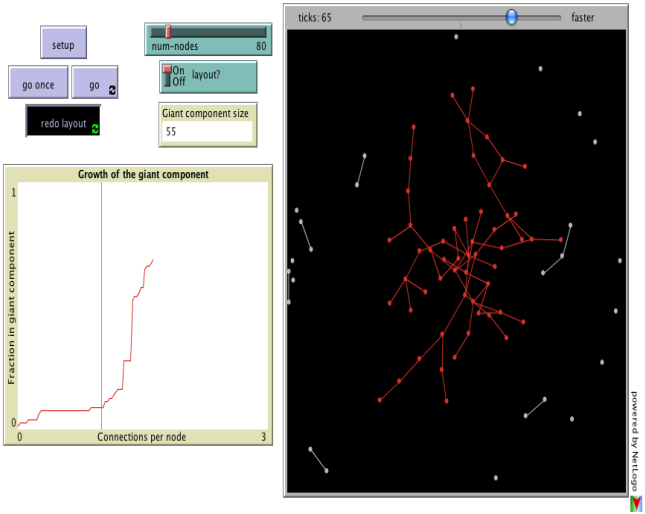
Erdös and Rényi Random Graph

¿Qué pasa con el tamaño del componente gigante (giant component) como la densidad de la red aumenta?

Componente gigante (giant component) :

En una red, un "componente" es un **grupo de nodos (personas) que están todos conectados entre sí, directa o indirectamente.**

Así que si una red tiene un "**componente gigante**", que significa casi **cada nodo es accesible desde casi todos los demás.**
Este modelo muestra lo rápido que surge un componente gigante si usted crece una red aleatoria.



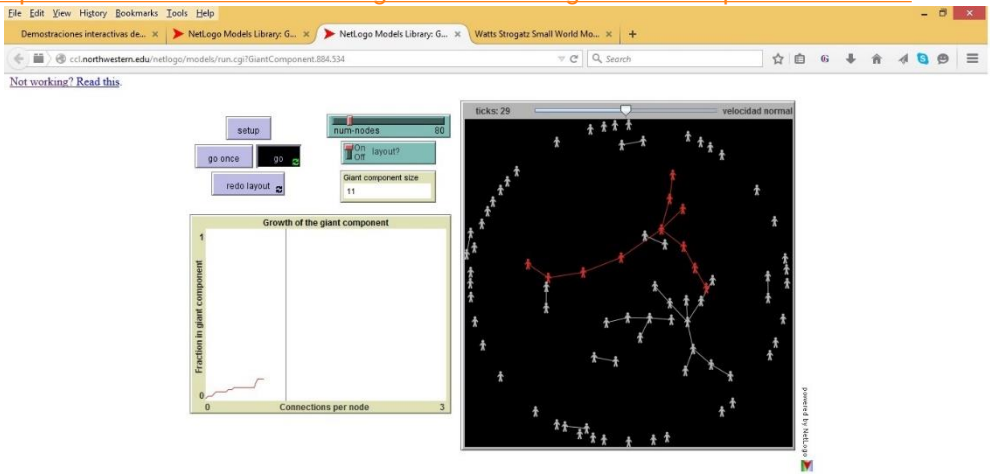
Ver simulacion en internet , esto fue realizado con el software Netlogo:
http://ccl.northwestern.edu/netlogo/models/run.cgi?GiantComponent.884.534

Erdös and Rényi Random Graph

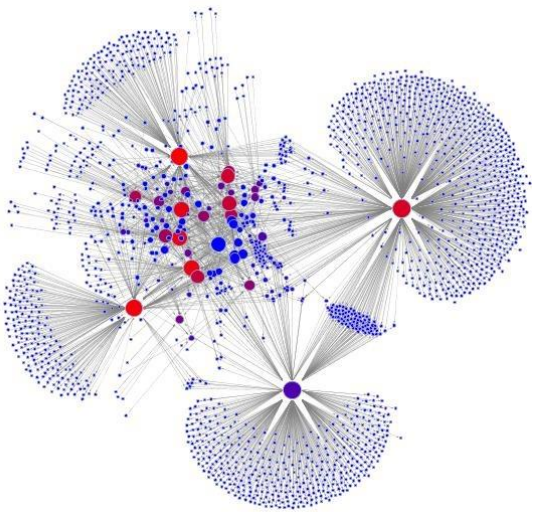
¿Qué pasa con el tamaño del componente gigante (giant component) como la densidad de la red aumenta?

PAGINA WEB:

http://ccl.northwestern.edu/netlogo/models/run.cgi?GiantComponent.884.534

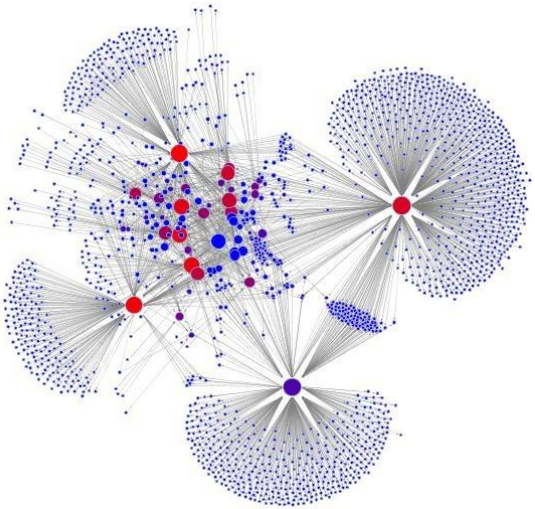


Real Networks (power-law)



- Nodos aparecen con el tiempo (growth model)
- Nodos prefieren unirse a nodos populares (preferential model)

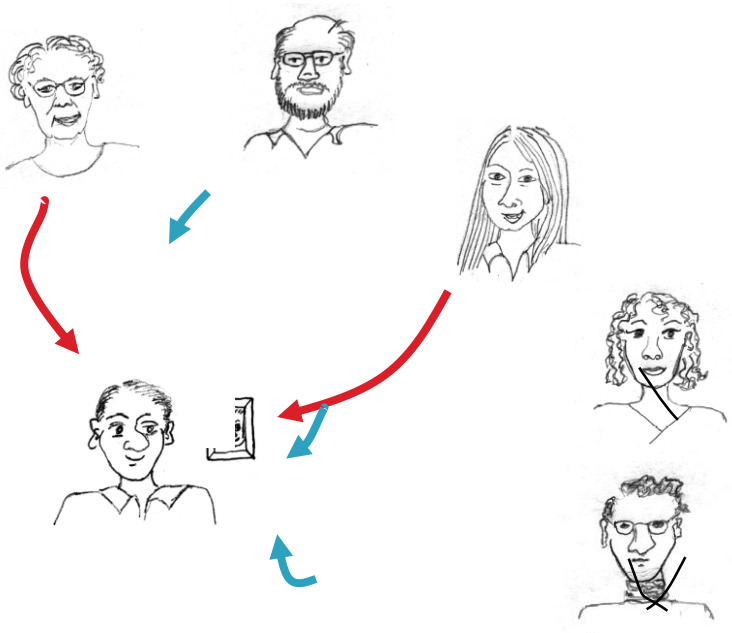
Real Networks (power-law)



- Nodos viejos Mueren
- Algunos nodos son mas sociables
- Las amistades pueden desaparecer

