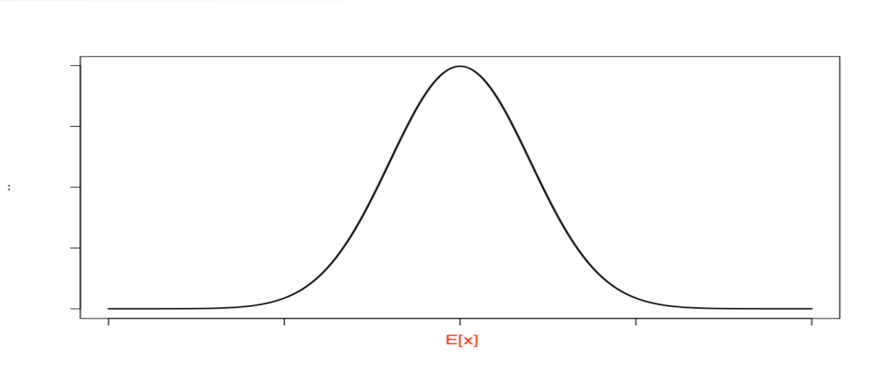
Questioni di popolarità

Il materiale esposto in questi lucidi corrisponde al Cap. 18 del testo

Ricapitolando

- Abbiamo introdotto un modello di grafi aleatori – il modello di Erdös-Renyi
- E abbiamo visto che, in un grafo generato in accordo a tale modello, al crescere di k, il numero di nodi che hanno grado k decresce molto velocemente – decresce esponenzialmente in k
- Cerchiamo di capire perché la decrescita è così veloce…
- **Teorema del limite centrale** (*enunciato intuitivo e informale*): la somma di (molti) valori aleatori indipendenti si distribuisce approssimativamente come una distribuzione normale (o gaussiana) intorno al suo valore atteso



... che decresce molto velocemente quando ci si allontana dal valore atteso

Ricapitolando

- Che c'entra il Teorema del limite centrale?!
- Nel grafo $G_{n,p}$ gli archi vengono aggiunti come una sequenza di oggetti, indipendentemente gli uni dagli altri
- Perciò, il grado di un nodo è la somma di eventi indipendenti:

$$\delta_i = \sum_{j \in [n]-i} e_{ij}$$

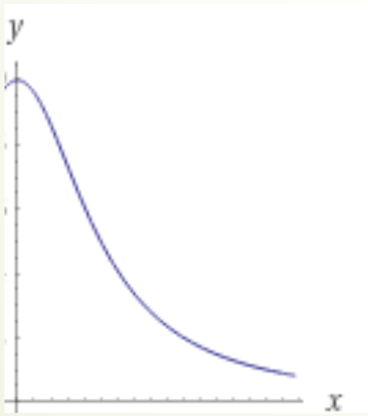
  - dove e_{ij} è la variabile aleatoria che vale 1 se esiste l'arco (i,j), 0 altrimenti
- ci troviamo, allora, nella situazione illustrata dal Teorema del limite centrale
- Il punto cruciale è che *la porzione di grafo che già è stata costruita (gli archi che già sono stati inseriti in $G_{n,p}$) non ha alcuna influenza sul prossimo arco che verrà creato*

Ma nella realtà…

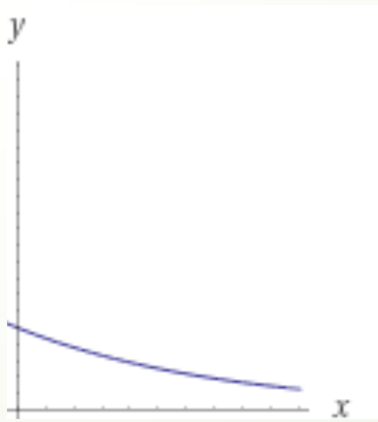
- Consideriamo il grafo orientato del web: i nodi sono le pagine e, se la pagina a contiene un hyperlink alla pagina b, allora (a,b) è un arco (diretto) del grafo
- È stato condotto uno studio [Broder et al., 2000] al fine di analizzare la distribuzione dei gradi (archi entranti) nei nodi del grafo del web
- Quel che è stato osservato differisce ***sostanzialmente*** da quanto indicato nel Teorema del limite centrale: *la frazione di pagine web che ha grado entrante k è proporzionale a $1/k^c$ – per qualche **costante** c - invece che a $1/k^k$*
  - una funzione che decresce come l'inverso di un polinomio è chiamata ***power law***
- ossia*, il numero di nodi del grafo del web che hanno grado entrante elevato è molto maggiore di quello che ci si aspetterebbe assumendo che gli archi si formino indipendentemente gli uni dagli altri*
- Ma perché?!

Riconoscere una Power Law

- È stato osservato che la frazione di pagine web che ha grado entrante k è proporzionale a $1/k^c$ – per qualche costante c - invece che a $1/k^k$
- Prima di cercare di capire le ragioni della power law, cerchiamo di capire come hanno fatto, gli studiosi che si sono occupati della questione, ad accorgersi di questo fenomeno
  - ossia, partendo da un grafico (discreto) di un grafo che, per ogni intero k in ascissa riporta in ordinata il numero di nodi del grafo che hanno grado k
  - come fai ad accorgerti che quella funzione decresce come l'inverso di un polinomio invece che come l'inverso di una esponenziale?!



