POWER LAWS

Una funzione f(k) che diminuisce come $k^{-\alpha}$ è detta **power law**

Perché ci interessano

Abbiamo visto fenomeno small-world e come possa funzionare la ricerca decentralizzata

$$\text{Prob}[v \text{ è amico di } u] \approx d(u,v)^{-1}$$

Abbiamo anche considerato la distanza geografica/sociale, osservando la correlazione con l'amicizia

$$\text{Prob}[v \text{ è amico di } u] \approx (\text{rank}_u(v))^{-1}$$

Le reti del mondo reale sono dinamiche!

Esempi:

- Il Web nel 1991 aveva un unico nodo, oggi ce ne sono miliardi
- Le reti di citazioni di articoli scientifici e le reti di collaborazione tra scienziati continuano a crescere grazie alla pubblicazione di nuovi documenti
- La rete di collaborazione degli attori continua a crescere grazie all'uscita di nuovi film
- La rete di interazione delle proteine è cresciuta nel corso di 4 miliardi di anni: da pochi geni a oltre 20.000

Le reti del mondo reale sono dinamiche!

Esempi:

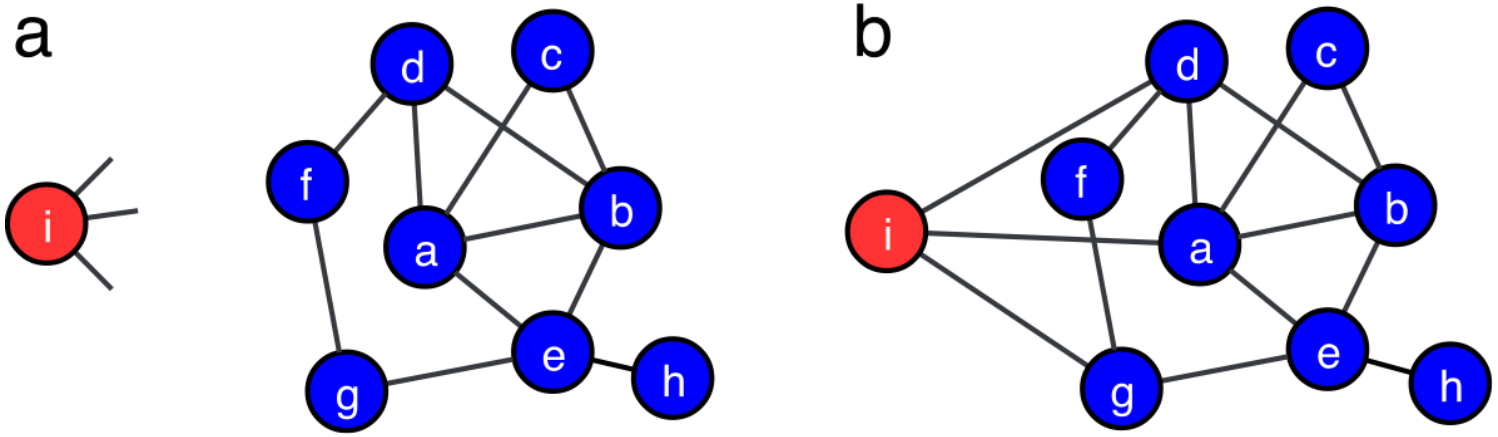
- Il Web nel 1991 aveva un unico nodo, oggi ce ne sono miliardi
- Le reti di citazioni di articoli scientifici e le reti di collaborazione tra scienziati continuano a crescere grazie alla pubblicazione di nuovi documenti
- La rete di collaborazione degli attori continua a crescere grazie all'uscita di nuovi film
- La rete di interazione delle proteine è cresciuta nel corso di 4 miliardi di anni: da pochi geni a oltre 20.000

Come si formano i collegamenti?

Procedura generica

Un nuovo nodo arriva con un dato numero di futuri vicini del nodo (grado)

I vicini sono scelti secondo qualche *regola*



Come sono scelti i vicini?