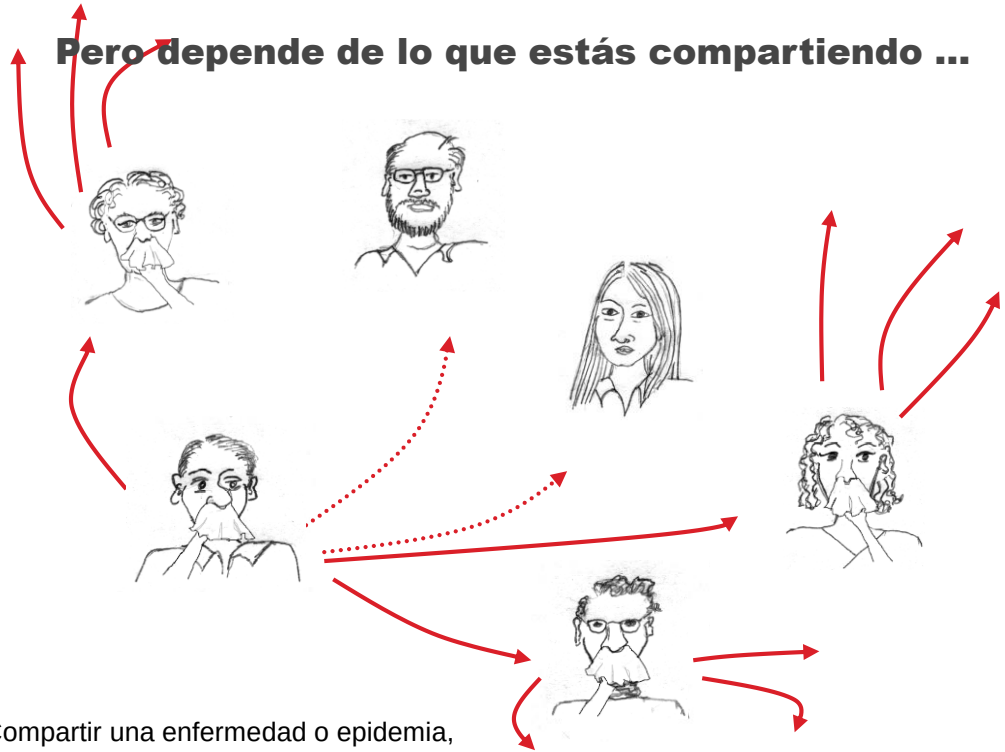
Propagacion de epidemias:

En las redes sociales, es bueno ser un hub



Propagacion de epidemias

Pero depende de lo que estás compartiendo …



Compartir una enfermedad o epidemia,
Seria fatal la facilidad de contagio.

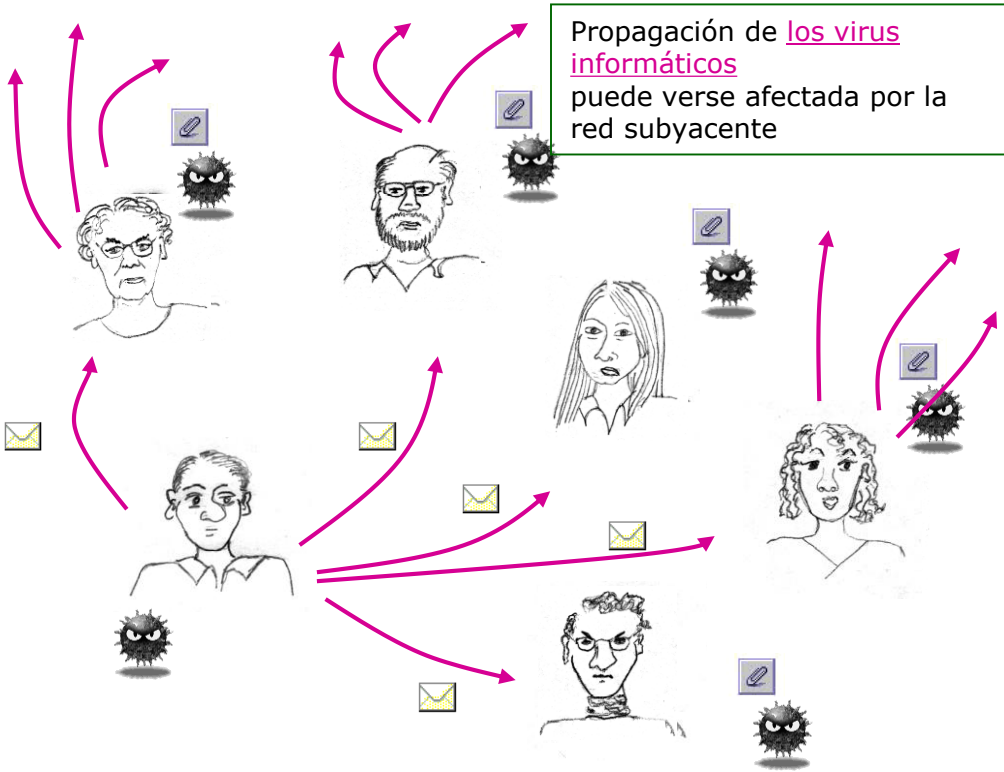
Propagacion de epidemias

El papel de los centros en las epidemias

- En una red de poder de la ley, un virus puede persistir por bajo que su capacidad de infección

- Muchas redes del mundo real hacen exhibir power-leyes:
  - compartir agujas
  - contactos sexuales
  - redes de correo electrónico

Propagacion de epidemias

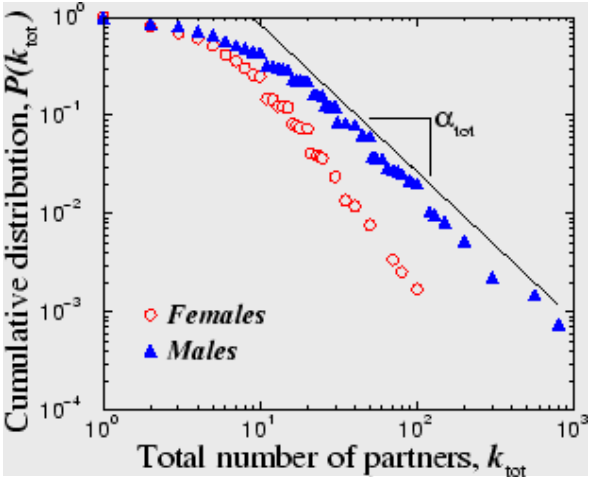


Propagación de los virus informáticos
puede verse afectada por la red subyacente

Propagacion de epidemias:
RED DE CONTACTOS SEXUALES

Nodos: individuos
Links: relaciones sexuales

'La red influye sobre la dinámica'



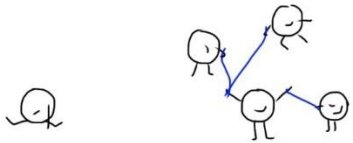
RED LIBRE DE ESCALA

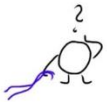
HAY UNOS POCOS NODOS CON MUCHA MAYOR PROBABILIDAD DE CONTAGIAR QUE OTROS (HUBS)

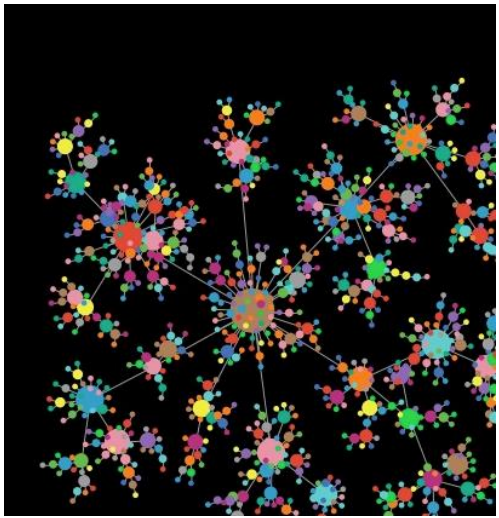
ESTRATEGIAS DE PREVENCION DE EPIDEMIAS

Barabasi-Albert model

- Cada nodo nuevo se conecta a otro con una probabilidad proporcional al grado del nodo
- Los nodos viejos son más ricos







