UNIVERSITAT DE BARCELONA

Facultat de Matemàtiques
i Informàtica

GRAU DE MATEMÀTIQUES

Treball final de grau

Grafs aleatoris i la Web

Autor: Jordi Artús Suárez

Director: Dr. Jordi Marzo

Realitzat a: Departament de Matemàtiques i Informàtica

Barcelona, 20 de juny de 2019

Abstract

In this end-of-grade project, we will study the random graphs defined by Erdös and Rényi, with the objective of using them to generate models that help us to study the behavior of the Web. We will also do a brief analysis on another model called LCD PA. To carry out this work, we will review both the theory of graphs and probabilities, as well as describing the properties and characteristics of the graph W, which represents the Web.

Resum

En aquest treball de final de grau estudiarem els grafs aleatoris que van definir Erdös i Rényi, amb l'objectiu d'utilitzar-los per generar models que ens serveixin per estudiar el comportament de la Web. També farem un breu anàlisi a un altre model anomenat LCD PA. Per dur a terme aquest treball, farem un repàs tant de la teoria de grafs com de probabilitats, a més a més de descriure les propietats i característiques del graf W, el qual representa la Web.

2010 Mathematics Subject Classification. 68R10, 94C15, 05C80, 05C82, 60B99

Agraïments

Vull agrair al meu tutor per haver acceptat dirigir el meu treball i haver dedicat el seu temps en ajudar-me i guiar-me per dur-lo a terme. També fer menció a la meva família i amics per donar-me suport durant a tot el grau de Matemàtiques. No ha estat un camí fàcil, però sí satisfactori.

All models are wrong, but some models are useful.

-G.E.P. Box

Índex

1 Introducció

Actualment vivim en una societat on la tecnologia està a tot arreu. És difícil imaginar-se un dia a dia sense l'ús de l'ordinador, el telèfon mòbil o la televisió. Aquest fet ha traspassat fronteres fins arribar al punt de vendre automòbils amb servei de wifi. El món digital ens envolta arreu on es vagi, i és normal ja que aporta multitud d'avantatges i comoditats als usuaris. La immediatesa que proporciona internet és fonamental en una societat cada vegada més globalitzada, on ara mateix dos persones que viuen en continents separats poden treballar juntes gairebé com si estiguessin una al costat de l'altra.

La invenció d'internet ha estat una de les grans revolucions per entrar al món de la globalització. Però internet és un terme força general que engloba diferents elements com: diverses components físiques i hardware, la Web, el correu electrònic o els jocs on-line.

En aquest treball de final de grau volem introduir-nos en la modelització matemàtica de la Web, la qual és usual confondre-la amb el propi internet. La Web, o *World Wide Web* (WWW), consisteix en informació emmagatzemada i disponible a internet, en llocs que anomenarem pàgines web, o simplement pàgines. La majoria d'aquestes són documents en *Hypertext Markup Language (HTML)* identificats amb una cadena de caràcters anomenada *Uniform Resource Locators (URLs)*. Els documents HTML estan units entre ells pels *links (o hyperlinks)*, que relacionen les diferents pàgines creant així una xarxa, informació que ja hauríem intuït si ens haguéssim quedat amb el significat literal de la paraula *web*.

Per realitzar aquest estudi utilitzarem la teoria de grafs, que és una branca de les matemàtiques i la informàtica que es dedica a l'estudi dels grafs, estructures matemàtiques utilitzades per a modelitzar relacions entre parelles d'objectes. Nosaltres identificarem aquest objectes com les pàgines que hem descrit anteriorment. Ens centrarem en un graf, que anomenarem graf web W, ens introduirem en la seva modelització per estudiar la seva evolució i el seu comportament.Per dur a terme aquesta modelització utilitzarem grafs aleatòris.

Existeixen molts altres exemples de xarxes que són diferents a la Web i que també poden ser estudiats mitjançant grafs. Un dels exemples podria ser la *xarxa de regulació genética (genetic network)*, que és una col·lecció de segments de ADN dins una cèl·lula que interactuen entre si i amb altres substàncies de la pròpia cèl·lula, amb el que regulen les taxes a les que els gens de la xarxa transcriuen l'ARNm. Una altra possible xarxa a és la que descriu el número d'Erdös, el qual és un nombre natural que defineix la *distància de col·laboració* de qualsevol matemàtic respecte a Erdös pel que fa a articles publicats. Cada una de les xarxes que poguem descriure compliran una sèrie de propietats i comportaments que s'hauran pogut descriure mitjançant proves empíriques. Per exemple, poden seguir una evolució que marqui una *llei de potència* a la distribució dels seus vèrtexs on la potència β canvia entre les diferents xarxes, o directament pot no tenir aquest comportament. Més endavant descriurem detalladament a que ens referim amb llei de potència.

Hi ha moltes preguntes interessants que ens podem fer sobre l'estructura del graf Web. Alguns exemples són: quina és la mitja d'enllaços que té una pàgina arbitrària, quina és la distància mitjana entre dos pàgines qualssevol, o quina és la probabilitat, de que una pàgina escollida a l'atzar tingui exactament k enllaços. Aquests temes els tractarem més detalladament a mesura que anem definint conceptes.

Dins d'aquest treball explicarem la modelització de W amb l'ajuda dels grafs aleatoris.

Hi ha diferents mètodes per a generar grafs aleatoris, en veurem alguns més endavant. Però dintre d'aquesta diversitat de maneres, ens concentrarem en el mètode per generar-los que van establir els matemàtics Paul Erdös i Alfréd Rényi. Veurem les propietats que proporciona aquest model al graf Web i si genera una aproximació prou acurada a la realitat o podria haver millors mètodes per estudiar el comportament de W.

Finalment, donarem una descripció d'un altre model per estudiar el comportament i l'evolució de la web, diferent al dels grafs aleatòris d'Erdös i Rényi. Realitzarem un anàlisi descriptiu d'aquest nou model per tenir una prespectiva més àmplia i obtenir un altre punt de vista sobre la modelització del graf W. Així podrem realitzar un contrast de resultats dels dos models i comparar les propietats que compleixen cada un. Per últim, remarcar que la nostra referència principal serà el llibre [3].

2 Conceptes bàsics i notació

2.1 Teoria de Grafs

En aquest treball estudiarem i utilitzarem els grafs per poder modelitzar el que avui en dia anomenem la *Web* (*WWW = World Wide Web*). Per poder-ho fer introduirem ara alguns conceptes i notacions que farem servir al llarg d'aquest estudi.

2.1.1 Notació per a les funcions

Començarem aclarint alguna notació respecte el tracte amb les funcions. Molts resultats que veurem seran asimptòtics. Siguin f i g funcions i el seu domini algun subconjunt fixat de $\mathbb{R}$. Escriurem que $f \in O(g)$ si

$$\limsup_{x\to\infty} \frac{f(x)}{g(x)}$$

existeix i és finit. Això significa que $\exists c > 0$ constant i independent de x i $\exists x_0 > 0$ tal que per $x > x_0$, $f(x) \leq cg(x)$. També ho escriurem $f = O(g)$. D'aquí treiem dos notacions més: $f = \Omega(g)$ si $g = O(f)$; i $f = \Theta(g)$ si $f = O(g)$ i $f = \Omega(g)$.

Escriurem que $f = o(g)$ si $\lim_{x\to\infty} \frac{f(x)}{g(x)} = 0$. Així $f = o(1)$ significa que f tendeix a 0.

2.1.2 Notació de grafs

Ara introduirem els conceptes i notacions principals que utilitzarem de la teoria de grafs. Un *graf* G és un conjunt no buit de *vèrtexs* $V(G)$, i un conjunt d'*arestes* $E(G)$, les quals uneixen dos vèrtexs. Sovint representarem una aresta entre els vèrtexs $u, v \in V(G)$ com $uv \in E(G)$. Si l'aresta uneix el vèrtex amb ell mateix, es diu que el graf té un *loop* (bucle). Normalment els cardinals $n = |V(G)|$ com l'ordre de G i $m = |E(G)|$ a la mida. Definirem el *grau* d'un vèrtex u com el número d'arestes que el connecten amb els altres, que escriurem $deg(u)$. D'aquests conceptes que acabem de definir, podem obtenir el següent teorema:

Teorema 2.1. *Si G és un graf, aleshores*

$$2|E(G)| = \sum_{u\in V(G)} deg(u).$$

Demostració: Si sumem els graus de tots els vèrtexs, estem comptant dos cops cada aresta.

□

Recordem el lema de l'encaixada de mans que va demostrar Euler, que es prova directament del teorema anterior.

Lema 2.2. *Si G és un graf, aleshores hi ha un nombre parell de vèrtexs amb grau senar.*

Demostració: Dividim els vèrtexs entre els que tenen grau parell P i els que tenen grau senar S. Llavors el teorema anterior diu:

$$\sum_{v\in P} deg(v) + \sum_{u\in S} deg(u) = 2|E(G)|.$$

Per tant, aquesta suma ha de ser parell. El primer sumand és suma de números parells, llavors es parell. El segon sumand és una suma de nombres imparells, i per a què es compleixi la igualtat, aquest subconjunt ha de tenir un cardinal parell.

□

Definició 2.3. *Sigui G un graf, direm que G' és un subgraf de G si $V(G') \subseteq V(G)$ i $E(G') \subseteq E(G)$.*

Entendrem com un *passeig* (*walk*) una successió que alterna entre vèrtexs i arestes

$$x_0, e_1, x_1, ..., e_t, x_t$$

on $\forall i \in [1, t]$, $e_i = x_{i-1}x_i$. Els vèrtexs i arestes es poden repetir. Si $x_0 = x_t$ direm que és *tancat*, i serà *obert* altrament. El nombre d'arestes que hi ha a la successió la definim com a *llargada* del passeig. Per altra banda, un *camí* (*path*), el denotarem P_n, és un passeig on no es poden repetir els vèrtexs i per tant tampoc les arestes. Si un camí de llargada n és tancat, direm que tenim un *cicle d'ordre n*, l'anomenarem C_n.

Definició 2.4. *La cintura (girth) d'un graf G és el cicle més petit que està contingut en G, l'escriurem $g(G)$.*

Aquesta mesura pot arribar a valdre ∞, ja que si per exemple $G = P_n$ un camí obert, no hi ha cap cicle dins de G.

Definició 2.5. *Es diu que un graf és connex si existeix un camí que connecta cada parella de vèrtexs.*

Aquesta relació entre els vèrtexs del graf actua com a relació d'equivalència al conjunt V, i les classes d'equivalència serien les *components connexes* de G. Un *arbre* (*tree*), que escriurem com T_n, és un graf sense cicles, connex amb n vèrtexs.

Definició 2.6. *Descrivim la distància entre u i v, $d(u, v)$, com la llargada del camí més curt entre ells.*

Si $u = v$ serà 0, i si pertanyen a diferents components connexes, serà ∞. El *diàmetre*, que escriurem $diam(G)$, és el suprem de totes les distàncies de G.

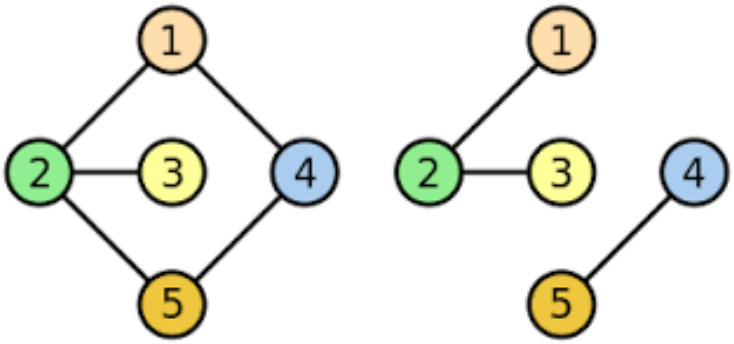

Figura 1: Exemples de grafs, que anomenarem graf A i B, d'esquerra a dreta.

Mitjançant la figura 1, podem ensenyar exemples dels conceptes que acabem d'introduir. Respecte el graf A, és connex i és fàcil de veure que $g(A) = 4$, que en aquest cas coincideix amb el $diam(A) = 4$. Per altra banda, el graf B no és connex, per tant $g(B) = diam(B) = \infty$.

Definició 2.7. *El complementari d'un graf G, el notarem com a $\overline{G}$, és un graf que comparteix el mateix conjunt de vèrtexs que G, però $u, v \in V$ estaran units a $\overline{G}$ si, i només si no estan units a G.*

Un graf serà *complet* si té totes les arestes possibles, l'anomenarem K_n. Direm que S és un *subconjunt de vèrtexs independent* si no estan connectats entre ells.