(a)

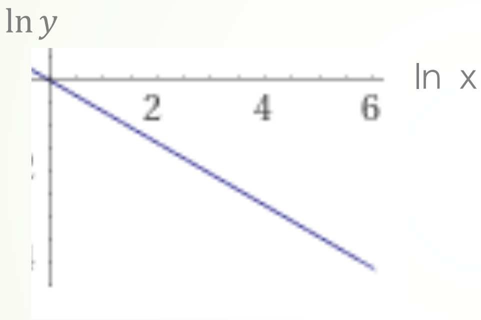


???

(b)

Riconoscere una Power Law

- Facile: è sufficiente considerare il grafico log-log
  - un grafico i cui assi rappresentano ln x e ln y
  - e invece di rappresentare y = f(k), rappresentiamo **ln y = ln f(k)**
    - dove con f(k) indichiamo la funzione che esprime il numero di nodi di grado k



(a)

- In questo modo, se f(k) = $1/k^c$, il grafo sarà, grosso modo una retta:
  - ln y ≈ - c ln k
  - infatti: (a) è y = $\frac{1}{(x^2+1)}$
- Possiamo ora tornare a cercare di comprendere perché si verifica la Power Law

Power law e fenomeno rich get richer

- Supponiamo di stare a costruire una pagina web.
  - ad esempio, una pagina in cui descriviamo un certo modello generativo di grafi aleatori
  - mentre scriviamo, ci accorgiamo che è opportuno inserire un hyperlink a qualche pagina che descrive un altro modello generativo
- Quale pagina scegliamo di puntare dalla nostra pagina?
  - Ce ne saranno a bizzeffe che trattano quell'argomento!
- Beh, fra tutte quelle pagine, scegliamo la pagina che ci sembra più autorevole…
- E come facciamo a capire l'autorevolezza di una pagina?
- Beh, ad esempio perché facendo una ricerca su un motore di ricerca, quella pagina appare in prima posizione
  - e, come vedremo, questo accade quando quella pagina ha tanti link entranti
- Cioè, **aumentiamo il grado di pagine che hanno già un grado elevato!**
  - **E scegliamo quale arco aggiungere sulla base degli archi già presenti**

Power law e fenomeno rich get richer

- Nel creare pagine web **aumentiamo il grado di pagine che hanno già un grado elevato!**
- Un fenomeno analogo si verifica nelle reti sociali
  - la rete ci mostra i i post dei nostri contatti
  - e quando i nostri contatti commentano post di pagine che loro seguono – magari blog di cucina (per dire), o di politica, o di libri…
  - anche in noi può nascere l'interesse a seguire quelle pagine
  - che diventano sempre più popolari!!
- Oppure, possiamo pensare alle pagine di persone "famose"
  - divi del cinema o dello spettacolo
  - personaggi politici
  - …
- I nodi corrispondenti a queste entità hanno tutti grado molto elevato
- E il numero di queste entità non è certo irrilevante

Un modello per la Power Law

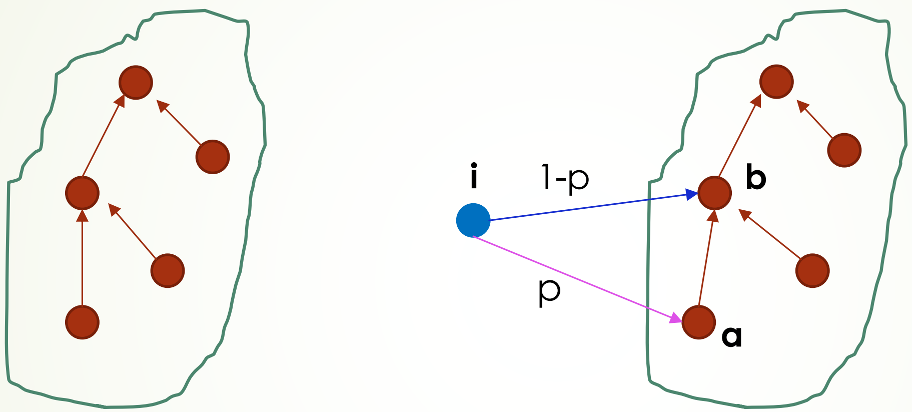
- Nel creare pagine web **scegliamo quale arco aggiungere sulla base degli archi già presenti**
- Vogliamo, a questo punto, definire un modello generativo di grafi nel quale, a differenza del modello di Erdös-Renyi, l'aggiunta di un nuovo arco dipenda dagli archi già presenti nel grafo
  - e che tale modello esibisca una power law!
- Il nostro modello sarà basato sulla semplice assunzione che ***gli individui tendono a copiare il comportamento degli altri individui***
  - che abbiamo già visto valere nel processo di creazione delle pagine web e nelle reti sociali
  - e che vedremo diffusamente più avanti quando parleremo di herding
- Dunque, consideriamo un processo di creazione di un grafo che ricorda il meccanismo di creazione delle pagine web:
  - le pagine vengono create una alla volta – ossia, ***in sequenza***
  - quando viene creata una pagina, si decide a quali pagine essa debba puntare
  - ciascun puntatore è un arco diretto