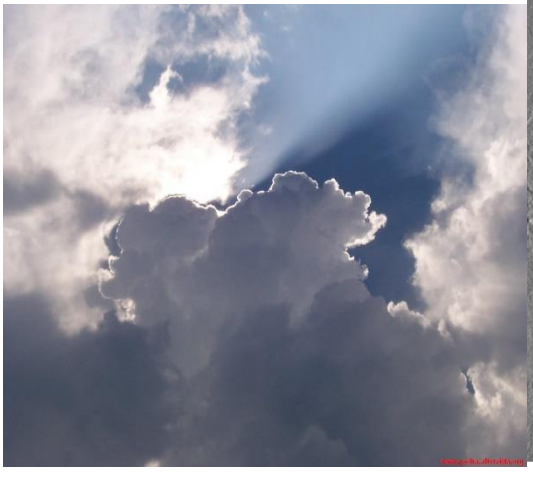
Barabasi-Albert model

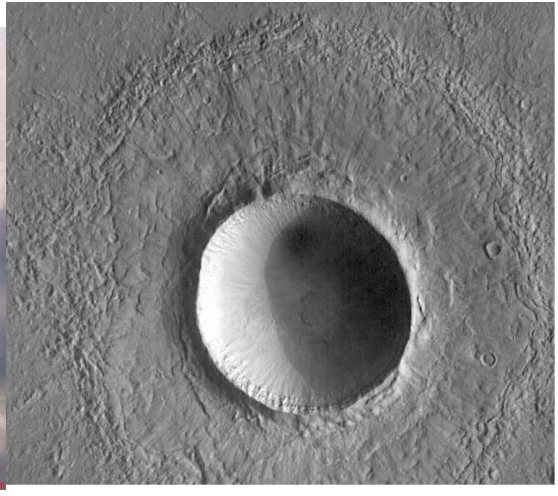
Laszlo Barabasi : Redes libres de escala

¿ QUE TIENEN EN COMÚN ESTAS DOS FOTOGRAFIAS ?

NO PODEMOS ASEGURAR EL TAMAÑO !!!

...EXISTEN MUCHOS CUERPOS QUE NO TIENEN UN TAMAÑO o ESCALA CARACTERÍSTICA...



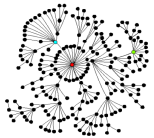


Barabasi-Albert model

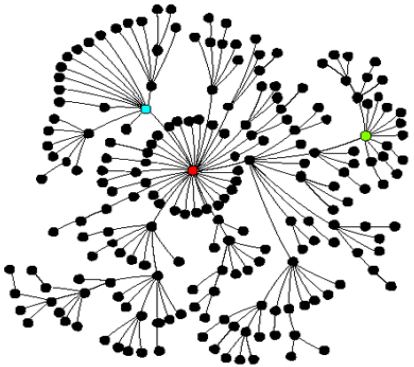
¿QUÉ ES UNA RED LIBRE DE ESCALA?

1. Ausencia de escala: ley de potencias:

$$p(x) = Cx^{-\alpha}$$



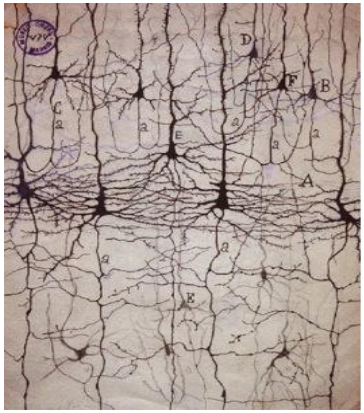
$P(k) \sim k^{-g}$



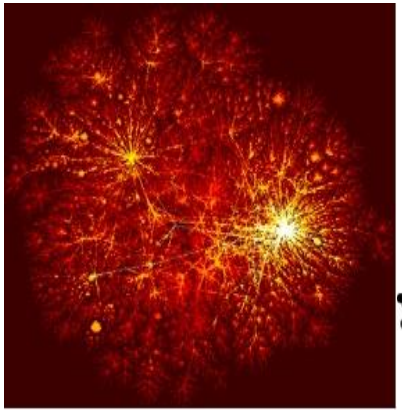
Barabasi-Albert model

¿QUÉ ES UNA RED LIBRE DE ESCALA?

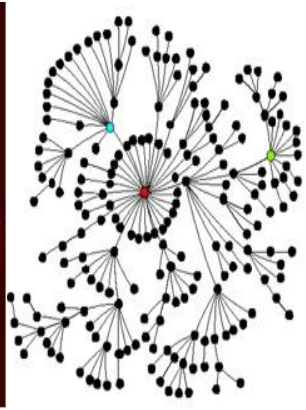
Las red que conforma Internet tienen la misma topología que la red neuronal o la red de mercados, llamadas redes Libres de Escala.



red neuronal : libre de escala



simulación de la red Internet, libre de escala



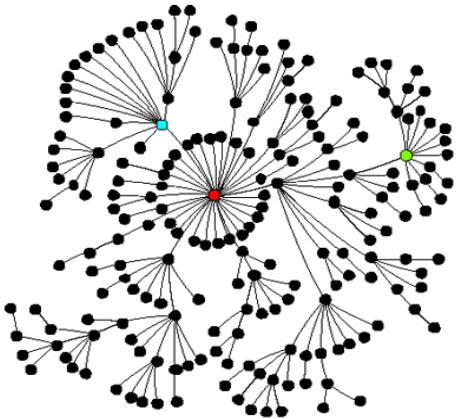
forma típica de la red libre de escala

Barabasi-Albert model

¿QUÉ ES UNA RED LIBRE DE ESCALA?

1) UNA RED ES UN CONJUNTO DE NODOS Y ENLACES.
NODOS -> COMPONENTES DE UN SISTEMA.
ENLACES -> RELACIÓN ENTRE COMPONENTES.

2) UNA RED LIBRE DE ESCALA ES AQUELLA QUE POSEE UNA DISTRIBUCIÓN DE CONECTIVIDAD DE TIPO LEY DE POTENCIAS : $P(k) \sim k^{-g}$



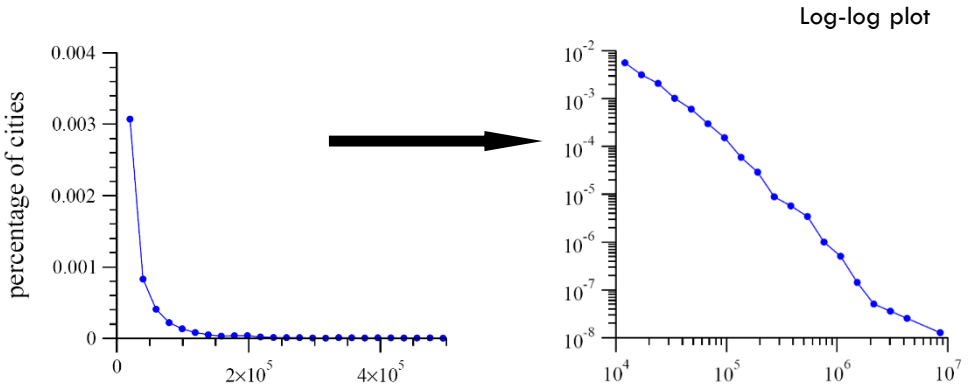
- HAY MUCHOS NODOS CON POCOS ENLACES.

* PERO TAMBIÉN HAY ALGUNOS NODOS CON MUCHOS ENLACES (HUBS).

Barabasi-Albert model

LEY DE POTENCIAS O 'LIBRE DE ESCALA' $p(x) = Cx^{-\alpha}$

Ejemplo: Distribución de tamaños de ciudades



NO EXISTE UN `TAMAÑO MEDIO´ DE CIUDAD BIEN DEFINIDO:
hay muchas ciudades pequeñas pero también hay ciudades muy grandes

Modelo de Wattz-Strogatz

(1998) Transforma un grafo regular en una red aleatoria al recablear enlaces añadiendo o moviendo los ya existentes.

