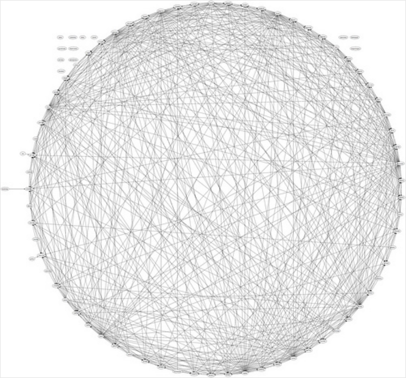
Apunts del Taller de Nous Usos de la Informàtica

Jordi Vitrià

Universitat de Barcelona

10 de setembre de 2019

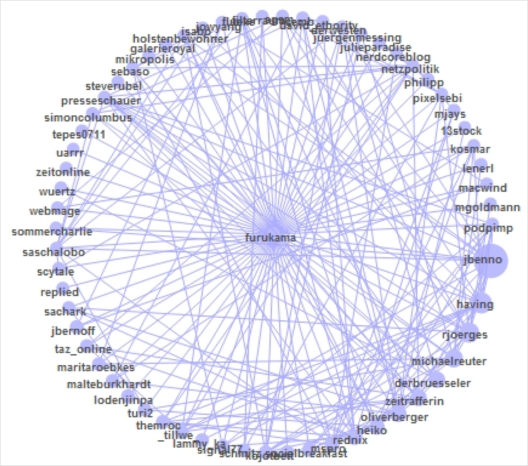
Lliçó 9: Anàlisi de Xarxes Socials



Perquè estudiar les xarxes socials?

- L'**anàlisi de xarxes socials** és una àrea d'investigació que estudia les relacions entre les persones des del punt de vista del graf implícit que formen les seves relacions.
- Una *xarxa social* és un conjunt d'actors vinculats entre si. Els actors poden ser persones o grups de persones (empreses, països, ciutats, etc.)
- Els vincles són qualsevol cosa que relacioni els actors (amor, poder, aliances, contactes del correu electrònic, etc).

Perquè estudiar les xarxes socials?



Perquè estudiar les xarxes socials?

- Les aplicacions de l'estudi de les xarxes socials són moltes. Per exemple:
  - L'estudi de la web i de les seves propietats. Fins i tot l'algorisme PageRank es pot veure com una aplicació de les xarxes socials.
  - L'estudi de la ciència com a disciplina (*scientometria*).
  - Sociologia, antropologia, ciències polítiques, gestió organizacional, mitjans de comunicació, etc.
  - Etc.

Com estudiar una xarxa social?

- El primer pas és recollir la informació i expressar-la en forma de *graf*.
- El segon pas és analitzar el graf per determinar les propietats de la xarxa social original.
- Les propietats de la xarxa social es poden estudiar des de quatre punts de vista diferents:
    - Estudiant les seves característiques generals (p.e. xarxes de món petit, longituds de camins, etc).
    - Estudiant característiques locals o les posicions dels actors (p.e. centralitat).
    - Estudiant els grups que té.
    - Visualitzant-la.

Com obtenim una xarxa social?

- La web es pot veure com una xarxa social formada per pàgines i links. Una de les mesures que veurem és la centralitat del node d'una xarxa social.
- Els llocs web que allotgen xarxes socials o 2.0: facebook, Wikipedia, Fotolog, Youtube, MyHeritage, Classmates, Twitter, etc.
- Altres registres digitals, o aquells llocs on deixem el nostre rastre digital, com e-mails, comerç electrònic, telefonia, etc.
- Enquestes.
- Simulació, o l'estudi de comunitats artificials. Exemple: la creació de comunitats socialment segregades a partir d'individus no racistes.

Estructura

- Les xarxes es representen com *grafs*, G, que és una entitat matemàtica formada per un conjunt no buit V d'elements que anomenem vèrtexs i una col·lecció E de parelles desordenades de vèrtexs que anomenem arestes o enllaços i que simbòlicament es representa com $G = (V, E)$.
- En general ens referim a un vèrtex particular pel seu ordre i en el conjunt V i dos vèrtexs, i i j, rebran el nom d'*adjacents* (o connectats) si existeix l'aresta (i, j) que els uneix.
- Un graf es pot representar per la seva corresponent *matriu d'adjacència* $A = \{a_{ij}\}$ que, en el cas més simple, serà una matriu simètrica de $V \times V$ elements.
- L'element a_{ij} de la matriu A prendrà un valor 1 si l'aresta (i, j) pertany a E, i 0 en qualsevol altra cas.

Estructura

- La matriu és simètrica, doncs si existeix una aresta que uneix i amb j, també existirà l'aresta que connecti j amb i. Llavors $a_{ij} = a_{ji}$.
- Les arestes d'una xarxa poden ser també *dirigides* (quan apunten una direcció determinada de l'enllaç entre els dos vèrtexs), *múltiples* (quan més d'una aresta connecta els mateixos nodes), *autoarestes* (quan surten de i arriben a un mateix node) i molts altres tipus diferents que es poden introduir a mida que la correcta definició d'una xarxa així ho requereixi.