Un modello per la Power Law

- Come nel modello di Erdös-Renyi, fissiamo un parametro $p \in [0,1]$
- I nodi vengono aggiunti al grafo in time-step discreti
  - al passo 1 viene creato il nodo 1
  - al passo 2 vengono creati il nodo 2 e l'arco (diretto) (2,1)
  - finita questa fase di inizializzazione, i nodi e gli archi vengono creati in accordo alla seguente regola:
    - al passo i vengono creati il nodo i e un arco uscente da i: viene scelto uniformemente a caso un nodo $a < i$ e
      - con probabilità p viene creato l'arco (i,a)
      - con probabilità 1-p viene creato l'arco (i,b), dove (a,b) è l'arco uscente da a
- Intuitivamente: man mano che il numero di archi entranti in un nodo aumenta, cresce la probabilità che quel nodo venga selezionato come estremo di un arco uscente da un nodo di nuova creazione
  - Lo stesso fenomeno che abbiamo osservato nel web
  - Abbiamo costruito un modello che descrive il fenomeno **Rich Get Richer**
- Si osservi che ogni nodo $i > 1$ ha esattamente un arco uscente

Un modello per la Power Law

Vediamo graficamente cosa accade quando il nodo i viene aggiunto al grafo



grafo al passo i-1

grafo al passo i

alla creazione del nodo i:

- viene scelto (uniformemente a caso il nodo a

- con probabilità p viene creato l'arco (i,a)

- con probabilità 1-p viene creato l'arco (i,b)

Il modello generativo che abbiamo descritto è stato proposto da A.L. Barabási e R. Albert nel 1999 ed è perciò noto come ***modello di Barabasi-Albert***

Un modello per la Power Law

- Il passo iterativo del processo di costruzione del grafo che abbiamo appena introdotto può essere anche descritto in un altro modo
- Dunque, fissiamo un parametro $p \in [0,1]$
- e aggiungiamo i nodi al grafo in time-step discreti
  - al passo 1 viene creato il nodo 1
  - al passo 2 vengono creati il nodo 2 e l'arco (diretto) (2,1)
  - finita questa fase di inizializzazione, i nodi e gli archi vengono creati in accordo alla seguente regola:
    - al passo i vengono creati il nodo i e un arco uscente da i
      - con probabilità p viene scelto uniformemente a caso un nodo $a < i$ e viene creato l'arco (i,a)
      - con probabilità 1-p viene scelto uniformemente a caso un nodo $a < i$ e, detto (a,b) l'arco uscente da a, viene creato l'arco (i,b)
- è facile convincersi che le due descrizioni sottintendono lo stesso modello di grafo
  - e questo secondo modo ci sarà utile nella formalizzazione del modello

Un modello per la Power Law

- Abbiamo costruito un modello che descrive il fenomeno **Rich Get Richer**
- Resta da verificare se i grafi generati in accordo a questo modello esibiscono una Power Law
  - ossia, che la funzione che descrive il numero atteso di nodi di grado k si comporta come l'inverso di un polinomio
- Per farlo, occorre formalizzare: per ogni coppia di interi i e j tali che i > j e i ≥ 2, introduciamo la variabile aleatoria

$$d_{ij} = \begin{cases} 1 & \text{se } (i,j) \in E \\ 0 & \text{altrimenti} \end{cases}$$

- Allora, in accordo alle regole che abbiamo descritto
  - $P(d_{ij}=1)$ = p · probabilità che venga scelto il nodo j + (1-p) · probabilità che venga scelto un nodo h tale che $d_{hj}=1$

$$= \frac{p}{i-1} + (1-p)\, P\left(\bigcup_{h<i:\,(h,j)\in E} \{\text{viene scelto h}\} \right)$$

$$= \frac{p}{i-1} + (1-p) \sum_{1 \le h < i:\,(h,j)\in E} P(\text{ viene scelto h }) \quad \textit{perché eventi disgiunti}$$

$$= \frac{p}{i-1} + (1-p) \sum_{1 \le h < i:\,(h,j)\in E} \frac{1}{i-1} = \frac{p}{i-1} + \frac{1-p}{i-1} \sum_{1 \le h < i:\,(h,j)\in E} 1 = \frac{p}{i-1} + \frac{1-p}{i-1} \sum_{1 \le h < i} d_{hj}$$