Definició 2.8. *El nombre cromàtic, el qual anomenarem $\chi(G)$, és el mínim cardinal n tal que $V(G)$ pot ser partit en n subconjunts independents. Si $\chi(G) = 2$, tenim un graf bipartit.*

Definició 2.9. *Dos conceptes més serien el nombre complet (clique number), que denotarem com $\omega(G)$, que és l'ordre del subgraf complet més gran dins de G; i l'estabilitat de G, que anomenarem com a $\alpha(G)$, que és l'ordre més gran que pot tenir un subconjunt independent de G.*

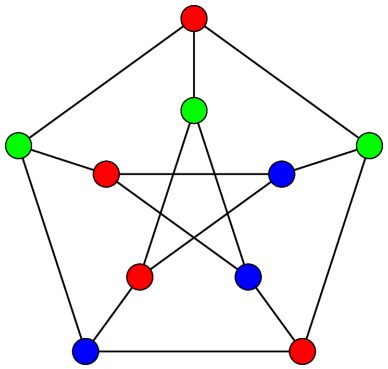

Figura 2: El graf de Petersen.

Com podem veure en la figura 2, ja tenim feta una coloració del graf, per tant el nombre cromàtic d'aquest graf G serà $\chi(G) = 3$. És fàcil de veure que l'estabilitat de G és $\alpha(G) = 4$, i el nombre complet és $\omega(G) = 2$.

La següent desigualtat relaciona el nombre cromàtic i el l'estabilitat d'un graf G i ens serà útil més endavant:

Teorema 2.10. *Si G és un graf d'ordre $n = |V(G)|$, aleshores:*

$$\alpha(G) \geq \frac{|V(G)|}{\chi(G)}$$

Demostració: Ja que $\frac{n}{\chi(G)}$ és la mitja de la mida de cada classe de la coloració $\chi(G)$, i com que per definició $\alpha(G)$ és el màxim dels conjunts independents, directament obtenim: $\alpha(G) \geq \frac{n}{\chi(G)}$.

□

Un altre tipus de grafs diferents als que hem vist fins ara són els grafs *dirigits*. Aquests es diferencien en que les arestes no són bidireccionals, o sigui que u estigui unit amb v no significa que v ho estigui amb u. Els vèrtexs tindran dos classes de grau: grau d'*entrada* i de *sortida*. El grau d'entrada de u, $deg^-(u)$ és el nombre d'arestes que arriben a u. En canvi el de sortida, $deg^+(u)$ és el nombre d'arestes que surten de u.

Aquests grafs són especialment útils a l'hora d'estudiar el comportament del graf web, que denotarem com W al llarg de tot aquest treball. I també serveixen als buscadors d'internet, com Google, a modelitzar el seu algorisme per ordenar les diferents pàgines web.

Com als grafs usuals, ara també es poden definir camins i cicles dirigits de forma anàloga a com ho hem definit anteriorment. Dins un graf dirigit G, definim la relació d'equivalència R tal que xRy si, i només si existeix un camí dirigit de x a y i un altre de y a x. Les classes d'equivalència s'anomenen *components fortament connexes* (*SCC, strongly connected components*) de G.

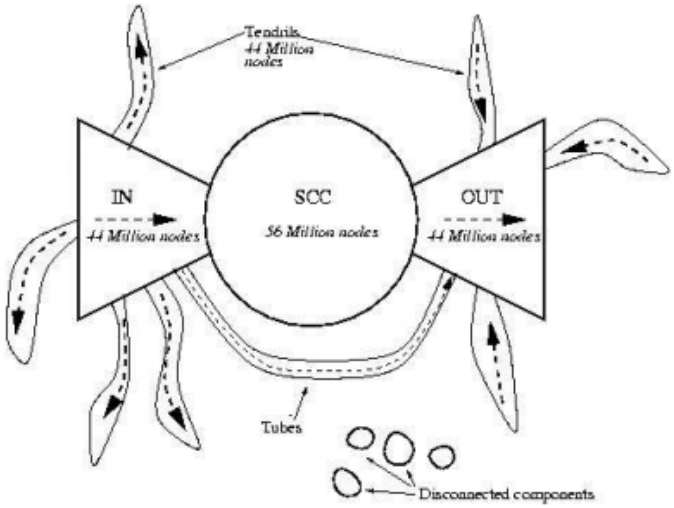

Figura 3: Esquema simplificat de l'estructura de la web

Respecte de la topologia de la Web, a [6], s'ha proposat la figura 3. Té una estructura peculiar que ens pot recordar un organisme biològic. Al centre es situa la component SCC. Al seu costat, apareixen dos peces: la IN està formada per les pàgines que tenen links cap a SCC, i la OUT està constituida per les pàgines a les que apunten les de SCC. Cal deixar clar, que en la nostra modelització, per contra, no considerarem grafs dirigits.

2.2 Teoria de Probabilitats

A continuació repassarem alguns conceptes de la teoria de probabilitats que ens seran útils més endavant per complementar-se amb la teoria de grafs. Tots aquests conceptes i definicions presentats a continuació serveixen a més per fixar la notació que utilitzarem al llarg del treball. Només considerarem els espais de probabilitats discrets.

Recordem que un *espai de probabilitat discret* Ω és una terna $(\Omega, F, \mathbb{P})$, on Ω és l'*espai mostral*, F és la *σ-àlgebra* de subconjunts de Ω (els *esdeveniments*), i $\mathbb{P}$ és la *mesura de*

probabilitat, $\mathbb{P} : F \to \mathbb{R}$, que satisfà les següents propietats:

1. Per tot esdeveniment A, $\mathbb{P}(A) \in [0,1]$, i $\mathbb{P}(\Omega) = 1$.
2. Si $\{A_i : i \in I\}$ és un conjunt numerable d'esdeveniments disjunts dos a dos, llavors:

$$\mathbb{P}\left(\bigcup_{i\in I} A_i\right) = \sum_{i\in I} \mathbb{P}(A_i).$$

Donat un espai de probabilitat finit Ω amb $n > 0$ elements, definim la *probabilitat uniforme* com: $\mathbb{P}(A) = \frac{|A|}{n}$, per a $A \subseteq \Omega$. Al triar un element amb probabilitat $\frac{1}{n}$ de Ω direm que ha estat triat uniformement a l'atzar o *u.a.r.* (*uniformly at random*).

Si $\mathbb{P}(B) > 0$ podem definir la *probabilitat condicionada* de que passi A donat B com:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

Direm que dos esdeveniments, $A, B \in F$, són *independents* si $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. Equivalentment i utilitzant la definició de probabilitat condicionada, també podem definir la independència de dos esdeveniments com: $\mathbb{P}(A|B) = \mathbb{P}(A)$.

Una *variable aleatòria* X a l'espai de probabilitat Ω és una funció $X : \Omega \to \mathbb{R}$. Donat $x \in \mathbb{R}$, definim:

$$\mathbb{P}(X = x) = \sum_{\omega\in\Omega, X(\omega)=x} \mathbb{P}(\{\omega\}).$$

Les probabilitats $\mathbb{P}(X \geq x)$, $\mathbb{P}(X > x)$, $\mathbb{P}(X \leq x)$, i $\mathbb{P}(X < x)$ es defineixen de forma anàloga. L'*esperança* o *primer moment* d'una variable aleatòria X és:

$$\mathbb{E}(X) = \sum_{\omega\in\Omega} X(\omega)\mathbb{P}(\{\omega\}) = \sum_{y\in X(\Omega)} y\mathbb{P}(X = y).$$

Si Ω és finit, $\mathbb{E}(X)$ sempre serà finita; si no ho fos, $\mathbb{E}(X)$ podria ser infinita. Dos propietats molt importants i bàsiques de l'esperança són:

1. És una funció lineal.
2. És una funció monòtona.

Les variables X i Y són independents si per tots els nombres reals x i y tenim que $X \leq x$ i $Y \leq y$ són independents. Un altre concepte molt important i lligat al de l'esperança, és el de *variància* d'una variable aleatòria. Si l'esperança ens dona el valor esperat de la variable, la variància ens mesura la dispersió respete d'aquest valor.

$$Var(X) = \mathbb{E}((X - \mathbb{E}(X))^2) = \mathbb{E}(X^2) - \mathbb{E}(X)^2.$$

Introduirem el concepte de *convergència en probabilitat* que ens servirà més endavant. Una successió de variables aleatòries X_n convergeix en probabilitat a una constant c si $\forall \varepsilon > 0$ tenim que

$$\lim_{n\to\infty} \mathbb{P}(|X_n - c| > \varepsilon) = 0.$$

Ara podem donar uns exemples clàssics de variables aleatòries que ens acompanyaran al llarg del treball:

1. *Binomial*: té com a paràmetres n i p, $X \in Bi(n,p)$, amb la funció:

$$\mathbb{P}(X=i) = \binom{n}{i} p^i (1-p)^{n-i}.$$

La seva esperança i variància són p i $p(1-p)$ respectivament. Un cas particular és quan $n=1$, la variable s'anomena *Bernoulli*. Una suma de variables Bernoulli independents té distribució binomial.

2. *Poisson*: la variable pren valors als nombres naturals $\mathbb{N}$ i té com a paràmetre un valor $\lambda > 0$. La seva funció és:

$$\mathbb{P}(X=i) = \frac{\lambda^i}{i!} \mathrm{e}^{-\lambda}$$

Aquesta variable té la peculiaritat de què l'esperança i la variància coincideixen i són igual a λ.

Recordem el teorema que relaciona en el límit aquestes dues distribucions. És l'anomenada llei dels esdeveniments rars.

Teorema 2.11. *Sigui* $\{X_n : n \in \mathbb{N}\}$ *una successió de variables aleatòries tals que per tot* n, $X_n \in Bi(n, p(n))$ *on* $\lim_{n\to\infty} np(n) = \lambda > 0$. *Llavors,* $\forall i \in \mathbb{N}$*:*

$$\lim_{n\to\infty} \mathbb{P}(X_n = i) = \frac{\lambda^i}{i!} \mathrm{e}^{-\lambda}.$$

Demostració: Partim de la probabilitat binomial i ja considerant que $np(n) \approx \lambda$:

$$\begin{aligned}
\mathbb{P}(X_n = i) &\approx \binom{n}{i} p(n)^i (1-p(n))^{n-i} \approx \frac{n!}{i!(n-i)!} \frac{\lambda^i}{n^i} \left(1 - \frac{\lambda}{n}\right)^{n-i} \\
&= \frac{n(n-1)(n-2)...(n-i+1)}{i!} \frac{\lambda^i}{n^i} \left(1 - \frac{\lambda}{n}\right)^{n} \left(1 - \frac{\lambda}{n}\right)^{-i} \\
&= \frac{n}{n}\left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)...\left(1 - \frac{i-1}{n}\right) \frac{\lambda^i}{i!} \left(1 - \frac{\lambda}{n}\right)^{n} \left(1 - \frac{\lambda}{n}\right)^{-i} \\
&= \left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)...\left(1 - \frac{i-1}{n}\right) \frac{\lambda^i}{i!} \left(1 - \frac{\lambda}{n}\right)^{n} \left(1 - \frac{\lambda}{n}\right)^{-i} = G(n)
\end{aligned}$$

Si ara fem el límit de $n \to \infty$:

$$\begin{aligned}
\lim_{n\to\infty} G(n) &= \lim_{n\to\infty} \frac{\lambda^i}{i!} \left(1 - \frac{\lambda}{n}\right)^{n} = \lim_{n\to\infty} \frac{\lambda^i}{i!} \left(1 + \frac{1}{-n/\lambda}\right)^{n} \\
&= \lim_{n\to\infty} \frac{\lambda^i}{i!} \left(\left(1 + \frac{1}{-n/\lambda}\right)^{\frac{-n}{\lambda}}\right)^{\frac{-n\lambda}{n}} = \frac{\lambda^i}{i!} \mathrm{e}^{-\lambda}
\end{aligned}$$

□

2.2.1 Esperança condicionada

Primerament, definirem el concepte d'esperança condicionada d'una variable aleatòria respecte un esdeveniment:

Definició 2.12. *Sigui X una variable aleatòria i B un esdeveniment qualsevol en el mateix espai de probabilitats. L' esperança condicionada de X respecte de B és:*

$$\mathbb{E}[X|B] = \frac{1}{\mathbb{P}(B)} \int_B X d\mathbb{P} = \frac{\mathbb{E}(X\mathbb{1}_B)}{\mathbb{P}(B)}.$$

Ara presentarem el concepte d'*esperança condicionada entre dues variables aleatòries.* Siguin X, Y variables aleatòries al mateix espai de probabilitat finit. La funció de massa condicional de X donat $Y = y$, escrita $f_{X|Y}(\cdot|y)$, es defineix com:

$$f_{X|Y}(x|y) = \mathbb{P}(X = x|Y = y),$$

$\forall y$ tal que $\mathbb{P}(Y = y) > 0$. Donat $Y = y$, podem pensar $f_{X|Y}(x|y)$ com una funció de x. L'esperança d'aquesta distribució és l'esperança condicionada de X quan $Y = y$, i s'escriu:

$$\mathbb{E}[X|Y = y] = \sum_x x f_{X|Y}(x|y) = \sum_x \frac{x f_{X,Y}(x, y)}{f_Y(y)},$$

on $f_{X,Y}(x, y)$ és la funció de densitat conjunta de les dos variables aleatòries X i Y.