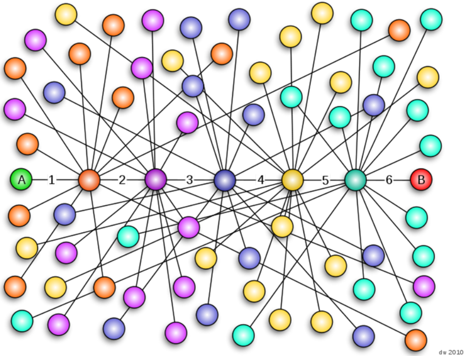
Xarxes de món petit.

- La teoria dels *sis graus de separació* és una teoria que intenta provar que qualsevol persona pot estar connectat a qualsevol altra persona del planeta a través d'una cadena de coneguts que no té més de cinc intermediaris (connectant a dues persones amb només sis enllaços).
- Segons aquesta teoria, cada persona coneix de mitjana, entre amics, familiars i companys de feina o escola, a unes 100 persones. Si cada un d'aquests amics o coneguts propers es relaciona amb altres 100 persones, qualsevol individu pot passar un encàrrec a 10.000 persones més només demanant a un amic que passi el missatge als seus amics.

Xarxes de món petit.

- En 1967, el psicòleg nord-americà Stanley Milgram va idear una nova manera de provar la teoria, que ell va anomenar "el problema del petit món". L'experiment del món petit de Milgram va consistir en la selecció a l'atzar de diverses persones del mig oest nord-americà, perquè enviessin targetes postals a un estrany situat a Massachusetts, situat a diversos milers de milles de distància. Els remitents coneixien el nom del destinatari, la seva ocupació i la localització aproximada. Se'ls va indicar que enviessin el paquet a una persona que ells coneguessin directament i que pensessin que fos la que més probabilitats tindria, de tots els seus amics, de conèixer directament al destinatari. Aquesta persona hauria de fer el mateix i així successivament fins que el paquet fos lliurat personalment al seu destinatari final. Encara que els participants esperaven que la cadena inclogués almenys centenars d'intermediaris, el lliurament de cada paquet només va portar, com a mitjana, entre cinc i set intermediaris.

Xarxes de món petit.



Xarxes de món petit.

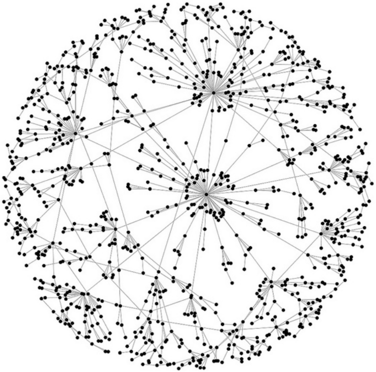
- Aquesta propietat s'ha detectat a moltes xarxes diferents i es considera un fenomen natural (que va des de la xarxa neural del cuc C.Elegans fins a les xarxes de distribució elèctrica).
- En termes generals, el fenòmen del *small world effect* es deineix com la tendència dels elements d'un sistema gran a estar connectats per un conjunt reduït de passos.
- L'anàlisi d'aquestes xarxes ens dóna informació sobre la xarxa analitzada, que té tendència a:
  - Estar formada per cliques o sub-xarxes que estan quasi totalment connectades.
  - Estar poblada per un nombre apreciable de *hubs* o nodes amb un gran nombre de connexions.
  - Ser molt robustes a la pèrdua de nodes.

Com detectem una xarxa de món petit?

- Una xarxa de món petit està caracteritzada pel fet que la distància típica L entre dos nodes escollits aleatòriament, creix proporcionalment amb el logaritme del nombre de nodes N de la xarxa:

$$L \propto log N$$

Xarxa de món petit: exemple



Longitud de camí característica.

- Una de les mesures globals tradicionalment més utilitzades per caracteritzar una xarxa ha estat la de la *longitud de camí característica* L.
- Aquest concepte deriva del de *camí d'una xarxa*, que es defineix com la seqüència de nodes que, units per arestes, porten del node i al node j. La longitud de camí característica es defineix com la mitja dels *camins més curts* calculada per totes les parelles de nodes de la xarxa:

$$L(G) = \frac{1}{N(N-1)} \sum_{i \neq j \in N} l_{ij} = \frac{2}{N(N-1)} \sum_{i > j} l_{ij}$$

on el camí més curt l_{ij}, en termes d'arestes utilitzades per anar del node i al node j, rep també el nom de *camí geodèsic*.

Longitud de camí característica.

- Per a grafs que tenen més d'una component (xarxes no connexes), la definició de L és problemàtica perquè ens podem trobar amb parelles de nodes desconnectats pels quals l hagi de prendre un valor infinit.
- Per evitar aquest resultat, s'ha proposat el càlcul de la *mitjana harmònica* o *eficiència global* en la que els valors infinits de l_{ij} no contribueixen a la suma:

$$E_{glob}(G) = \frac{1}{N(N-1)} \sum_{i \neq j \in N} \frac{1}{l_{ij}} = \frac{2}{N(N-1)} \sum_{i>j} \frac{1}{l_{ij}}$$

- L'eficiència global es correlaciona amb $1/L$, doncs una longitud de camí característica elevada correspondrà a una baixa eficiència.