Un modello per la Power Law

- **P((i,j) ∈ E) = P(d $_{ij}$ =1) = $\frac{p}{i-1} + \frac{1-p}{i-1} \sum_{1 \le h < i} d_{hj}$**
- Osservazione 1: ha senso calcolare P((i,j) ∈ E) per i > 1 (il nodo 1 non ha archi uscenti!)
- Osservazione 2: ogni nodo i > 1 ha esattamente un arco uscente. Quindi, deve essere P(∃ j < i: (i,j) ∈ E) = 1. Infatti:
  - P(∃ j < i: (i,j) ∈ E) = $\sum_{1 \le j < i}$ P((i,j) ∈ E) e, procedendo per induzione:
  - per i =2: $\sum_{1 \le j < 2}$ P((2, j) ∈ E) = P((2, 1) ∈ E)= 1 per costruzione
  - assumendo $\sum_{1 \le j < a}$ P((a, j) ∈ E) = 1 per ogni a ≤ i-1, si ha che

$$\sum_{1 \le j < i} P((i,j) \in E) = \sum_{1 \le j < i} \left[\frac{p}{i-1} + \frac{1-p}{i-1} \sum_{1 \le h < i} d_{hj}\right]$$

$$= \sum_{1 \le j < i} \frac{p}{i-1} + \sum_{1 \le j < i} \frac{1-p}{i-1} \sum_{1 \le h < i} d_{hj}$$

$$= \frac{(i-1)\,p}{i-1} + \frac{1-p}{i-1} \sum_{1 \le h < i} \sum_{1 \le j < i} d_{hj} =$$

per l'ipotesi induttiva, ogni h < i ha esattamente un arco uscente e quindi $\sum_{1 \le j < i} d_{hj} = 1$

$$= p + \frac{1-p}{i-1} \sum_{1 \le h < i} 1 = p + (1-p) = 1$$

Un modello per la Power Law

- **P((i,j) ∈ E) = P(d_{ij}=1) = $\frac{p}{i-1} + \frac{1-p}{i-1} \sum_{1 \le h < i} d_{hj}$**

- Osservazione 3. Possiamo così interpretare la regola per costruire il grafo:

  il nodo j a cui connettere i è scelto
  uniformemente a caso con probabilità p,
  con probabilità proporzionale al grado di j con probabilità (1-p)

  - che esprime chiaramente il fenomeno Rich get Richer
  - (ed è il terzo modo di descrivere il modello)

Dimostriamo la Power Law

- Siamo pronti a dimostrare che il modello che abbiamo definito esibisce una Power Law.

- La dimostrazione procede attraverso i seguenti passi intermedi:
- 1) definizione della legge aleatoria che esprime la variazione del grado di un nodo nel tempo
- 2) *approssimazione* deterministica e continua della legge al punto 1), che porterà ad una equazione differenziale
- 3) risoluzione dell'equazione differenziale: la soluzione costituirà un'approssimazione della funzione che esprime il grado di un nodo nel tempo
- 4) individuazione della Power Law

1) Legge aleatoria per la variazione del grado

- Sia $D_j(t)$ la variabile aleatoria che esprime il numero di archi entranti nel nodo j al passo t di generazione del grafo
  - naturalmente, la variabile $D_j(t)$ è definita per $t \geq j$
  - e per ogni $j \geq 1$
- Al passo t=j, il grado entrante di j è 0: $D_j(j)= 0$
  - ciò esprime la condizione iniziale della variabile
- Esprimiamo ora la variazione attesa nel tempo della variabile $D_j(t)$:
  - al passo t+1, il grado di j può essere invariato rispetto al passo t oppure può essere aumentato di una unità, e
  - $D_j(t+1) = D_j(t) +1$ se e solo se è stato creato l'arco (t+1, j)
- quindi,

$$P(D_j(t+1) - D_j(t) =1) = P(d_{t+1\,j} =1) = \frac{p}{t} + \frac{1-p}{t} \sum_{1 \leq h < t+1} d_{hj} = \frac{p}{t} + \frac{1-p}{t} D_j(t)$$

- Ossia: $\mathbf{P(D_j(t+1) - D_j(t) =1) = \frac{p}{t} + \frac{1-p}{t} D_j(t)}$

2) Approssimazione deterministica e continua

- **$P(D_j(t+1) - D_j(t) = 1) = \frac{p}{t} + \frac{1-p}{t} D_j(t)$** **con condizione iniziale $D_j(j)=0$**
- Per ogni $j \geq 1$ definiamo, innanzi tutto, una funzione <u>discreta deterministica</u> $X_j(t)$ che "assomiglia" a $D_j(t)$
  - la funzione $X_j(t)$ è definita per $t \geq j$ e
  - **$X_j(j) = 0$**
  - **$X_j(t+1) - X_j(t) = \frac{p}{t} + \frac{1-p}{t} X_j(t)$**
- A questo punto, approssimiamo il comportamento di $X_j(t)$ con una funzione $x_j(t)$, definita su un dominio continuo, e ancora definita per $t \geq j$:
  - **$x_j(j) = 0$**
  - **$\frac{d}{dt} x_j(t) = \frac{p}{t} + \frac{1-p}{t} x_j(t)$**
  - che è un'equazione differenziale
- **non è detto che l'andamento della funzione $x_j(t)$ sia effettivamente vicino al comportamento della variabile aleatoria $D_j(t)$** – ma le analogie fra le due sono forti
  - Vedremo più avanti quanto le due si assomiglino