Definim $g(y) = \mathbb{E}[X|Y = y]$. La funció g és *l'esperança condicionada de X sobre Y*, escrit $\mathbb{E}[X|Y]$. Notem que $\mathbb{E}[X|Y]$ és una variable aleatòria, i que per tant té esperança. Intuïtivament, és el valor de X suposant que coneixem Y. De la definició podem concloure el següent lema:

Lema 2.13. *Siguin X, Y variables aleatòries tal que X tingui esperança finita, llavors:*

$$\mathbb{E}(\mathbb{E}[X|Y]) = \mathbb{E}(X).$$

Demostració: Sempre que els sumatoris siguin absolutament convergents tenim:

$$\begin{aligned}
\mathbb{E}(\mathbb{E}[X|Y]) &= \sum_y \mathbb{E}[X|\{Y = y\}] f_Y(y) \\
&= \sum_y \left(\sum_x \frac{x f_{X,Y}(x, y)}{f_Y(y)} \right) f_Y(y) \\
&= \sum_x x f_X(x) = \mathbb{E}(X).
\end{aligned}$$

□

A continuació donarem la definició d'esperança condicionada respecte d'una σ-àlgebra $\mathcal{G}$, $\mathbb{E}[X|\mathcal{G}]$.

Definició 2.14. *Sigui X una variable aleatòria a l'espai de probabilitat $(\Omega, F, \mathbb{P})$, i sigui $\mathcal{G}$ una σ-àlgebra continguda a F. Llavors l'esperança condicionada de X donat $\mathcal{G}$ és una variable aleatòria $\mathbb{E}[X|\mathcal{G}]$ tal que:*

1. $\mathbb{E}[X|\mathcal{G}]$ *és* $\mathcal{G}$*-mesurable.*

2. *Per qualsevol* $A \in \mathcal{G}$

$$\int_A \mathbb{E}[X|\mathcal{G}]d\mathbb{P} = \int_A X d\mathbb{P}. \tag{2.1}$$

A partir de la definició anterior podem deduir el següent lema:

Lema 2.15. *Per tot* $B \in \mathcal{G}$*:*

$$\mathbb{E}(\mathbb{E}[X|\mathcal{G}]|B) = \mathbb{E}[X|B].$$

Demostració: Per la definició 2.1:

$$\int_B \mathbb{E}[X|\mathcal{G}]d\mathbb{P} = \int_B X d\mathbb{P},$$

per $B \in \mathcal{G}$. Es segueix que:

$$\mathbb{E}(\mathbb{E}[X|\mathcal{G}]|B) = \frac{1}{\mathbb{P}(B)} \int_B \mathbb{E}[X|\mathcal{G}]d\mathbb{P} = \frac{1}{\mathbb{P}(B)} \int_B X d\mathbb{P} = \mathbb{E}[X|B].$$

□

Una σ-àlgebra $\mathcal{G} \subseteq F$ qualsevol pot estar generada pels conjunts $B_i = \{Y = y_i\}$ que formen una partició d'Ω d'una variable aleatòria Y discreta. Amb la definició i el lema que acabem de donar es pot veure que:

$$\mathbb{E}[X|Y] = \mathbb{E}[X|\sigma(Y)],$$

i ara podem considerar $\mathbb{E}[X|Y_1, ..., Y_s]$ com a condicionar respecte de la σ-àlgebra generada per les variables aleatòries $Y_1, ..., Y_s$. Repetint el mateix argument d'abans tindriem que:

$$\mathbb{E}(\mathbb{E}[X|Y_1, ..., Y_s]) = \mathbb{E}(X). \tag{2.2}$$

Aquests conceptes ens permeten donar un caràcter més general de les següents propietats de l'esperança condicionada.

Proposició 2.16. *Siguin* X *i* Y *variables aleatòries discretes i* $\mathcal{G}, \mathcal{G}_1, \mathcal{G}_2$ σ*-àlgebres. Es compleixen les següents propietats:*

1. *Si* a *i* b *són nombres reals, aleshores* $\mathbb{E}[aX + bY|\mathcal{G}] = a\mathbb{E}[X|\mathcal{G}] + b\mathbb{E}[Y|\mathcal{G}]$.

2. *Si* $X \leq Y$, $\mathbb{E}[X|\mathcal{G}] \leq \mathbb{E}[Y|\mathcal{G}]$.

3. *Si* $\mathcal{G}_1 \subseteq \mathcal{G}_2 \subseteq F$, *tenim*

$$\mathbb{E}(\mathbb{E}[X|\mathcal{G}_1]|\mathcal{G}_2) = \mathbb{E}(\mathbb{E}[X|\mathcal{G}_2]|\mathcal{G}_1) = \mathbb{E}[X|\mathcal{G}_1].$$

Demostració:

1. Anomenem $z = a\mathbb{E}[X|\mathcal{G}] + b\mathbb{E}[Y|\mathcal{G}]$. Llavors, per qualsevol $G \in \mathcal{G}$:

$$\begin{aligned}\mathbb{E}[z\mathbb{1}_G] &= \mathbb{E}([a\mathbb{E}[X|\mathcal{G}] + b\mathbb{E}[Y|\mathcal{G}]]\mathbb{1}_G) = \\ &= a\mathbb{E}[(\mathbb{E}[X|\mathcal{G}]\mathbb{1}_G) + b\mathbb{E}(\mathbb{E}[Y|\mathcal{G}]\mathbb{1}_G) = \\ &= a\mathbb{E}[X\mathbb{1}_G] + b\mathbb{E}[Y\mathbb{1}_G] = \mathbb{E}([aX + bY]\mathbb{1}_G)\end{aligned}$$

   Per tant,

$$z = \mathbb{E}[aX + bY|\mathcal{G}].$$

2. Suposem que $\mathcal{G}$ està generada per una partició $\{G_i : i \in I\}$ finita o numerable d'esdeveniments de Ω. Si $X \leq Y$, és clar que $\mathbb{E}[X|G_i] \leq \mathbb{E}[Y|G_i]$ per a tot G_i i per tant, $\mathbb{E}[X|G] \leq \mathbb{E}[Y|G]$. És a dir, l'esperança condicionada és un operador creixent, com ho és l'esperança.

3. Primer demostrarem

$$\mathbb{E}(\mathbb{E}[X|\mathcal{G}_1]|\mathcal{G}_2) = \mathbb{E}[X|\mathcal{G}_1].$$

   Sigui $z = \mathbb{E}[X|\mathcal{G}_1]$. Llavors z és $\mathcal{G}_1$-mesurable, i al ser $\mathcal{G}_1 \subseteq \mathcal{G}_2$ implica que és $\mathcal{G}_2$-mesurable. Significa que $z = \mathbb{E}[z|\mathcal{G}_2]$, és a dir:

$$\mathbb{E}(\mathbb{E}[X|\mathcal{G}_1]|\mathcal{G}_2) = \mathbb{E}[X|\mathcal{G}_1].$$

   Ara demostrarem l'altra igualtat:

$$\mathbb{E}(\mathbb{E}[X|\mathcal{G}_2]|\mathcal{G}_1) = \mathbb{E}[X|\mathcal{G}_1].$$

   Sigui $z = \mathbb{E}(\mathbb{E}[X|\mathcal{G}_2]|\mathcal{G}_1)$. Llavors, z és $\mathcal{G}_1$-mesurable, té esperança finita, i per qualsevol $G \in \mathcal{G}_1$:

$$\mathbb{E}[z\mathbb{1}_G] = \mathbb{E}(\mathbb{E}(\mathbb{E}[X|\mathcal{G}_2]|\mathcal{G}_1)\mathbb{1}_G) = \mathbb{E}(\mathbb{E}[X|\mathcal{G}_2]\mathbb{1}_G) = \mathbb{E}[X\mathbb{1}_G].$$

   Per tant, $z = \mathbb{E}[X|\mathcal{G}_1]$.

□

2.2.2 Martingales

Les martingales són útils en la teoria dels grafs aleatoris per provar la concentració dels valors d'una variable aleatòria al voltant del valor esperat. Les desigualtats de Chernoff s'apliquen a les sumes de variables aleatòries independents; la concentració de martingala és útil quan hi ha dependència, o bé la concentració que utilitzi la variància no és aplicable d'una altra manera. Per descriure la concentració de martingala, necessitem els coneixements d'esperança condicionada explicats prèviament.

Definició 2.17. *Una successió de variables aleatòries $X_0, X_1, ...$ és una martingala respecte $Z_0, Z_1, ...$ si, per tot $n \geq 0$, es compleixen les següents condicions:*

1. *X_i és una funció de $Z_0, Z_1, ..., Z_n$.*
2. *$\mathbb{E}(|X_n|) < \infty$.*

3. $\mathbb{E}[X_n|Z_0, ..., Z_{n-1}] = X_{n-1}$.

Una successió de variables aleatòries $X_0, X_1, ...$ *s'anomena martingala quan és una martingala respecte d'ella mateixa. Això significa que:* $\mathbb{E}(|X_n|) < \infty$, *i* $\mathbb{E}[X_n|X_0, ..., X_{n-1}] = X_{n-1}$.

Al treballar amb variable discretes i considerar F_i la σ-àlgebra generada per la pròpia variable, les condicions (1) i (2) soldran ser immediates.

De fet, es poden calcular martingales de qualsevol successió de variables aleatòries. Per ser més precisos, sigui A i $\{Z_i : 1 \leq i \leq t\}$ variables aleatòries al mateix espai de probabilitat; definim $X_0 = \mathbb{E}[A]$ i sigui $X_i = \mathbb{E}[A|Z_1, ..., Z_i]$, on $1 \leq i \leq t$.

Lema 2.18. *La successió de variables aleatòries* $\{X_i : 0 \leq i \leq t\}$ *definida anteriorment, és una martingala respecte* Z_i.

Demostració: Com hem dit abans, les propietats (1) i (2) són immediates de la definició d'aquesta successió de variables aleatòries. Per tant passem a demostrar la tercera propietat. Llavors tenim:

$$\mathbb{E}[X_n|Z_1, ..., Z_{n-1}] = \mathbb{E}[\mathbb{E}[A|Z_1, ..., Z_n]|Z_1, ..., Z_{n-1}] = \mathbb{E}[A|Z_1, ..., Z_{n-1}] = X_{n-1}.$$

Per dur a terme la prova hem aplicat la tercera propietat de l'esperança condicionada que hem demostrat a l'apartat anterior.

□

La martingala $\{X_i : 0 \leq i \leq t\}$ s'anomena *martingala de Doob de A respecte a* Z_i. Podem pensar en A com una funció

$$f(Z_1, ..., Z_t)$$

determinada per les variables aleatòries $\{Z_i : 1 \leq i \leq t\}$. Llavors $X_0 = \mathbb{E}[A]$ i $X_t = A$. La martingala de Doob ens mostra més informació sobre f a mesura que passa el temps, per això al final tota la informació ja està disponible i obtenim el valor actual de f.

2.2.3 Resultats de concentració

L'objectiu d'aquest apartat és presentar alguns dels principals teoremes i desigualtats de concentració, és a dir, cotes per la probabilitat de que una variable aleatòria es desviï del su valor central.

Teorema 2.19. *Desigualtat de Markov: Sigui* $X \geq 0$ *una variable aleatòria en un espai de probabilitat. Si* c *és un número real positiu es compleix:*

$$\mathbb{P}(X \geq c) \leq \frac{\mathbb{E}(X)}{c}.$$

Demostració: Partim de la formula de l'esperança:

$$\mathbb{E}(X) = \sum_{\omega \in \Omega} X(\omega)\mathbb{P}(\{\omega\}) \geq \sum_{\omega \in \Omega, X(\omega) \geq c} X(\omega)\mathbb{P}(\{\omega\}) \geq$$
$$\geq \sum_{\omega \in \Omega, X(\omega) \geq c} c\mathbb{P}(\{\omega\}) = c\mathbb{P}(X \geq c)$$

Si passem la c a l'altra banda i que divideixi l'esperança, ja obtenim el que voliem veure.