Grau d'un node.

- Una mesura local útil per caracteritzar una xarxa és la de *grau* d'un node, definit pel nombre d'arestes que s'hi connecten i que enllacen, a la vegada, amb els nodes primers veïns. De forma matemàtica, el grau k_i d'un node i és

$$k_i = \sum_{j \in N} a_{ij}$$

- A partir d'ella podem calcular la distribució estadística del grau en una xarxa, això és la *distribució de grau*. Si definim p_k com la fracció de nodes d'una xarxa G que tenen grau k, p_k serà, al mateix temps, la probabilitat que un node escollit aleatòriament tingui un grau k.

Grau d'un node.

- Per facilitar la interpretació de les distribucions de grau, una alternativa pot ser presentar-les mitjançant la funció de distribució acumulada com

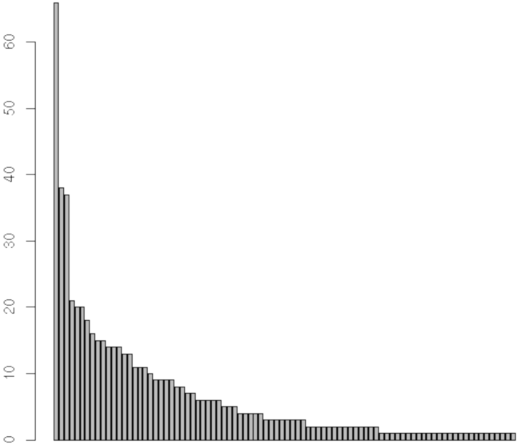
$$P(k) = \sum_{k'=k}^{\infty} p_{k'}$$

que és la probabilitat que el grau d'un node sigui major o igual que k.

Grau d'un node.

- La importància del coneixement de la distribució de grau d'una xarxa es comença a posar de relleu amb els treballs sobre xarxes de món petit.
- Aquesta investigació destaca un aspecte molt característic de nombroses xarxes: la presència d'una distribució de grau en forma de funció potencial (o functions de cua gruixuda), molt lluny de qualsevol funció de variable aleatòria com seria d'esperar a priori.
- La funció potencial, que caracteritza $P(k)$ com $P(k) = k^{-\lambda}$, implica, per una banda, que els vertex molt connectats (k elevat) tinguin mes probabilitat d'apareixer que en el cas de xarxa aleatòria, dominant la connectivitat global de la xarxa i, per una altra, que la distribució de grau sigui monòtonament decreixent, caracteritzant-se per un únic exponent λ i no tenint, en conseqüència, una escala definida.

Grau d'un node



Grau d'un node.

- Després de comprovar com la distribució de grau decau com a llei potencial per l'estructura d'Internet, les xarxes de reaccions metabòliques, els grafs creats per les trucades telefòniques i la World-Wide Web entre d'altres (sorprenentment, amb un mateix exponent λ per a totes elles entre 2 i 3), s'ha batejat aquest tipus de xarxes amb el nom de *xarxes lliures d'escala* (an. scale-free networks) pel fet que no existeix en elles un node caracteristic que en permeti definir una escala de connectivitat.
- Donada una xarxa que volem analitzar, direm que és una xarxa de món petit si la seva distribució de grau es pot aproximar prou bé amb una funció potencial.

Caracerització de xarxes.

- A part de la caracterització global d'una xarxa, és interessant caracteritzar nodes i grups de nodes.
- Més concretament, és interessant caracteritzar el rol dels nodes a la xarxa (per la seva funció d'elements que poden controlar o influir la informació que corre per la xarxa) o els grups de nodes de la xarxa.

Mitjania.

- L'anomenada mitjania (*betweenness*) d'un node sorgeix al comprovar com la relació existent entre dos nodes no adjacents pot dependre d'un tercer, especialment si aquest es troba en el camí d'unió dels dos primers.
- En conseqüència, els nodes existents en els camins d'unió entre d'altres, poden exercir una influència notable i un control estratègic sobre la resta.

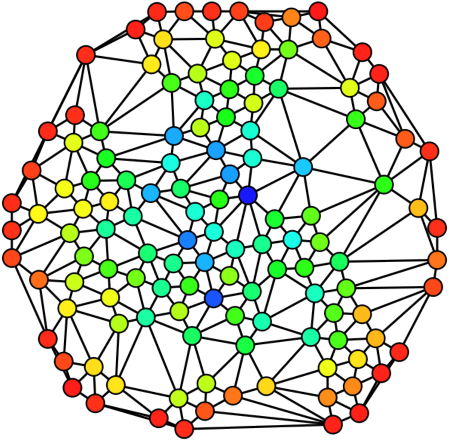
Mitjania.

- Assumint que un element és important si es troba enmig de la resta d'elements i que la informació (o qualsevol altre producte) viatja tan sols a través de camins geodèsics, si n_{jk} és el nombre de camins geodèsics que uneixen els dos vèrtexs j i k, i $n_{jk}(i)$ és el nombre de camins geodèsics que uneixen els vèrtexs j i k que contenen el vèrtex i, la mitjania del vèrtex i es pot definir com

$$C_i^B = \frac{1}{(N-1)(N-2)} \sum_{j,k \in N} \frac{n_{jk}(i)}{n_{jk}}$$

on N és el nombre de nodes de la xarxa (utilitzats a l'equació anterior a efectes normalitzadors). C^B pren valors entre 0 i 1, essent igual a 1 en el cas que el vèrtex i estigui present en tots els camins geodèsics del graf.

