Recuperació de la Informació (REIN)

Grau en Enginyeria Informàtica

Departament de Ciències de la Computació (CS)

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

Escola Politècnica Superior d'Enginyeria
de Vilanova i la Geltrú

Índex

Anàlisi de xarxes, Part I

1. Exemples de xarxes reals.

Anàlisi de xarxes, Part I

1. Exemples de xarxes reals.
2. Com són les xarxes reals?

Anàlisi de xarxes, Part I

1. Exemples de xarxes reals.
2. Com són les xarxes reals?
   - les xarxes reals tenen un diàmetre petit (model Erdös-Rényi o aleatori)
   - les xarxes reals tenen un alt coeficient d'agrupació o de *clustering* (model Watts-Strogatz)
   - la distribució de grau de les xarxes reals segueixen una Llei de potència o *power law* (model Barabási-Albert o d'adjunció preferent)

Índex

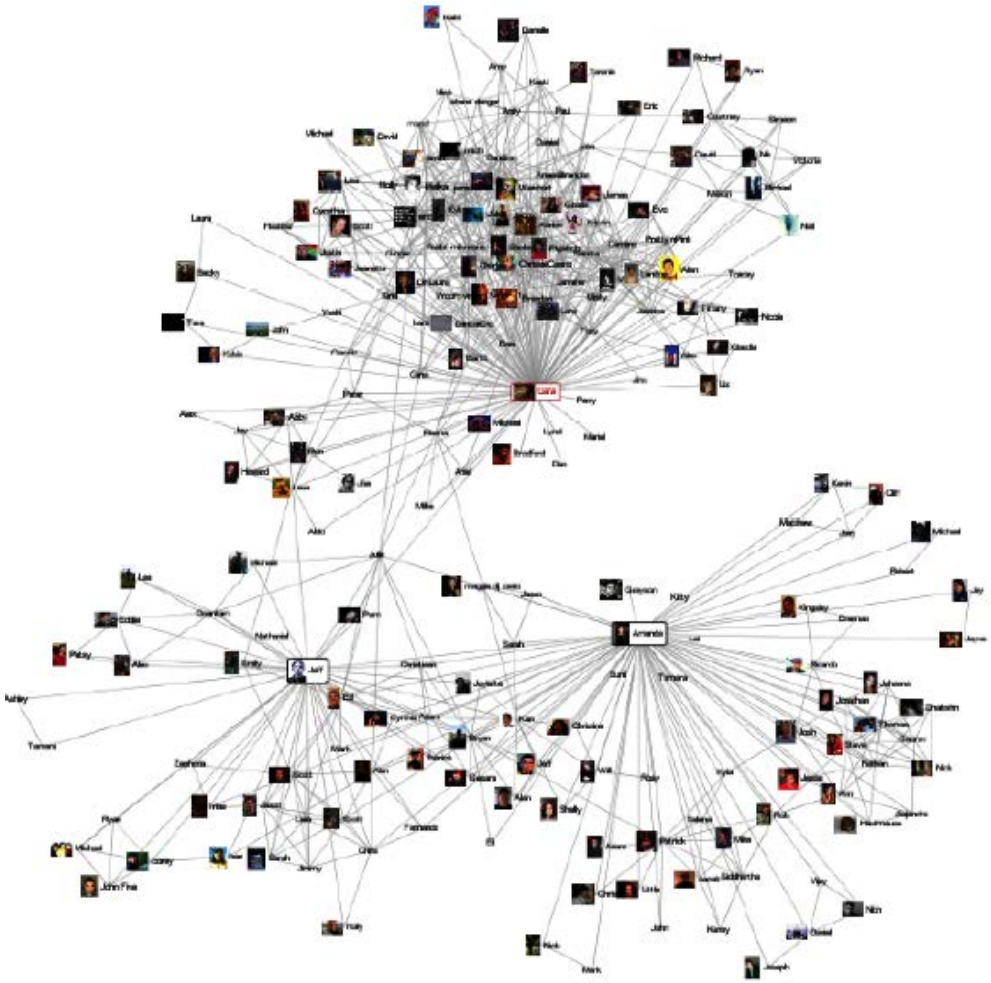
Anàlisi de xarxes, I

Exemples de xarxes reals

- Xarxes socials
  - Grup de persones(agents) amb algun patró d'interrelació entre elles (amistat, col·laboracions, etc).
- Xarxes d'informació
  - Informació amb referències creuades (*links*) per permetre una cerca eficient (www, bases de dades de citacions, etc).
- Xarxes tecnològiques
  - Dissenyades per distribuir ("transportar") béns, persones o recursos (xarxes elèctriques, d'autopistes, d'aeroports, internet, etc).
- Xarxes biològiques
  - Corresponen a sistemes biològics (xarxes genètiques, metabòliques, alimentàries, etc).

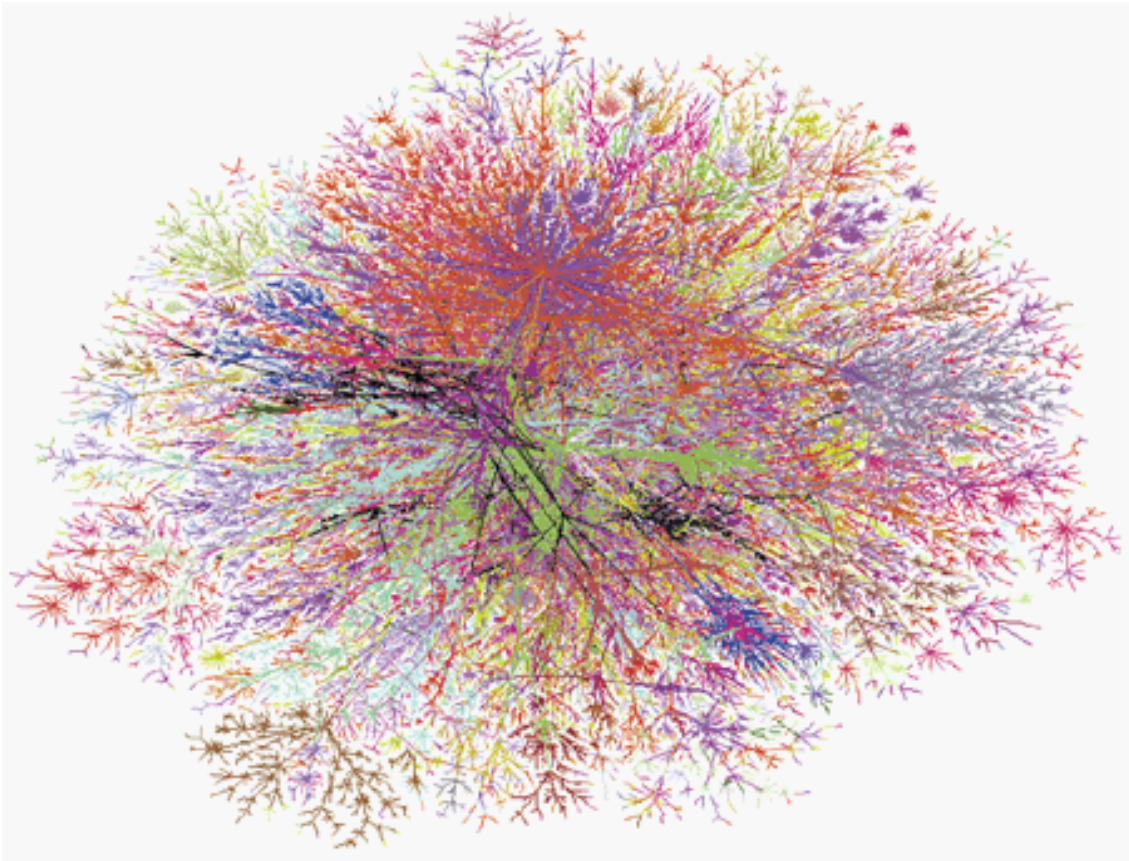
Xarxes socials

Els enllaços representen “interaccions” socials: amistat, col·laboració, correu-e, etc.



Xarxes d'informació

Els nodes emmagatzemen informació i els enllaços relacionen la informació: internet, xarxes p2p, xarxes de citacions, etc.

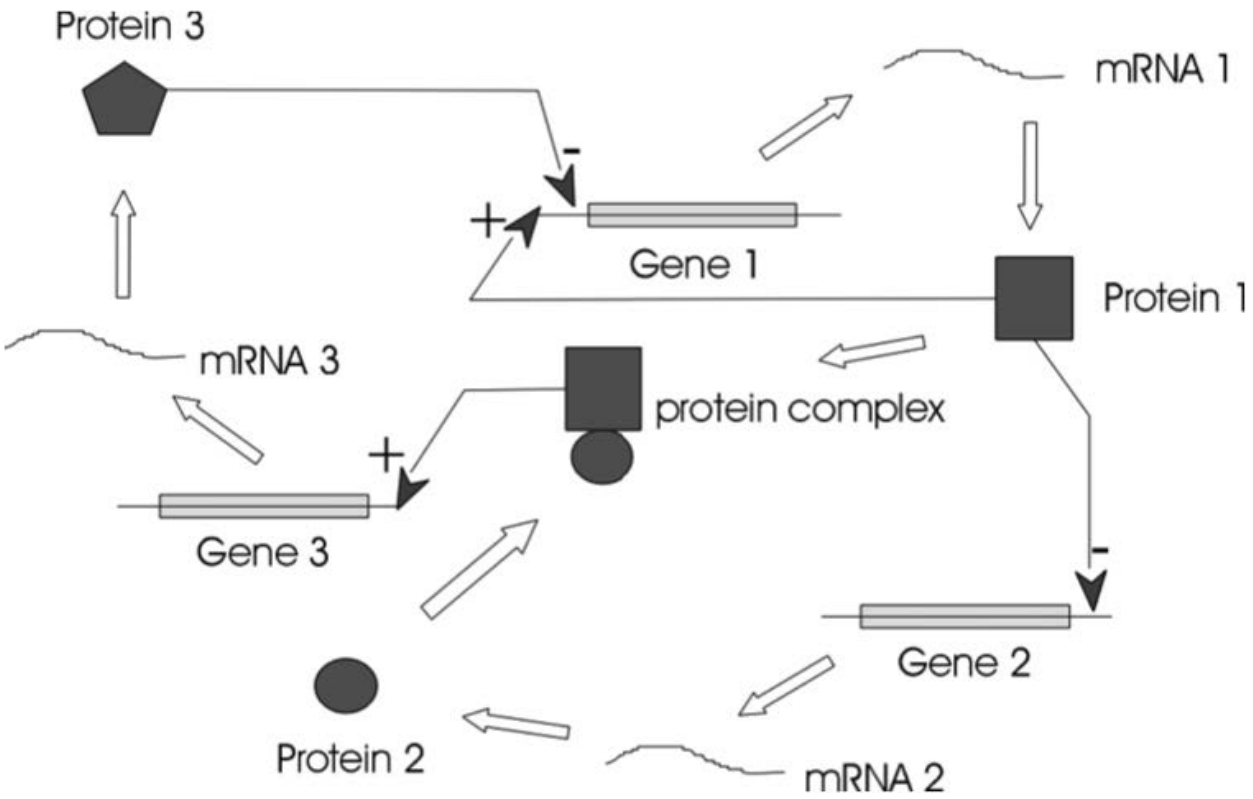


Xarxes tecnològiques

Construïdes pels humans per a la distribució d'un producte bàsic: telefonia, transport, electricitat, etc.

Xarxes biològiques

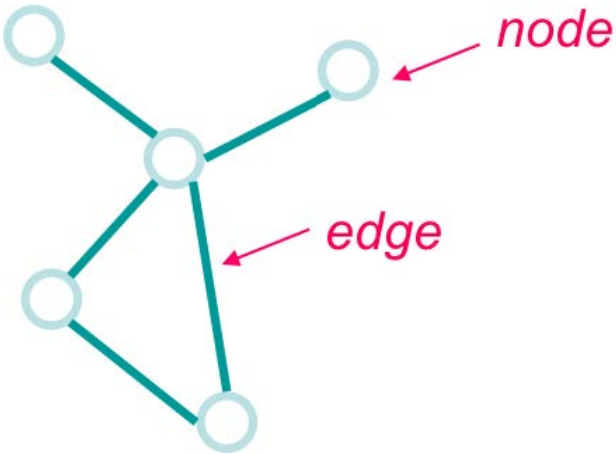
Representen sistemes biològics: interaccions entre proteïnes, regulació genòmica, vies metabòliques, etc.



Índex

Representació de les xarxes

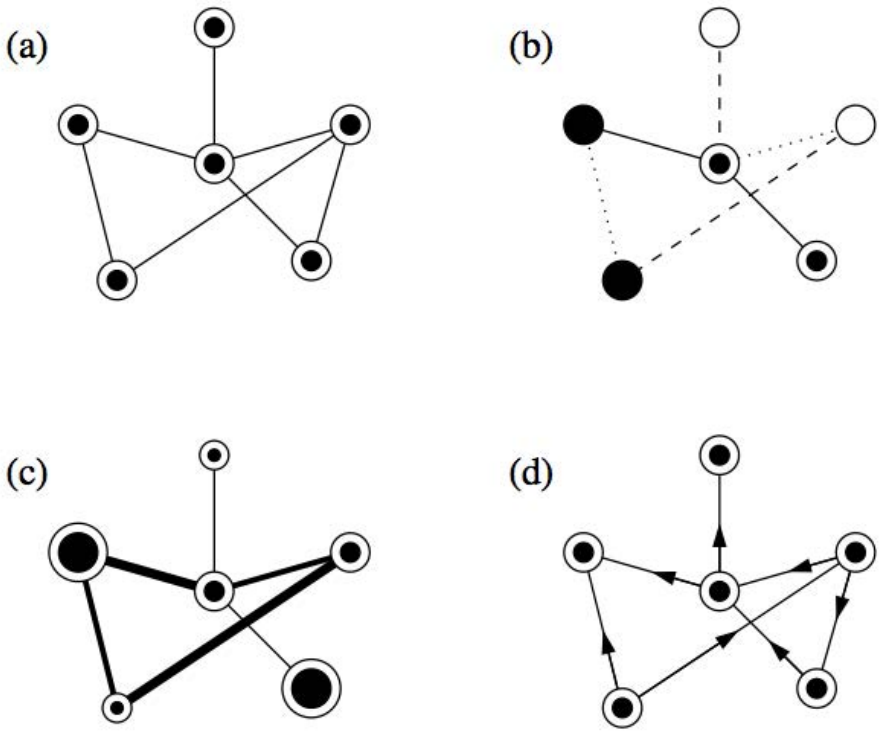
- Xarxa $\equiv$ graf
- Les xarxes són col·leccions de “punts” units per “línies”



punts	línies	
vèrtexs	arestes, arcs	mates
nodes	enllaços	informàtica
actors	llaços, relacions	sociologia

Tipus de xarxes

[Newman, 2003]



(a) no dirigida, sense pesos, només un tipus de vèrtex

(b) no dirigida, tipus de vèrtexs i d'arestes poc variat

(c) no dirigida, vèrtexs i arestes amb pesos força variats

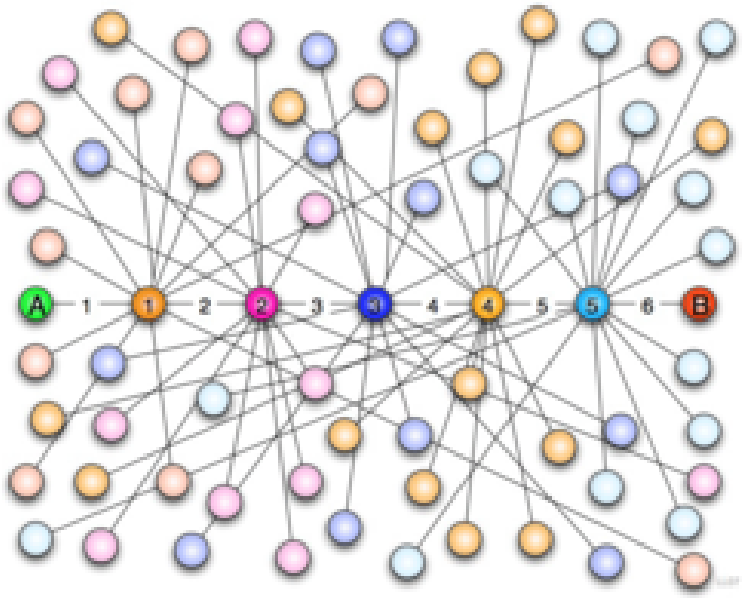
(d) dirigida

Xarxes de món petit (*small-world*)

- Un amic d'un amic és sovint també un amic.