I nodi preferiscono collegarsi ai nodi più connessi

- La nostra conoscenza del Web è orientata verso pagine popolari, che sono altamente collegate, quindi è più probabile che il nostro sito Web punti a siti Web altamente collegati

- Gli scienziati hanno più familiarità con gli articoli altamente citati (che sono spesso i più importanti), quindi tenderanno a citarli, nei propri articoli, più spesso rispetto a quelli meno citati

- Più film fa un attore, più diventa popolare e maggiori sono le possibilità di essere scelto per un nuovo film

Popolarità

Il comportamento/le decisioni di una persona dipendono dalle scelte fatte da altre persone

- Queste dipendenze possono portare a risultati molto diversi da ciò che troviamo quando gli individui prendono decisioni indipendenti

La **popolarità** è un fenomeno caratterizzato da squilibri estremi

- quasi ognuno è noto alle persone nella propria cerchia sociale immediata
- qualcuno ottiene una visibilità più ampia
- pochissimi raggiungono la visibilità globale

- Come possiamo modellare questi squilibri?

Modello

Caratteristiche:

Crescita: il numero di nodi cresce nel tempo in seguito all'aggiunta di nuovi nodi.

Preferential attachment (Attaccamento preferenziale): i nuovi nodi tendono ad essere collegati ai nodi più connessi.

I modelli finora considerati stabiliscono collegamenti tra coppie di nodi, indipendentemente dal loro grado

Popolarità sul WEB

Il Web è un dominio concreto in cui è possibile misurare la popolarità in modo molto accurato

Possiamo prendere il numero di in-link come misura della popolarità di una pagina

- Facciamo un'istantanea del Web completo e contiamo il numero di in-link per ogni pagina

Popolarità sul WEB

Il Web è un dominio concreto in cui è possibile misurare la popolarità in modo molto accurato

Possiamo prendere il numero di in-link come misura della popolarità di una pagina

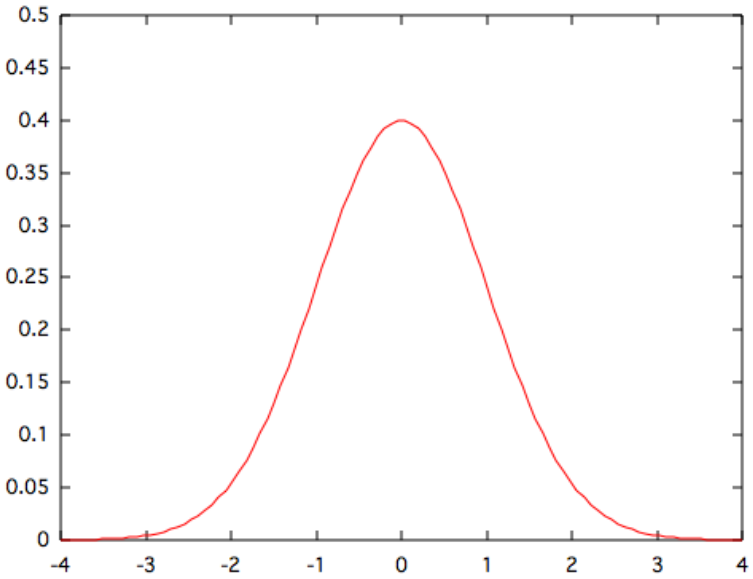
- Facciamo un'istantanea del Web completo e contiamo il numero di in-link per ogni pagina

Distribuzione della popolarità sulle pagine Web:

- In funzione di k, quale frazione di pagine sul Web ha k in-link?

Ipotesi semplice: Distribuzione normale

- Un'ipotesi naturale per la distribuzione della popolarità è la distribuzione normale (gaussiana)
  - usata ampiamente in probabilità e statistiche
  - caratterizzato da due quantità: un valore medio e una deviazione standard attorno a questo valore



densità di valori nella distribuzione normale con media 0 e deviazione standard 1

Ipotesi semplice: Distribuzione normale

- La distribuzione normale è onnipresente nelle scienze naturali

- Il **teorema del limite centrale (TLC)** dice che la somma (o la media) di ogni sequenza di quantità casuali indipendenti e dotate della stessa media, sarà distribuita al limite secondo la distribuzione normale, indipendentemente dalla distribuzione soggiacente.