Exemple



Mitjania.

- Calcular la mitjania de tots els nodes implica calcular tots els camins mínims de totes les parelles del graf.
- Això té un cost $O(|V|^3)$ si fem servir l'algorisme de Floyd-Marshall, convenientment modificat per contar no només un dels camins geodèsics, sinó tots els camins geodèsics entre dos nodes.
- Si el graf és *sparse*, l'algorisme de Johnson baixa aquesta complexitat fins a $O(|V|^2 log|V| + |V||E|)$.

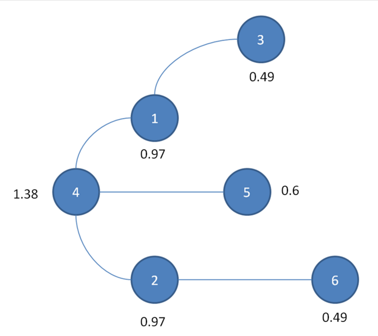
Mesures de centralitat.

- Les mesures de centralitat mesuren la posició d'un actor en una xarxa segons uns certs criteris.
- La *centralitat de grau* d'un actor és la més simple de les mesures i es calcula com el nombre dels veins que té.
- Hi ha gent, la posició favorable de la qual a la xarxa els hi permet ser els iniciadors de *processos d'influència* com la transmissió de creences, publicitat viral, etc. En aquest cas, el procés comença en un actor i es va transmetent successivament. Una mesura de centralitat que mesuri això ha d'estar basada en contar camins.

Mesures de centralitat.

- Les xarxes amb cicles són problemàtiques per contar camins (en tenen infinits!).
- La solució pràctica és *atenuar* els camins usant una tassa de descompte: els camins més llargs es conten com valors més petits i els infinits zero.
- Tenint en compte això, definim la centralitat $c(\beta)$ de *Bonacich* com $c(\beta) = (\sum_{k \geq 1} \beta^k A^k) \cdot \vec{1}$ on β és la tassa de descompte i A^k conta els camins de longitud k entre qualsevol parella d'actors.
    - Recordeu que per contar el nombre de camins de longitud n que van d'un node i a un altre node j només cal calcular A^n i el valor és $A^n(i,j)$.
    - Per tant, un node que pot arribar amb camins curts a la majoria dels altres tindrà un valor alt segons aquesta mesura.

Exemple



Al graf es veu la centralitat per cada actor, avaluat amb $\beta = 0.5$.

Mesures de centralitat.

- La centralitat també es pot definir de manera recursiva: un actor central és aquell que té un veïnatge de bona centralitat.
- Aquest concepte ens porta al model del PageRank!

Cerca amb ranking basat en els links

- L'algorisme **PAGERANK**, inventat pels fundadors de Google, calcula un *score* per cada pàgina web indicant la seva 'importància'.
- La 'importància' es calcula a partir de la importància de les pàgines que l'apunten i del nombre total de links que surten d'elles.

L'algorisme PAGERANK

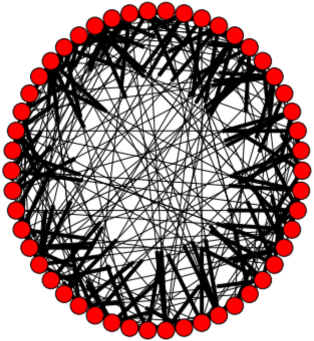
- Els `out-links` de la pàgina *i* són els hyperlinks que apunten des de la pàgina *i* a una altra pàgina. Els hyperlinks del mateix *site* no es consideren.
- Els `in-links` de la pàgina *i* són els hyperlinks que apunten des d'altres pàgines a la pàgina *i*. Els hyperlinks del mateix *site* no es consideren.
- Un hyperlink d'una pàgina apuntant a una altra pàgina s'entén com una **concessió implícita d'autoritat** a la pàgina receptora. Com més `in-links` té una pàgina *i*, més prestigi té la pàgina *i*.
- Les pàgines que apunten a la pàgina *i* també tenen el seu prestigi. Una pàgina amb prestigi alt apuntant a *i* és més important que una pàgina amb un prestigi menor apuntant a *i*. En altres paraules, una pàgina és important si és apuntada per pàgines importants.

L'algorisme PAGERANK

- La **concessió implícita d'autoritat**, que s'ha de calcular "recursivament", determinarà un *PageRank Score* per cada pàgina.
- Segons el model d'ordenació per prestigi en xarxes socials, la importància d'una pàgina i (o *PageRank Score*) està determinat per la suma dels *PageRank Scores* de les pàgines que apunten a i.
- Com que una pàgina pot apuntar a moltes altres pàgines, el seu prestigi s'ha de compartir entre les pàgines a les que apunta.