□

Teorema 2.20. *Desigualtat de Chebyshev: Sigui X una variable aleatòria a un espai de probabilitat. Si c és un número real positiu es compleix:*

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq c) \leq \frac{Var(X)}{c^2}.$$

Demostració: Si considerem $Y = |X - \mathbb{E}(X)|$ i apliquem la desigualtat de Markov sobre Y, obtenim:

$$\mathbb{P}(Y \geq c) = \mathbb{P}(Y^2 \geq c^2) \leq \frac{Y^2}{c^2}.$$

Com que $Y^2 = Var(X)$, ja tenim el que voliem veure. Cal remarcar que la primera igualtat es pot fer degut a que c i Y són positiu.

□

Teorema 2.21. *Desigualtats de Chernoff: Siguin $X_1, ..., X_m$ variables aleatòries independents i idènticament distribuïdes tals que X_i sempre cau dins l'interval $[0, 1]$ per tot i. Definim $X = \sum_i X_i$. Sigui $p_i = \mathbb{E}[X_i]$ i $\mu = \mathbb{E}[X] = \sum_i p_i$. Llavors $\forall \delta \in [0, 1]$ tenim*

$$\mathbb{P}(X \geq (1 + \delta)\mu) \leq \exp(\delta\mu - (1 + \delta)\ln(1 + \delta)\mu). \tag{2.3}$$

$$\mathbb{P}(X \leq (1 - \delta)\mu) \leq \exp(\delta\mu - (1 + \delta)\ln(1 + \delta)\mu). \tag{2.4}$$

Demostració: La demostració de les desigualtats de Chernoff no seria possibles sense la suposició de que les X_i són independents entre elles. Una de les propietats bàsiques d'aquest fet és que:

$$\mathbb{E}[X_i X_j] = \mathbb{E}[X_i]\mathbb{E}[X_j] \tag{2.5}$$

per tot $i \neq j$.

Per la demostració també necessitem enunciar una altra variable aleatòria derivada de les X_i i un corol·lari.

Sigui $t > 0$ un paràmetre fix. Definim:

$$Y_i = \exp(tX_i)$$

$$Y = \exp(tX) = \exp\left(t \sum_i X_i\right) = \prod_i \exp(tX_i) = \prod_i Y_i.$$

Com que X_i són independents per hipòtesi, això comporta que les variables Y_i també ho siguin. Per tant, per 2.5 tenim:

$$\mathbb{E}[Y] = \prod_i \mathbb{E}[Y_i]. \tag{2.6}$$

Aquesta equació és útil perquè podem analitzar $\mathbb{E}[Y]$ analitzant per separat els termes $\mathbb{E}[Y_i]$. A més, Y_i està molt relacionat amb X_i. Ara enunciarem i demostrarem un corol·lari que relaciona $\mathbb{E}[Y_i]$ amb $\mathbb{E}[X_i]$.

Corol·lari 2.22. *$\forall i$ podem acotar:* $\mathbb{E}[Y_i] \leq \exp((e^t - 1)p_i)$.

Demostració corol·lari: Volem relacionar $\mathbb{E}[Y_i] = \mathbb{E}[e^{tX_i}]$ i $\mathbb{E}[X_i]$. Com que la funció exponencial no és lineal, no podem aplicar la linealitat de les esperances. Però com que sabem que $X_i \in [0,1]$, podem utilitzar una aproximació lineal de e^{tx} en aquest interval.

Per convexitat, tenim que:

$$e^{tx} \leq 1 + (e^t - 1)x.$$

Per tant,

$$\mathbb{E}[e^{tX_i}] \leq \mathbb{E}[1 + (e^t - 1)X_i] = 1 + (e^t - 1)p_i \leq \exp((e^t - 1)p_i),$$

on a l'última desigualtat hem utilitzat que: $1 + x \leq e^x$.

□

Ara podem continuar amb la demostració de Chernoff. Només provarem la cota superior ja que la inferior es fa de forma anàloga.

$$\begin{aligned}\mathbb{P}(X \geq (1+\delta)\mu) &= \mathbb{P}(\exp(tX) \geq \exp(t(1+\delta)\mu)) \leq \\ &\leq \frac{\mathbb{E}[\exp(tX)]}{\exp(t(1+\delta)\mu)} \leq \frac{\prod_i \exp((e^t-1)p_i)}{\exp(t(1+\delta)\mu)}.\end{aligned}$$

A la primera igualtat hem utilitzat la monotonicitat de l'esperança, a la primera desigualtat hem aplicat la desigualtat de Markov, i en l'última desigualtat hem aplicat 2.6 i el corol·lari 2.22. Si expressem la darrera fracció dins d'una exponencial, ens queda:

$$\mathbb{P}(X \geq (1+\delta)\mu) \leq \exp((e^t - 1)\sum_i p_i - t(1+\delta)\mu).$$

Finalment, si substituïm $t = \ln(1+\delta)$ i $\mu = \sum_i p_i$, obtenim el que volíem veure.

$$\mathbb{P}(X \geq (1+\delta)\mu) \leq \exp(\delta\mu - (1+\delta)\ln(1+\delta)\mu).$$

□

Si $\delta \in [0,1]$, llavors podem arribar a una millor aproximació:

$$\mathbb{P}(X \geq (1+\delta)\mu) \leq \exp\left(\frac{-\mu\delta^2}{3}\right). \tag{2.7}$$

$$\mathbb{P}(X \leq (1-\delta)\mu) \leq \exp\left(\frac{-\mu\delta^2}{3}\right). \tag{2.8}$$

Per aconseguir aquesta nova aproximació s'utilitza el polinomi de Taylor de la funció

$$\ln(1+x) = \sum_{n=1}^{\infty} (-1)^{n+1} \frac{x^n}{n}.$$

La successió $\{X_i : 0 \leq i \leq t\}$ satisfà la *condició c-Lipschitz*, on $c > 0$ és un enter, si per tot $0 \leq i \leq t-1$

$$|X_{i+1} - X_i| \leq c.$$

Posat d'una altra manera, les diferències entre variables aleatòries successives no són molt grans.

Teorema 2.23. (Desigualtat d'Azuma-Hoeffding) *Si* $\{X_i : 0 \leq i \leq t\}$ *és una martingala que satisfà la condició de c-Lipschitz, llavors* $\forall \lambda > 0$ *real,*

$$\mathbb{P}(|X_t - X_0| \geq \lambda\sqrt{t}) \leq 2\exp\left(-\frac{\lambda^2}{2c^2}\right). \tag{2.9}$$

En altres paraules, si es compleix la condició c-Lipschitz, aleshores els valors de X_t estan *concentrades al voltant de* X_0. Per ser més precisos, si $t \to \infty$ i $\lambda \to \infty$ molt lentament com una funció de t, aleshores amb probablitat 1-2exp(-$\frac{\lambda^2}{2}$), X_t està en un interval al voltant de X_0 de longitud $2\lambda\sqrt{t}$. Aquest mètode per provar la concentració per variables aleatòries ha estat anomenat el *mètode de diferències limitades.*

Demostració: És suficient provar la desigualtat

$$\mathbb{P}(X_t \geq X_0 + \lambda\sqrt{t}) \leq \exp\left(-\frac{\lambda^2}{2c^2}\right). \tag{2.10}$$

Llavors per provar la desigualtat del teorema, apliquem la desigualtat anterior a la martingala $\{-X_i : 0 \leq i \leq t\}$

Per $1 \leq i \leq t$, definim $Y_i = X_i - X_{i-1}$. Llavors $|Y_i| \leq c$. A més a més, $\forall i$ tal que $1 \leq i \leq t$,

$$\mathbb{E}(Y_i|X_0, ..., X_{i-1}) = 0,$$

ja que $\{X_i : 0 \leq i \leq t\}$ és una martingala. Per la convexitat de la funció $f(x) = \mathrm{e}^{nx}$ per $n > 0$, tenim que $\forall 1 \leq i \leq t$

$$\begin{aligned}\mathbb{E}(\mathrm{e}^{nY_i}|X_0, ..., X_{i-1}) &\leq \mathbb{E}\left(\frac{1}{2c}(\mathrm{e}^{nc} - \mathrm{e}^{-nc})Y_i + \frac{1}{2}(\mathrm{e}^{nc} - \mathrm{e}^{-nc})|X_0, ..., X_{i-1}\right) \\ &= \frac{1}{2}(\mathrm{e}^{nc} + \mathrm{e}^{-nc}) \leq \mathrm{e}^{\frac{n^2c^2}{2}},\end{aligned} \tag{2.11}$$

on la primera igualtat es dona per la linealitat de l'esperança i la segona considerant les sèries de Maclaurin per e^{nc} i e^{-nc}.

Per tant tenim que

$$\begin{aligned}\mathbb{E}(\mathrm{e}^{n(X_t - X_0)}) &= \mathbb{E}\left(\prod_{i=1}^{t} \mathrm{e}^{nY_i}\right) \\ &= \mathbb{E}\left(\left(\prod_{i=1}^{t-1} \mathrm{e}^{nY_i}\right)\mathbb{E}(\mathrm{e}^{nY_t}|X_0, ..., X_{t-1})\right) \\ &\leq \mathbb{E}\left(\prod_{i=1}^{t-1} \mathrm{e}^{nY_i}\right)\mathrm{e}^{\frac{n^2c^2}{2}} \leq \mathrm{e}^{\frac{n^2c^2t}{2}},\end{aligned} \tag{2.12}$$

on la primera desigualtat es segueix de 2.11, i l'última desigualtat s'obté per inducció.

Sigui $n = \frac{\lambda}{c^2\sqrt{t}}$. Aleshores,

$$\begin{aligned}\mathbb{P}(X_t \geq X_0 + \lambda\sqrt{t}) &\leq \mathbb{P}(\mathrm{e}^{n(X_t - X_0)} \geq \mathrm{e}^{n\lambda\sqrt{t}}) \\ &\leq \mathrm{e}^{-n\lambda\sqrt{t}}\mathbb{E}(\mathrm{e}^{n(X_t - X_0)}) \\ &\leq \mathrm{e}^{-n\lambda\sqrt{t}}\mathrm{e}^{\frac{n^2c^2t}{2}} \\ &= \mathrm{e}^{-\frac{\lambda^2}{2c^2}},\end{aligned}$$

on la segona desigualtat ve per la desigualtat de Markov, i la tercera és gràcies a 2.12 .

□

2.3 El Graf Web

Arribats a aquest punt, hem esmentat ja algun cop el graf web, que denotarem com W, però no hem acabat de concretar exactament que és. Això és el que explicaré en aquest apartat. Entendrem com a W, el graf on els seus vèrtexs representen les diferents pàgines web, i les arestes els *links* que hi ha entre elles. Les arestes podran ser dirigides o no (en els nostres models no considerarem grafs dirigits). No existeix només un model pel graf W, ja que basant-se en comportaments empírics de W s'han establert un seguit de propietats que hauria de seguir. A més a més, també es poden crear models molt similars de W que no estiguin estudiant el comportament de la web sinó que estudien altres camps, els quals podem trobar a [1]. Remarquem que hi ha diferents classes de models per estudiar W.

Ara explicarem quines són algunes de les propietats que es consideren més rellevants a l'hora de que un model sigui representatiu. Remarquem que encara que un model no compleixi totes les propietats, continua sent un model, que pot ser interessant encara que menys representatiu. A part de les que explicaré a continuació, existeixen més propietats que poden ajudar a l'estudi de W, però aquestes, gràcies a diferents estudis empírics, s'han considerat que són de gran importància. Aquestes són: el fet de que W segueixi un model *on-line*; que els diferents graus dels vèrtexs segueixi la *llei de potència*; i per últim, que tingui la propietat de *small world*. Passo a descriure aquestes propietats.

2.3.1 Propietat On-line

Que un model segueixi la propietat on-line també es pot denotar com que el model és *dinàmic*. Això significa que a cada pas que hi hagi, el nombre de vèrtexs i/o arestes s'ha de modificar, ha d'anar variant. Aquesta propietat pot semblar una obvietat però és molt important tenir-la en compte. Considerar models *estàtics*, en certes ocasions pot arribar a concloure models que no s'ajustin al que es vol estudiar, ja que podem entendre la web com un organisme viu, on es van creant i esborrant noves pàgines i interaccions a cada moment. Un altre punt de vista seria considerar W com a un graf infinit, com podem veure al capítol 6 de [3]. Però això no ho estudiarem en aquest treball de final de grau.