3) Risoluzione dell'equazione differenziale

- Dobbiamo risolvere l'equazione differenziale $\frac{d}{dt}\, x_j(t) = \frac{p}{t} + \frac{1-p}{t}\, x_j(t)$
- ossia $\frac{d}{dt}\, x_j(t) = \frac{1}{t}\,[\, p + (1\text{-}p)\, x_j(t)\,]$
- ossia $\frac{1}{p+(1-p)\, x_j(t)}\, \frac{d\, x_j(t)}{dt} = \frac{1}{t}$
- integrando: $\int \frac{1}{p+(1-p)\, x_j(t)}\, \frac{d\, x_j(t)}{dt}\, dt = \int \frac{1}{t}\, dt$
- ossia $\int \frac{d\, x_j(t)}{p+(1-p)\, x_j(t)} = \int \frac{1}{t}\, dt$ da cui, moltiplicando ambo i membri per (1-p)

$\int \frac{(1-p)\, d\, x_j(t)}{p+(1-p)\, x_j(t)} = (1\text{-}p) \int \frac{1}{t}\, dt$ sostituiamo: y = p+(1-p) $x_j(t)$, da cui dy = (1-p) d $x_j(t)$

$\int \frac{dy}{y} = (1\text{-}p) \int \frac{1}{t}\, dt$

3) Risoluzione dell'equazione differenziale

- ossia $\int \frac{d\,[p+(1-p)\,x_j(t)\,]}{p+(1-p)\,x_j(t)}$ = (1-p) $\int \frac{\mathbf{1}}{\mathbf{t}}$ dt
- otteniamo **ln [p + (1-p) x $_j$(t)] = (1-p) ln t +c**
- ossia, **ln [p + (1-p) x $_j$(t)] = ln t $^{1-p}$ + c**
- da cui, esponenziando, **p + (1-p) x $_j$(t) = t $^{1-p}$ e ^{c}**
- ossia, p + (1-p) x $_j$(t) = **H** t $^{1-p}$
- E, per calcolare H, imponiamo la condizione iniziale **x $_j$(j) = 0** ottenendo p = H j $^{1-p}$
- da cui **H =** $\frac{\mathbf{p}}{\mathbf{j^{1-p}}}$
- e quindi **x $_j$(t) =** $\frac{\boldsymbol{p}}{\mathbf{1}-\boldsymbol{p}}\left[\left(\frac{t}{j}\right)^{\mathbf{1}-\boldsymbol{p}} - \mathbf{1}\right]$

4) Individuazione della Power Law

- Calcoliamo, infine, nel modello deterministico continuo, dati k e t, quale frazione dei nodi, al passo t, ha "grado" k
  - poiché al passo t l'insieme dei nodi è [t] = { 1, … , t }
  - sia **A_t(k) l'insieme { j ≤ t : x_j(t) ≥ k }**
  - vogliamo calcolare $\frac{1}{t}$ | A_t(k) – A_t(k+1) | = $\frac{1}{t}$ **[| A_t (k)| – | A_t (k+1)|]**

    in quanto A_t(k+1) ⊆ A_t(k)

- Per definizione, **j ∈ A_t(k) se e solo se j ≤ t e x_j(t) ≥ k**
- ma $x_j(t) = \frac{p}{1-p}\left[\left(\frac{t}{j}\right)^{1-p} - 1\right]$ e quindi **x_j(t) ≥ k** quando $\frac{p}{1-p}\left[\left(\frac{t}{j}\right)^{1-p} - 1\right] \geq k$

ossia: $\left(\frac{t}{j}\right)^{1-p} \geq \frac{(1-p)}{p}k + 1$

ossia: $\frac{t}{j} \geq \left[\frac{(1-p)}{p}k + 1\right]^{\frac{1}{1-p}}$

ossia: $j \leq t\left[\frac{(1-p)}{p}k + 1\right]^{-\frac{1}{1-p}}$

4) Individuazione della Power Law

- Calcoliamo, infine, nel modello deterministico continuo, dati k e t, quale frazione dei nodi, al passo t, ha "grado" k
  - poiché al passo t l'insieme dei nodi è [t]={1, … , t}
  - sia **$A_t(k)$ l'insieme { j ≤ t : $x_j(t) \geq k$ }**
  - vogliamo calcolare $\frac{1}{t}\,|\,A_t(k) - A_t(k+1)\,|\ = \frac{1}{t}\left[\,|A_t(k)| - |A_t(k+1)|\,\right]$
- Per definizione, **j ∈ A_t(k) se e solo se j ≤ t e $x_j(t) \geq k$**
- ossia, **j ∈ A_t(k) se e solo se j ≤ t e j ≤ $t\left[\frac{(1-p)}{p}\,k+1\right]^{-\frac{1}{1-p}}$**
  - ma poiché $\frac{(1-p)}{p}\,k+1 > 1$ e $-\frac{1}{1-p} < 0$, allora $\left[\frac{(1-p)}{p}\,k+1\right]^{-\frac{1}{1-p}} < 1$
  - ossia, $t\left[\frac{(1-p)}{p}\,k+1\right]^{-\frac{1}{1-p}} \leq t$
- e quindi, in conclusione, **j ∈ A_t(k) se e solo se j ≤ $t\left[\frac{(1-p)}{p}\,k+1\right]^{-\frac{1}{1-p}}$**