L'algorisme PAGERANK

Per formalitzar les idees de PageRank, considerem la web com un graf dirigit $G = (V, E)$, on V és el conjunt de n nodes (o sigui, el conjunt de pàgines web) i E el conjunt d'arcs dirigits (o sigui, els hyperlinks).



L'algorisme PAGERANK

- Sigui $n = |V|$ el nombre total de pàgines web. Llavors, el *PageRank Score* de la pàgina i es defineix com:

$$P(i) = \sum_{(j,i) \in E} \frac{P(j)}{O_j} \tag{1}$$

  on $O(j)$ és el nombre de `out-links` de la pàgina j.
- Matemàticament, això és un sistema de n equacions lineals amb n incògnites.

L'algorisme PAGERANK

- Podem usar una matriu per representar les equacions.
- Sigui p un vector columna n dimensional de *PageRank Scores*,

$$P = (P(1), P(2), \ldots, P(n))^{\mathrm{T}} \tag{2}$$

- Sigui A la matriu d'adjacència del graf:

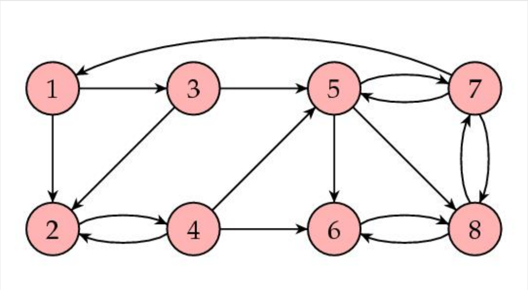
$$A_{ij} = \begin{cases} \frac{1}{O_i} & \text{if}(i,j) \in E \\ 0 & \text{Otherwise} \end{cases} \tag{3}$$

- Llavors podem escriure el sistema de n equacions com $\mathbf{P} = \mathbf{A}^{\mathrm{T}}\mathbf{P}$.

L'algorisme PAGERANK

- La matriu **A** té un parell de característiques importants:
  - Totes les seves entrades són no negatives: $A_{ij} \geq 0$.
  - La suma de les entrades d'una columna són 1 excepte en el cas que la pàgina corresponent a aquella columna no contingui cap link: $\sum_j A_{ij} = 1$.
- El vector **P** s'anomena el *vector estacionari* de **A** i correspon a l'eigenvector de **A** amb eigenvalor 1.

Exemple

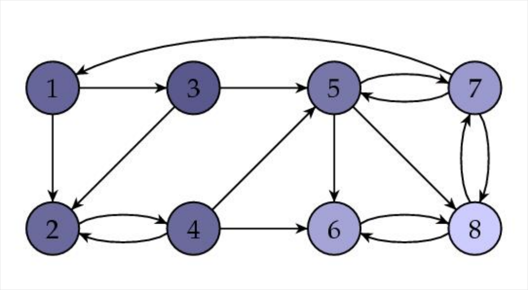


- La matriu és:

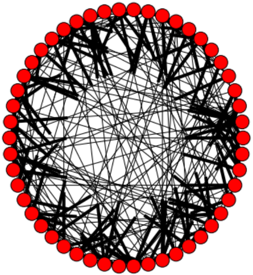
$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1/3 & 0 \\ 1/2 & 0 & 1/2 & 1/3 & 0 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/2 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 1/3 & 1/3 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 1/3 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 1/3 & 1 & 1/3 & 0 \end{bmatrix}$$

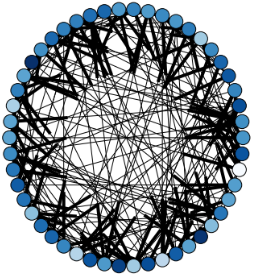
- El vector de probabilitats estacionàries és:

$$(0.06, 0.0675, 0.03, 0.0675, 0.0975, 0.02025, 0.18, 0.295)$$



- La pàgina 8 és la pàgina més important.





L'algorisme PAGERANK

- Aquesta formulació es pot interpretar amb el model conegut com el **random surfer**: un usuari que navega seguint links de forma aleatòria i ininterrompuda per la web té una probabilitat estacionària $P(i)$ d'estar en un determinat moment a la pàgina i.
- Com podem calcular la distribució $P(i)$?
- Cal tenir en compte que la matriu $\mathbf{A}$ té milers de mil·lions de files i columnes, però és *sparse* atès que les pàgines web tenen una mitja de 10 links de sortida.
- Per aquest tipus de matrius és adequat usar un mètode iteratiu de càlcul del vector estacionari conegut amb el nom del **mètode de la potència**.