- Il TLC implica che se abbiamo un campione ‘’grande’’, allora la distribuzione della somma di n variabili casuali indipendenti sarà ‘’quasi’’ normale

Ipotesi semplice: Distribuzione normale

- La distribuzione normale è onnipresente nelle scienze naturali

- Il **teorema del limite centrale (TLC)** dice che la somma (o la media) di ogni sequenza di quantità casuali indipendenti e dotate della stessa media, sarà distribuita al limite secondo la distribuzione normale, indipendentemente dalla distribuzione soggiacente.

- Il TLC implica che se abbiamo un campione ‘’grande’’, allora la distribuzione della somma di n variabili casuali indipendenti sarà ‘’quasi’’ normale

Es. Supponiamo di *eseguire misurazioni ripetute di una quantità fisica fissa*

Si presume che variazioni nelle misurazioni tra le prove sono il risultato cumulativo di molte fonti indipendenti di errore in ogni prova

quindi la distribuzione dei valori misurati dovrebbe essere approssimativamente normale

➔ la distribuzione della media del campione è nota anche se non si conosce la distribuzione della popolazione da cui è tratto il campione

Ipotesi semplice: Distribuzione normale

Normal

- The cumulative density function is given by the formula

$$\Pr[X \geq x] = \int_{z=x}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(z-\mu)^2}{2\sigma^2}} dz$$

and the probability density function is given by the formula

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),$$

where μ is the mean, σ is the standard deviation (a measure of the "width of the bell"), and exp denotes the exponential function.

- For a mean of 0 and a standard deviation of 1, this formula

$$p(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right),$$

is known as the standard normal distribution

Distribuzione Normale e Popolarità sul WEB

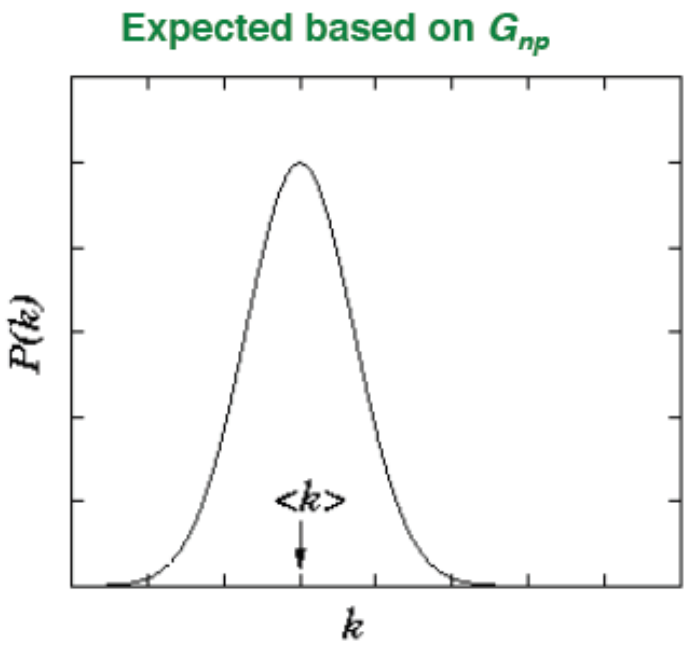
La distribuzione normale si applica nel caso di pagine Web?

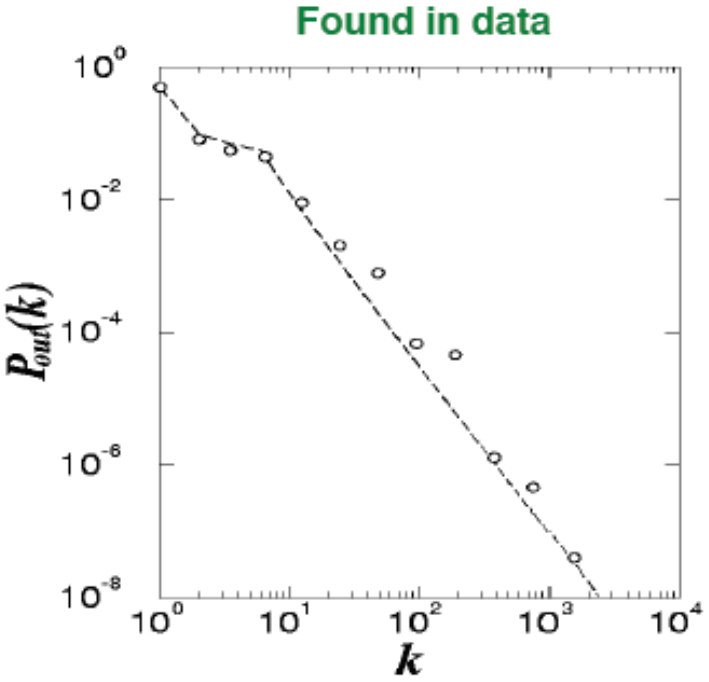
La distribuzione normale modella la struttura di collegamento del Web presumendo che ogni pagina decida in modo indipendente a caso se collegarsi a una data pagina