Xarxes de món petit (*small-world*)

- Un amic d'un amic és sovint també un amic.
- En el món, dues persones estan separades per només 6 "salts".



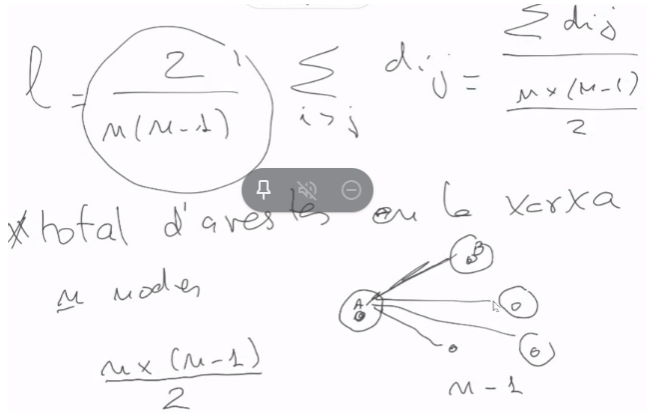
Mesurant el fenomen de món petit, I

- Sigui d_{ij} la longitud del camí més curt entre els nodes i i j.
- Per comprovar si "qualsevol parell de nodes estan a 6 salts" fem servir les definicions següents:
  - El diàmetre (longitud màxima dels camins més curts) és:

$$d = \max_{i,j} d_{ij}$$

- La mitjana de la longitud dels camins més curts és:

$$l = \frac{2}{n(n-1)} \sum_{i>j} d_{ij}$$



  - La mitjana harmònica[1] de la longitud dels camins més curts és:

$$l^{-1} = \frac{2}{n(n-1)} \sum_{i>j} d_{ij}^{-1}$$

[1] Usada en xarxes amb més d'un component.

[Newman, 2003]

	network	type	n	m	z	ℓ	α	$C^{(1)}$	$C^{(2)}$	r	Ref(s).
social	film actors	undirected	449 913	25 516 482	113.43	3.48	2.3	0.20	0.78	0.208	20, 416
	company directors	undirected	7 673	55 392	14.44	4.60	–	0.59	0.88	0.276	105, 323
	math coauthorship	undirected	253 339	496 489	3.92	7.57	–	0.15	0.34	0.120	107, 182
	physics coauthorship	undirected	52 909	245 300	9.27	6.19	–	0.45	0.56	0.363	311, 313
	biology coauthorship	undirected	1 520 251	11 803 064	15.53	4.92	–	0.088	0.60	0.127	311, 313
	telephone call graph	undirected	47 000 000	80 000 000	3.16		2.1				8, 9
	email messages	directed	59 912	86 300	1.44	4.95	1.5/2.0		0.16		136
	email address books	directed	16 881	57 029	3.38	5.22	–	0.17	0.13	0.092	321
	student relationships	undirected	573	477	1.66	16.01	–	0.005	0.001	−0.029	45
	sexual contacts	undirected	2 810				3.2				265, 266
information	WWW `nd.edu`	directed	269 504	1 497 135	5.55	11.27	2.1/2.4	0.11	0.29	−0.067	14, 34
	WWW Altavista	directed	203 549 046	2 130 000 000	10.46	16.18	2.1/2.7				74
	citation network	directed	783 339	6 716 198	8.57		3.0/–				351
	Roget's Thesaurus	directed	1 022	5 103	4.99	4.87	–	0.13	0.15	0.157	244
	word co-occurrence	undirected	460 902	17 000 000	70.13		2.7		0.44		119, 157
technological	Internet	undirected	10 697	31 992	5.98	3.31	2.5	0.035	0.39	−0.189	86, 148
	power grid	undirected	4 941	6 594	2.67	18.99	–	0.10	0.080	−0.003	416
	train routes	undirected	587	19 603	66.79	2.16	–		0.69	−0.033	366
	software packages	directed	1 439	1 723	1.20	2.42	1.6/1.4	0.070	0.082	−0.016	318
	software classes	directed	1 377	2 213	1.61	1.51	–	0.033	0.012	−0.119	395
	electronic circuits	undirected	24 097	53 248	4.34	11.05	3.0	0.010	0.030	−0.154	155
	peer-to-peer network	undirected	880	1 296	1.47	4.28	2.1	0.012	0.011	−0.366	6, 354
biological	metabolic network	undirected	765	3 686	9.64	2.56	2.2	0.090	0.67	−0.240	214
	protein interactions	undirected	2 115	2 240	2.12	6.80	2.4	0.072	0.071	−0.156	212
	marine food web	directed	135	598	4.43	2.05	–	0.16	0.23	−0.263	204
	freshwater food web	directed	92	997	10.84	1.90	–	0.20	0.087	−0.326	272
	neural network	directed	307	2 359	7.68	3.97	–	0.18	0.28	−0.226	416, 421

Però...

- Podem imitar aquest fenomen en xarxes simulades ("models")?
- La resposta és SÍ.

Xarxes aleatòries (model bàsic)

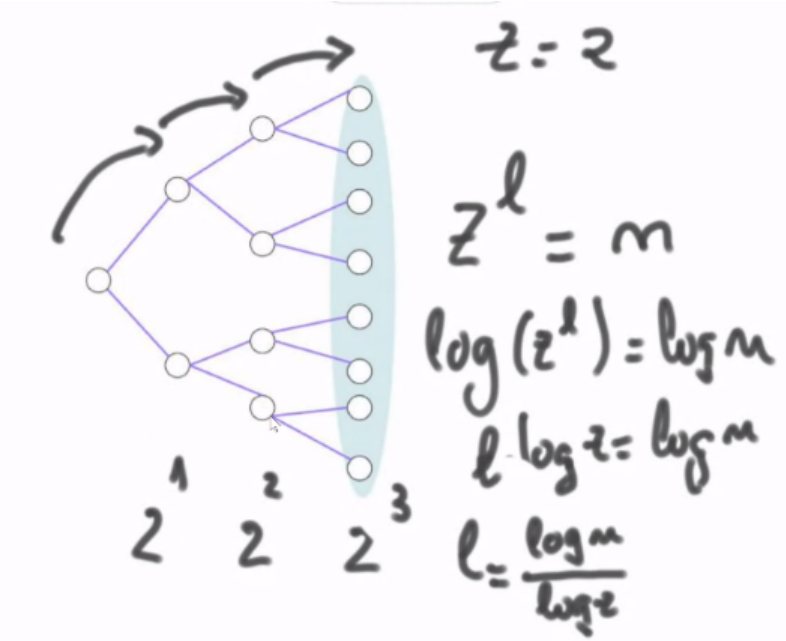
Erdös-Rényi (ER)

El model aleatori bàsic d'Erdös-Rényi és un graf $G_{n,p}$ on:

- el paràmetre n és el nombre de vèrtexs,
- el paràmetre p és tal que $0 \leq p \leq 1$,
- genera una aresta (i, j) de forma aleatòria amb probabilitat p independentment de les arestes ja existents.

Mesura del diàmetre de les xarxes ER

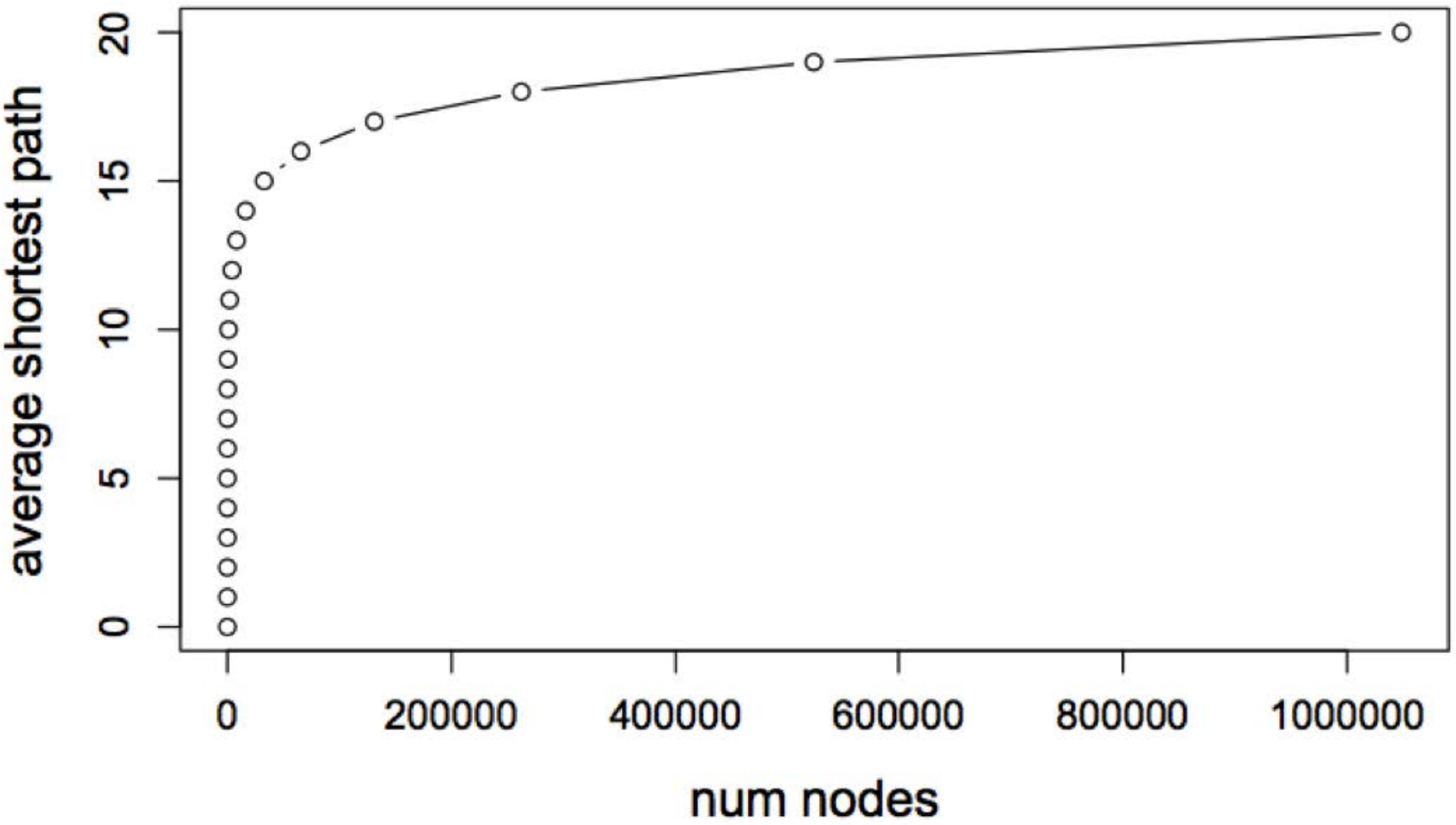
Volem demostrar que el diàmetre de les xarxes ER és petit.



- Anomenem z al grau mitjà.
- A distància l, es pot arribar a z^l nodes.
- A distància $\frac{\log(n)}{\log(z)}$, es pot arribar a n nodes.
- Per tant, el diàmetre és (aproximadament) $O(\log n)$.

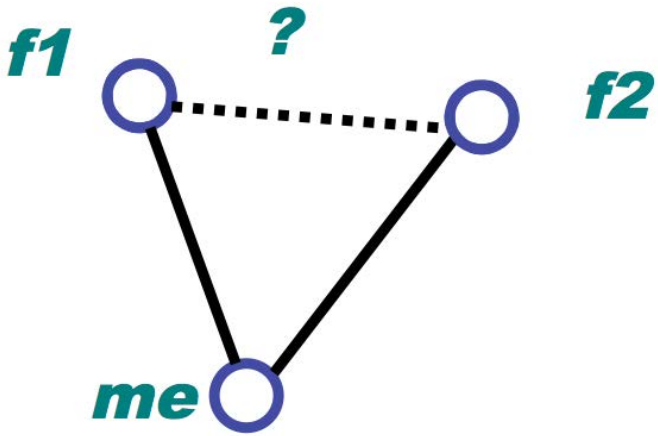
Les xarxes ER tenen un diàmetre petit

Simulació



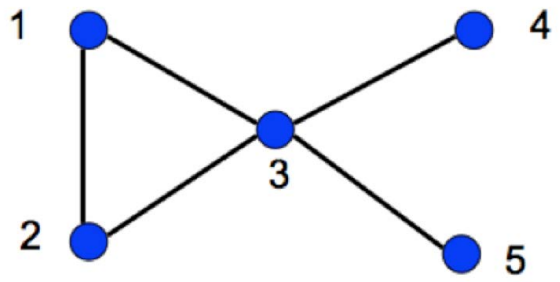
Mesurant el fenomen de món petit, II

- Per comprovar si “l’amic d’un amic és sovint també un amic”, fem servir:
  - La transitivitat o coeficient de *clustering*, que bàsicament mesura la probabilitat de que dos dels meus amics també siguin amics entre ells.