L'algorisme PAGERANK

Algorithm 1 PageRank

for $i = 1$ to n **do**
  $P(i,0) = 1/n$
end for
$j \to 1$
while no hi hagi convergència **do**
  $P(i,j) = \sum_{(k,i)\in E} \frac{P(k,j-1)}{O_k}$
  $j \to j+1$
end while
return $P(i,j)$

L'algorisme PAGERANK

- El vector $P(i,0)$ es pot inicialitzar de diverses maneres.
- Per l'exemple anterior, si inicialitzem amb $P(1,0) = 1$ i la resta a 0, les iteracions donarien:

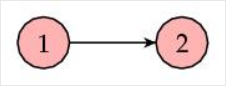
P(i,1)=	(1	0	0	0	0.0278	...	0.06	0.06)
P(i,2)=	(0	0.5	0.25	0.1667	0.0833	...	0.0675	0.0675)
P(i,3)=	(0	0.5	0	0	0	...	0.03	0.03)
P(i,4)=	(0	0	0.5	0.25	0.1667	...	0.0675	0.0675)
P(i,5)=	(0	0	0.25	0.1667	0.1111	...	0.0975	0.0975)
P(i,6)=	(0	0	0	0.25	0.1806	...	0.2025	0.2025)
P(i,7)=	(0	0	0	0.0833	0.0972	...	0.18	0.18)
P(i,8)=	(0	0	0	0.0833	0.3333	...	0.295	0.295)

L'algorisme PAGERANK

- Considerant la forma de resoldre el problema, ens podem fer dues preguntes rellevants:
  1. Convergeix sempre l'algorisme?
  2. És independent la solució final dels valors inicials?
- I la resposta a les dues preguntes és no, tot i que una petita modificació de l'algorisme les fa positives.
- La xarxa formada per dos nodes i un únic link és un exemple simple on no funciona.

L'algorisme PAGERANK

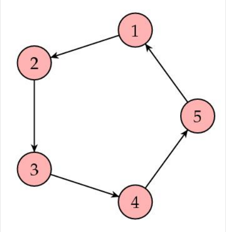
- En aquest cas l'algorisme convergeix al vector $(0, 0)$, que no té sentit.



- El problema ve causat per l'existència de nodes sense links.

L'algorisme PAGERANK

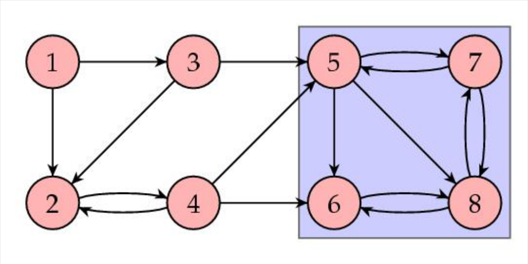
- En aquest altre cas l'algorisme tampoc no convergeix i va generant una seqüència infinita: $(1, 0, 0, 0, 0)$, $(0, 1, 0, 0, 0)$,$(0, 0, 1, 0, 0)$, $(0, 0, 0, 1, 0)$,$(0, 0, 0, 0, 1)$, $(1, 0, 0, 0, 0)$, ..., que no té sentit.



- El problema és que el segon eigenvector d'aquesta matriu és ≥ 1.0 i sota aquestes condicions el mètode la potència no funciona.

L'algorisme PAGERANK

- En aquest altre cas l'algorisme convergeix al vector $(0, 0, 0, 0, 0.12, 0.24, 0.24, 0.4)$, que no té sentit.



- El problema ve causat per l'existència d'un grup de nodes sense links de sortida.

L'algorisme PAGERANK

- Matemàticament, podem assegurar l'existència d'un vector propi estacionari amb eigenvalor 1 si la matriu és *estocàstica* i *primitiva*.
- Una matriu no negativa $\mathbf{A}$ tal que la suma de les entrades d'una columna són 1 s'anomena *matriu estocàstica*.
- Una matriu $\mathbf{A}$ és *primitiva* si per algun m, $\mathbf{A}^m$ té totes les seves entrades positives. En altres paraules, si donades dues pàgines qualsevol, és possible arribar d'una pàgina a l'altre després de seguir m links.

L'algorisme PAGERANK

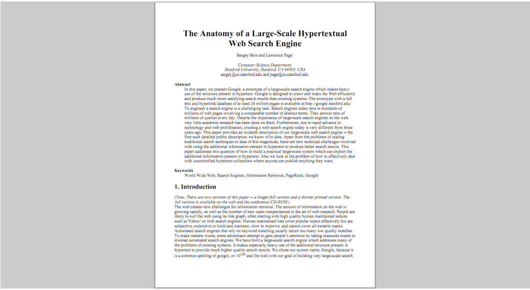
- El model considerat per l'algorisme es pot modificar si considerem que l'usuari, amb una probabilitat determinada $1 - d$, pot decidir no seguir cap link de la pàgina on està i seleccionar la següent pàgina de forma aleatòria d'entre les N possibles, i també considerar que quan arribem a una pàgina sense out-links, saltem a la següent pàgina de forma aleatòria d'entre les N possibles.
- Llavors la probabilitat de visitar una pàgina depèn tant de la probabilitat d'arribar-hi aleatòriament com de la d'arribar-hi pels links:

$$P(i) = \frac{1-d}{N} + d \sum_{(j,i)\in E} \frac{P(j)}{O_j} \tag{4}$$

- Amb aquesta petita modificació es pot demostrar que l'algorisme convergeix sempre a la solució desitjada atès que tenim una matriu estocàstica i primitiva.

Exemple

- La probabilitat d s'ha estimat experimentalment i s'ha fixat a $d = 0.85$.