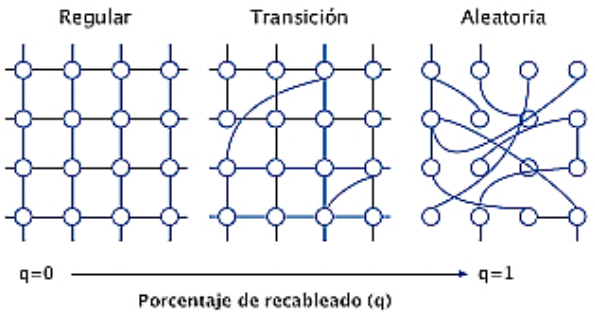
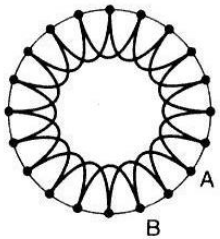


Fig. 13. Recableado de una malla regular hasta llegar a una aleatoria modelo de Watts y Strogatz. En una red regular existe un 0% de enlaces recacableados en forma aleatoria mientras que en una red aleatoria un 100% de los enlaces son recableados aleatoriamente. Entre ambos extremos se ubican una serie de topologías de transición.

Modelo de Wattz-Strogatz

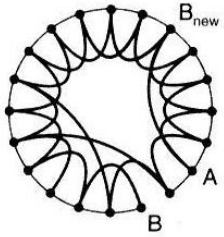
Duncan Watts y Steven Strogatz :

algunos enlaces aleatorios de otra manera en un grafo estructurado hacen la red de un **pequeño mundo**: la ruta más corta media es corta.



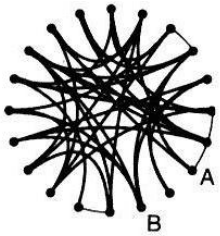
regular lattice o enrejado regular: P=0

el amigo de mi amigo es siempre mi amigo



small world o pequeño mundo: 0<P<<1

principalmente estructurada con unos pocas conexiones aleatoria



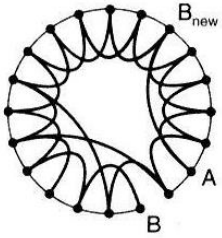
random graph o grafo aleatorio:P=1

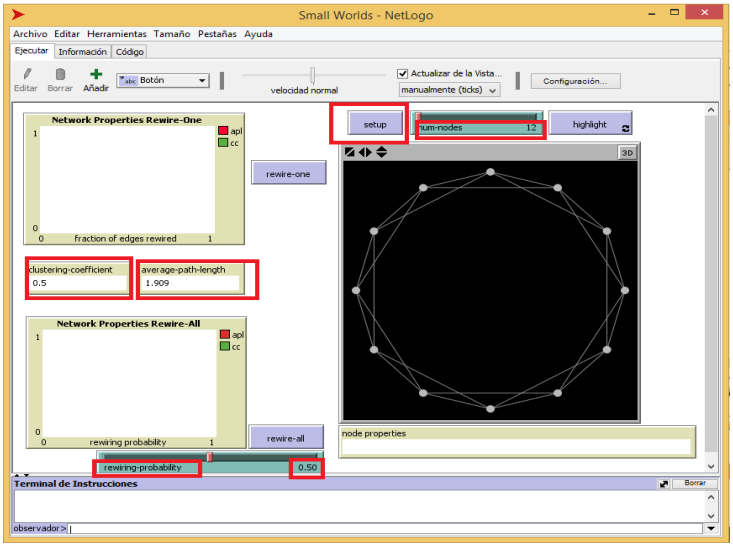
todas las conexiones aleatorio

Fuente: Watts, DJ, Strogatz, SH (1998) Las dinámicas colectivas de las redes 'small-world'. Nature 393: 440-442.

Modelo de Wattz-Strogatz

PAGINA DE DESCARGA SOFTWARE NETLOGO: http://ccl.northwestern.edu/netlogo/





small world o pequeño mundo:

0<P<<1

P=0.5

C=0.5 L=1.909

principalmente estructurada con unos pocas conexiones aleatoria

- **L = longitudes entre dos nodos arbitrarios**
- **C= altos coeficientes de agrupamiento o clustering**
- **P= probabilidad recableada**

Ver simulación en internet , esto fue realizado con el software Netlogo:

http://projects.si.umich.edu/netlearn/NetLogo4/SmallWorldWS.html

Modelos Dinámicos

- Consideran las redes como sistemas con interacciones que varían en el tiempo según determinadas leyes.
- Se llaman también *Modelos de Crecimiento* o de *Evolución* ya que imitan los procesos de crecimiento mediante la adición gradual de nodos o enlaces.
- Logran reproducir la heterogeneidad en el conexionado, la *Distribución de Grados*, el *Coeficiente de Clustering y el efecto Small World* observados en *sistemas complejos*.

Modelo de Enlace Preferencial

Asume que la conexión de los nuevos nodos añadidos al sistema está regulada por la cantidad de conexiones de los ya presentes. Es decir, los nuevos elementos se unirán con mayor probabilidad a los más conectados ya ubicados en la red: “el rico se vuelve más rico”.(Barabási y Albert, 1999)

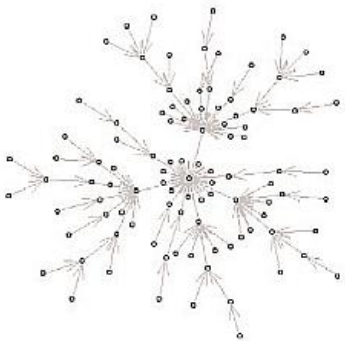


Fig. 14. Modelo de Enlace Preferencial. Los nodos más conectados tiene mayor probabilidad de atrapar a los recién llegados. Las flechas indican estos enlaces.

Modelo de Duplicación

Asume que el origen de los nuevos elementos añadidos a la red es interno. Los nodos se duplican y se unen a los ya existentes según determinada probabilidad (Pastor Santorras et al, 2003).

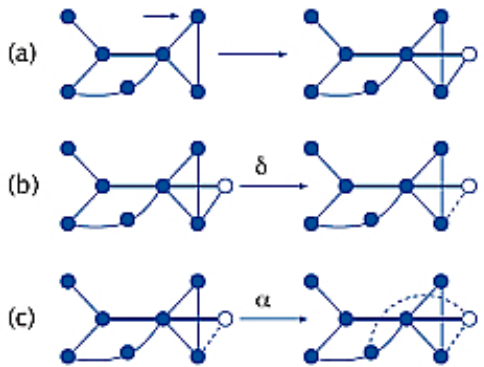


Fig. 15. Modelo de Duplicación en tres etapas: a) copia del nodo señalado con una flecha, b) eliminación de nexión según una probabilidad ‰ y c) generación de nueva conexión según una probabilidad ·

Aplicaciones

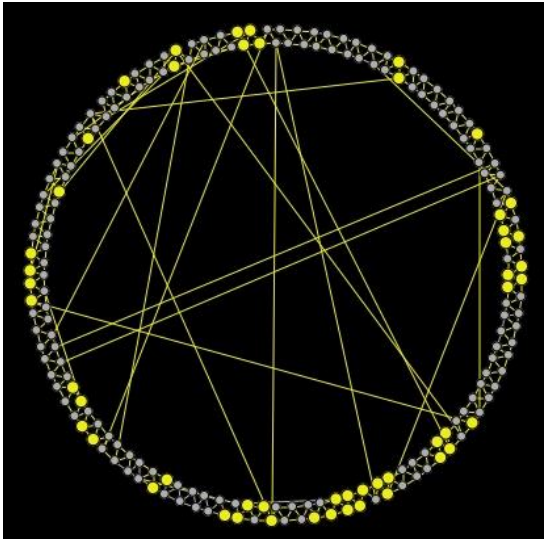
- Epidemiología (propagación de virus).

- Robustez, tolerancia frente a ataques (deliberados).

- Procesos de optimización (publicidad).

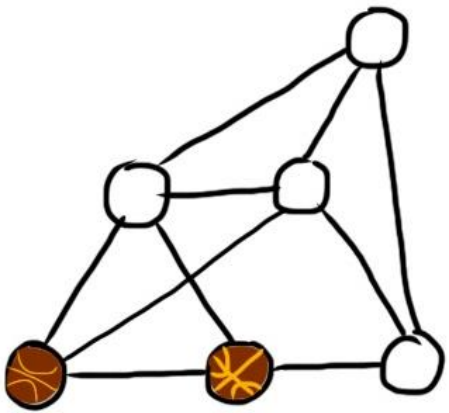
- . . .