4) Individuazione della Power Law

- Calcoliamo, infine, nel modello deterministico continuo, dati k e t, quale frazione dei nodi, al passo t, ha “grado” k
  - poiché al passo t l’insieme dei nodi è [t]={1, … , t}
  - sia A_t(k) l’insieme { j ≤ t : x_j(t) ≥ k }
  - vogliamo calcolare $\frac{1}{t}\,|\,A_t(k) - A_t(k+1)\,|\; = \frac{1}{t}\,[\;|A_t(k)| - |A_t(k+1)|\;]$
- Abbiamo visto j ∈ A_t(k) se e solo se $j \leq t\left[\frac{(1-p)}{p}\,k+1\right]^{-\frac{1}{1-p}}$
- allora, $A_t(k) = \{\,j:\; j \leq t\left[\frac{(1-p)}{p}\,k+1\right]^{-\frac{1}{1-p}}\,\}$
- e, quindi, $|A_t(k)| = t\left[\frac{(1-p)}{p}\,k+1\right]^{-\frac{1}{1-p}}$

4) Individuazione della Power Law

- Vogliamo calcolare $\frac{1}{t}$ **[$|A_t$(k)$|$ – $|A_t$(k+1)$|$]** e $|A_t$ (k)$|$ = $t\left[\frac{(1-p)}{p}\text{ k }+1\right]^{-\frac{1}{1-p}}$
- Sia **F(k) =** $\left[\frac{(1-p)}{p}\textbf{ k }+\mathbf{1}\right]^{-\frac{1}{1-p}}$
- allora, $\frac{1}{t}$ **[$|$ A_t(k)$|$– $|A_t$(k+1) $|$] = F(k) - F(k+1) = f(k)**
- approssimiamo f(k) con - $\frac{\mathrm{dF(k)}}{\mathrm{dk}}$
  - infatti f(k) = F(k) - F(k+1) = – [– $\frac{\text{F(k)} - \text{F(k+1)}}{1}$] = – [$\frac{\text{F(k+1)} - \text{F(k)}}{1}$] $\approx -\frac{\mathrm{dF(k)}}{\mathrm{dk}}$
- Quindi: $\frac{1}{t}$ **[$|$ A_t(k)$|$– $|A_t$(k+1) $|$]** = f(k)

$$\approx -\frac{\mathrm{dF(k)}}{\mathrm{dk}} = -\frac{(1-\mathrm{p})}{p}\left[-\frac{1}{1-p}\right]\left[\frac{(1-p)}{p}\text{ k+1}\right]^{-\frac{1}{1-p}-1}$$

- ossia, $\frac{1}{t}$ **[$|$ A_t(k)$|$– $|A_t$(k+1) $|$] =** $\frac{1}{p}\left[\frac{(1-p)}{p}\textbf{ k+1}\right]^{-\frac{1}{1-p}-1}$
- che è una Power Law con esponente 1 + $\frac{1}{1-\mathrm{p}}$

Conclusioni sull'analisi del modello

- Per analizzare il nostro modello Rich get Richer, abbiamo approssimato mediante una funzione (deterministica e) continua la legge probabilistica che esprime la variazione del grado di un nodo fra due passi consecutivi

- ma non abbiamo idea di quanto tale approssimazione si avvicini al modello discreto e aleatorio

- In [Bollobas, Riordan – 2005] è stato dimostrato che

  nel modello Rich get Richer che abbiamo proposto, con alta probabilità la frazione del numero di nodi di grado k è proporzionale a $\mathbf{k^{-\frac{1}{1-p}-1}}$

- esattamente come nell'approssimazione che abbiamo discusso!

Impredicibilità della popolarità

- **Esperimento di Salganik, Dodds e Watts [2006]**
  - creano un sito di download musicale e vi caricano 48 brani sconosciuti
  - meglio: creano 8 diverse liste, e caricano in ciascuna gli stessi 48 brani
  - quando qualcuno visita il sito, viene indirizzato ad una delle 8 liste scelta a caso: il visitatore può ascoltare i brani e può anche scaricarli [vedi più avanti]
  - per ciascuna delle 8 liste, viene tenuto un "contatore di download" che conta, appunto, per ciascun brano, quante volte quel brano è stato scaricato dai visitatori indirizzati a quella lista
  - Non appena un visitatore viene indirizzato ad una lista, gli viene mostrato il "contatore di download" di quella lista: solo successivamente il visitatore può ascoltare e scaricare i brani che desidera
- **Ingredienti dell'esperimento**:
  - 1) 8 copie inizialmente identiche di 48 brani (8 *liste* identiche)
  - 2) il "contatore di download" di una lista è un indice della popolarità di un brano fra i visitatori indirizzati a quella lista (ed è ad essi visibile)
  - 3) il "contatore di download" può incentivare i visitatori di una lista a scaricare i brani più popolari in quella lista [***effetti di feedback***]

Impredicibilità della popolarità

- Poiché i “contatori di download” di tutte le liste sono inizializzati al valore 0 per ogni brano, ci si aspetterebbe che, se tutte liste ricevono lo stesso numero di visitatori, i brani più popolari in tutte le 8 liste siano gli stessi

- Ma non è così!