➔ il numero di collegamenti in entrata a una determinata pagina è la somma di molte quantità casuali indipendenti e viene normalmente distribuito

➔ il numero di pagine con k in-link dovrebbe diminuire esponenzialmente in k

Diverse misurazioni hanno dimostrato che ciò è sbagliato

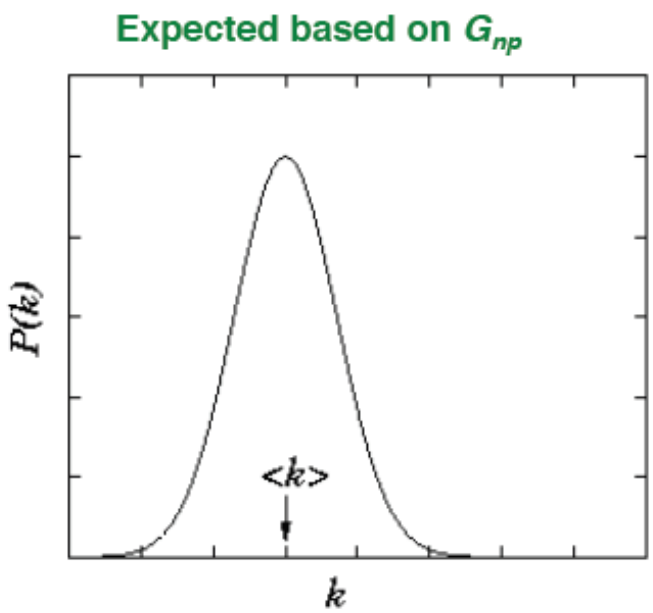


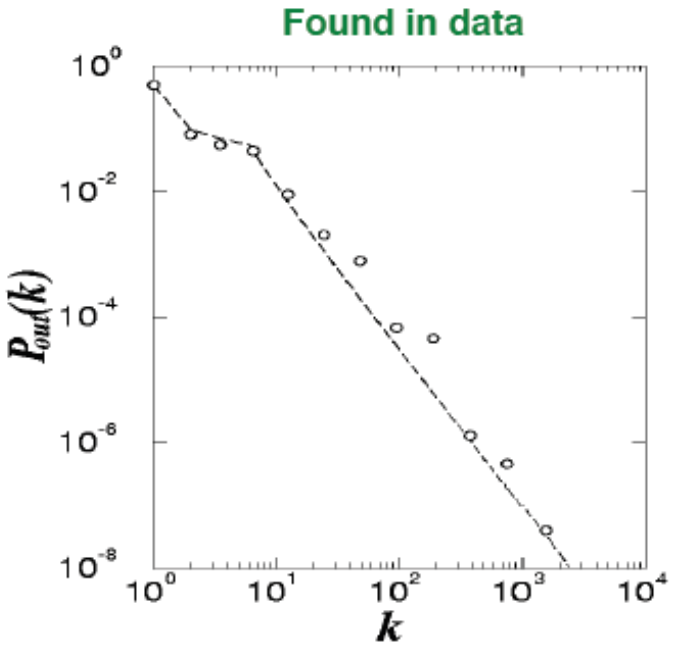


Power Laws

Gli studi su varie istantanee del Web hanno mostrato che

- le pagine con un numero molto elevato di link in entrata sono molto più comuni di quanto ci aspettiamo con una distribuzione normale.
- la frazione di pagine Web con k in-link è approssimativamente proporzionale a ***1/k²***