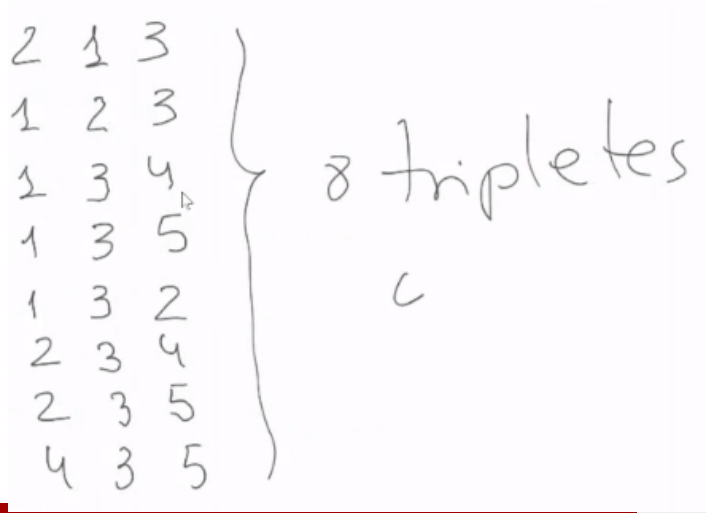


Coeficient de *clustering* global d'una xarxa

$$C = \frac{\text{3 x nombre triangles}}{\text{nombre de tripletes connectades}}$$



$$C = \frac{3 \text{ x } 1}{8} = 0,375$$



Coeficient de *clustering* local, I

- Per cada vèrtex i, anomenem n_i al nombre de veïns de i.
- Sigui C_i la fracció de parells de veïns que estan connectats entre si:

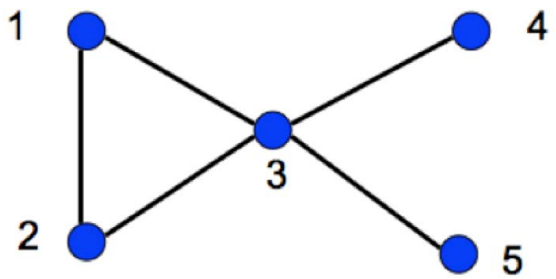
$$C_i = \frac{\text{\# connexions entre els veïns de } i}{\frac{1}{2} n_i (n_i - 1)}$$

- Finalment, calculem la mitjana dels C_i per tots els nodes i de la xarxa:

$$C = \frac{1}{n} \sum_i C_i$$

Coeficient de *clustering* local, II

Exemple



- $C_1 = C_2 = \frac{1}{1}$
- $C_3 = \frac{1}{6}$
- $C_4 = C_5 = 0$
- $C = \frac{1}{5}(1 + 1 + \frac{1}{6}) = \frac{13}{30} = 0,433$

[Newman, 2003]

	network	type	n	m	z	ℓ	α	$C^{(1)}$	$C^{(2)}$	r	Ref(s).
social	film actors	undirected	449 913	25 516 482	113.43	3.48	2.3	0.20	0.78	0.208	20, 416
	company directors	undirected	7 673	55 392	14.44	4.60	–	0.59	0.88	0.276	105, 323
	math coauthorship	undirected	253 339	496 489	3.92	7.57	–	0.15	0.34	0.120	107, 182
	physics coauthorship	undirected	52 909	245 300	9.27	6.19	–	0.45	0.56	0.363	311, 313
	biology coauthorship	undirected	1 520 251	11 803 064	15.53	4.92	–	0.088	0.60	0.127	311, 313
	telephone call graph	undirected	47 000 000	80 000 000	3.16		2.1				8, 9
	email messages	directed	59 912	86 300	1.44	4.95	1.5/2.0		0.16		136
	email address books	directed	16 881	57 029	3.38	5.22	–	0.17	0.13	0.092	321
	student relationships	undirected	573	477	1.66	16.01	–	0.005	0.001	−0.029	45
	sexual contacts	undirected	2 810				3.2				265, 266
information	WWW `nd.edu`	directed	269 504	1 497 135	5.55	11.27	2.1/2.4	0.11	0.29	−0.067	14, 34
	WWW Altavista	directed	203 549 046	2 130 000 000	10.46	16.18	2.1/2.7				74
	citation network	directed	783 339	6 716 198	8.57		3.0/–				351
	Roget's Thesaurus	directed	1 022	5 103	4.99	4.87	–	0.13	0.15	0.157	244
	word co-occurrence	undirected	460 902	17 000 000	70.13		2.7		0.44		119, 157
technological	Internet	undirected	10 697	31 992	5.98	3.31	2.5	0.035	0.39	−0.189	86, 148
	power grid	undirected	4 941	6 594	2.67	18.99	–	0.10	0.080	−0.003	416
	train routes	undirected	587	19 603	66.79	2.16	–		0.69	−0.033	366
	software packages	directed	1 439	1 723	1.20	2.42	1.6/1.4	0.070	0.082	−0.016	318
	software classes	directed	1 377	2 213	1.61	1.51	–	0.033	0.012	−0.119	395
	electronic circuits	undirected	24 097	53 248	4.34	11.05	3.0	0.010	0.030	−0.154	155
	peer-to-peer network	undirected	880	1 296	1.47	4.28	2.1	0.012	0.011	−0.366	6, 354
biological	metabolic network	undirected	765	3 686	9.64	2.56	2.2	0.090	0.67	−0.240	214
	protein interactions	undirected	2 115	2 240	2.12	6.80	2.4	0.072	0.071	−0.156	212
	marine food web	directed	135	598	4.43	2.05	–	0.16	0.23	−0.263	204
	freshwater food web	directed	92	997	10.84	1.90	–	0.20	0.087	−0.326	272
	neural network	directed	307	2 359	7.68	3.97	–	0.18	0.28	−0.226	416, 421

Les xarxes ER no presenten transitivitat, I

- $C = p$, perquè les arestes s'afegeixen de forma independent.
- Donat un graf de n nodes i e arestes, podem "estimar" p com:

$$\tilde{p} = \frac{e}{\frac{1}{2}n(n-1)}$$

- Diem que el *clustering* és gran si $C \gg \tilde{p}$,
  - per tant, les xarxes ER no tenen un coeficient de *clustering* gran, perquè per a elles $C \approx \tilde{p}$.

Les xarxes ER no presenten transitivitat, II

Table 1: Clustering coefficients, C, for a number of different networks; n is the number of node, z is the mean degree. Taken from [146].

Network	n	z	C measured	C for random graph
Internet [153]	6,374	3.8	0.24	0.00060
World Wide Web (sites) [2]	153,127	35.2	0.11	0.00023
power grid [192]	4,941	2.7	0.080	0.00054
biology collaborations [140]	1,520,251	15.5	0.081	0.000010
mathematics collaborations [141]	253,339	3.9	0.15	0.000015
film actor collaborations [149]	449,913	113.4	0.20	0.00025
company directors [149]	7,673	14.4	0.59	0.0019
word co-occurrence [90]	460,902	70.1	0.44	0.00015
neural network [192]	282	14.0	0.28	0.049
metabolic network [69]	315	28.3	0.59	0.090
food web [138]	134	8.7	0.22	0.065

Les xarxes ER no tenen un elevat coeficient de *clustering*, però...

- Podem imitar aquest fenomen en xarxes simulades ("models") mantenint el diàmetre petit?

Les xarxes ER no tenen un elevat coeficient de *clustering*, però...

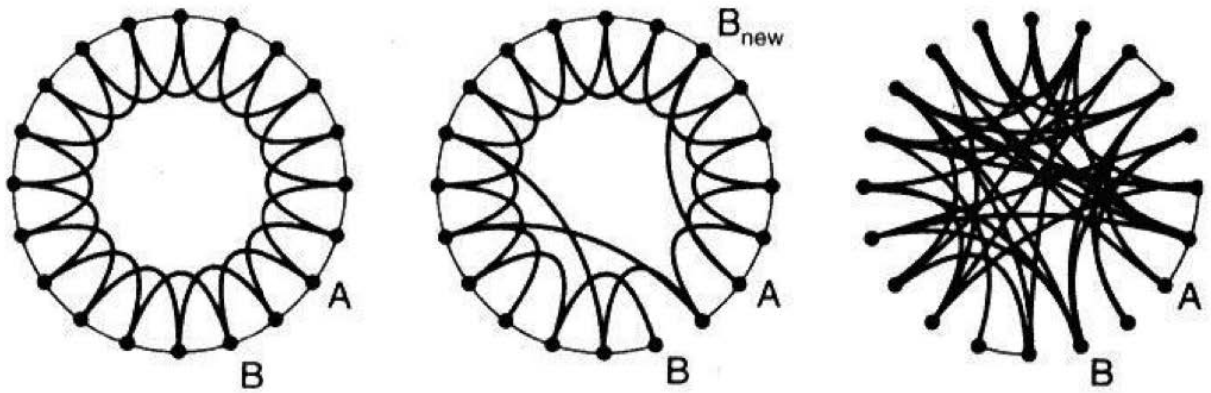
- Podem imitar aquest fenomen en xarxes simulades ("models") mantenint el diàmetre petit?
- La resposta és SÍ.

El model de Watts-Strogatz, I

[Watts and Strogatz, 1998]

Reprenem dos fets observats en les xarxes reals:

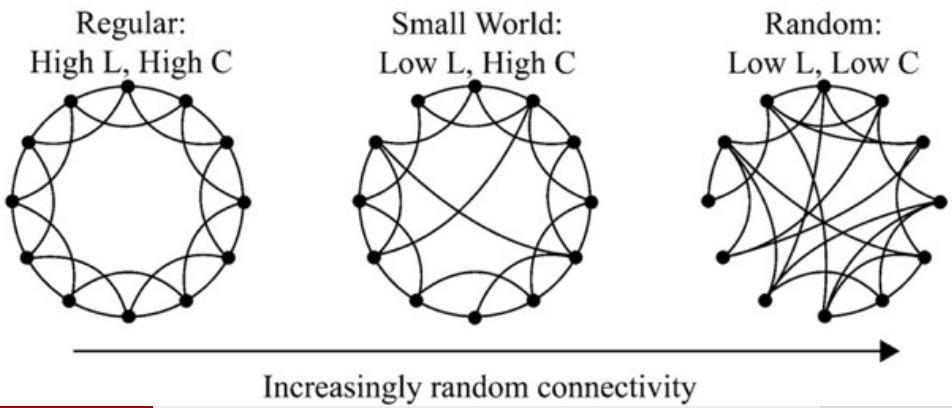
- Elevat coeficient de *clustering*: els amics dels meus amics també són amics meus.
- Diàmetre petit.



D'esquerra a dreta: xarxa regular (tots els nodes tenen el mateix grau), xarxa de món petit i xarxa aleatòria.

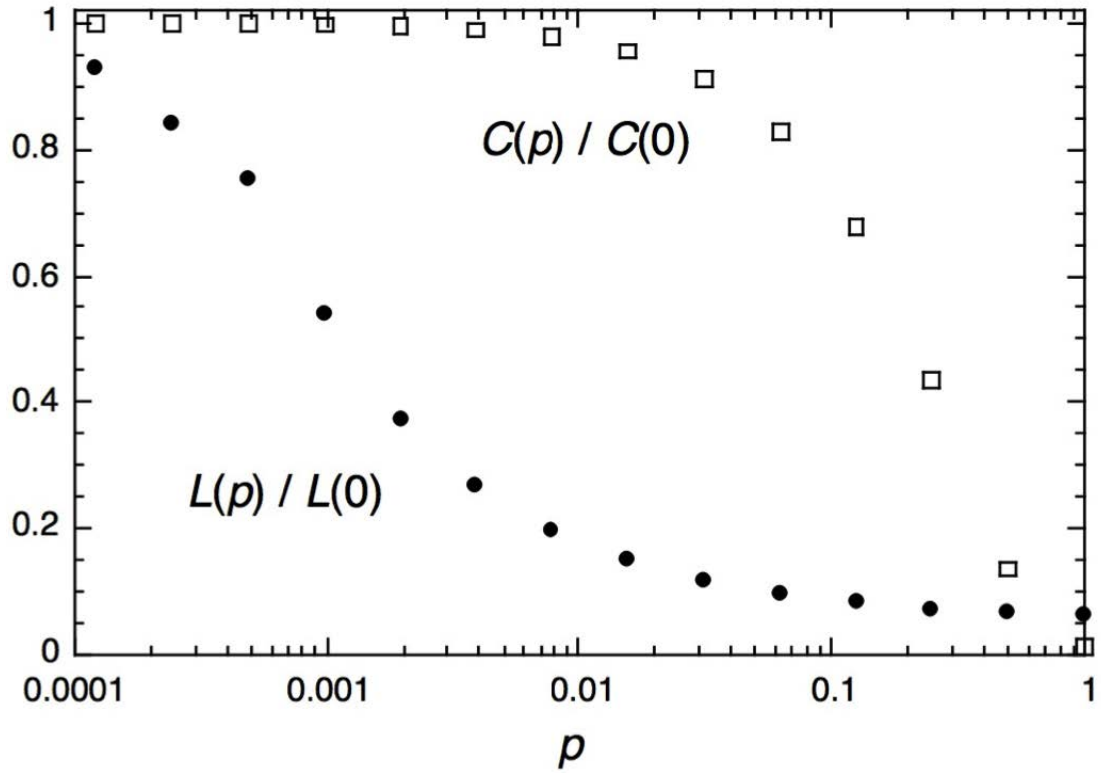
El model de Watts-Strogatz, II

- Comença amb tots els n vèrtexs situats sobre un anell.
- Cada vèrtex té, inicialment, 4 connexions amb els seus nodes més propers.
  - imita la connectivitat local o geogràfica
- Amb probabilitat p, reconnecta cada connexió local a un vèrtex a l'atzar.
  - $p = 0$ elevat *clustering*, diàmetre gran
  - $p = 1$ baix *clustering*, diàmetre petit (model d'ER)
- Què passa entremig?
  - A mida que incrementem p de 0 a 1
    - ★ La distància mitjana decreix ràpidament.
    - ★ El *clustering* decreix lentament.



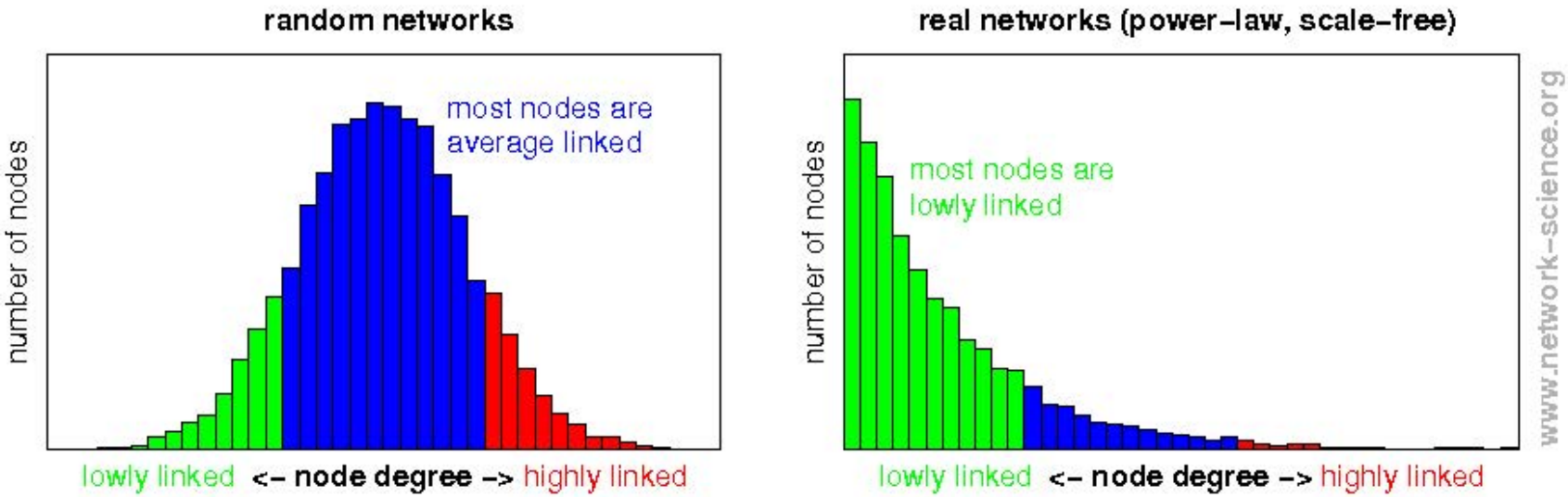
El model de Watts-Strogatz, III

Amb un valor adequat de $p \approx 0,01$ (1%), observem que el model assoleix un elevat coeficient de *clustering* i alhora un diàmetre petit.