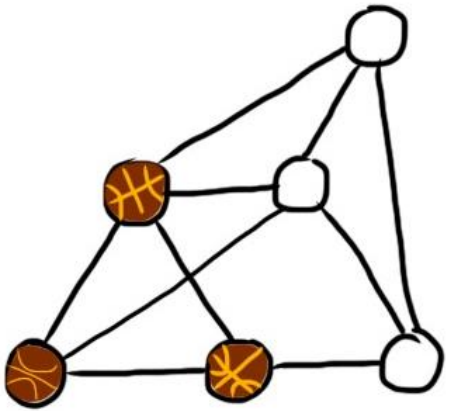
Difusión en Pequeños mundos. Por ejemplo, virus



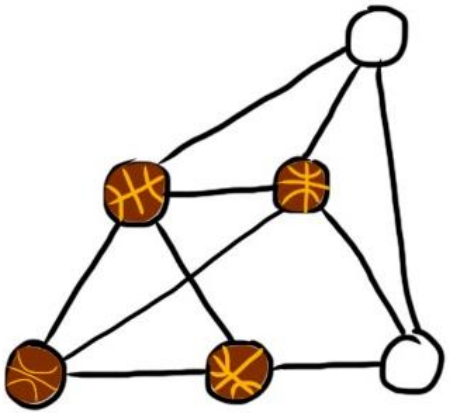
- Se pueden aislar por comunidades
- Si se tiene un recover rate la red tiende a estabilizarse por aislamiento de comunidades y en ciertos casos a recuperarse de la infeccion
- Si existe gran cantidad de hubs el virus se esparse mas rapidamente

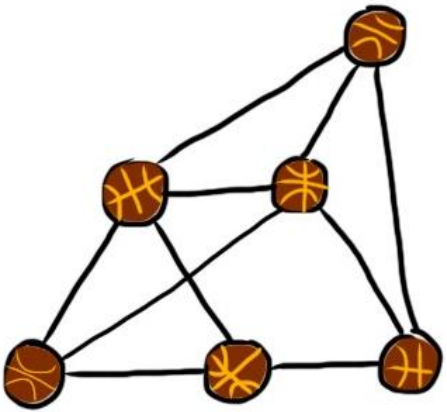
Difusión en Pequeños mundos



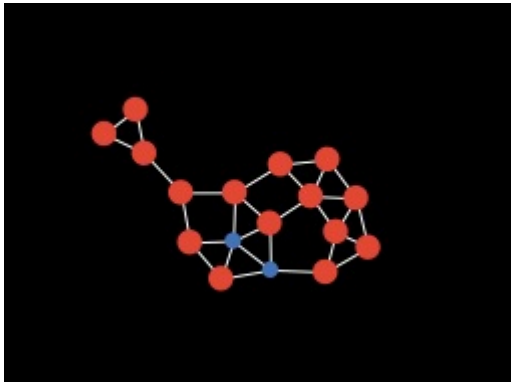


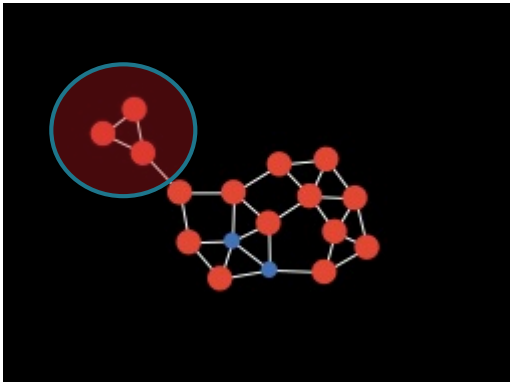
Difusión en Pequeños mundos



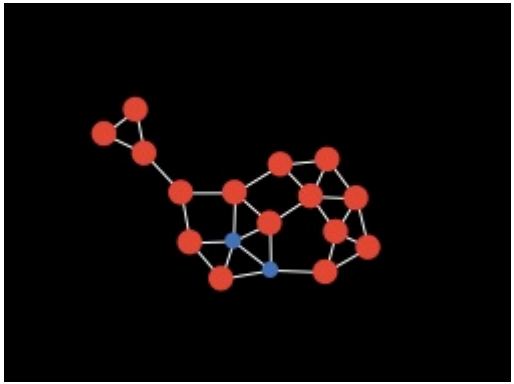


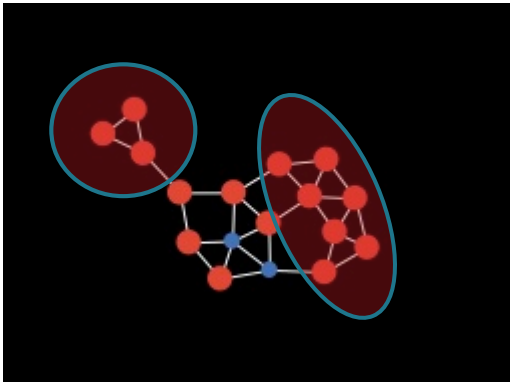
Difusión en Pequeños mundos



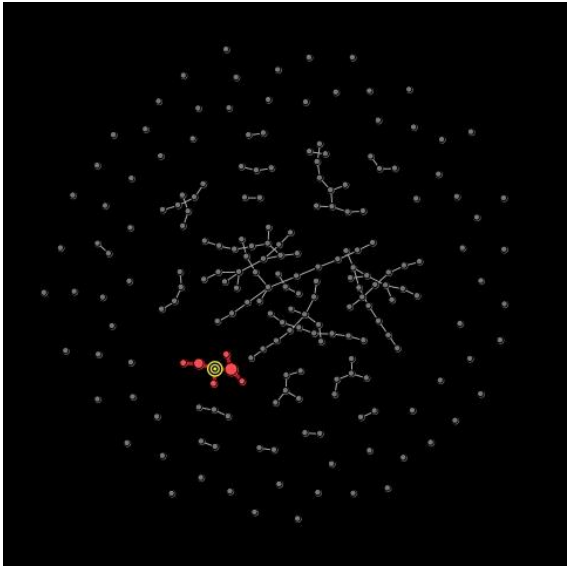


Difusión en Pequeños mundos





Difusión: ER random graph



ROBUSTEZ FRENTE A FALLOS / ATAQUES: RED ELÉCTRICA

	SISTEMA ELÉCTRICO COLOMBIANO	MODELO MATEMÁTICO (RED)
	Generadores	Nodos
	Líneas eléctricas	Enlaces
Fallo aleatorio	Calentamiento de un generador	Desaparición de un nodo
	Mal estado de un cable de tensión	Desaparición de un enlace
Consecuencia	Cortes en cadena Apagón general	Desmembración de la red

☐ PREVENCIÓN: ESTUDIO MATEMÁTICO DE LA ROBUSTEZ DE LA RED

(OTRO EJEMPLO: REDES ECOLÓGICAS y EXTINCIONES)

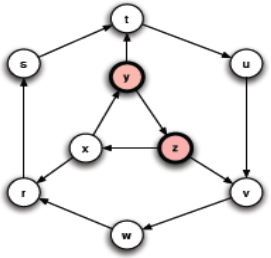
Epidemias a través de redes

- Probabilidad de que dos personas al azar entren en contacto << probabilidad de que una persona entre en contacto con colega o amigo.
- Nodo infectado *v* puede contagiar a un nodo susceptible *w* sólo si existe un lazo que los vincula directamente.
- Red direccional tendría sentido cuando se trata de una relación asimétrica.
- Por ejemplo: virus se transmite de un agente que tiene una mayor capacidad para generar anti-cuerpos a uno que tiene menor capacidad.