2.3.2 Distribuició del grau amb llei de potència

Començarem introduint el concepte de $N_{k,G}$ donat un enter no negatiu k i un graf G no dirigit:

$$N_{k,G} = |\{x \in V(G) : deg_G(x) = k\}|.$$

$N_{k,G}$ representa el número de vèrtexs de G amb grau k. Per simplicitat, suposarem que $|V(G)| = t$, aleshores $N_{k,G}$ és un enter dins de $[0, t]$. Podem veure la distribució dels graus de G com la successió, $\{N_{k,G} : 0 \leq k \leq t\}$. Un exemple seria:

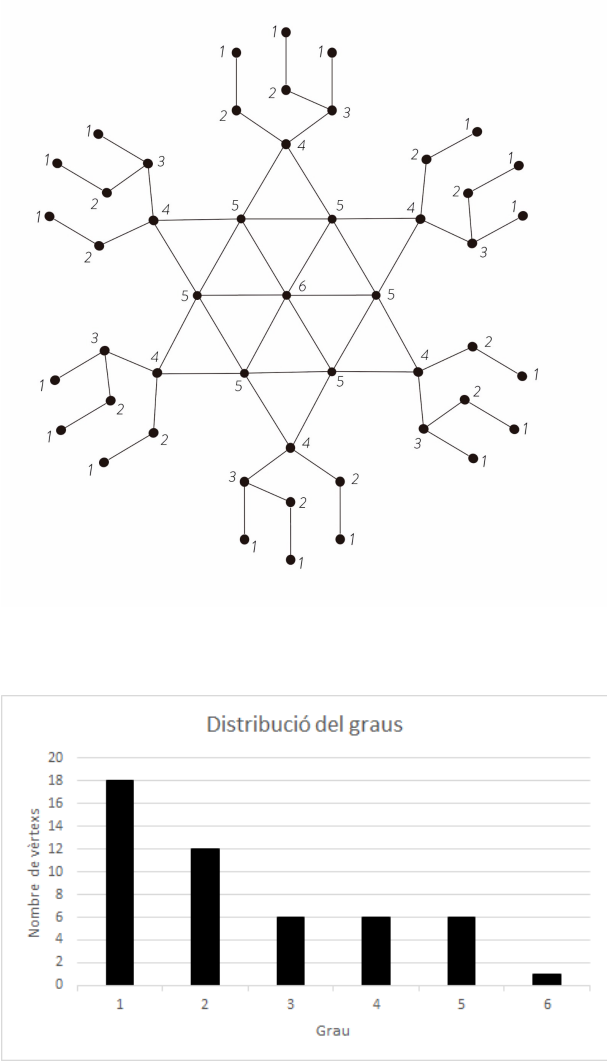

Figura 4: Un graf G i el gràfic de la distribució dels seus graus

Direm que la distribució del grau de G segueix una *llei de potència* si per cada grau k:

$$\frac{N_{k,G}}{t} \sim k^{-\beta}, \tag{2.13}$$

per una constant $\beta > 1$. La identitat 2.13 és assimptòtica ja que més endavant treballarem amb una successió infinita de grafs finits generats per un model de grafs aleatoris. Com que W serà un graf enorme, no estem tant interessats en valor exacte de $\frac{N_{k,G}}{t}$ sinó que ens interessa la seva aproximació.

Considerarem β com *l'exponent de la llei de potència*. Per simplicitat, si G compleix la propietat, direm que G és un *graf amb llei de potència*. Si apliquem logaritmes al dos costats de la identitat 2.13, obtenim la següent relació:

$$\log(N_{k,G}) \sim \log(t) - \beta\log(k).$$

Per tant, ens quedaria una expressió lineal de pendent $-\beta$. És obvi que aquesta propietat no es complirà per tota k. Serà normal que no respecti es compleixi en graus molt petits o molt grans, als extrems.

Si considerem que W és un graf dirigit, haurem de considerar si es compleix la llei de potència pels dos tipus de grau que tindrà cada vèrtex, el d'entrada i el de sortida.

2.3.3 Propietat Small World

Aquesta propietat simbolitza que al món actual, gràcies a les noves tecnologies s'han pogut trencar moltes barreres generades per la distància entre les persones. Per això el món sembla més petit ara amb invents com el telèfon mòbil.

El diàmetre d'un graf podria ser una bona mesura per decidir si satisfà la propietat. En teoria, els grafs d'ordre t haurien de complir que:

$$diam(G) = \Theta(\log t).$$

Encara que pugui semblar bona idea, a la vida real el diàmetre de W és infinit, ja que qualsevol persona pot crear una pàgina web que no estigui connectada a cap altra.

Per això, introduïm un altra mesura de la distància:

$$L(G) = \sum_{u,v \in S} \frac{d(u,v)}{|S|}, \tag{2.14}$$

on S és un conjunt de parelles de vèrtexs diferents u, v de G tals que $d(u,v)$ és finit. Si G és connex d'ordre t, podem simplificar la fòrmula 2.14 per:

$$L(G) = \sum_{u,v \in V(G)} d(u,v) \binom{t}{2}^{-1} = \frac{2}{t^2 - t} \sum_{u,v \in V(G)} d(u,v). \tag{2.15}$$

$L(G)$ és la distància mitja de G. Podem obtenir el mateix paràmetre per a grafs dirigits de forma anàloga, considerant la distància entre dos vèrtexs com el camí dirigit més curt.

Per a que un graf satisfaci la condició *small world* hauria de tenir un $L(G)$ més petit que el seu ordre. Per ser més precisos demanarem que

$$L(G) = \Theta(\log(\log t)).$$

Ara definirem el coeficient d'*agrupació* (*clustering coefficient*), el qual mesura la densitat local del graf. Sigui G d'ordre t i $x \in V(G)$, aleshores

$$C(x) = \frac{|E(G \backslash N(x))|}{\binom{deg(x)}{2}} = \frac{2|E(G \backslash N(x))|}{deg(x)(deg(x)-1)}. \tag{2.16}$$

Quan escrivim $|E(G \backslash N(x)|$ ens referim a al nombre d'arestes que queden a G quan eliminem les que estan unides a x. El coeficient d'agrupació, $C(G)$, serà una mitjana de tots els $C(x), \forall x \in G$:

$$C(G) = \frac{1}{t} \sum_{x \in V(G)} C(x) = \frac{2}{t} \sum_{x \in V(G)} \frac{|E(G \backslash N(x))|}{deg(x)(deg(x)-1)}. \tag{2.17}$$

Per considerar que un model satisfà adequadament la propietat small world, ha de ser gran. Com que $C(G) \in [0,1]$, gran significa que no ha de ser molt proper a 0. El $C(G)$ d'un graf complet val 1, això seria al cas en que totes les webs connectessin amb totes directament.

3 Grafs aleatoris

3.1 Introducció

Quan barregem la Teoria de Grafs amb la Teoria de Probabilitat, el resultant són els grafs aleatoris. Aquests són molt útils a l'hora d'estudiar el graf web W ja que ens proporcionen models. Avancem que donarem molta més importància als grafs aleatoris clàssics $G(n,p)$ que van definir Erdös i Rényi, els quals tenen una sèrie de propietats força interessants i dignes d'estudi.

La teoria dels grafs aleatoris l'inicien Paul Erdös i Alfréd Rényi. Analitzarem alguns mètodes fonamentals i útils per estudiar models del graf web.

Podem definir un espai de probabilitat dels grafs d'ordre n, on hi haurà un conjunt de n vèrtex, que direm V, i una probabilitat $p \in [0,1]$. Notarem com a $G(n,p)$ l'espai de grafs aleatoris d'ordre n i probabilitat p, on l'espai mostral és igual al conjunt de $2^{\binom{n}{2}}$ grafs amb vèrtexs V i

$$\mathbb{P}(G) = p^{|E(G)|}(1-p)^{\binom{n}{2}-|E(G)|}.$$

Això es pot entendre com a que cada parella de vèrtexs de G estan units independentment amb una probabilitat p. Per tant, el conjunt V no canvia però el conjunt E sí.

Una alternativa a $G(n,p)$ és considerar l'espai $G(n,M)$, que és l'espai amb probabilitat uniforme de tots els grafs amb n vèrtexs i exactament M arestes. Per tant, $\mathbb{P}(G)$ és igual a

$$\binom{\binom{n}{2}}{M}^{-1}.$$

Ara ensenyarem el potencial dels grafs aletòris mitjançant un problema clàssic dins la teoria de grafs. Per fer-ho definirem el *nombre de Ramsey.*

Definició 3.1. *Sigui un enter $n \geq 1$, el nombre de Ramsey és l'enter m més petit tal que qualsevol graf d'ordre m conté K_n o $\overline{K_n}$ com a subgrafs induïts.*

Una manera més senzilla d'entendre la definició és que per qualsevol coloració de les arestes amb només dos colors, hi ha un subgraf complet d'ordre n pintat d'un sol color. Els números de Ramsey, $R(n)$, existeixen pel següent teorema.

Teorema 3.2. *$\forall n \geq 1$, $R(n) \leq 2^{2(n-1)}$.*

Demostració: Per aquesta prova utilitzarem la definició del cas general del nombre de Ramsey, on els subgrafs induïts poden ser de diferent ordre. Ja no hi hauria la condició que K_r o $\overline{K_s}$ hagin de compartir el mateix número de vèrtexs, per tant r i s podrien ser números diferents. Utilitzarem la següent desigualtat, la qual està demostrada a [8], on hi ha un estudi més exhaustiu d'aquesta:

$$R(r,s) \leq \binom{r+s-2}{r-1}.$$

Si adaptem la desigualtat al cas que volem veure, que és $r = s = n$:

$$R(n) \leq \binom{2(n-1)}{n-1}.$$

Si fem servir el desenvolupament del binomi de Newton per

$$2^{2(n-1)} = (1+1)^{2(n-1)} =$$
$$= \binom{2(n-1)}{0}1^{2(n-1)}1^0 + ... + \binom{2(n-1)}{n-1}1^{n-1}1^{n-1} + ... + \binom{2(n-1)}{0} + 1^0 1^{2(n-1)},$$

com que $\binom{2(n-1)}{n-1}$ és un dels termes d'aquesta suma on tots els sumands són positius, això implica que $\binom{2(n-1)}{n-1} \leq 2^{2(n-1)}$. Ens queda que $R(n) \leq \binom{2(n-1)}{n-1} \leq 2^{2(n-1)}$.

□

És fàcil comprovar que $R(3) = 6$, però ja no es coneixen els valors exactes de $R(n)$ quan $n \geq 5$. Gràcies a l'aplicació dels grafs aleatòris, podem enunciar i demostrar el teorema següent, on obtenim una cota inferior del nombre de Ramsey.

Teorema 3.3. *Per tot* $n \geq 3, R(n) > 2^{\frac{n}{2}}$.

Demostració: Provarem que per un $m \leq 2^{\frac{n}{2}}$ fixat, per $G \in G(m, \frac{1}{2})$ satisfà $\alpha(G) < n$ i $\omega(G) < n$ amb probabilitat positiva. Recordem que $\alpha(G)$ i $\omega(G)$ són dos conceptes definits al tema anterior que signifiquen: l'ordre més gran que pot tenir un subconjunt independent de G, i l'ordre del subgraf complet més gran dins de G respectivament. Per tant, existiria un graf d'ordre m que no contindria K_n ni el seu complementari.

La probabilitat de que donat un conjunt S de n vèrtexs de $G \in G(m, \frac{1}{2})$ sigui un complet, és de $\left(\frac{1}{2}\right)^{\binom{n}{2}}$. Com que hi ha $\binom{m}{n}$ per S, tenim que:

$$\mathbb{P}(\omega(G) \geq n) \leq \binom{m}{n}\left(\frac{1}{2}\right)^{\binom{n}{2}} \leq \frac{m^n}{2^n}2^{-\frac{1}{2}(n^2-n)}$$
$$\leq 2^{\frac{n^2}{2}-n-\frac{1}{2}(n^2-n)} = 2^{-\frac{n}{2}} < \frac{1}{2},$$

amb $n \geq 3$.

Un càlcul anàleg mostra que

$$\mathbb{P}(\alpha(G) \geq n) < \frac{1}{2}.$$

Per tant,

$$\mathbb{P}(\omega(G) < n \text{ i } \alpha(G) < n) = 1 - \mathbb{P}(\omega(G) \geq n \text{ o } \alpha(G) \geq n) > 0.$$

□

La proposta més habitual per estudiar $G(n,p)$ és considerant quines propietats es mantenen a $G \in G(n,p)$ amb alta probabilitat quan n tendeix a infinit. Formalment:

Definició 3.4. *Sigui* $\mathcal{P}$ *una propietat del graf. Escriurem que* $G \in G(n,p)$ *satisfà* $\mathcal{P}$ *assimptòticament quasi segurament (a.q.s.) si*

$$\lim_{n\to\infty} \mathbb{P}(G \in G(n,p) \text{ satisfà } \mathcal{P}) = 1. \tag{3.1}$$

Si $\mathcal{P}$ satisfà la condició anterior, també podem dir que $G \in G(n,p)$ satisfà $\mathcal{P}$ *amb alta probabilitat (a.a.p.).*

La probabilitat de les arestes, p, normalment està en funció de n, per tant podem escriure-ho $p(n)$ o simplement p. L'estudi de quines propietats es compleixen a.q.s. en funció de p és un tema clau a la teoria de grafs aleatoris.