Distribució de grau

Histogrames del nombre de nodes que tenen un determinat grau:

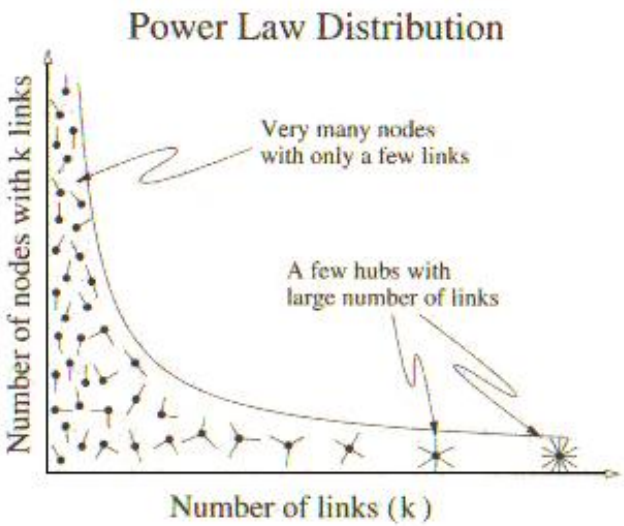


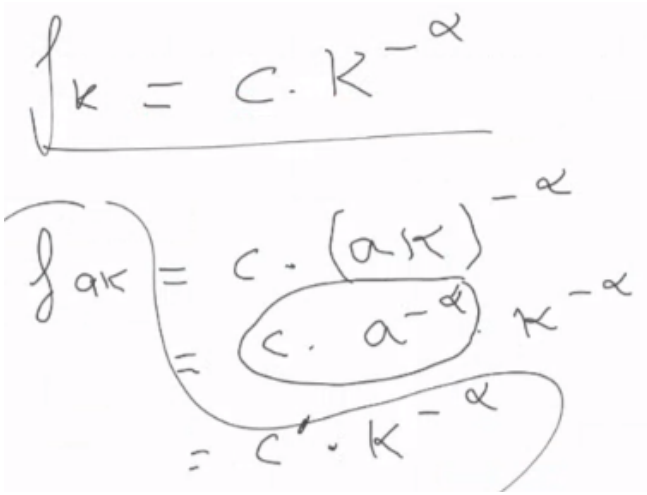
$f_k =$ proporció de nodes de grau k

Xarxes lliures d'escala

La distribució de grau de la majoria de xarxes del món real, segueix una distribució de Llei de potència (*power-law*).

$$f_k = ck^{-\alpha}$$





- La "llarga cua" implica l'existència de *hubs*.
- Els *hubs* són nodes amb moltes connexions (grau alt).

Les xarxes aleatòries no són lliures d'escala

La distribució de grau de les xarxes aleatòries segueix una distribució binomial (o Poisson si n és gran):

$$f_k = \binom{n}{k} p^k (1-p)^{n-k} \approx \frac{z^k e^{-z}}{k!}$$

- On $z = p(n-1)$ és el grau mig.
- La probabilitat dels nodes d'alt grau es fa exponencialment petita,

Les xarxes aleatòries no són lliures d'escala

La distribució de grau de les xarxes aleatòries segueix una distribució binomial (o Poisson si n és gran):

$$f_k = \binom{n}{k} p^k (1-p)^{n-k} \approx \frac{z^k e^{-z}}{k!}$$

- On $z = p(n-1)$ és el grau mig.
- La probabilitat dels nodes d'alt grau es fa exponencialment petita,
- per tant, no hi ha *hubs*.

Les xarxes ER no són lliures d'escala, però...

- Podem simular xarxes lliures d'escala?

Les xarxes ER no són lliures d'escala, però...

- Podem simular xarxes lliures d'escala?
- La resposta és SÍ.

Adjunció preferent (*preferential attachment*)

- Dinàmica “el ric es fa més ric”.
  - Cada node nou que s’afegeix a la xarxa té una major preferència o tendència a enllaçar amb els nodes de la xarxa que ja tenen més enllaços.

Adjunció preferent (*preferential attachment*)

- Dinàmica “el ric es fa més ric”.
  - Cada node nou que s’afegeix a la xarxa té una major preferència o tendència a enllaçar amb els nodes de la xarxa que ja tenen més enllaços.
- Exemples:
  - Com més amics tinguis, més fàcil serà de fer-ne de nous.

Adjunció preferent (*preferential attachment*)

- Dinàmica “el ric es fa més ric”.
  - Cada node nou que s’afegeix a la xarxa té una major preferència o tendència a enllaçar amb els nodes de la xarxa que ja tenen més enllaços.
- Exemples:
  - Com més amics tinguis, més fàcil serà de fer-ne de nous.
  - Com més negocis faci una empresa, més fàcil serà aconseguir-ne més.

Adjunció preferent (*preferential attachment*)

- Dinàmica “el ric es fa més ric”.
  - Cada node nou que s’afegeix a la xarxa té una major preferència o tendència a enllaçar amb els nodes de la xarxa que ja tenen més enllaços.
- Exemples:
  - Com més amics tinguis, més fàcil serà de fer-ne de nous.
  - Com més negocis faci una empresa, més fàcil serà aconseguir-ne més.
  - Com més gent va a un restaurant, més gent voldrà anar-hi.

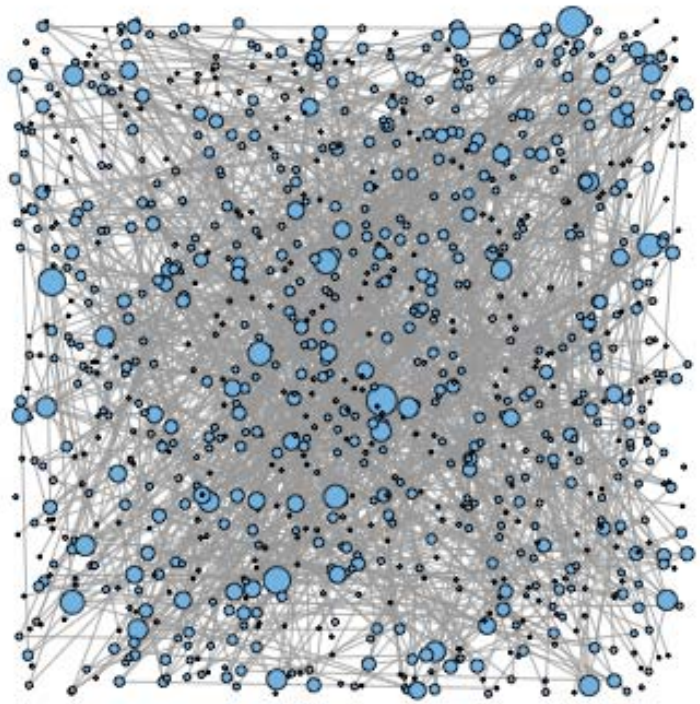
El model de Barabási-Albert

[Barabási and Albert, 1999]

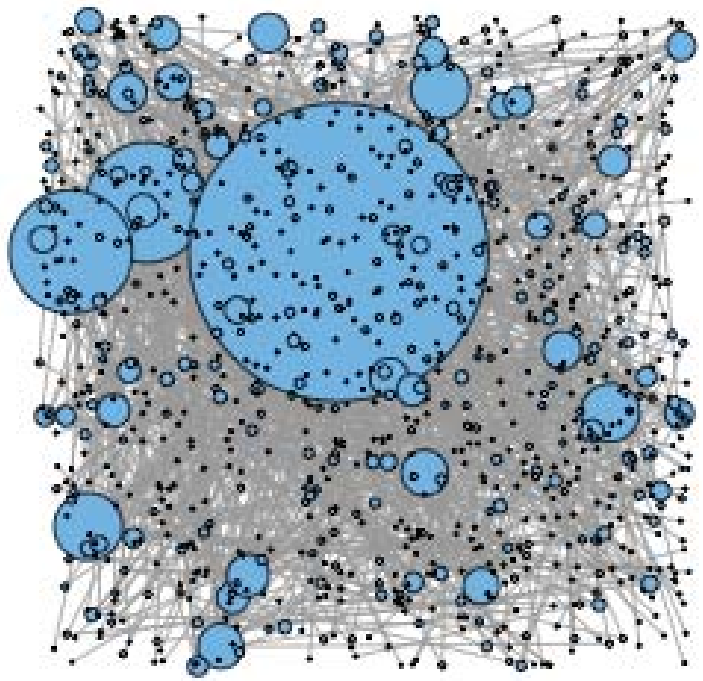
- "Model de creixement":
  - El model controla com creix la xarxa al llarg del temps.
- Usa el mecanisme d'adjunció preferent per guiar el creixement de la xarxa.
  - Els nodes nous prefereixen enllaçar-se amb nodes que ja estiguin molt connectats.
- Procés (simplificat):
  1. El procés comença amb un subgraf inicial.
  2. Cada node nou "arriba" amb m arestes.
  3. La probabilitat de connectar-se a un node existent i és proporcional al grau de i.
  4. La xarxa resultant té una distribució de grau que segueix una Llei de potència amb exponent $\alpha = 3$.

Erdös Renyi vs. Barabási-Albert

Experiment amb 1000 nodes, 999 arestes ($m_0 = 1$ en el model BA).



random



preferential attachment

Resum

fenomen	xarxes reals	ER	WS	BA
diàmetre petit	sí	sí	sí	sí
coef. *clustering* gran	sí	no	sí	sí[2]
lliure d'escala	sí	no	no	sí

[2] El coeficient de *clustering* és més gran que en les xarxes aleatòries però no tant com en les xarxes WS, per exemple.

Anàlisi de xarxes, Part II

1. Mesures de centralitat

Anàlisi de xarxes, Part II

1. Mesures de centralitat
   - Centralitat de grau (*Degree centrality*).
   - Centralitat de proximitat (*Closeness centrality*).
   - Centralitat d'intermediació (*Betweenness centrality*).

Anàlisi de xarxes, Part II

1. Mesures de centralitat
   - Centralitat de grau (*Degree centrality*).
   - Centralitat de proximitat (*Closeness centrality*).
   - Centralitat d'intermediació (*Betweenness centrality*).
2. Algorismes de detecció de comunitats

Anàlisi de xarxes, Part II

1. Mesures de centralitat
   - Centralitat de grau (*Degree centrality*).
   - Centralitat de proximitat (*Closeness centrality*).
   - Centralitat d'intermediació (*Betweenness centrality*).
2. Algorismes de detecció de comunitats
   - *Clustering* jeràrquic.
     - ★ D'aglomeració
     - ★ De Girvan-Newman
   - Maximització de la modularitat: mètode de Louvain.

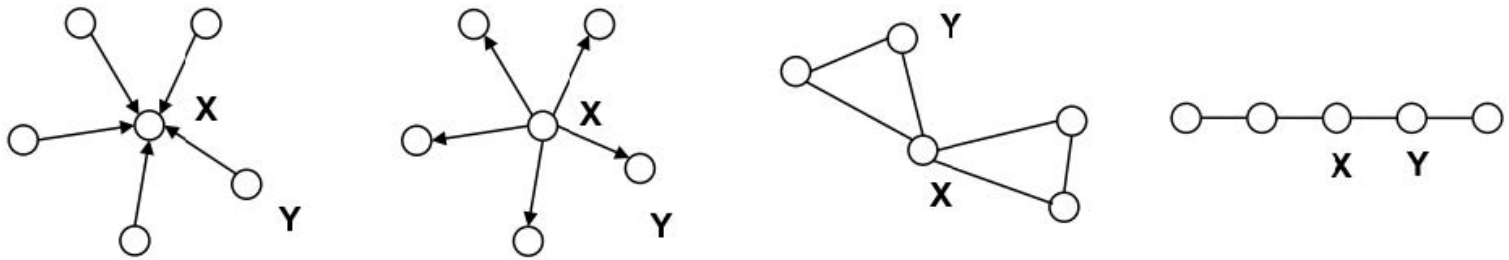
Índex

Centralitat en les xarxes

La centralitat és una mesura d'un node respecte els altres nodes de la xarxa.

- Un node central és *important* i/o *potent*.
- Un node central ocupa una *posició influent dins la xarxa*.
- Un node central té una *posició avantatjosa dins la xarxa*.

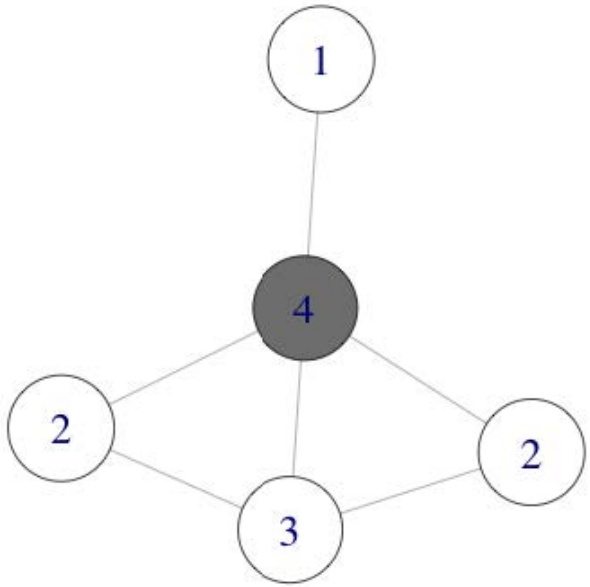
En aquestes xarxes, X té una centralitat més alta que Y d'acord amb una mesura concreta: grau d'entrada, grau de sortida, intermediació i proximitat.



Centralitat de grau (*degree centrality*), I

El que té més amics és el més important.

$$degree\_centrality(i) \stackrel{\text{def}}{=} k(i)$$

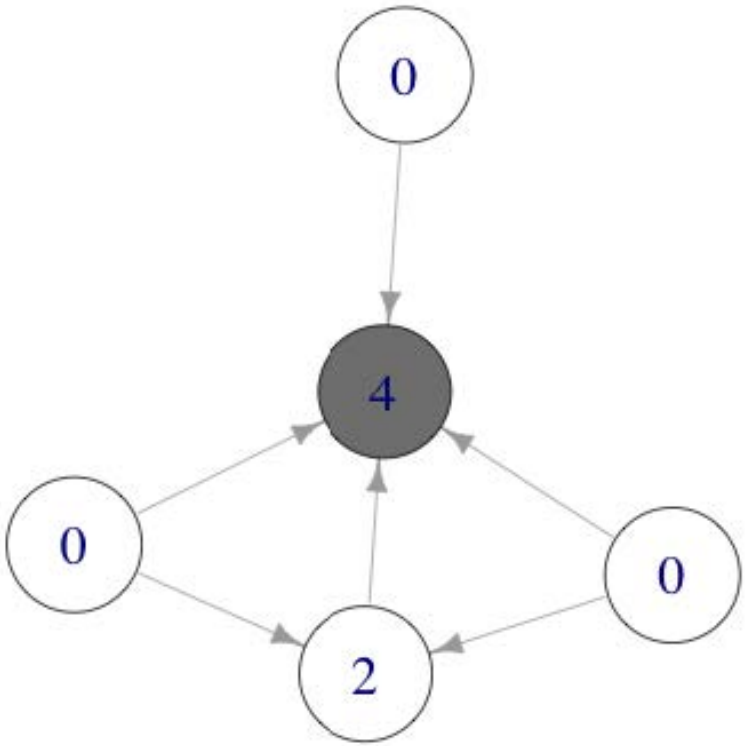


En quins casos és una bona mesura de centralitat?