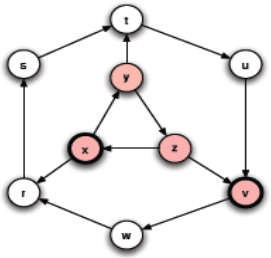
Algoritmo de contagio (Easley y Kleinberg)

- (i) Inicialmente un número determinado de nodos I están infectados y el resto de la población se encuentra en el estado S;
- (ii) Cada nodo *v* que entra en el estado I se mantiene infectado por un número fijo de periodos (t_1);
- (iii) A lo largo de estos t_1 periodos el nodo *v* tiene la probabilidad de pasar la enfermedad a alguno de sus vecinos que sea susceptible;
- (iv) Después de los t_1 periodos el nodo *v* se vuelve resistente y, por ende, no puede contagiar enfermedades ni ser contagiado a partir de ese momento.

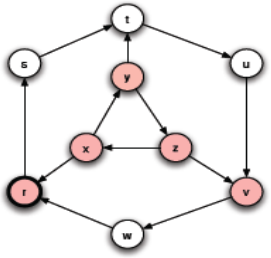
Transmisión de enfermedades a través de una red

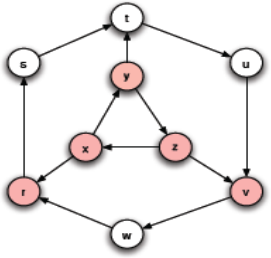


(a)



(b)





Propagación de un virus a través de red

- Tres estados: S, I, R
- Un nodo infectado (rojo) contagia a uno susceptible (verde) con el que está vinculado en función de una probabilidad.
- La infección de un nodo no es detectada inmediatamente
- Un nodo infectado es detectado con una cierta probabilidad
- A partir de otra probabilidad se establece si dicho nodo se vuelve resistente a contagios futuros.
- **Resultado:** infecciones se propagan pero crecimiento de nodos resistentes hace que los nodos infectados desaparezcan

factores que influyen en la difusión

- estructura de la red (no ponderada)
  - densidad
  - grado de distribución
  - clustering
  - componentes conectados
  - estructura de la comunidad
- fuerza de los lazos (ponderado)
  - frecuencia de la comunicación
  - la fuerza de influencia
- agente de extensión
  - atractivo y la especificidad de la información

Difusión de información en las redes

- factores que influyen en la difusión de información
  - estructura de la red: la que se conectan los nodos?
  - fuerza de los lazos: qué tan fuerte son las conexiones?

- estudios en la difusión de la información:
  - Granovetter: la fuerza de los lazos débiles
  - J-P Onnela et al: fuerza de los lazos intermedios
  - Kossinets et al: fuerza de los lazos de backbone
  - Davis: enclavamientos de mesa y adopción de practices

- posición de la red y el acceso a la información
  - Burt: agujeros estructurales y buenas ideas
  - Redes y aprovechar información: Aral y Van Alstyne

- redes y la innovación
  - Lazer y Friedman: la innovación

Difusión de la innovación

Usuarios de la innovación
(Rogers, 1995)

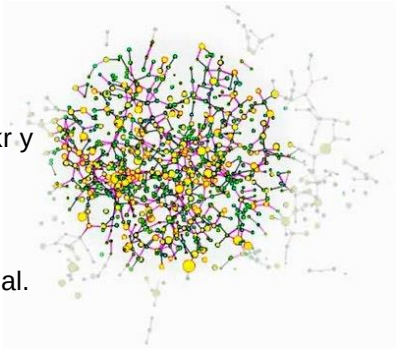
La teoría de **difusión de innovaciones** explora las redes sociales y su rol en la influencia de **la difusión de nuevas ideas y prácticas.**

El cambio en los agentes y en la opinión del líder a menudo tienen un papel más importante en el estímulo a la adopción de innovaciones, a pesar de que también intervienen factores inherentes a las innovaciones.

Difusión de la innovación

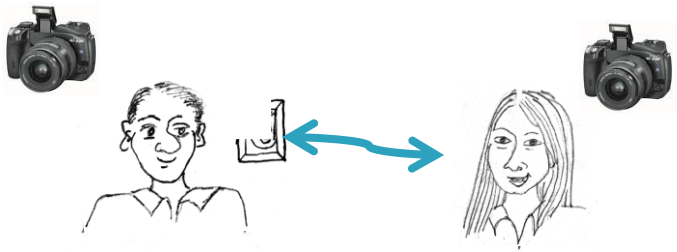
- encuestas:
  - los agricultores que adoptan nuevas variedades de maíz híbrido mediante la observación de lo que sus vecinos estaban plantando (Ryan y Gross, 1943)
  - los médicos que prescriben nuevo medicamento (Coleman et al 1957). (ver laboratorio para jugar con el conjunto de datos)
  - Christakis y Fowler (propagación de la obesidad y la felicidad en las redes sociales) 2008

- datos sobre el comportamiento en línea:
  - Lerman (propagación de las fotos de Flickr y cuentos Digg) 2007
  - Backstrom et al. (Unirse a grupos de LiveJournal y conferencias CS) 2006
  - + otros, por ejemplo, Anagnostopoulos et al. 2008



fuente de la imagen: Christakis y Fowler, 'La propagación de la obesidad en una gran red social de más de 32 años », NEJM 357 (4): 370-379, 2007

Pregunta abierta: ¿cómo le decimos a la influencia de la correlación?



- enfoques:
  - datos resuelto tiempo: si se baraja tiempo adopción, ¿se dan los mismos patrones?
  - si los bordes se dirigen: hace invirtiendo el rendimiento dirección del borde menos poder predictivo?

Redes en el tiempo

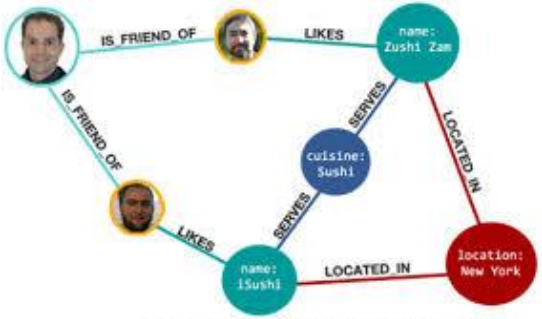
- dinámica aparición / desaparición de nodos y enlaces individuales
  - nuevos enlaces (red de correo electrónico universitario a través del tiempo)
  - ensamblaje equipo (redes coautor y colaborador)
  - evolución del programa de afiliados en relación con la red social (grupos en línea, conferencias CS)
- evolución de las estadísticas totales:
  - diámetros de densificación y contracción (Internet, citación, la autoría, patentes)
  - modelos:
    - estructura de la comunidad
    - modelo de archivo de bosque

Universidad: red de correo electrónico

- propiedades tales como la distribución de grado, camino más corto promedio, y el tamaño del componente gigante tienen variación estacional (vacaciones de verano, inicio de semestre, etc.)
  - ventana de suavizado apropiado (τ) Necesario

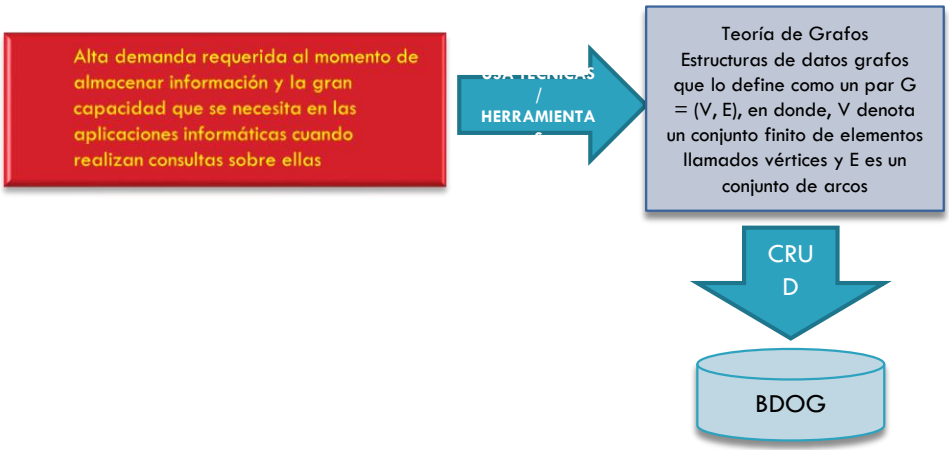
- coeficiente de agrupamiento, la forma de la constante distribución de grado
  - *pero* rango de individuos cambia con el tiempo

BASES DE DATOS ORIENTADAS A GRAFOS (BDOG)



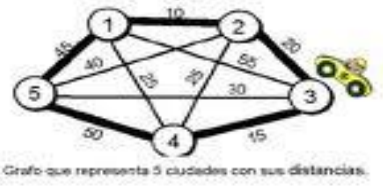
- **BDOG:**
- Representan la información como nodos de un grafo y sus relaciones con las aristas del mismo.
- Una BDOG debe estar absolutamente normalizada.