3.2 Esperança i Mètode del primer moment

Un paràmetre X dels grafs esdevé una variable aleatòria a $G(n,p)$. Per exemple, la mida (el nombre d'arestes) o el número cromàtic són variables aleatòries als grafs aleatoris. Calcular els valors exactes d'aquests paràmetres normalment resulta difícil per a certs grafs, però calcular les seves mitges a vegades és més senzill. La utilització de l'esperança és una eina important per entendre les propietats de les variables aleatòries a $G(n,p)$.

La linealitat de les esperances, encara que sigui una de les propietats bàsiques, té conseqüències interesants pels grafs aleatoris. Un exemple, sigui X el nombre d'arestes a $G(n,p)$. Per cada parella desordenada $\{i,j\}$ de $\{1,...,n\}$, considerem X_u *l'indicador de variable aleatòria* per l'aresta u. Això és:

$$X_u = \begin{cases} 1 & \text{si } ij \in E \\ 0 & \text{altrament.} \end{cases}$$

Llavors, $\mathbb{E}(X_u) = p$. Com

$$X = \sum_u X_u,$$

per la linealitat de l'esperança tenim que:

$$\mathbb{E}(X) = \binom{n}{2} p.$$

La desigualtat de Markov, enunciada i demostrada a l'anterior tema, s'usa en diversos resultats sobre $G(n,p)$. A continuació recordem dos conceptes que vam definir al tema passat: *la cintura* d'un graf G era la longitud del cicle més petit contingut a G, i el *nombre cromàtic* és el número mínim n tal que el conjunt $V(G)$ queda dividit en n subconjunts independents. Aquests dos conceptes els denotem com $g(G)$ i $\chi(G)$, respectivament. A primera vista, podem pensar que un graf sense cicles petits hauria de tenir un nombre cromàtic baix. El següent teorema ens demostra just el contrari.

Teorema 3.5. *Per tot enter $m \geq 2$, existeix un graf G tal que $\chi(G), g(G) \geq m$.*

Demostració: Considerem $G \in G(n,p)$ per un enter positiu n, i $p = p(n)$ a determinar. Com

$$\chi(G) \geq \frac{|V(G)|}{\alpha(G)},$$

el nostre enfocament per mostrar que $\chi(G)$ és gran, és provar que $\alpha(G)$ és petit en comparació amb n. Per $2 \leq i \leq n$:

$$\begin{aligned} \mathbb{P}(\alpha(G) \geq i) &\leq \binom{n}{i}(1-p)^{\binom{i}{}} \leq \\ &\leq \left(n(1-p)^{\frac{i-1}{2}}\right)^i \leq \left(n\exp\left(-p\frac{i-1}{2}\right)\right)^i. \end{aligned} \tag{3.2}$$

Ara escollim $p = n^{\frac{-m}{m+1}}$. Triem una n suficientment gran tal que $n^{\frac{1}{m+1}} \geq 6m\log n$. Per tant, amb $i = \lceil \frac{n}{2m} \rceil$ obtenim que

$$\begin{aligned} \mathbb{P}(\alpha(G) \geq i) &\leq \left(n\exp\left(\frac{-pi}{2} + \frac{p}{2}\right)\right)^i \leq \\ &\leq \left(n\exp\left(-\frac{3}{2}\log n + \frac{1}{2}\right)\right)^i = \left(\left(\frac{\mathrm{e}}{n}\right)^{\frac{1}{2}}\right)^i = o(1), \end{aligned}$$

on la primera desigualtat es una continuació de 3.2, i la segona del fet que $pi \leq 3\log n$. Si escollim un enter n_1 tal que per $n \geq n_1$:

$$\mathbb{P}\left(\alpha(G) \geq \frac{n}{2m}\right) < \frac{1}{2}. \tag{3.3}$$

Ara considerarem la cintura de G. Sigui X una variable aleatòria que conta el nombre de cicles de longitud com a màxim m a $G(n,p)$. Llavors és fàcil comprovar que l'esperança de X és:

$$\mathbb{E}(X) = \sum_{i=3}^{m} \binom{n}{i} \frac{(i-1)!}{2} p^i.$$

D'aquí es segueix que:

$$\mathbb{E}(X) = \sum_{i=3}^{m} \binom{n}{i} \frac{(i-1)!}{2} p^i \leq \frac{1}{2} \sum_{i=3}^{m} n^i p^i \leq \frac{1}{2}(m-2)(np)^m$$

ja que per l'elecció de p, $np \geq 1$. Per la desigualtat de Markov, tenim que

$$\mathbb{P}\left(X \geq \frac{n}{2}\right) \leq \frac{\mathbb{E}(X)}{\frac{n}{2}} \leq (m-2)\frac{(np)^m}{n} = (m-2)n^{-\frac{1}{m+1}} = o(1),$$

on la segona desigualtat ve de l'elecció de p.

Escollim un enter n_2 tal que per els $n \geq n_2$,

$$\mathbb{P}\left(X \geq \frac{n}{2}\right) < \frac{1}{2}. \tag{3.4}$$

Ara per $n \geq n_1 + n_2$, tenim per 3.3 i 3.4:

$$\mathbb{P}\left(\alpha(G) \geq \frac{n}{2m} \text{ o } X \geq \frac{n}{2}\right) < 1. \tag{3.5}$$

Per tant, per 3.5 existeix un graf G' de n vèrtexs amb $\alpha(G') < \frac{n}{2m}$ i $X(G') < \frac{n}{2}$. A G', eliminem un vèrtex de cada cicle que tingui llargada m com a màxim, per obtenir el graf G d'ordre n. Aleshores $g(G) \geq m$, $|V(G) \geq \frac{n}{2}$, i $\alpha(G) < \frac{n}{2m}$. Llavors amb la fòrmula enunciada al principi,

$$\chi(G) \geq \frac{\frac{n}{2}}{\frac{n}{2m}} = m.$$

□

Aquest teorema que acabem d'enunciar i demostrar, ens mostra quin comportament podria seguir un graf $G \in G(n,p)$ respecte la tercera propietat mencionada en el tema anterior, la small world.

En relació a la propietat de la distribució del grau com una llei de potència, ara considerem el grau dels vèrtexs de $G(n,p)$. Utilitzarem una notació més forta que a.q.s. (recordem que la vam definir a 3.4), que serà la de a.e.p., ja que simplifica algunes demostracions.

Definició 3.6. *Direm que un esdeveniment succeeix amb extrema probabilitat (a.e.p.) si succeeix amb una probabilitat d'almenys* $1 - \mathrm{e}^{-\Theta(\log^2 n)}$ *quan* $n \to \infty$.

Observem que si considerem un nombre polinòmic d'esdeveniments tals que cada un es compleixi a.e.p., llavors a.e.p. tots els esdeveniments es compleixen. El següent resultat és la nostra primera exposició cap a un resultat de concentració als grafs aleatoris, indicant que el grau d'un vèrtex és assimptòticament proper al seu grau esperat.

Teorema 3.7. *Si $p \in (0,1)$ és fix, llavors a.e.p. cada vèrtex de $G \in G(n,p)$ té grau igual a:*

$$pn + O(\sqrt{pn}\log n) = (1 + o(1))pn.$$

Demostració: Sigui Y el número de vèrtexs amb grau més gran que $pn + \sqrt{pn}\log n$ o de grau més petit que $pn - \sqrt{np}\log n$. Per provar el teorema, és suficient veure que el número esperat $\mathbb{E}(Y)$ tendeix a 0 més ràpid que la funció $\mathrm{e}^{-c\log^2 n}$ quan $n \to \infty$, per alguna constant $c > 0$. La conclusió segueix a continuació de la desigualtat de Markov:

$$\mathbb{P}(Y \geq t) \leq \frac{\mathbb{E}(Y)}{t}.$$

En efecte,

$$\mathbb{P}(Y = 0) = 1 - \mathbb{P}(Y \geq 1) \geq 1 - \mathbb{E}(Y) > 1 - \mathrm{e}^{-\Theta(\log^2 n)}.$$

Ara considerem una variable aleatòria amb distribució binomial $X \in Bi(n,p)$ amb $\mathbb{E}(X) = np$. A continuació utilitzarem les desigualtats de Chernoff explicades al tema 2.

Farem servir l'estimació a dos bandes, la qual és un corol·lari immediat de les desigualtats anteriors:

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq \delta\mathbb{E}(X)) \leq 2\exp\left(-\frac{1}{3}\delta^2\mathbb{E}(X)\right), \tag{3.6}$$

on $\delta \in [0,1]$.

Fixem un vèrtex arbitrari v de G. Llavors $\mathbb{E}(\deg(v)) = pn - O(1)$. Utilitzant la desigualtat anterior amb $\delta = \frac{1}{2}\log n/\sqrt{\mathbb{E}(\deg(v))} = o(1)$, tenim que

$$\mathbb{P}(|\deg(v) - \mathbb{E}(\deg(v))| \geq \delta\mathbb{E}(\deg(v))) \leq \mathrm{e}^{-\Omega(\log^2 n)},$$

la qual tendeix a 0 més ràpidament que qualsevol polinomi fixat. Es segueix que $\mathbb{E}(Y) \leq n\mathrm{e}^{-\Omega(\log^2 n)} = \mathrm{e}^{-\Omega(\log^2 n)}$. Així, a.e.p. tots els vèrtexs tenen grau $pn + O(\sqrt{pn}\log n)$.

□

Aquest teorema que acabem de veure, es pot estendre amb el cas que p està en funció de n, i la situació seria més complicada. Si $\lim_{n\to\infty} np > 0$ és constant, llavors la distribució del grau de $G(n,p)$ és assimptòticament una Poisson, com indica el Teorema 2.6.

3.3 Variància i mètode del segon moment

La distribució d'una variable aleatòria pot variar molt del valor esperat. L'ús de la variància (o bé mètode del segon moment) ens permet mesurar quant una variable aleatòria es desvia de la seva esperança. A més de l'esperança d'una variable aleatòria X, la variància de X, escrita $\mathrm{Var}(X)$, és de gran importància a la teoria dels grafs aleatoris. Per exemple, gràcies a la desigualtat de Chebyshev, tenim:

Teorema 3.8. *Sigui X una variable aleatòria no negativa i que pren valors enters a $G(n,p)$. Si*

$$\mathrm{Var}(X) = o(\mathbb{E}(X)^2),$$

llavors a.q.s. $X > 0$ i $X \sim \mathbb{E}(X)$.

Demostració: Amb la desigualtat de Chebyshev obtenim que

$$\mathbb{P}(X=0) \leq \mathbb{P}(\{X \geq 2\mathbb{E}(X)\} \cup \{X=0\}) = \mathbb{P}(|X - \mathbb{E}(X)| \geq \mathbb{E}(X)) \leq \frac{\mathrm{Var}(X)}{\mathbb{E}(X)^2},$$

la qual prova que a.q.s. $X > 0$. $\forall \varepsilon > 0$, la desigualtat de Chebyshev implica que

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq \varepsilon \mathbb{E}(X)) \leq \frac{\mathrm{Var}(X)}{\varepsilon^2 \mathbb{E}(X)^2}$$

i per això a.q.s. $X \sim \mathbb{E}(X)$.

□

El teorema 3.8 és un exemple de *resultat de concentració*: la variable aleatòria X i $\mathbb{E}(X)$ estan més propers com més gran sigui la n. Com a exemple més il·lustratiu del mètode del segon moment, considerem el nombre de triangles a $G(n,p)$ per diferents valors de p.

Teorema 3.9. *(1) Si $np = o(1)$, llavors $\lim_{n\to\infty} \mathbb{P}(G \in G(n,p)$ conté un $K_3) = 0$.*

(2) Si $\frac{1}{np} = o(1)$, aleshores $\lim_{n\to\infty} \mathbb{P}(G \in G(n,p)$ conté un $K_3) = 1$.