Power Laws

Una funzione che diminuisce come $k^{-\alpha}$ è detta **power law**

Una funzione power law per una variabile casual discreta *X* soddisfa

$$Prob[X=k] \approx b\, k^{-\alpha}$$

(ci concentriamo solo sull'esponente α)

Power Laws

Una funzione che diminuisce come $k^{-\alpha}$ è detta **power law**

Una funzione power law per una variabile casual discreta *X* soddisfa

$$Prob[X=k] \approx b\, k^{-\alpha}$$

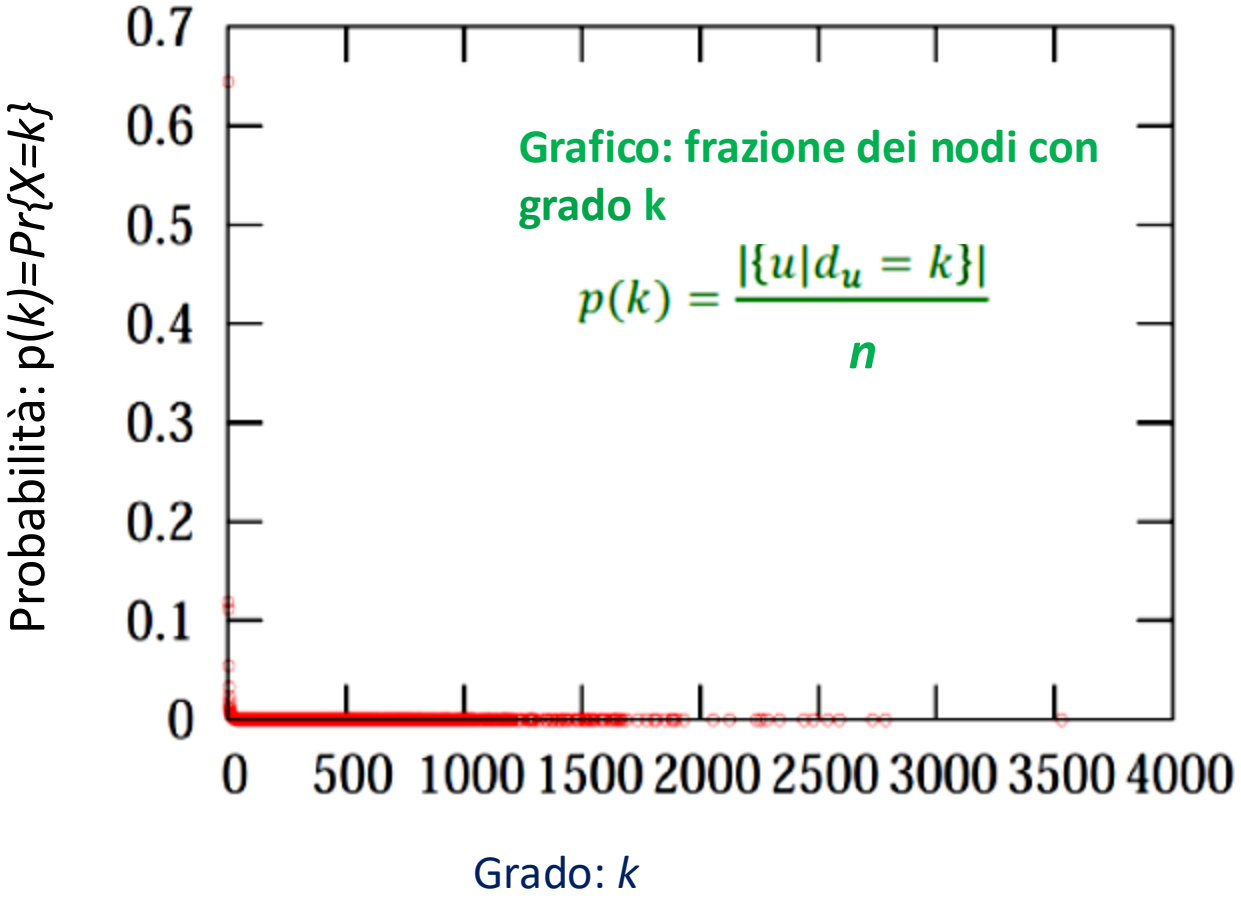
(ci concentriamo solo sull'esponente α)

- Le funzioni power law sembrano dominare nei casi in cui la quantità misurata può essere vista come un tipo di popolarità.