Centralitat de grau, II

Xarxes dirigides

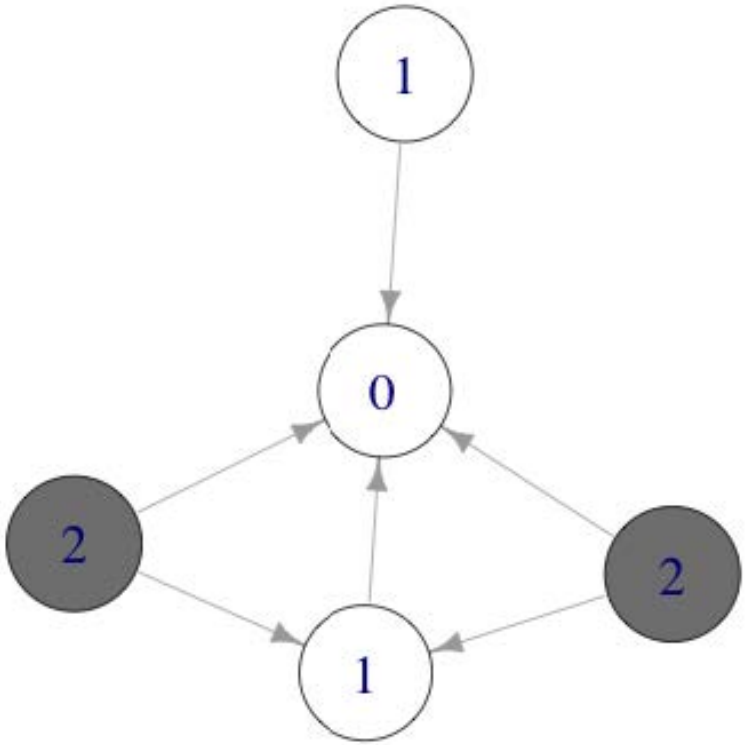
$$in\_degree\_centrality(i) \stackrel{\text{def}}{=} k_{in}(i)$$



Centralitat de grau, III

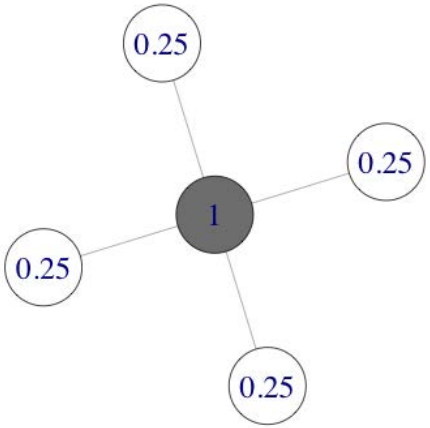
Xarxes dirigides

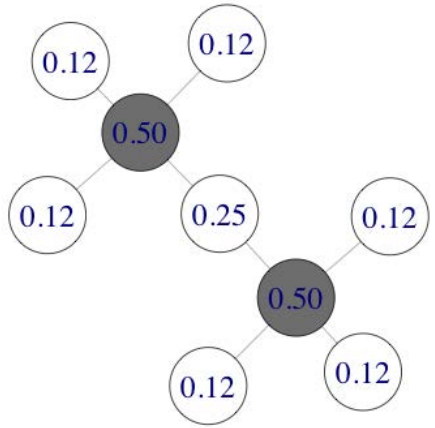
$$out\_degree\_centrality(i) \stackrel{\text{def}}{=} k_{out}(i)$$



Centralitat de grau, IV

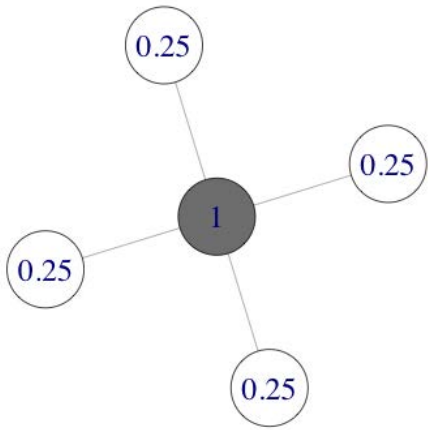
Existeix una versió *normalitzada* que divideix la centralitat de cada grau pel valor de centralitat màxim possible, és a dir, $n - 1$ (per tant, tots els valors quedaran entre 0 i 1).

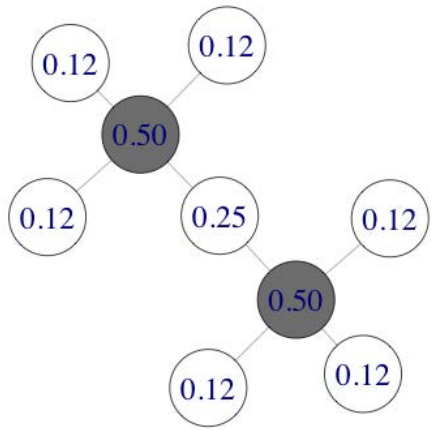




Centralitat de grau, IV

Existeix una versió *normalitzada* que divideix la centralitat de cada grau pel valor de centralitat màxim possible, és a dir, $n - 1$ (per tant, tots els valors quedaran entre 0 i 1).

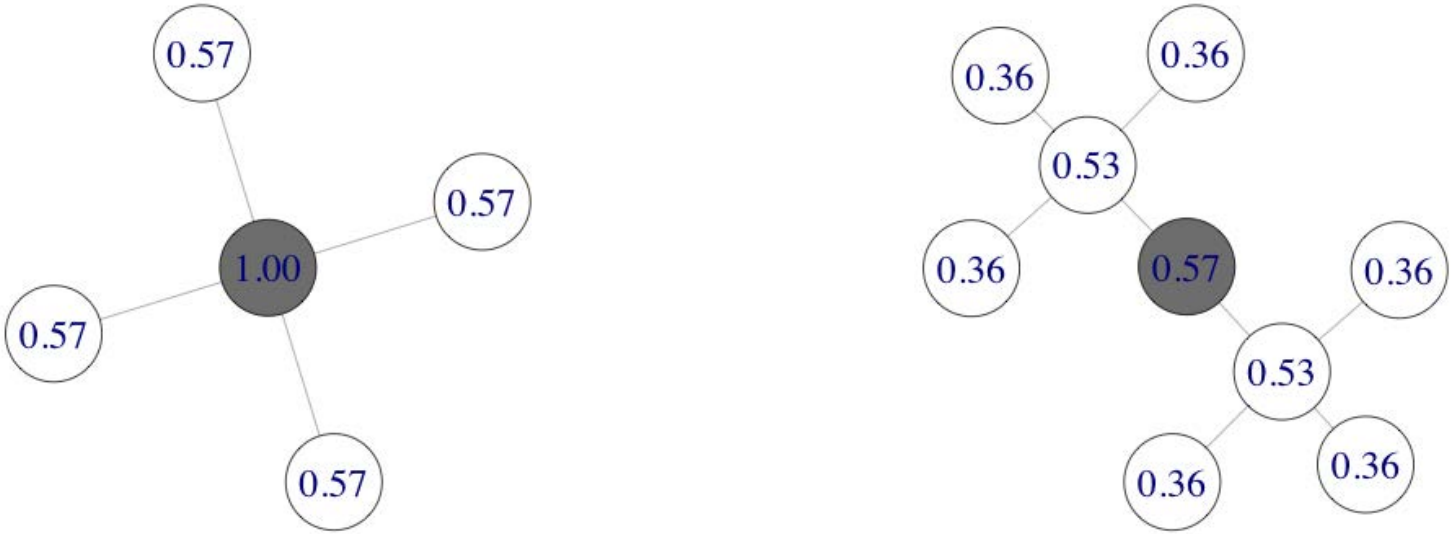




Mireu aquests exemples, us semblen correctes els valors de centralitat de grau?

Centralitat de proximitat (*closeness centrality*)

$$closeness\_centrality(i) \stackrel{\text{def}}{=} \left(\frac{\sum_{j \neq i} d(i,j)}{n-1} \right)^{-1} = \frac{n-1}{\sum_{j \neq i} d(i,j)}$$

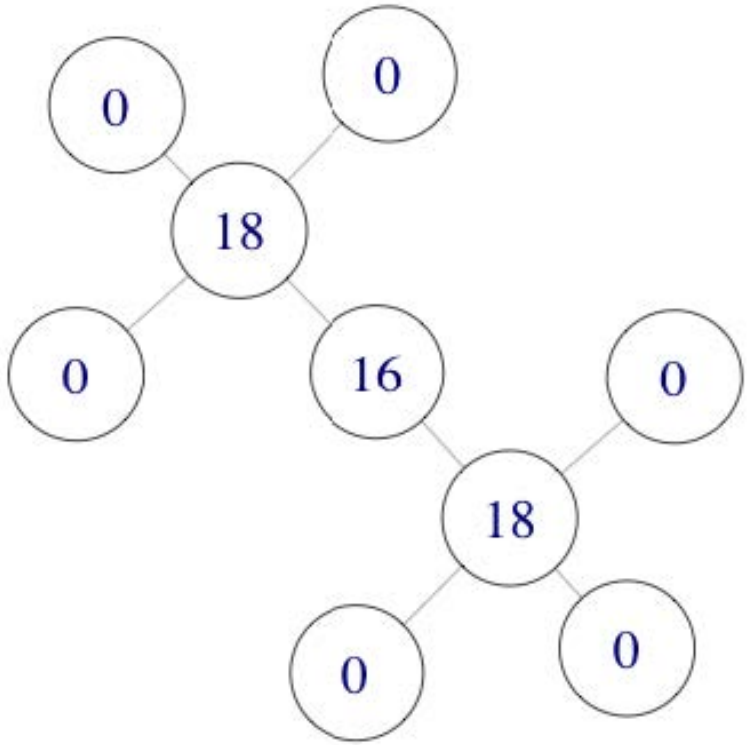


Ara el que importa és estar a prop de tothom, és a dir, ser de fàcil accés o tenir el poder d'arribar ràpidament als altres.

Centralitat d'intermediació (*betweenness centrality*), I

Un node és important si es troba en molts dels camins més curts

- per tant, és essencial per passar informació a través de la xarxa.



Centralitat d'intermediació, II

$$betweenness\_centrality(i) \stackrel{\text{def}}{=} \sum_{j<k} \frac{g_{jk}(i)}{g_{jk}}$$

On

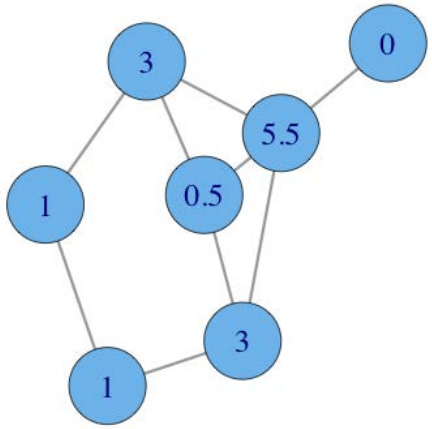
- g_{jk} és el nombre de camins més curts entre j i k, i
- $g_{jk}(i)$ és el nombre de camins curts que passen per i.

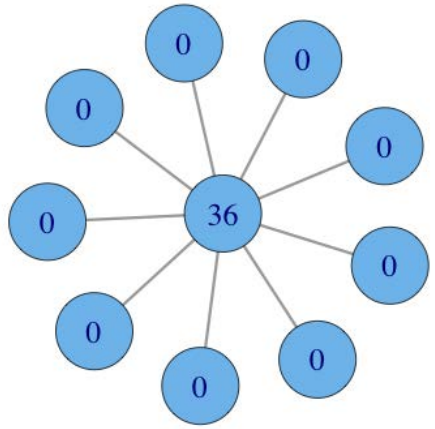
Sovint es dóna normalitzada:

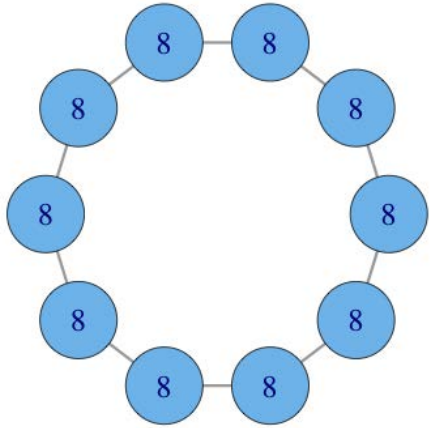
$$norm\_betweenness\_centrality(i) \stackrel{\text{def}}{=} \frac{betweenness\_centrality(i)}{\binom{n-1}{2}}$$

Centralitat d'intermediació, III

Exemples (sense normalitzar)





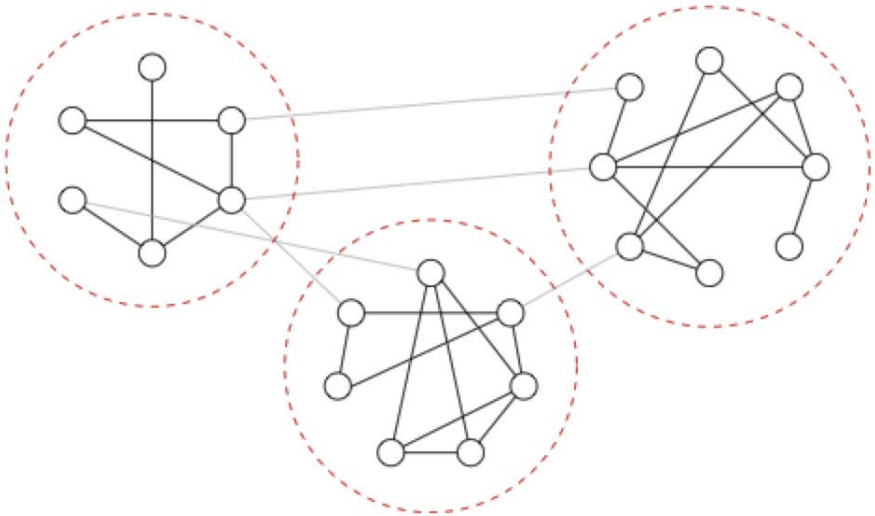


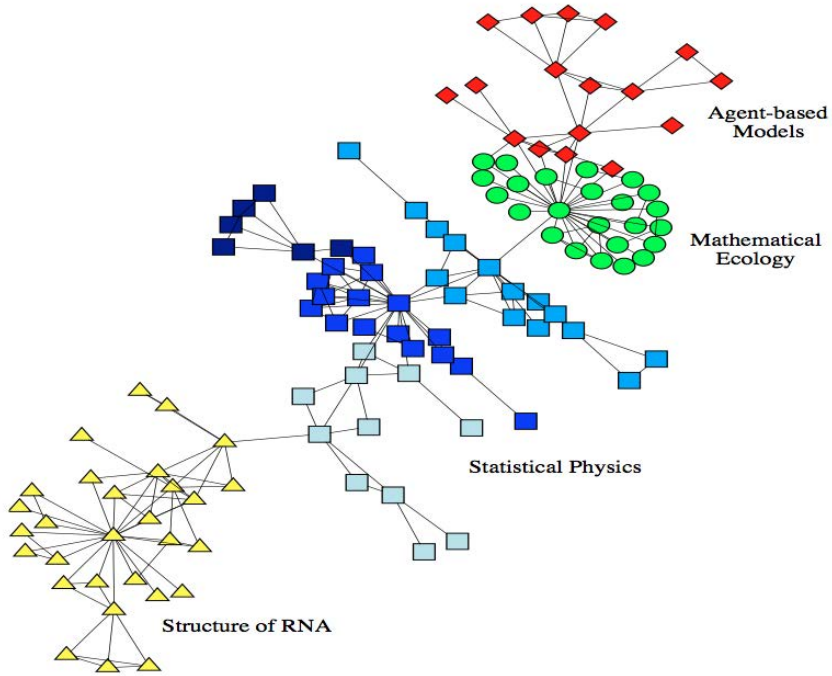
Índex

Què és una comunitat?