Demostració: Sigui X una variable aleatòria a $G(n,p)$ que compta el nombre de triangles diferents. Definirem X i X^2 com:

$$X = \sum_A \mathbb{1}_{\{\text{A triangle}\}},$$

$$X^2 = \sum_{A,B} \mathbb{1}_{\{\text{A,B triangles}\}}.$$

Llavors

$$\mathbb{E}(X) = \binom{n}{3} p^3 = O(n^3 p^3). \tag{3.7}$$

Per la condició (1), notem que si $np = o(1)$, llavors per la igualtat que acabem d'enunciar, tenim que $\mathbb{E}(X) = o(1)$. L'afirmació de (1) llavors segueix ràpidament de la desigualtat de Markov

$$\mathbb{P}(X \geq 1) \leq \mathbb{E}(X).$$

Per (2), calculem $\mathrm{Var}(X)$. La feina ve de calcular $\mathbb{E}(X^2)$. Amb la utilització d'indicadors de variables, obtenim

$$\mathbb{E}(X^2) = \sum_{(A,B)} \mathbb{P}(A, B \text{ són triangles}). \tag{3.8}$$

Hi ha tres casos de termes en la suma anterior:

i) A i B tenen 3 elements en comú. Aquest cas es produeix amb una probabilitat de p^3 i es dona $\binom{n}{3}$ cops a la suma anterior.

ii) A i B tenen 2 elements en comú. Aquest cas es produeix amb una probabilitat de p^5 i es dona $6\binom{n}{4}$ cops a la suma anterior. Això significa que donats 4 punts es poden formar 6 combinacions de triangles amb un costat comú.

iii) A i B tenen com a molt 1 element en comú. Aquest cas es produeix amb una probabilitat de p^6 i es dona $\binom{n}{3}^2 - \binom{n}{3} - 6\binom{n}{4}$ cops a la suma anterior.

Amb aquest 3 apunts obtenim:

$$\begin{aligned}\mathbb{E}(X^2) &= \binom{n}{3}p^3 + 6\binom{n}{4}p^5 + p^6\left(\binom{n}{3}^2 - \binom{n}{3} - 6\binom{n}{4}\right)\\ &= \mathbb{E}(X)^2 + O(n^3p^3(1-p^3)) + O(n^4p^5(1-p)).\end{aligned}$$

D'aquí tenim que:

$$\mathrm{Var}(X) = O(n^3p^3(1-p^3)) + O(n^4p^5(1-p)) = o(\mathbb{E}(X)^2).$$

I pel teorema 3.8, a.q.s. $X > 0$.

□

El teorema 3.9 és un fragment d'una classe de resultats molt més gran. Donada una propietat $\mathcal{P}$ dels grafs (la qual pot ser pensada formalment com una classe de grafs tancats sota isomorfisme), diem que una funció real no-zero $t(n)$ és una *funció límit* per $\mathcal{P}$ si:

$$\lim_{n\to\infty} \mathbb{P}(G \in G(n,p) \text{ satisfà } \mathcal{P}) = \begin{cases} 1 & \text{si } \frac{t(n)}{p} = o(1),\\ 0 & \text{si } \frac{p}{t(n)} = o(1).\end{cases}$$

Al teorema 3.9 tenim $t(n) = \frac{1}{n}$. El següent resultat va ser provat per Erdös i Rényi. Un graf G és equilibrat si per tot subgraf G' de G,

$$\frac{|\mathbb{E}(G')|}{|V(G')|} \leq \frac{|E(G)|}{|V(G)|}.$$

Un exemple de graf equilibrat seria un arbre o un cicle.

Podem pensar en $G(n,p)$ com un organisme en creixement, evolucionant des de un graf buit fins a un complet a mesura que p augmenta de 0 a 1. Per exemple, amb $p = \frac{1}{n}$ té lloc un canvi macroscòpic a $G(n,p)$. Quan $p = \frac{c}{n}$ amb $c < 1$, a.q.s. $G \in G(n,p)$ consisteix amb components petites, la més gran de cardinalitat $\Theta(\log n)$. D'altra banda, quan $p = \frac{c}{n}$ amb $c > 1$, emergeix una component gegant de cardinalitat $\Theta(n)$, que absorbeix la majoria de les components més petites. Aquest fenomen remarcable va ser anomenat el *doble salt*, per Erdös i Rényi. Totes aquestes dades que diem en aquest paràgraf es poden trobar a [9].

3.4 Martingales i grafs aleatòris

Les martingales també poden arribar a ser útils en l'estudi dels grafs aleatòris. Una aplicació de les marginales i del mètode de les diferències limitades, explicat a apartats anteriors, als grafs aleatòris podria ser el següent. Shamir i Spencer van provar com el nombre cromàtic de $G(n,p)$ es concentra al voltant de la seva esperança, però no ens diu quin valor és.

Una aplicació útil de la martingala de Doob ve amb la martingala d'exposició d'arestes (o vèrtexs) sobre $G(n,p)$. Fixem la disposició de $t = \binom{n}{2}$ arestes de K_n. Per $1 \leq i \leq t$, sigui Z_i la variable aleatòria de l'indicador de si està l'aresta i-èssima. Sigui $A = f(Z_1, ..., Z_t)$ la funció del graf teòric fixat definit a $G \in G(n,p)$, com ara el número cromàtic de G. Llavors La martingala de Doob $\{X_i : 0 \leq i \leq t\}$ en aquest cas seria la *martingala d'exposició d'arestes*. Al començament no tenim cap informació sobre G. Obtenim més informació a mesura que passa el temps; al temps i sabem quines de les primeres i arestes estan a G.

La *martingala d'exposició de vèrtexs* és semblant, però s'exposa un vèrtex cada vegada. Fixem un ordre lineal de $\{1, ..., n\}$. Sigui Z_i una $(n-i)$-successió de 0's i 1's, indicant si hi ha una aresta entre el vèrtex i i els vèrtexs $j > i$. Llavors la martingala de Doob $\{X_i : 0 \leq i \leq t\}$ en aquest cas s'anomena la martingala d'exposició de vèrtexs.

Teorema 3.10. *Si $G \in G(n,p)$, aleshores $\forall \lambda > 0$ real,*

$$\mathbb{P}(|\chi(G) - \mathbb{E}(\chi(G))| \geq \lambda\sqrt{n}) \leq 2\mathrm{e}^{-\frac{\lambda^2}{2}}.$$

Demostració: Considerem la martingala d'exposició de vèrtexs $\{X_0, ..., X_n\}$ que acabem d'enunciar. Siguin G_i els subgrafs aleatoris de G induïts pel conjunt de vèrtexs $1, ..., i$, sigui $Z_0 = \mathbb{E}(\chi(G))$, i

$$Z_i = \mathbb{E}[\chi(G)|G_1, ..., G_i].$$

Ja que un vèrtex no utilitza més d'un nou color, les variables Z_i són 1-Lipschitz. Per tant, la martingala d'exposició de vèrtexs és 1-Lipschitz. Si apliquem el teorema 2.23, ja haurem acabat amb la demostració. Aquest resultat es compleix encara que no es conegui $\mathbb{E}(\chi(G))$.

□

Observem que el teorema anterior va ser el primer en utilitzar martingales dins la teoria de grafs.

Teorema 3.11. *A.q.s. per $G \in G(n,p)$, tenim que*

$$\mathbb{E}(\chi(G)) = (1 + o(1))\frac{n}{2\log\frac{1}{1-p}n}.$$

Es saben més coses del nombre cromàtic dels grafs aleatoris. Per exemple, Alon i Krivelevich van provar que per tota constant $c > 0$, el nombre cromàtic de $G(n,p)$ amb $p = n^{-\frac{1}{2}-c}$ és a.q.s. concentrat en dos valors. Per si s'està més interessat en aquest tema, el resultat es troba a [2].

4 Models per el Graf Web

4.1 Introducció

Deixarem ara el model de la Web que ens proporciona $G(n,p)$ i introduirem alguns models que s'han demostrat que sí que compleixen les propietats de l'apartat 2.3. Els bons models per W solen tenir pocs paràmetres que siguin importants per W. L'equilibri per saber quin dels paràmetres són més representatius fa que la modelització de W sigui dura i complicada.

L'aproximació que explicarem en aquest capítol será mitjançant *models estocàstics*. Això significa que els nostres grafs es generaran sobre una successió infinita de temps discret amb una llei de probabilitat. El repte actual és dissenyar un model matemàtic rigurós que pugui simular algunes característiques de W. Per analitzar aquests models necessitarem tècniques del tema anterior i de noves.

No hi ha una resposta clara a quines són les característiques que fan bo un model, només un consens basat en les propietats observades de W. A continuació recordarem ràpidament tres de les propietats desitjables que hauria de complir el model i que ja vam avançar al capítol 2:

1. *Propietat on-line* El graf generat pel model haurá de canviar a mesura que passi el temps, tant el nombre d'arestes com el de vèrtexs.
2. *Distribució del grau amb llei de potència.* A.q.s. la distribució del grau dels grafs generats pels models segueix una llei de potència amb exponent $\beta > 2$
3. Propietat *Small world* . El model a.q.s. genera grafs dispersos amb diàmetres i distàncies mitges "baixes". Per exemple, el diàmetre hauria de ser a.q.s. $\log t$ aproximadament si hi ha t vèrtexs, mentre que la distància mitja hauria de ser $\log(\log t)$ aproximadament.

4.2 Models On-line pel Graf Web

Ens centrarem en els models on-line per descriure W, que són els que tenen un número variable de vèrtexs a mesura que passa el temps. La idea general per tots els models és considerar tant el comportament assimptòtic com l'aproximació dels resultats. La justificació per aquestes dos consideracions és que W és un graf gegant amb una gran quantitat de vèrtexs i arestes, i de mitjana, petits canvis marquen la diferència a l'estructura de tot el graf. Les tècniques aleatòries assimptòtiques són les més adients per analitzar aquests models.

Els models sempre tindran un conjunt finit de paràmetres reals, i tindran un graf H finit fixat com a paràmetre addicional. El model generarà per algun procés de grafs aleatòris una successió de grafs G_t finits per $\{t : t \in \mathbb{N}\}$. Si no s'indica el contrari, $\forall t \in \mathbb{N}$, tenim:

1. $G_0 \cong H$.
2. G_t és un subgraf induït de G_{t+1}.
3. $|V(G_{t+1})| = |V(G_t)| + 1$

En tots els models, G_t estarà definit inductivament. En el pas inductiu, l'únic vèrtex a $V(G_{t+1})\backslash V(G_t)$ serà el *nou vèrtex*, i l'escriurem v_{t+1}. Els vèrtexs de $V(G_t)$ seran els *vèrtexs existents.* Generalment l'elecció dels paràmetres afecta a l'exponent β de la llei de potència. A més, el nombre d'arestes a $E(G_{t+1})\backslash E(G_t)$ és constant. Tot i que existeixen casos on el número d'arestes i vèrtexs de G_t són variables aleatòries.

Per enters $k, t \geq 0$, definim $N_{k,t}$ com el número de vèrtexs amb grau k al moment t. Per tant $N_{k,t}$ és una variable aleatòria. Ja que normalment $|V(G_t)|$ és aproximadament t, sovint s'estudia la proporció $\frac{N_{k,t}}{t}$. La tècnica més freqüent per provar que $\frac{N_{k,t}}{t}$ segueix una llei de potència implica primer el càlcul de $\mathbb{E}\left(\frac{N_{k,t}}{t}\right)$, i després provar que $\frac{N_{k,t}}{t}$ no es desvia molt lluny de $\mathbb{E}\left(\frac{N_{k,t}}{t}\right)$. Això és, veurem que la variable aleatòria *es concentra* al seu valor esperat.

4.2.1 Models d'adherència preferents.

Sens dubte els models més importants del graf web són aquells que incorporen adherència preferent. La idea darrera aquests models és que el nou vèrtex és més propens a unir-se als vèrtexs existents amb grau més alt. Ens referirem a aquests models com: *models d'adherència preferents o models PA*, (de l'anglès *preferential attachment*).

4.2.2 El model LCD PA

El primer anàlisi rigurós d'un model PA va ser utilitzant el *diagrama de cordes linealitzat o model LCD*, (de l'anglès Linearized Chord Diagram), ja que una formulació equivalent del model és mitjançant aparellament aleatoris a un conjunt finit i constant d'enters. L'únic paràmetre d'aquest model és un enter m, on H és una còpia de K_1 amb un loop. Primer descriurem aquest model pel cas $m = 1$. Per formar G_t des de G_{t-1} afegim una aresta des de v_t a v_i, on v_i serà triat de forma aleatòria a partir dels vèrtexs existents de la forma:

$$\mathbb{P}(i = s) = \begin{cases} \frac{deg_{G_{t-1}}(v_s)}{2t-1} & \text{si } 1 \leq s \leq t-1, \\ \frac{1}{2t-1} & \text{si } s = t. \end{cases} \tag{4.1}$$

Notem que d'aquesta manera, un vèrtex que tingui grau elevat té més probabilitat d'adquirir una nova aresta. El graf G_t no conté cicles no-trivials, tot i que pot contenir loops propis.