- **Conclusioni**: sono gli effetti di feedback che hanno prodotto esiti diversi
  - ossia, i gusti e le scelte dei visitatori iniziali di una lista hanno influenzato le scelte successive – e, dunque, la popolarità dei brani

- Cosa ci mostra questo esperimento?
- Che la fase iniziale del processo di diventare popolari è molto fragile
- Ovvero, **il processo che conduce alla popolarità è pesantemente sensibile alle condizioni iniziali**

La lunga coda

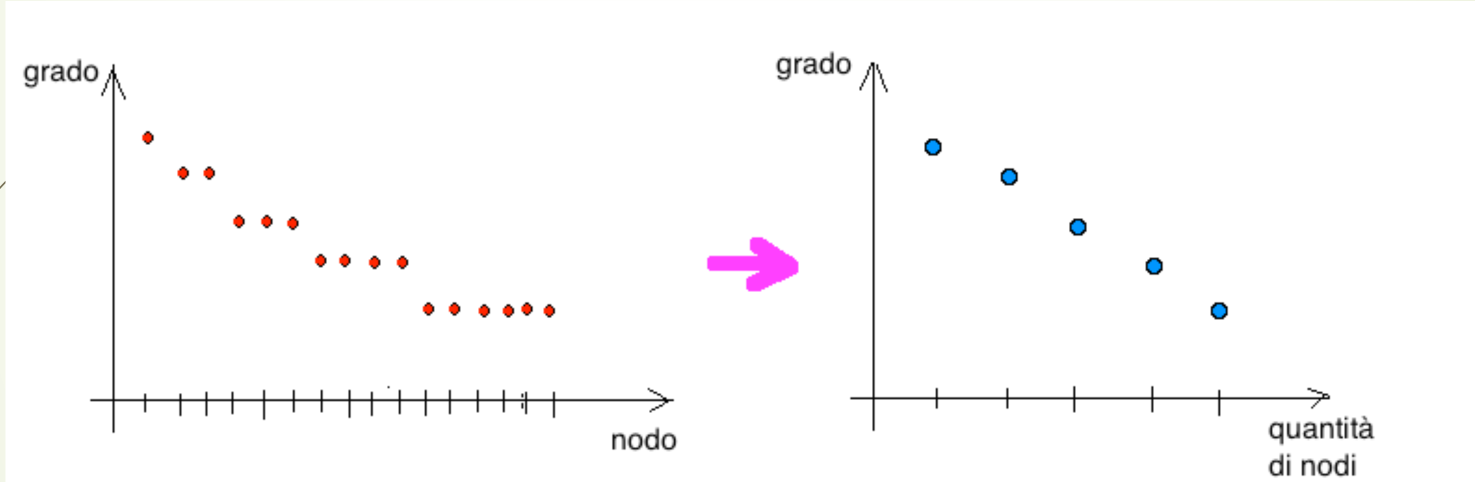
- La distribuzione della popolarità può avere importanti effetti sulle attività commerciali – in particolar modo sui "magazzini virtuali"
  - ossia, sulle società di vendita on-line
  - che dispongono di "magazzini" molto vasti
  - che, cioè, commercializzano elevate quantità di prodotti - che non potrebbero essere gestite da magazzini fisici
- Il punto è: uno di detti venditori (diciamo, una libreria on-line) guadagna di più vendendo tante copie di pochi libri molto popolari oppure vendendo poche copie di tantissimi libri poco popolari?
- In questo modo, ci stiamo sempre interessando di questioni di popolarità, ma da un altro punto di vista
- Invece che interessarci a quanti sono i nodi di grado k, ci interessa ora sapere se, complessivamente, il volume di affari derivante dai nodi poco popolari, ossia, i nodi che hanno grado basso (e che sono tanti) equivale al volume di affari dei nodi molto popolari, ossia, quelli che hanno grado elevato (che sono pochi)

La lunga coda

- Per gestire questa questione, quello che possiamo fare è ordinare i nodi per popolarità non crescente
  - un po' come fare una "hit parade" dei nodi…
- Creiamo un grafico
  - lungo l'asse delle ascisse elenchiamo i nodi, ordinati per popolarità non crescente: il primo nodo sarà il più popolare, seguito dal secondo nodo più popolare, e così via
  - lungo l'asse delle ordinate elenchiamo le popolarità dei nodi: in corrispondenza al nodo i, indicheremo il grado del nodo i
- I nodi poco popolari costituiscono la ***coda*** del grafico appena costruito
- E il volume di affari che ci interessa è, grosso modo, l'area della regione compresa fra la coda del grafico e l'asse x
- Ma come è fatto questo grafico?
- Possiamo, in qualche modo, riutilizzare quanto studiato circa la distribuzione dei gradi dei nodi per ottenere questo grafico?