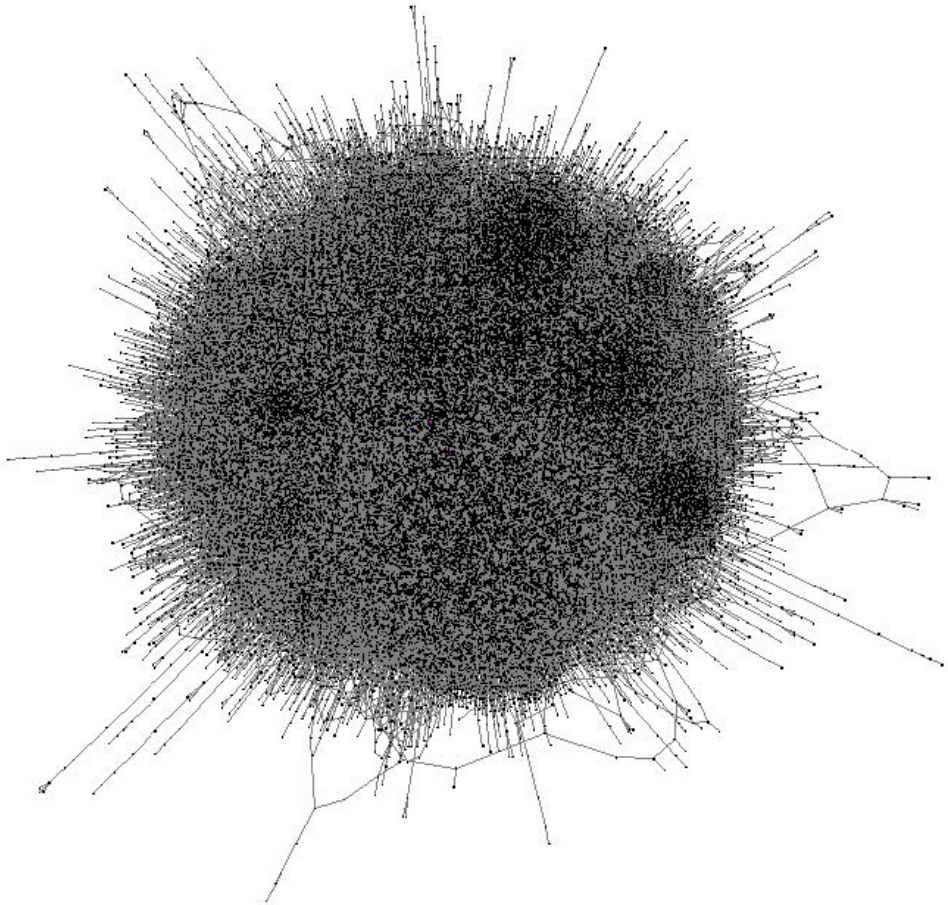
Comunitat: conjunt de nodes que estan fortament lligats entre si però poc lligats amb la resta de la xarxa.

Trobar una comunitat $\approx$ fer particions en un graf.





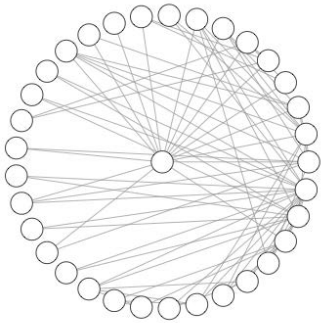
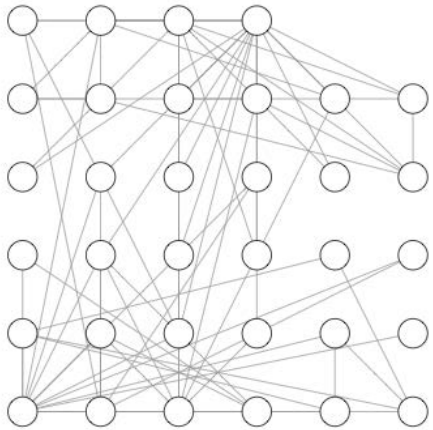
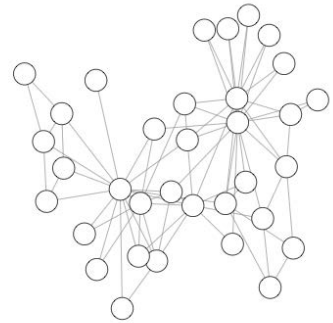
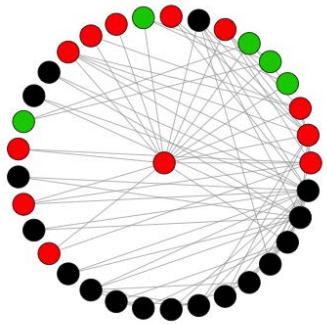
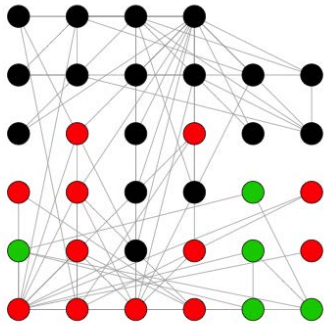
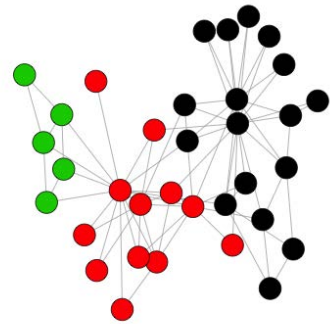
Per què és important detectar comunitats?



- P.e., a internet, poden representar pàgines de temes relacionats.
- En general, poden revelar propietats col·lectives sense mostrar informacions individuals privades.

La percepció visual no és fiable

Millor usar algorismes objectius

Idea principal

No hi ha una definició universal però algunes idees són:

- Una comunitat hauria d'estar *densament connectada*.
- Una comunitat hauria d'estar *ben separada* de la resta de la xarxa.
- Els membres d'una comunitat haurien de ser *més similars* entre ells que amb la resta.

El més habitual:

- # d'arestes *intra-cluster* $>$ # d'arestes *inter-cluster*

Algunes definicions, I

Sigui $G = (V, E)$ una xarxa amb $|V| = n$ nodes i $|E| = m$ arestes. Sigui C un subconjunt de nodes de la xarxa (un "clúster" o "comunitat") de mida $|C| = n_c$. Aleshores:

- densitat *intra-cluster*:

$$\delta_{int}(C) = \frac{\text{\# d'arestes internes de } C}{n_c(n_c - 1)/2}$$

- densitat *inter-cluster*:

$$\delta_{ext}(C) = \frac{\text{\# d'arestes entre els clústers de } C}{n_c(n - n_c)}$$

Una comunitat hauria de tenir $\delta_{int}(C) > \delta(G)$, on $\delta(G)$ és la densitat mitjana de les arestes de tot el graf G, és a dir:

$$\delta(G) = \frac{\text{\# d'arestes de } G}{n(n-1)/2}$$

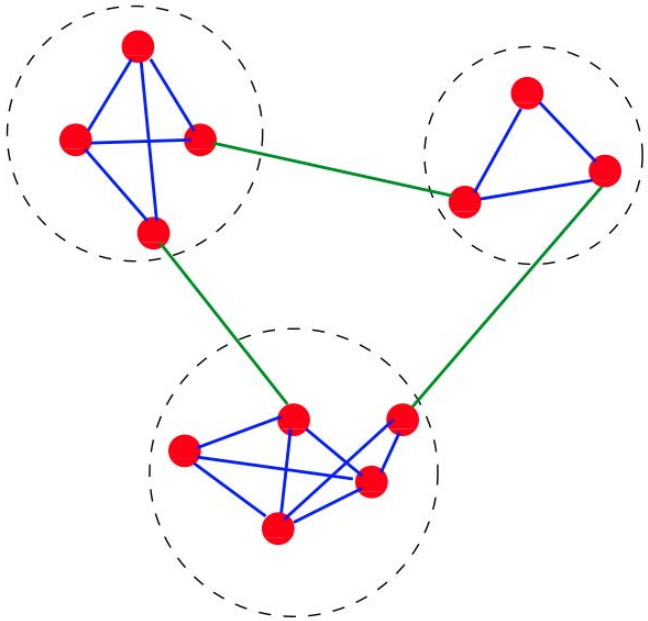
Algunes definicions, II

La majoria d'algorismes busquen un equilibri entre $\delta_{int}(C)$ *gran* i $\delta_{ext}(C)$ *petita*

- per exemple, optimitzant $\sum_C \delta_{int}(C) - \delta_{ext}(C)$ entre totes les comunitats C

Definint-ho amb major precisió:

- $m_c =$ # d'arestes dins el clúster $C = |\{(u, v)|u, v \in C\}|$
- $f_c =$ # d'arestes a la frontera $C = |\{(u, v)|u \in C, v \notin C\}|$



- $n_{c1} = 4, m_{c1} = 5, f_{c1} = 2$
- $n_{c2} = 3, m_{c2} = 3, f_{c2} = 2$
- $n_{c3} = 5, m_{c3} = 8, f_{c3} = 2$

Criteris de qualitat d'una comunitat

Sovint, tenen en compte més d'un factor. Tots recompensen els conjunts de nodes amb moltes arestes internes i poques d'externes.

- **conductància**: proporció d'arestes que surten del clúster $\frac{f_c}{2m_c+f_c}$
- **expansió**: # d'arestes per node que surten del clúster $\frac{f_c}{n_c}$
- **densitat interna** o densitat intra-cluster: $\frac{m_c}{n_c(n_c-1)/2}$
- ***cut ratio*** o densitat inter-cluster: $\frac{f_c}{n_c(n-n_c)}$
- **modularitat**: diferència entre el # d'arestes de C i el # esperat d'arestes $E[m_c]$ d'una xarxa aleatòria amb la mateixa distribució de grau

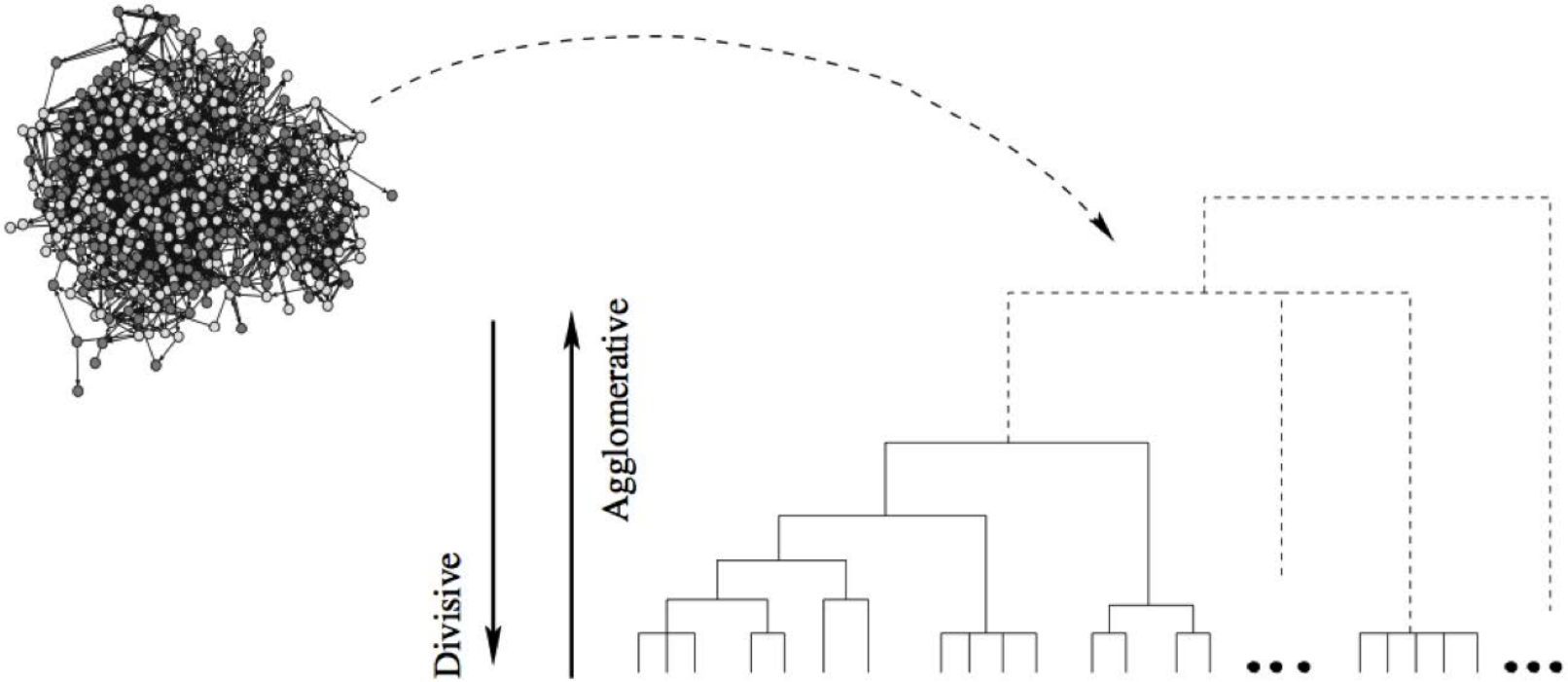
$$\frac{1}{4m}(m_c - E[m_c])$$

Mètodes que veurem

- *Clustering* jeràrquic:
  - D'aglomeració.
  - De divisió (algorisme de Girvan-Newman).
- Algorismes de maximització de la modularitat.
  - Mètode de Louvain.

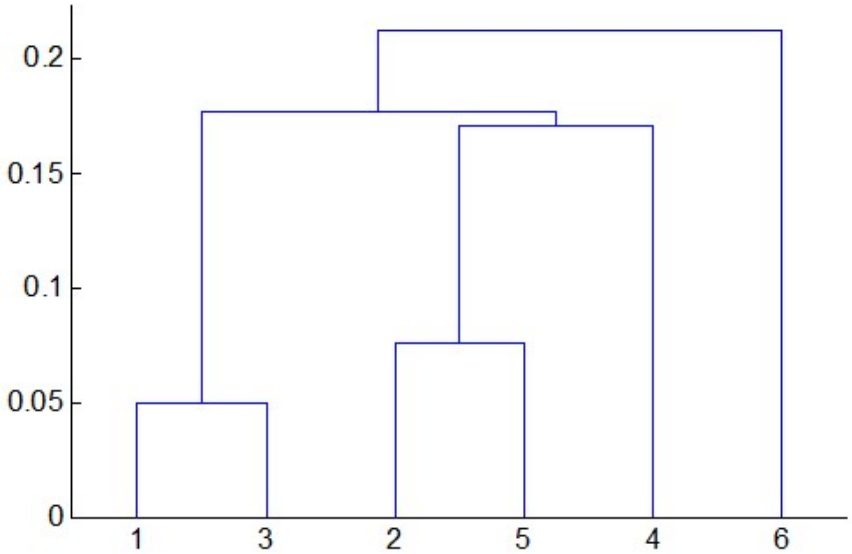
Clustering jeràrquic, I

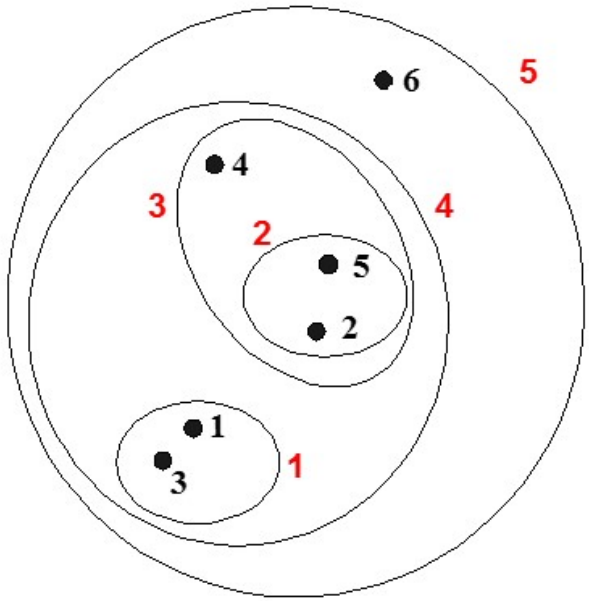
De la bola de pèls al dendrograma



Clustering jeràrquic, II

El dendrograma, diagrama en forma d'arbre, registra la seqüències de fusions o escissions.