Si $m > 1$, definim el procés $\{G_t^m : t \geq 0\}$ primer generant la successió $\{G_t : t \in \mathbb{N}\}$ de grafs utilitzant el cas $m = 1$ a la successió de vèrtexs $\{v_i' : i \in \mathbb{N}\backslash\{0\}\}$. El graf G_t^m es forma gràcies al G_{mt} identificant els vèrtexs $v'_{(i-1)m+1}, ..., v'_{im}$ per formar v_i.

Es defineix el LCD d'ordre t com una partició de $\{1, ..., 2t\}$ cap a n conjunts de diferents parelles. Llavors hi ha $(2t-1)!! = \frac{(2t)!}{t!2^t}$ combinacions de LCD a $[2t]$.

Podem pensar un LCD amb t cordes semi-circulars entre els $2n$ diferents punts que estaran sobre l'eix de les abscisses del semiplà de $\mathbb{R}^2$. Cada corda tindrà un punt final tant esquerra com dret. Per cada LCD L podem formar un graf no dirigit $G(L)$ d'ordre t. Per formar v_1, considerarem des del primer punt que serà un punt final esquerra fins al primer punt que tingui el final dret d'una corda, encara que no sigui el final de primera corda que hem considerat al primer punt. Tots aquests punts compresos entre el final esquerra i el final dret s'identificaran en un sol vèrtex v_1. Les cordes dels diferents punts

es convertiran en les arestes de v_1. Un cop format el primer vèrtex, s'ha de repetir el procés inductivament. Si un punt és el final esquerra i dret alhora, s'identifica amb ell mateix i es passa al següent pas.

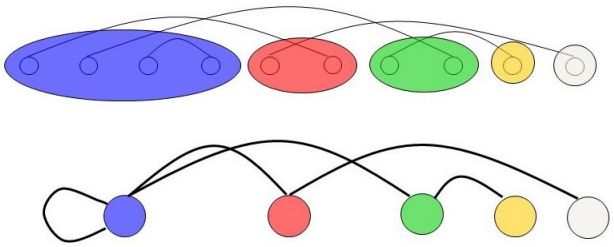

Figura 5: Esquema del model LCD.

La relació entre els LCD i el model LCD prové del següent teorema.

Teorema 4.1. *Sigui m=1. Suposem que una LCD L es tria u.a.r. de les* $\frac{(2t)!}{t!2^t}$ *combinacions de LCD a* $[2t]$*. Llavors la probabilitat amb la que els vèrtex* v_i *i* v_s *estiguin units és:*

$$\mathbb{P}(i=s) = \begin{cases} \frac{deg_{G_{t-1}}(v_s)}{2t-1} & si\ 1 \leq s \leq t-1, \\ \frac{1}{2t-1} & si\ s = t. \end{cases}$$

És la mateixa distribució de probabilitat que hem enunciat fa un moment.

Els autors d'aquest model utilitzen el següent teorema:

Teorema 4.2. *Al model LCD, fixem un enter positiu m i un* $\varepsilon > 0$ *i considerem k un enter no negatiu, i definim:*

$$\alpha_{m,k} = \frac{2m(m+1)}{(k+m)(k+m+1)(k+m+2)}.$$

Llavors a.q.s. per tot k que compleixi $0 \leq k \leq t^{\frac{1}{15}}$*,*

$$(1-\varepsilon)\alpha_{m,k} \leq \frac{N_{k,t}}{t} \leq (1+\varepsilon)\alpha_{m,k}. \tag{4.2}$$

Aquest teorema demostra que per una t molt gran, la distribució del grau dels grafs al model LCD segueix una llei de potència amb exponent $\beta = 3$ amb una probabilitat molt elevada. Un problema que podem trobar en aquest teorema és que la k no és un número qualsevol, sinó que està entre 0 i $t^{\frac{1}{15}}$, els dos inclosos. La solució a aquest inconvenient es troba a [5]. Allà trobarem una prova de que el grau k es pot extendre per números més grans que $t^{\frac{1}{15}}$. Una observació important és que $\beta = 3$ és independent de m. Per tant, compliria la segona condició per a ser un bon model per W.

La demostració del teorema anterior es complica molt a l'hora d'estimar $\mathbb{E}(N_{k,t})$ per a que sigui $\alpha_{m,k}t$. Així que en comptes de donar una demostració del Teorema 4.2, provarem un resultat similar.

Teorema 4.3. *Per tot k que satisfaci $0 \leq k \leq t^{\frac{1}{15}}$, la successió $\{\frac{N_{k,t}}{t} : t \in \mathbb{N}\}$ convergeix en probabilitat a $\mathbb{E}(\frac{N_{k,t}}{t})$.*

Donarem una demostració pel cas $m = 1$, la demostració del cas general està a [5]. Utilitzarem el mètode de les diferències limitades, descrit al capítol dels Grafs aleatòris.

Demostració: Definim una martingala de Doob considerant com a $A = N_{k,t}$ i per $Z_i = G_i$, considerem els primers grafs de la successió com a variables aleatòries. Com que un vèrtex nou pot afectar als graus de com a molt 2 vèrtexs ja existents, podem afirmar que per a $1 \leq i \leq t$:

$$|X_i - X_{i-1}| \leq 2. \tag{4.3}$$

Això és veritat perquè si al període i, el vèrtex v_i s'uneix als vèrtexs v_r o v_s, no afecta als futurs graus que pot tenir un tercer vèrtex v_u (amb u diferent de s i r). La distribució conjunta de tots els altres graus és la mateixa en cada cas. A mesura que es conten els vèrtexs d'un cert grau, encara que els graus de v_r i v_s canviïn, la desigualtat 4.3 es manté. Ara recuperem un teorema del capítol anterior. Podem aplicar el teorema 2.23 amb $\lambda = \sqrt{\log t}$ i $c = 2$:

$$\mathbb{P}(|X_t - X_0| \geq \sqrt{\log t}\sqrt{t}) \leq 2\mathrm{e}^{-\frac{\log t}{2*2^2}} = 2\left(\mathrm{e}^{\log t}\right)^{-\frac{1}{8}} = 2t^{-\frac{1}{8}}. \tag{4.4}$$

Per tant, de 4.4 tenim la següent desigualtat:

$$\mathbb{P}\left(|\frac{X_t}{t} - \frac{X_0}{t}| \geq \left(\frac{\log t}{t}\right)^{\frac{1}{2}}\right) \leq 2t^{-\frac{1}{8}},$$

amb $X_0 = \mathbb{E}(N_{k,t})$ i $X_t = N_{k,t}$. Com que $\frac{\log t}{t}$ tendeix a 0 quan $t \to \infty$, donat $\varepsilon > 0$ qualsevol, per a un t suficientment gran tenim que:

$$\mathbb{P}\left(|\frac{X_t}{t} - \frac{X_0}{t}| \geq \varepsilon\right) \leq \mathbb{P}\left(|\frac{X_t}{t} - \frac{X_0}{t}| \geq \left(\frac{\log t}{t}\right)^{\frac{1}{2}}\right) \overset{t\to\infty}{\longrightarrow} 0$$

Per tant, X_t convergeix cap a X_0 en probabilitat.

□

Ara només falta comprovar l'última de les propietats que ha de complir un bon model del graf W, la propietat de *small world*. La demostració del següent teorema és molt complicada i per això no la veurem aquí. En el cas d'estar-hi interessats, es pot trobar a [4].

Teorema 4.4. *Sigui $m \geq 2$ un enter fix i un nombre real positiu ε. Aleshores a.q.s. G_t^m satisfà*

$$(1-\varepsilon)\frac{\log t}{\log(\log t)} \leq \mathrm{diam}(G_t^m) \leq (1+\varepsilon)\frac{\log t}{\log(\log t)}.$$

Com abans, el resultat d'aquest teorema no depèn de m. En el cas de $m = 1$, la cota superior no es compleix generalment.

Per tant, acabem de veure que el model LCD compleix les tres propietats principals que hem dit al principi d'aquest capítol: és on-line ja que cada pas es van afegint vèrtexs i arestes, compleix que $\beta > 2$ pel Teorema 4.2, i per últim satisfà la propietat de small world pel Teorema 4.4.

5 Conclusions

Com a tram final d'aquest treball, ens dedicarem a extreure les conclusions del que hem anat veient al llarg de l'estudi. Com podríem haver vist a primera instància, els grafs aleatoris clàssics G(n,p), definits per Erdös i Rényi, no satisfan les propietats que hem definit al principi i que hauria de seguir un model del graf W.

Primerament, per definició de $G(n,p)$, si seguim aquest model estem fixant un nombre n de vèrtexs, fet que contradiu totalment la primera de les propietats descrites, la de ser on-line. Per molt gran que considerem que sigui n, la Web és comporta com un organisme viu, el qual successivament es van afegint noves pàgines. Per tant, considerar un model off-line o estàtic com aquest podria no ser el més adequat.

Seguidament, si ens fixem en la segona propietat, la distribució del grau amb llei de potència, el model de $G(n,p)$ es comporta com una distribució binomial (o, al límit, una distribució de Poisson). Llavors, de mitjana, totes les pàgines tenen segons aquest model un nombre similar d'enllaços. No obstant, estudis empírics suggereixen que això no és així sinó que, com hem mencionat anteriorment en aquest treball, el grau segueixi una distribució amb llei de potència $k^{-\beta}$, on β és d'ordre més gran o igual 2. Aquest fet, el que implicaria és que hi hauria una quantitat molt reduïda de pàgines amb molts enllaços, alhora que una gran majoria de pàgines amb pocs links. Tot i ser la minoria de les pàgines, les que tenen molts links són les més significatives. Aquest comportament no l'observem en el model de $G(n,p)$, per tant segons estudis empírics això podria produir un model no prou acurat.

Finalment, acabem amb la últma propietat, la small world. Per comprovar si el model dels grafs $G(n,p)$ la compleix o no, ens hem de remuntar fins el teorema 3.5. El resultat d'aquest ens indica que tampoc satisfà amb la propietat small world. Per tant, al no complir-la, segons comprovacions empíriques, el model podria generar una aproximació no realista del comportament de la Web.

En conclusió, els grafs aleatoris clàssics $G(n,p)$ no satisfan propietats observades de W, tot i que tenen distribucions binomial del grau i resulta interessant estudiar el seu comportament. A més a més, ens proporcionen un context matemàtic pels nous mètodes de modelització del graf W.

Per acabar el treball, hem fet un breu anàlisi del model LCD PA. Aquest és un model on-line per definició, i com mostren els teoremes 4.2 i 4.4, satisfà les propietats de distribució del grau com a llei de potència $\beta > 2$ i small world, respectivament. Aleshores, segons els estudis empírics, podríem concloure que si és un model representatiu de la Web, a diferència del model de $G(n,p)$.

Com a aspectes a seguir estudiant en un futur treball, podríem mencionar: realitzar un anàlisi rigurós del model LCD; estudiar altres models per W mitjançant l'adherència preferent (PA) diferent al LCD, com podria ser el model ACL PA; considerar models on-line que no només afegeixin vèrtexs i/o arestes, sinó que també en puguin eliminar; i per acabar, si consideréssim models amb grafs dirigits seria un punt a favor per la modelització de grafs dins l'estudi de W, on es podrien generar models més acurats, ja que a la realitat les relacions entre pàgines no solen ser simètriques.

Referències

[1] R. Albert, H. Jeong, A. Barabási, Diameter of the world-wide web, *Nature* **401** (1999).

[2] N. Alon, M.Krivelevich, The concentration of the chromatic number of random graphs, *Combinatorica* **17** (1997).

[3] A. Bonato: *A Course on the Web Graph*, Graduate Studies in Mathematics, Volume 89, American Mathematical Society (2008).

[4] B.Bollobás, O. Riordan, The diameter of a scale-free random graph, *Combinatorica* **24** (2004).

[5] B.Bollobás, O. Riordan, J. Spencer, G. Tusnády, The degree sequence of a scale-free random graph process, *Random Structures and Algorithms* **1**8 (2001).

[6] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R.Stata, A. Tomkins, J. Wiener: *Graph structure in the web*, Computer Networks **33** (2000).

[7] Z. Brzeźniak, T. Zastawniak *Basic Stochastic Processes*, Springer Undergraduate Mathematics Series, (2002).

[8] F. Comellas, J. Fàbrega, A. Sànchez, O. Serra: *Matemática discreta*, Edicions UPC, (2001).

[9] P. Erdös, A. Rényi, On random graphs I, *Publicationes Mathematicae Debrecen* **6** (1959).

[10] P. Fernández, *El secreto de Google i el álgebra lineal*, (2004).

[11] M. Mitzenmacher, E. Upfal: *Probability and Computing. Randomized algorithms and probabilistic analysis* Cambridge University Press, (2005).