  Es.:

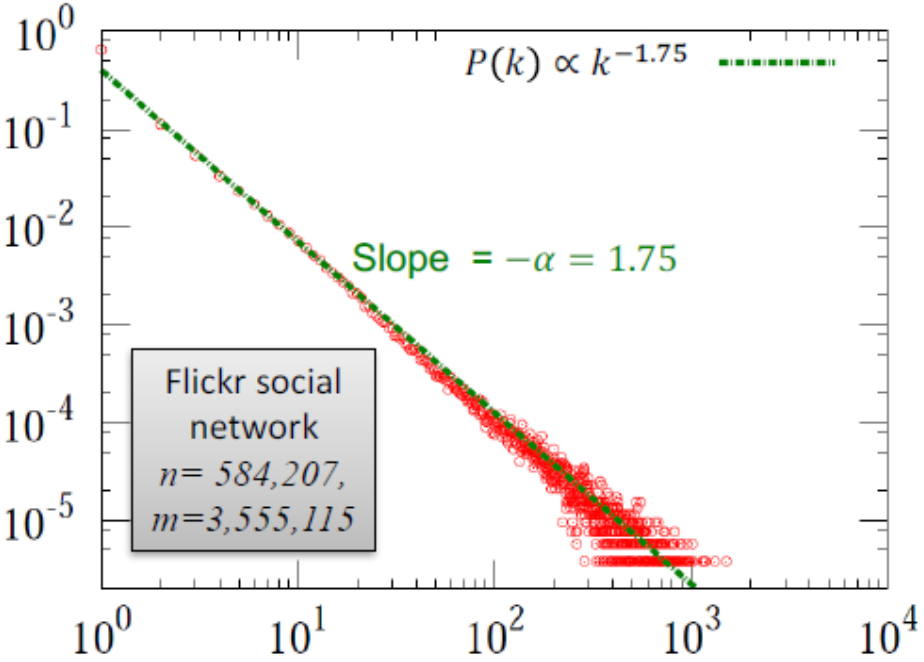
  - La frazione di **numeri di telefono** che ricevono k chiamate al giorno è approssimativamente proporzionale a $1/k^2$
  - la frazione di **libri** acquistati da k persone è approssimativamente proporzionale a $1/k^3$
  - la frazione di articoli scientifici che ricevono k **citazioni** in totale è approssimativamente proporzionale a $1/k^3$

Come riconoscere una funzione Power Law?



Flickr social network
n= 584,207,
m=3,555,115

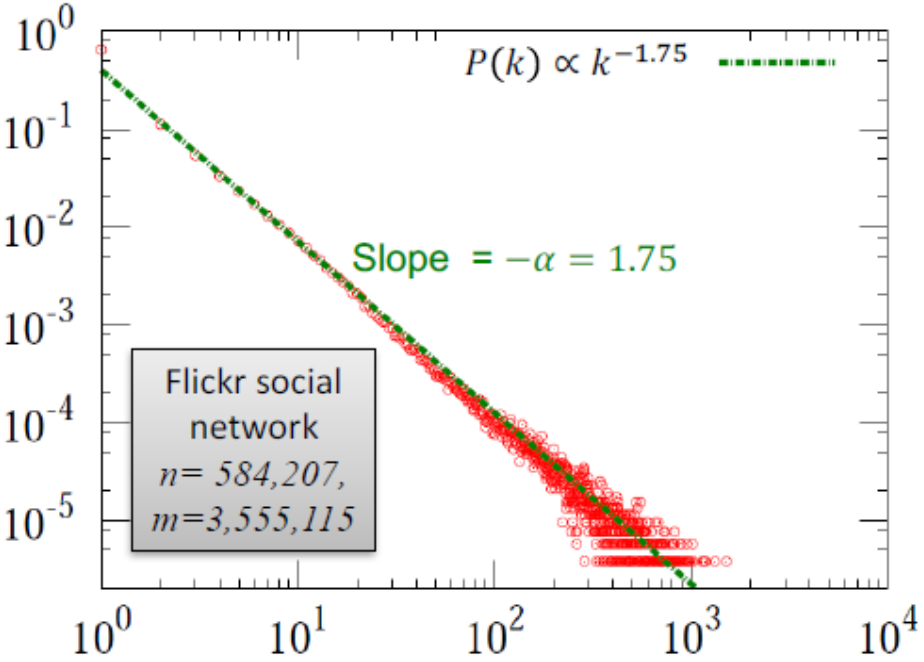
Come riconoscere una funzione Power Law?