La lunga coda

- Possiamo fare una cosa: invece di elencare lungo l'asse delle ascisse, uno per uno, tutti i nodi, elenchiamo con un solo punto tutti i nodi che hanno la stessa popolarità: l'ascissa di tale punto è il numero di nodi che hanno quella popolarità



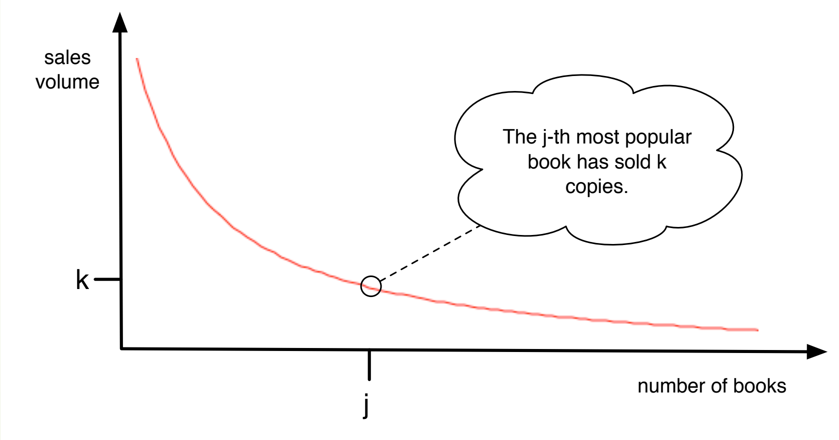
  - naturalmente, per compiere questa operazione dobbiamo assumere che a gradi di popolarità diversi corrispondano numeri di nodi diversi che hanno quelle popolarità
    - ossia, ***non*** deve accadere che, ad esempio, 5 nodi abbiano popolarità p_1 e 5 nodi abbiano popolarità $p_2 \neq p_1$

La lunga coda

- Invece di elencare lungo l'asse delle ascisse, uno per uno, tutti i nodi, elenchiamo con un solo punto tutti i nodi che hanno la stessa popolarità
- Pensandoci bene, questo grafico è, grosso modo, ottenuto dal grafico che esprime il numero n di nodi che hanno grado k scambiando gli assi coordinati!
- se d(k) è la funzione che esprime il ***numero*** di nodi di grado k, allora il grafico che ci occorre rappresenta, grosso modo, la funzione $k = d^{-1}(n)$
  - ammesso che d sia una funzione invertibile
- È possibile dimostrare che
  - se, come appare dai dati sperimentali, la funzione f che esprime la **frazione** del numero di nodi che hanno grado k è una Power Law
  - ossia, per qualche costante $c > 0$, $f(k) \approx k^{-c}$
- allora **anche d esprime una Power Law**
  - ossia, esiste una costante $a > 0$ tale che $d(k) \approx k^{-a}$
- e dunque, se $\mathbf{n = d(k) \approx k^{-a}}$, abbiamo che $\mathbf{k \approx n^{-1/a}}$

La lunga coda

- Se n = d(k), abbiamo che $k \approx n^{-1/a}$
- che decresce *moooolto* lentamente!
- Ossia, è una ***lunga coda***
- E il volume di affari che ci interessa
  - ossia, l'area della regione compresa fra la coda del grafico e l'asse delle ascisse
  - a partire da qualche j, abbastanza grande
- è tutt'altro che trascurabile



Gli effetti della web-search

- Ci domandiamo ora: l'utilizzo dei motori di ricerca, per la ricerca su internet, ha impatto sul fenomeno rich get richer?
- Da un lato, **il fenomeno rich get richer è amplificato** in virtù dell'utilizzo di motori di ricerca basati sul ranking
  - ossia, che elencano le pagine risultanti dalla ricerca in un ordine derivante dall'assegnazione di un punteggio a sua volta derivante, ad esempio, dal numero di in-link delle pagine
- D'altra parte, le richieste ad un motore di ricerca possono essere sottoposte in una enorme varietà di forme
  - dipendentemente dalle parole chiave scelte per effettuare una ricerca
- e questo aumenta la possibilità di essere indirizzati a pagine che non sono nella "top list"
  - ad esempio, usando parole chiave Inusuali…
- **e questo mitiga gli effetti rich get richer**

Gli effetti della web-search

- Per sfruttare la lunga coda
  - per vendere i prodotti "di nicchia" (quelli poco popolari)
- i venditori on-line provano ad indirizzare gli utenti, ad esempio ricorrendo all'utilizzo di *recommendation systems*
  - quando i motori di ricerca invitano gli utenti ad esplorare pagine poco popolari sulla base delle loro ricerche pregresse
- In conclusione: il progetto di strumenti di web-search è un esempio di utilizzo degli effetti feedback, volto ad indirizzare gli utenti nell'elaborare le loro richieste
  - in alcuni casi amplificando l'effetto rich get richer
  - in altri attenuandolo