BASES DE DATOS ORIENTADAS A GRAFOS (BDOG)



VENTAJAS DE UNA BDOG

Grafo que representa las redes de comunicación en España.



Grafo que representa 5 ciudades con sus distancias.

VENTAJAS

- Consultas más amplias y no demarcadas por tablas
- No hay que definir un número determinado de atributos
- Los registros también son de longitud variable
- Se puede recorrer directamente la base de datos de forma jerárquica

MOTORES DE MODELAMIENTO GRAFICO DE UNA BDOG

AllegroGraph - Escalable y de alto rendimiento.
Bigdata - RDF/base de datos orientada a grafo.
CloudGraph - .NET usa tanto los grafos como clave/valor para almacenar los datos.
Cytoscape - Bioinformática
DEX/Sparksee - De alto rendimiento, permite escalar billones de objetos. Comercializada por Sparsity Technologies.
Filament
GraphBase
Graphd, backend de Freebase
Horton
HyperGraphDB - Base de datos opensource basada en la idea de hipergrafo.
InfiniteGraph
InfoGrid - Open Source
Neo4j - Open Source.
OrientDB - Base de datos orientada a grafos y documental.
OQGRAPH
sones GraphDB
VertexDB
Virtuoso Universal Server
R2DF

Comparativa de algunas redes

Red	Tipo	L	•	C
Actores de cine	Social	3.48	2.3	0.2
Mensajes de e-mail	Social	4.95	1.5–2.0	Sin datos
Internet	Tecno-social	3.31	2.5	0.035
Paquetes de Software	Tecnológica	2.42	1.6	0.07
Redes metabólicas	Biológica	2.56	2.2	0.090
Interacción proteica	Biológica	6.8	2.4	0.072

- Estructura Libre de Escala, (e, exponente)
- Longitud Promedio (efecto Small World) entre nodos, L
- Coeficiente de Clustering, C.

Minería de Grafos

- Es el problema de descubrir subgrafos repetitivos (patrones) que se producen en él.

- **Motivación:**
  - Comprimir los datos mediante la abstracción de las instancias de las subestructuras.
  - Identificar patrones conceptualmente interesantes

Minería de Grafos

Objetivo: Desarrollar algoritmos para extraer y analizar grafos.

- Búsqueda de patrones en ellos
- Búsqueda de grupos de grafos similares (clustering)
- Construcción de modelos de predicción para las grafos (clasificación)

- Aplicaciones
  - descubrimiento motivo estructural
  - reconocimiento de proteínas
  - ingeniería inversa en VLSI
  - Mucho más …

Por qué Minería de Grafos?

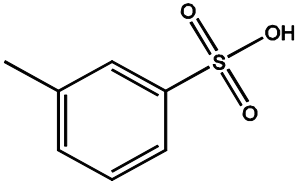
- **Los grafos son ubicuos**
  - Compuestos químicos (quimio-informática)
  - Estructuras de las proteínas, las vías/redes biológicas (Bioinformactica)
  - Flujo de programas, flujo de tráfico, flujo de trabajo
  - bases de datos XML, Web, de redes sociales
- **Grafos es un modelo general**
  - Árboles, secuencias, lazos, etc.
- **Diversidad de grafos**
  - Dirigidos vs. no dirigidos, etiquetados vs. no etiquetados (arcos y vértices), ponderados, con ángulos y geometrías (topológicos en 2-D/3-D)
- **La complejidad de los algoritmos: muchos problemas son de alta complejidad**

Minería de Patrones de Grafos

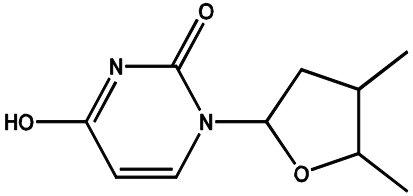
Minería subgrafo frecuentes

- **Encontrar subgrafos frecuentes dentro de un grafo**
  - SUBDUE (DOMINAR)
- **Encontrar (sub)grafos frecuentes en un conjunto de grafos**
  - *Support* (frecuencia de ocurrencia) no inferior a un umbral mínimo
  - Enfoque basado en A priori
  - Enfoque de crecimiento del patrón (Pattern-growt)
- **Aplicaciones de la minería de patrones de grafos**
  - Minería de estructuras bioquímicas, de flujos de programas, de estructuras XML y comunidades de la Web
  - Construcción de sistemas de clasificación, agrupación, compresión, comparación y análisis de correlación para grafos

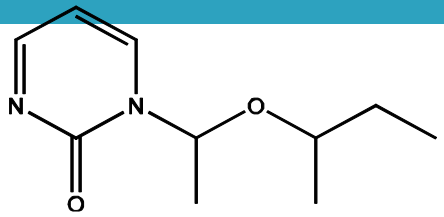
Descubrir subgrafos repetitivos (patrones) en Compuestos químicos



(A)



(B)



(C)

El soporte de un patrón dado, P, es : $$sup_P = \sum_{G' \in D_P} S(P, G')/|D|$$

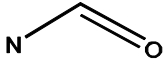
D colección de grafos

D_P representa el subconjunto de grafos en D que contienen un subgrafo aproximadamente isomorfo a P, dado el umbral de similitud τ.

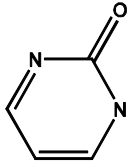
Un patrón P es frecuente si su soporte es mayor o igual que un umbral de frecuencia σ dado por el usuario.

FRECUENTES PATRONES (MIN SOPORTE ES 2)

(1)

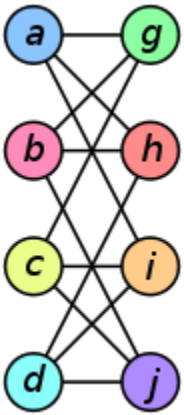


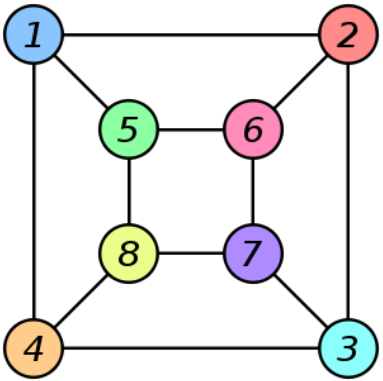
(2)



Descubrir subgrafos repetitivos (patrones)

Isomorfismo entre dos grafos *G* y *H* es una biyección *F* entre los conjuntos de sus vértices F: V (G) → V (H) que preserva la relación de adyacencia.