un grafico di tipo **log-log** è un grafico che usa la scala logaritmica su entrambi gli assi

- $f(k) = bk^{-\alpha}$ ➔ $\log f(k) = \log b - \alpha \log k$

Come riconoscere una funzione Power Law?



un grafico di tipo **log-log** è un grafico che usa la scala logaritmica su entrambi gli assi

- $f(k) = bk^{-\alpha}$ ➔ $\log f(k) = \log b - \alpha \log k$
- Una distribuzione power law si presenta come una **retta** su un grafico di tipo **log-log**
  - **- α** è la pendenza della linea e ***log b*** è l'intersezione con l'asse y

Funzione Power Law?

- A random variable X attains certain values, say for some constants β, α

$$\Pr[X = k] = \frac{\beta}{k^{\alpha}}, \text{ for } k = 1, 2, \ldots, n$$

- Then

$$\Pr[X \geq k] = \sum_{x=k}^{n} \Pr[X = x] = \sum_{x=k}^{n} \frac{\beta}{x^{\alpha}} \approx \beta \int_{k}^{n} \frac{1}{x^{\alpha}} dx$$

$$= \beta \left[-\frac{x^{1-\alpha}}{1-\alpha} \right]_{k}^{n} = \beta \frac{k^{1-\alpha}}{1-\alpha} - \beta \frac{n^{1-\alpha}}{1-\alpha}$$

- Note that β is a normalizing coefficient

$$1 = \sum_{k=1}^{n} \frac{\beta}{k^{\alpha}} \approx \beta \int_{1}^{n} \frac{1}{x^{\alpha}} dx$$

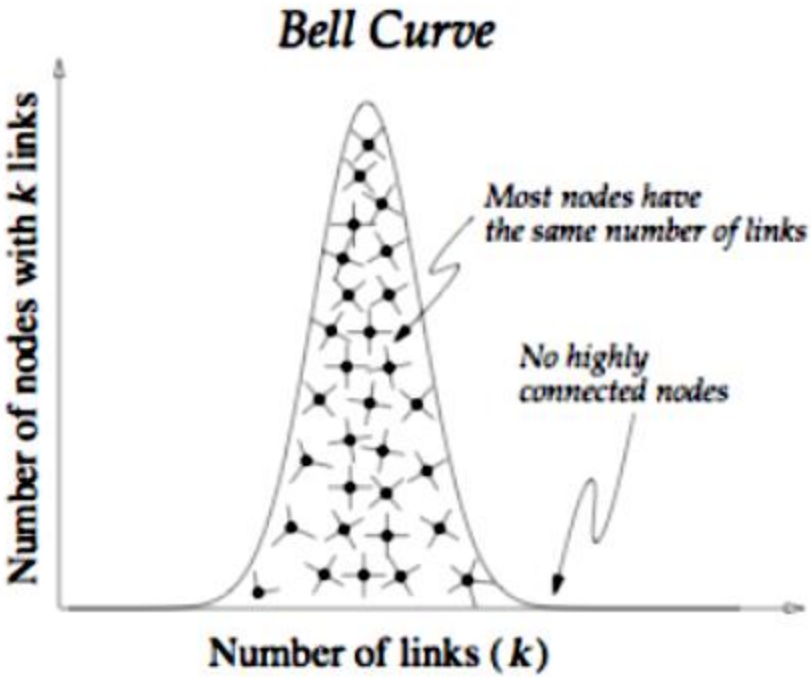
- A non-negative random variable X is said to have a *power law* distribution if