Clustering jeràrquic, III

D'aglomeració [Newman, 2010]

Cal definir:

- Una mesura de similitud entre els nodes.
- Una mesura de similitud entre *conjunts de nodes*.

Pseudocodi:

1. Assignar a cada node el seu propi clúster.
2. Trobar el parell de clústers de major similitud i ajuntar-los formant un sol clúster.
3. Recalcular les similituds entre el nou clúster ajuntat i la resta.
4. Tornar al pas 2 fins que tots els nodes formin un sol clúster.

Clustering jeràrquic, IV

Mesures de similitud entre nodes w_{ij}

Sigui A la matriu d'adjacència de la xarxa, i.e. $A_{ij} = 1$ si $(i, j) \in E$ i 0 altrament.

- **Índex Jaccard:**

$$w_{ij} = \frac{|\Gamma(i) \bigcap \Gamma(j)|}{|\Gamma(i) \bigcup \Gamma(j)|}$$

  on $\Gamma(i)$ és el conjunt de veïns del node i.

- **Similitud cosinus:**

$$w_{ij} = \frac{\sum_k A_{ik} A_{jk}}{\sqrt{\sum_k {A^2}_{ik}} \sqrt{\sum_k {A^2}_{jk}}} = \frac{n_{ij}}{\sqrt{k_i k j}}$$

  on:

  - $n_{ij} = |\Gamma(i) \bigcap \Gamma(j)| = \sum_k A_{ik} A_{jk}$ i
  - $k_i = \sum_k A_{ik}$ és el grau del node i.

Clustering jeràrquic, V

Mesures de similitud entre nodes w_{ij}

- **Distància euclidiana** (o de Hamming atès que A és binària):

$$d_{ij} = \sum_k (A_{ik} - A_{jk})^2$$

- **Distància euclidiana normalitzada:**

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

- **Coeficient de correlació de Pearson:**

$$r_{ij} = \frac{\sum_k (A_{ik} - \mu_i)(A_{jk} - \mu_j)}{n\sigma_i\sigma_j}$$

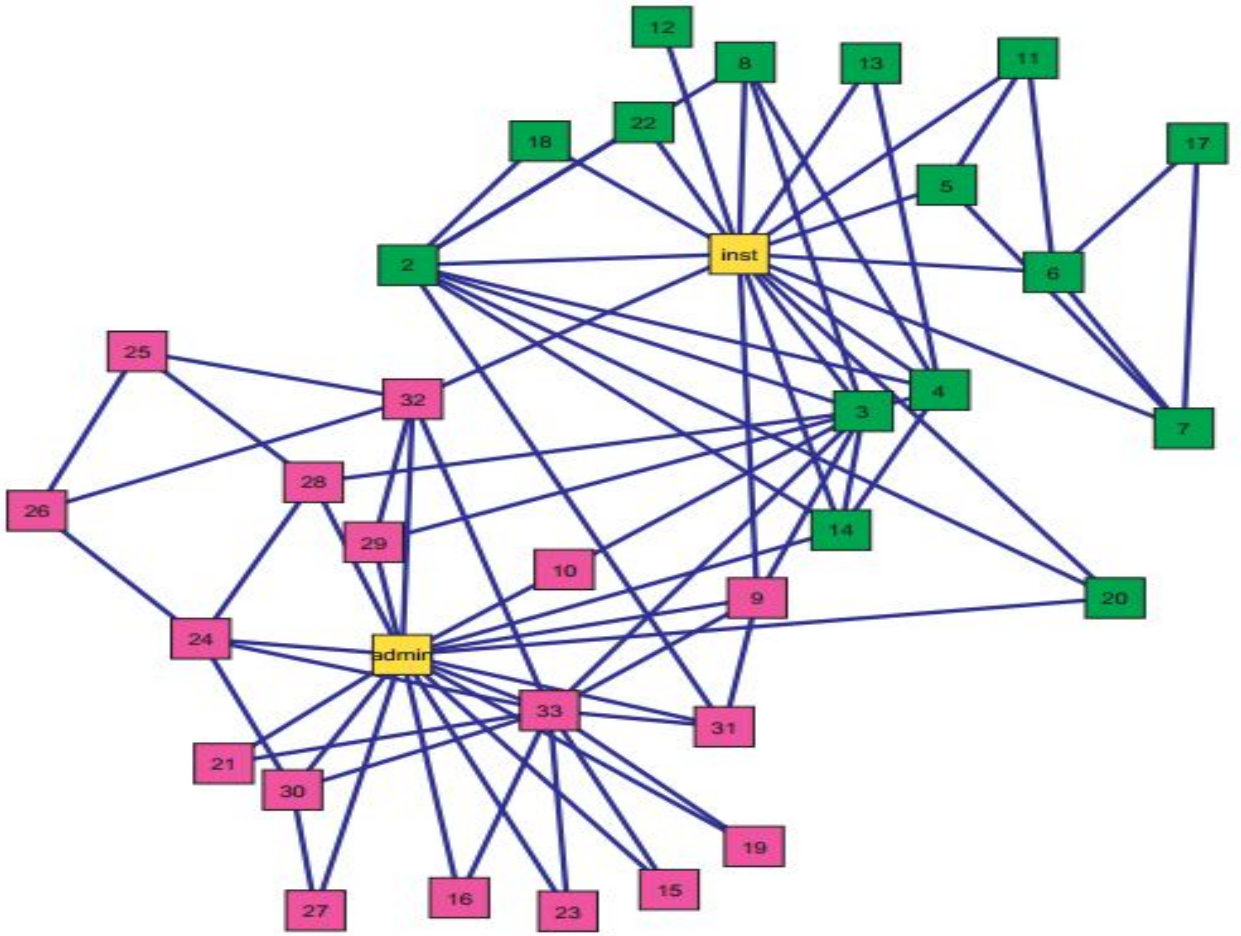
on $\mu_i = \frac{1}{n}\sum_k A_{ik}$ i $\sigma_i = \sqrt{\frac{1}{n}\sum_k (A_{ik} - \mu_i)^2}$

- **Enllaç simple:** $D_{XY} = \min_{x \in X, y \in Y} d_{xy}$
  Entre dos clústers X i Y, és la distància mínima entre qualsevol node de X i qualsevol node de Y.
- **Enllaç complet:** $D_{XY} = \max_{x \in X, y \in Y} d_{xy}$
  Entre dos clústers X i Y, és la distància màxima entre qualsevol node de X i qualsevol node de Y.
- **Enllaç mitjà:** $D_{XY} = \frac{\sum_{x \in X, y \in Y} d_{xy}}{|X| \times |Y|}$
  Entre dos clústers X i Y, és la distància mitjana entre els nodes de X i els nodes de Y.

Clustering jeràrquic d'aglomeració, VII

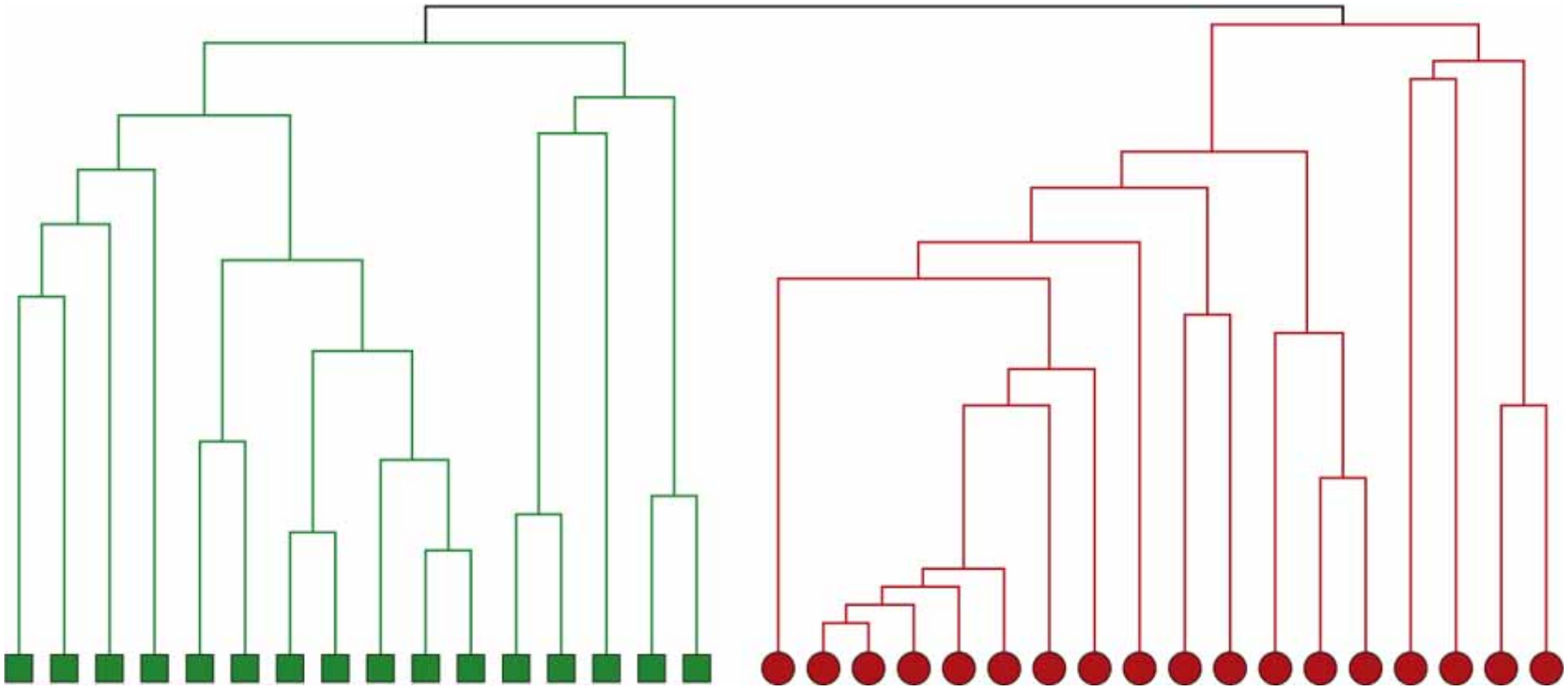
Zachary's karate club network

Dades: Interaccions entre els membres d'un club de karate. L'administrador i l'instructor van barallar-se i el club es va escindir formant dos clubs rivals.



Clustering jeràrquic d'aglomeració, VIII

Zachary's karate club network usant enllaç mitjà

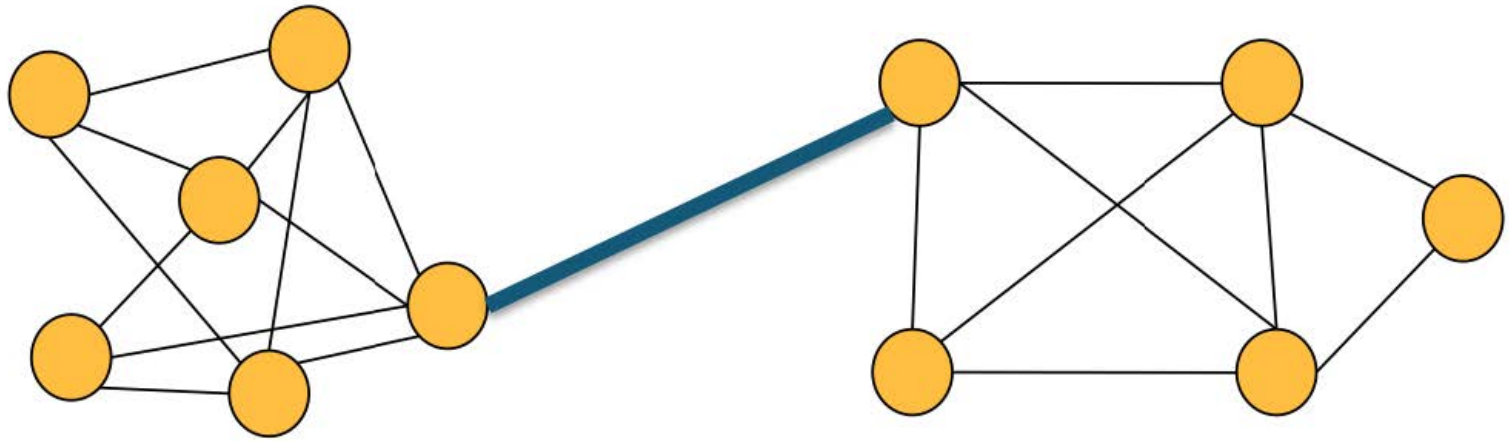


El resultat teòric va coincidir pràcticament amb el real tret d'un node.

Algorisme de Girvan-Newman, I

Algorisme jeràrquic **divisiu** [Girvan and Newman, 2002]

Intermediació (*betweenness*) en arestes El valor d'intermediació d'una aresta és el nombre de camins més curts de la xarxa que passen per aquesta aresta. Expressa la idea de que els "ponts" entre comunitats han de tenir un valor alt com a arestes d'intermediació.



Algorisme de Girvan-Newman, II

Pseudocodi:

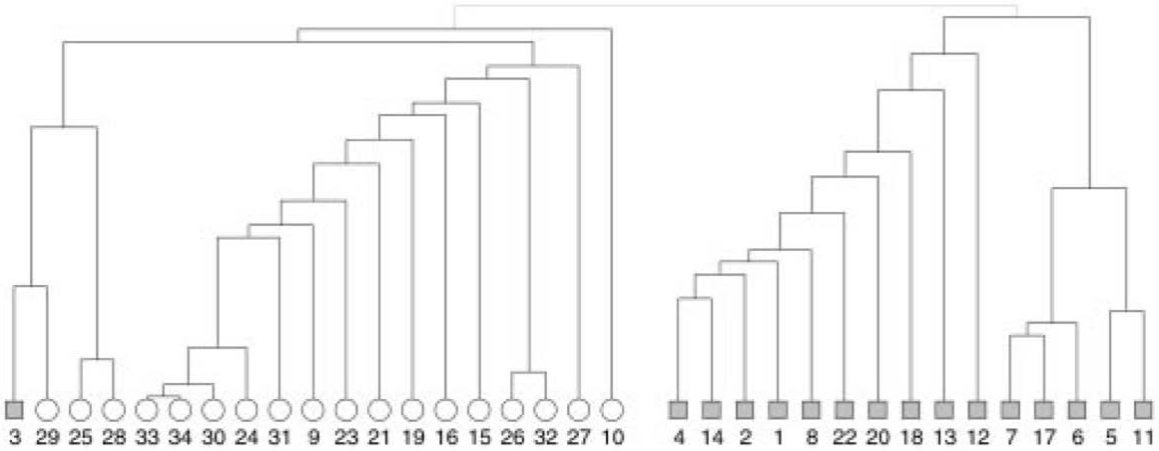
1. Calcular el valor d'intermediació de totes les arestes de la xarxa.
2. Eliminar l'aresta amb un valor d'intermediació més alt.
3. Tornar al pas 1 fins que no quedin arestes.