F(a)=1
F(b)=6
F(c)=8
F(d)=3
F(g)=5
F(h)=2
F(i)=4
F(j)=7

Enfoques de Minería subgrafo frecuentes

- **Enfoques basados en Apriori**
  - AGM: Inokuchi, et al. (PKDD'00)
  - FSG: Kuramochi and Karypis (ICDM'01)
  - PATH: Vanetik and Gudes (ICDM'02, ICDM'04)
  - FFSM: Huan, et al. (ICDM'03)
- **Enfoques de crecimiento del patrón (Pattern-growt)**
  - MoFa, Borgelt and Berthold (ICDM'02)
  - gSpan: Yan and Han (ICDM'02)
  - Gaston: Nijssen and Kok (KDD'04)
- **Minería de patrones cercanos**
  - CLOSEGRAPH: Yan & Han (KDD'03)

Medidas de similitud basadas en los patrones de grafos

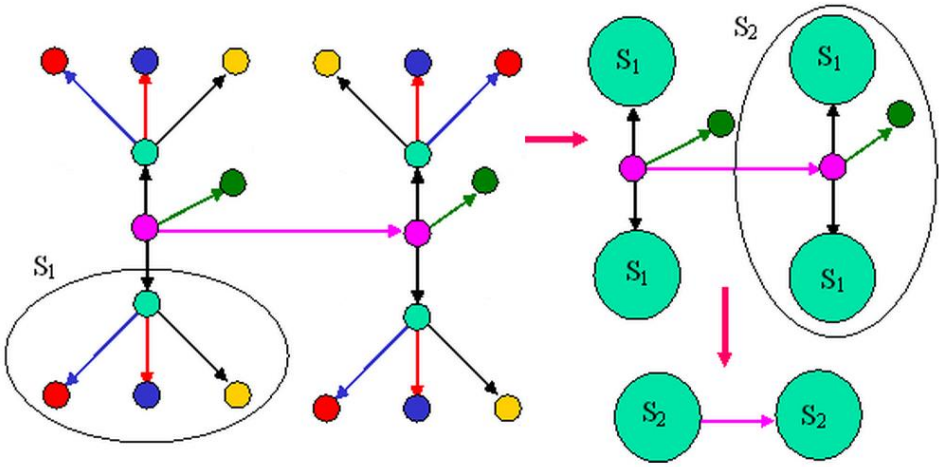
- Medidas de similitud basado en características
  - Cada grafo se representa como un vector de características
  - Subgrafos frecuentes se pueden utilizar como características
  - Vector de distancia
- Medida de similitud basada en la Estructura
  - Subgrafo común máximo
  - Grafo edita distancia: inserción, supresión, y re-etiquetado
- Subgrafos frecuentes y discriminativos son características de indexación de alta calidad

Buscando patrones frecuentes en Grafos

- Un patrón es una relación entre elementos del objeto que es **recurrente** una y otra vez.
  - Estructuras comunes en una familia de compuestos químicos o proteínas.
  - ...
- Hay dos formas generales para definir formalmente un patrón en el contexto de gráficos
  - Subgrafos arbitrarios (conectados o no)
  - Subgrafos inducidos (conectado o no)
- Descubrimiento de patrones frecuentes se traduce en el descubrimiento de subgrafos frecuentes ...

Algoritmo Descubrimiento Frecuentes subgrafos SUBDUE

- Se trata de identificar la subestructura que mejor comprime al grafo original,
- Se reemplaza dicha estructura por un solo vértice y
- Se aplica recursivamente el procedimiento al grafo resultante.



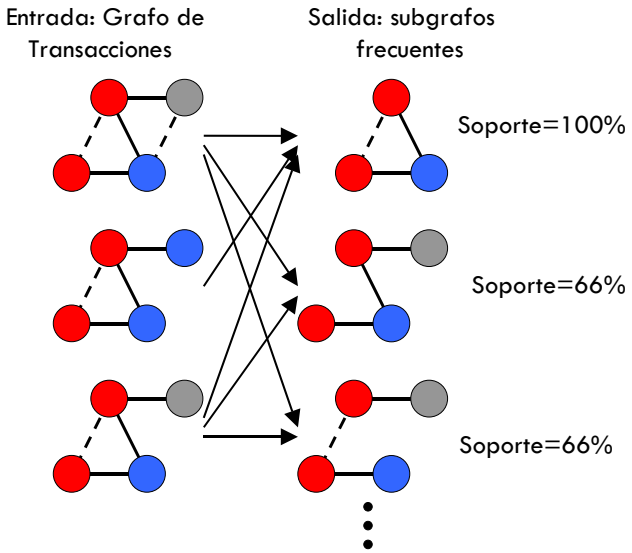
Buscando frecuentes subgrafos

Entrada

- Base de datos de transacciones de grafos.
- Grafo no dirigido simple (no hay bucles, no hay múltiples arcos).
- Cada grafo de transacción tiene etiquetas asociadas con sus vértices y aristas.
- Las transacciones pueden no estar conectadas.
- σ: umbral mínimo de soporte.

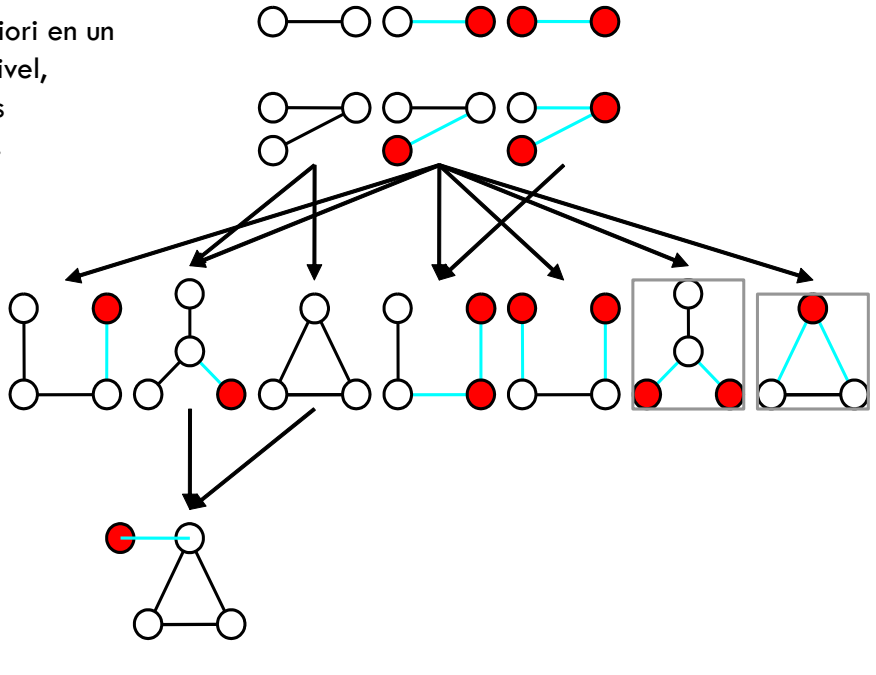
Salida

- subgrafos frecuentes que satisfacen la restricción mínima soporte.
- Cada subgrafo frecuente está conectado.



Algoritmo Descubrimiento Frecuentes subgrafos FSG

Sigue un estilo A-Priori en un enfoque nivel por nivel, y crece los patrones un arco-por-tiempo.



Algoritmo Descubrimiento Frecuentes subgrafos FSG

Notación: k-subgrafo es un subgrafo con k aristas.
Tarea: Analiza las operaciones para encontrar $\mathcal{F}_1$, el conjunto de todos los frecuentes 1-subgrafos y 2-subgrafos, junto con sus frecuencias;

Para (k = 3; $\mathcal{F}_{k-1} \neq \emptyset$; k ++)
Generación Candidato: C_k, el conjunto de candidatos k-subgrafos, desde $\mathcal{F}_{k-1}$, el conjunto de frecuentes (k-1)-subgrafos;
Poda de candidatos: una condición necesaria del candidato a ser frecuente es que cada uno de sus (k-1)-subgrafos es frecuente.
Medición de frecuencia: Analiza las operaciones para contar las ocurrencias de los subgrafos en C_k;
$\mathcal{F}_k = \{ c \in C_K \mid c$ tiene las cuentas no menores de #minSup$\}$
volver $\mathcal{F}_1 \cup \mathcal{F}_2 \cup \ldots\ldots \cup \mathcal{F}_k \quad (= \mathcal{F})$

Isoformismo