The Anatomy of a Large-Scale Hypertextual Web Search Engine

Sergey Brin and Lawrence Page

Computer Science Department,
Stanford University, Stanford, CA 94305, USA
sergey@cs.stanford.edu and page@cs.stanford.edu

Abstract

In this paper, we present Google, a prototype of a large-scale search engine which makes heavy use of the structure present in hypertext. Google is designed to crawl and index the Web efficiently and produce much more satisfying search results than existing systems. The prototype with a full text and hyperlink database of at least 24 million pages is available at http://google.stanford.edu/ To engineer a search engine is a challenging task. Search engines index tens to hundreds of millions of web pages involving a comparable number of distinct terms. They answer tens of millions of queries every day. Despite the importance of large-scale search engines on the web, very little academic research has been done on them. Furthermore, due to rapid advance in technology and web proliferation, creating a web search engine today is very different from three years ago. This paper provides an in-depth description of our large-scale web search engine -- the first such detailed public description we know of to date. Apart from the problems of scaling traditional search techniques to data of this magnitude, there are new technical challenges involved with using the additional information present in hypertext to produce better search results. This paper addresses this question of how to build a practical large-scale system which can exploit the additional information present in hypertext. Also we look at the problem of how to effectively deal with uncontrolled hypertext collections where anyone can publish anything they want.

Keywords

World Wide Web, Search Engines, Information Retrieval, PageRank, Google

1. Introduction

(Note: There are two versions of this paper -- a longer full version and a shorter printed version. The full version is available on the web and the conference CD-ROM.)
The web creates new challenges for information retrieval. The amount of information on the web is growing rapidly, as well as the number of new users inexperienced in the art of web research. People are likely to surf the web using its link graph, often starting with high quality human maintained indices such as Yahoo! or with search engines. Human maintained lists cover popular topics effectively but are subjective, expensive to build and maintain, slow to improve, and cannot cover all esoteric topics. Automated search engines that rely on keyword matching usually return too many low quality matches. To make matters worse, some advertisers attempt to gain people's attention by taking measures meant to mislead automated search engines. We have built a large-scale search engine which addresses many of the problems of existing systems. It makes especially heavy use of the additional structure present in hypertext to provide much higher quality search results. We chose our system name, Google, because it is a common spelling of googol, or 10^{100} and fits well with our goal of building very large-scale search

http://infolab.stanford.edu/pub/papers/google.pdf

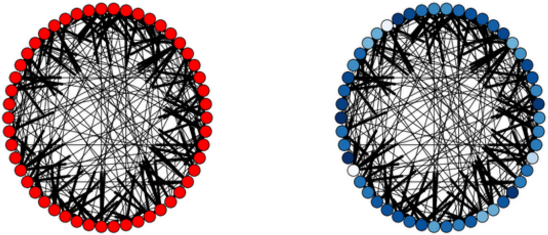
Exemple

In [6]:
```
import networkx as nx
import numpy
plt.figure(figsize=(8,8))

g=nx.fast_gnp_random_graph(50, 0.08, seed=None, directed=True)
pos=nx.circular_layout(g)

L = nx.to_numpy_matrix(g)

d = 0.85

# Nombre de files de la matriu
N = len(L)

# Normalitzem la matriu L per convertir-la en una matriu de trancisions.
for row in L:
    row /= sum(row)

# Apliquem el model de regularització de les trancisions
L = numpy.nan_to_num(L) * d + (1-d) / N

# Podem inicialitzar el vetor estacionari a qualsevol conjunt de
# valors que sumin 1
p = numpy.zeros(N)
p[0] = 1.0

# Apliquem el mètode d'iteració de potències per calcular el vector
# estacionari (per defecte farem 100 iteracions)
for i in range(100):
    p = numpy.dot (p, L)
pl = np.array(p)
print pl
```

```
[[ 4.31802221e-04   8.58899349e-05   3.89690164e-04   1.81102788e-04
   1.48325364e-04   4.52728077e-04   5.89916789e-04   1.51084993e-04
   2.42192944e-04   2.05064372e-04   1.82309714e-04   2.65188922e-04
   4.32189192e-04   4.00045187e-04   2.95267796e-04   1.49659695e-04
   8.75407111e-05   1.04445746e-03   5.68217310e-04   6.04213609e-04
   2.65904240e-04   4.28823894e-04   1.40676733e-04   2.37422218e-04
   1.12328014e-04   5.06665320e-05   3.01336319e-04   2.93855284e-04
   5.06665320e-05   1.09092707e-03   3.08588040e-04   3.31178409e-04
   2.12085916e-04   1.84141821e-04   1.86396283e-04   2.39771027e-04
   1.86812333e-04   5.21866215e-04   2.59104696e-04   3.62945457e-04
   2.13270262e-04   5.98956907e-04   5.63786620e-04   2.06809505e-04
   7.10487555e-05   3.54359373e-04   3.04073224e-04   8.10532856e-04
   3.53587400e-04   3.64659779e-04]]
```