Algorisme de Girvan-Newman, II

Pseudocodi:

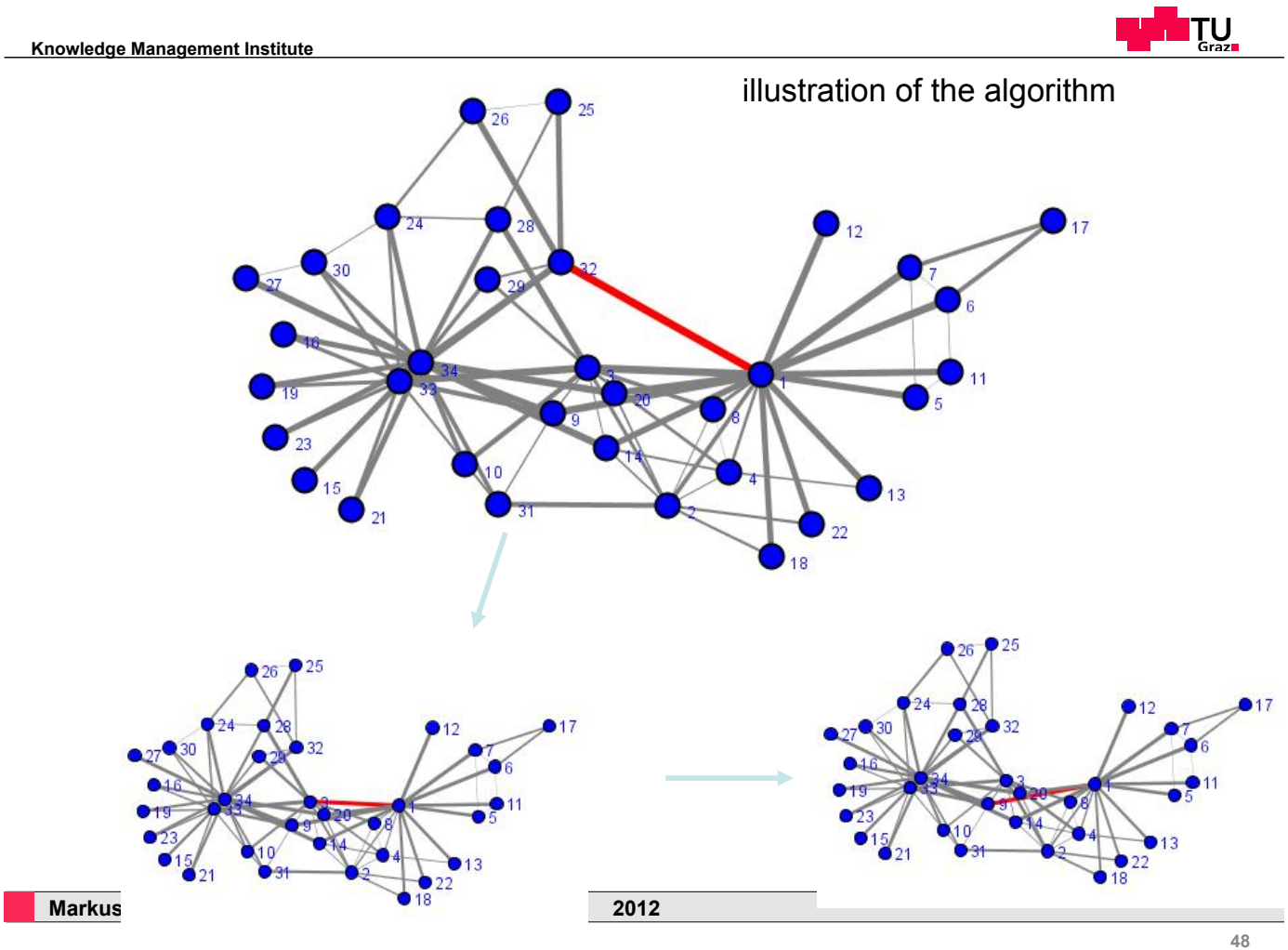
1. Calcular el valor d'intermediació de totes les arestes de la xarxa.
2. Eliminar l'aresta amb un valor d'intermediació més alt.
3. Tornar al pas 1 fins que no quedin arestes.

El resultat és un dendrograma:



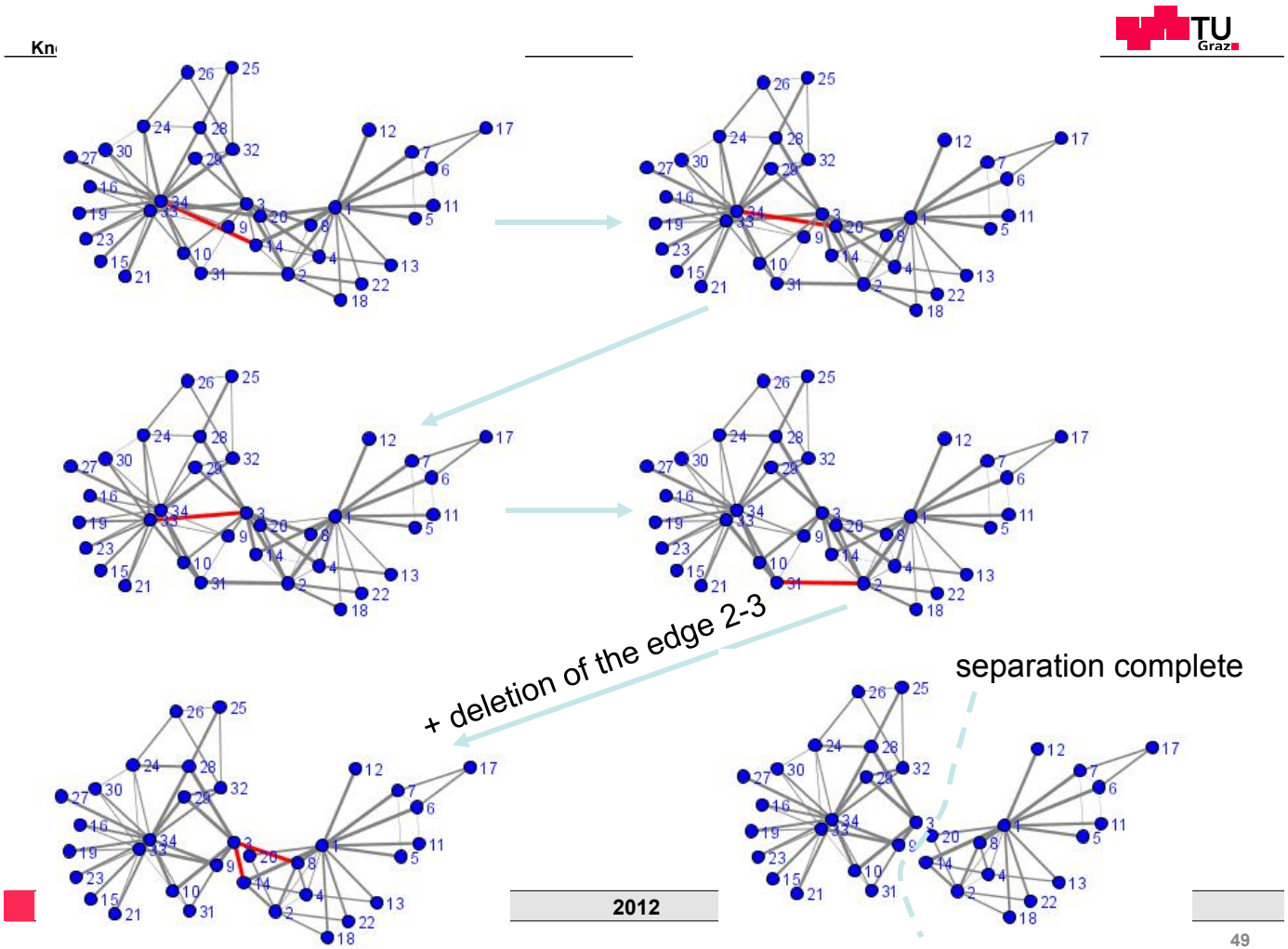
Algorisme de Girvan-Newman, III

Exemple



Algorisme de Girvan-Newman, IV

Exemple



Definició de modularitat [Newman, 2010]

Usant el model *nul*

Se suposa que les xarxes aleatòries no presenten estructures de comunitat, per tant, s'usen com a models nuls.

$Q =$ (# de comunitats *intra-cluster*) $-$ (# d'arestes esperat)

En particular:

$$Q = \frac{1}{2m} \sum_{ij} (A_{ij} - P_{ij}) \delta(C_i, C_j)$$

on P_{ij} és el nombre esperat d'arestes entre els nodes i i j en el model nul, C_i és la comunitat del vèrtex i, i $\delta(C_i, C_j) = 1$ si $C_i = C_j$ i 0 altrament.

Com es calcula P_{ij}?

Usant el model nul "configuració"

El model "configuració" d'un graf aleatori genera un graf amb la mateixa distribució de grau que el graf original uniformement a l'atzar.

- Calculem P_{ij}.
- Hi ha $2m$ *stubs* o arestes partides disponibles en el model de configuració.
- Sigui p_i la probabilitat d'agafar a l'atzar un *stub* que incideixi amb i.

$$p_i = \frac{k_i}{2m}$$

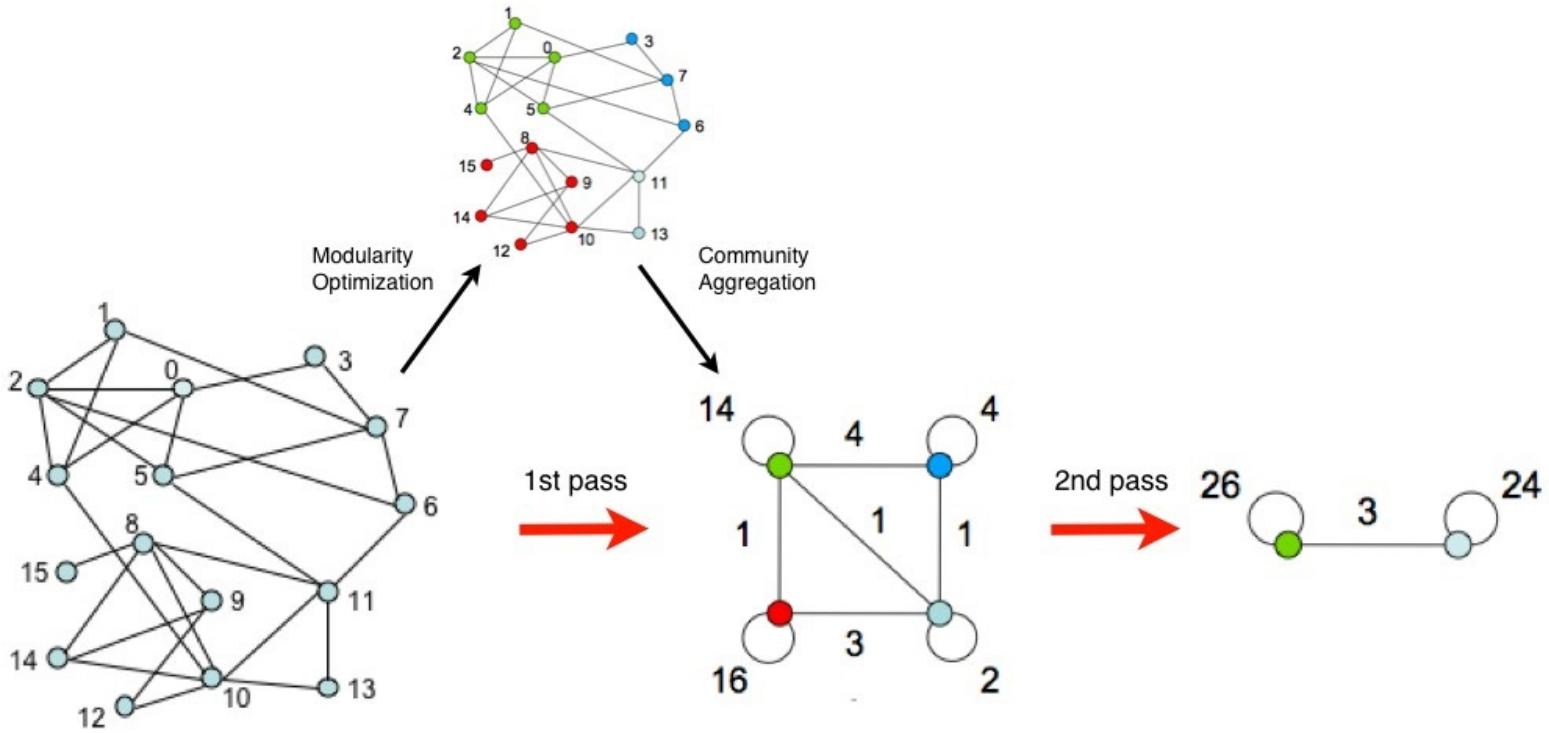
- La probabilitat de connectar i amb j és aleshores $p_i p_j = \frac{k_i k_j}{4m^2}$
- I, per tant, $P_{ij} = 2m p_i p_j = \frac{k_i k_j}{2m}$

Propietats de la modularitat

$$Q = \frac{1}{2m} \sum_{ij} (A_{ij} - \frac{k_i k_j}{2m}) \delta(C_i, C_j)$$

- Q només depèn dels nodes que es trobin en el mateix clúster.
- Una modularitat gran significa que les comunitats són millors (millors que la densitat *intra-cluster* del model aleatori).
- $Q \leq \frac{1}{2m} \sum_{ij} A_{ij} \delta(C_i, C_j) \leq \frac{1}{2m} \sum_{ij} A_{ij} \leq 1$
- Q pot ser negativa:
  - les particions amb valors molt negatius de Q impliquen l'existència de clústers amb una densitat d'arestes interna petita i moltes arestes entre comunitats.

El mètode de Louvain [Blondel et al., 2008], I



Pseudocodi:

1. Repetir fins aconseguir un valor local òptim:
   1. Fase 1: fer una partició (“greedy”) de la xarxa usant la modularitat.
   2. Fase 2: aglomerar els clústers trobats en nodes nous.

El mètode de Louvain, II

Fase 1: optimització de la modularitat

Pseudocodi per la fase 1:

1. Assignar una comunitat diferent a cada node.
2. Per cada node i:
   - Per cada veí j de i, considerar treure i de la seva comunitat i posar-lo a la comunitat de j.
   - De forma "greedy", escollir el posar a i a la comunitat del veí que aporti un guany més alt en modularitat.
3. Repetir fins que no hi hagi cap millora.

El mètode de Louvain, III

Fase 2: aglomeració dels clústers per formar un nova xarxa

Pseudocodi per la fase 2:

1. Fer que cada comunitat C_i passi a ser un node nou i.
2. Fer que les arestes entre els nodes i i j siguin la suma de les arestes entre els nodes de C_i i C_j en el graf previ (poden formar-se llaços).

El mètode de Louvain, IV

Observacions

- El resultat també és una jerarquia.
- Funciona per grafs amb pesos i, per tant, la modularitat s'ha de generalitzar a:

$$Q^w = \frac{1}{2W} \sum_{ij} (W_{ij} - \frac{s_i s_j}{2W}) \, \delta(C_i, C_j)$$

on W_{ij} és el pes de l'aresta no dirigida (i, j), $W = \sum_{ij} W_{ij}$ i $s_i = \sum_k W_{ik}$.

Índex

Referències, I

- Barabási, A.L. and Albert, R. (1999).
  Emergence of scaling in random networks.
  Science, 286(5439):509-512.
- Girvan, M. and Newman, M.E.J. (2002).
  Community structure in social and biological networks.
  Proceedings of the National Academy of Sciences of the United States of America, 99:7821-7826.
- Newman, M. (2010).
  Networks: An Introduction.
  Oxford University Press, USA, 2010 edition.
- Newman, M. E. (2003).
  The structure and function of complex networks.
  SIAM review, 45(2):167-256.

Referències, II

- Palla, G., Derényi, I., Farkas, I., and Vicsek, T. (2005).
  Uncovering the overlapping community structure of complex networks in nature and society.
  Nature, 435:814-818.
- Watts, D. J. and Strogatz, S. H. (1998).
  Collective dynamics of small-world networks.
  Nature, 393:440-442.