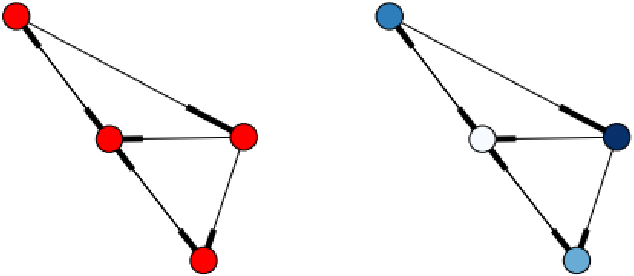
In [7]:
```
plt.figure(figsize=(16,8))

subplot(121)
plt.axis('off')
nx.draw_networkx_edges(g,pos)
nx.draw_networkx_nodes(g,pos,cmap=plt.cm.Reds_r)

subplot(122)
plt.axis('off')
nx.draw_networkx_edges(g,pos)
nx.draw_networkx_nodes(g,pos,node_color=pl[0],cmap=plt.cm.Blues_r)
```

Out[7]:
```
<matplotlib.collections.PathCollection at 0x495c630>
```



Exemple

$$(0.26232, 0.40279, 0.20868, 0.12619)$$



Mesures de centralitat.

- Una altra mesura de posició surt de considerar la distància mitja de cada node al resta de la xarxa: l'actor que està més a prop de tot altre element de la xarxa és el més central.
- Aquesta mesura és diu *centralitat de rodalies*.
- Matemàticament, es pot expressar com la inversa de la suma de les distàncies $c_i = \frac{1}{\sum_j d_{ij}}$ on d_{ij} és la distància entre l'actor i i l'actor j.

Mesures de centralitat.

- Els *actors claus* són els que si es desconnecten de la xarxa poden interrompre els fluxes que corren per ella.
- En aquest cas es pot calcular l'anomenada *centralitat d'intermediació* com el nombre de rutes mínimes en les que l'actor participa, que ja hem vist amb el nom de grau de mitjania.

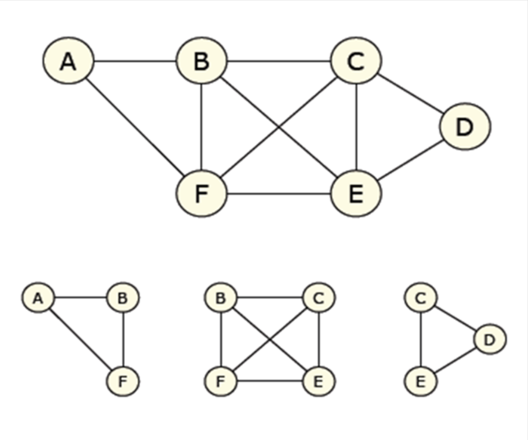
Estudi dels grups d'actors

- La detecció de comunitats, grups, cliques (grups exclusius), és un tema d'interès a les xarxes socials.
- Podem usar tècniques com el k-means o l'agrupació jeràrquica, sempre que tinguem clar sobre quines característiques.
- Si volem detectar grups usant teoria de grafs, hi ha tècniques el mètode dels *graph-cuts* que ens ho permeten.
- Aquesta tècnica adapta al concepte de graf la idea de minimitzar la distància entre els membres d'un grup i maximitzar la distància a la resta.
- La seva formulació resulta en un problema de càlcul dels valors i vectors propis de la matriu d'adjacència del graf.

Estudi dels grups d'actors

- Una *clique* és una agrupació social en la que un coneix a tots els altres.
- Una *clique maximal* és una clique que no és subconjunt de cap altra clique del graf.
- Una clique de mida igual o superior a qualsevol altra clique d'un graf s'anomena *clique maximal*.

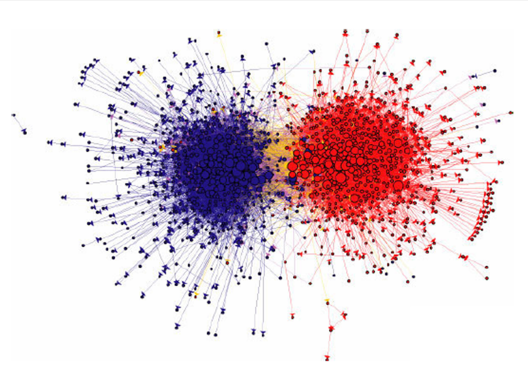
Exemple



Estudi dels grups d'actors

- Hi ha moltes definicions possibles per definir una agrupació d'actors:
  - Un grup de nodes que tenen característiques comunes o tenen un rol similar al graf.
  - Un subconjunt de nodes dins dels quals les connexions node-a-node són denses i en canvi les connexions a nodes d'altres comunitats són menys denses.

Exemple



Xarxes Face 2 Face

Xarxes Face 2 Face

- Aquest aparell permet enregistrar veu, moviment i proximitat.
- A partir de la veu i el moviment establim xarxes socials amb etiquetes a les arestes que corresponen a senyals socials establerts entre persones.
- Es pot analitzar la dinàmica d'aquestes xarxes.

Xarxes Face 2 Face

- Amb diversos estudis experimentals s'ha demostrat que és possible identificar:
    - Connectors i persones centrals en una xarxa social professional.
    - El cap d'un organitzacio.
    - Els líders d'equips.
    - El resultat d'una negociació.
    - El grau de persuasió d'una presentació.
    - Les afiliacions grupals.