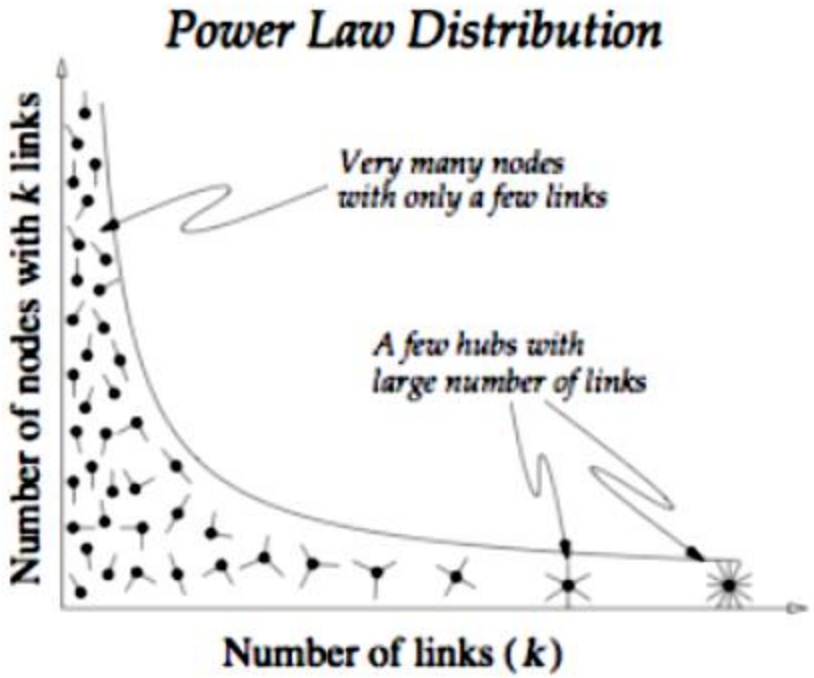
$$Pr[X \geq x] \sim cx^{-\alpha}$$

for constants $c > 0$ and $\alpha > 0$.

Esponenziale

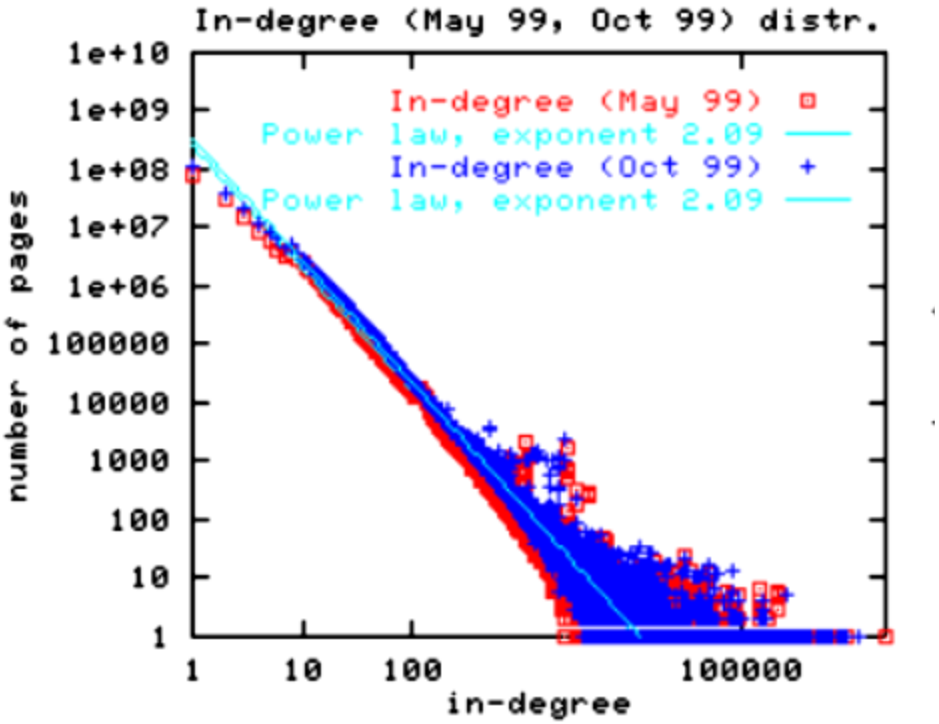
Power Law



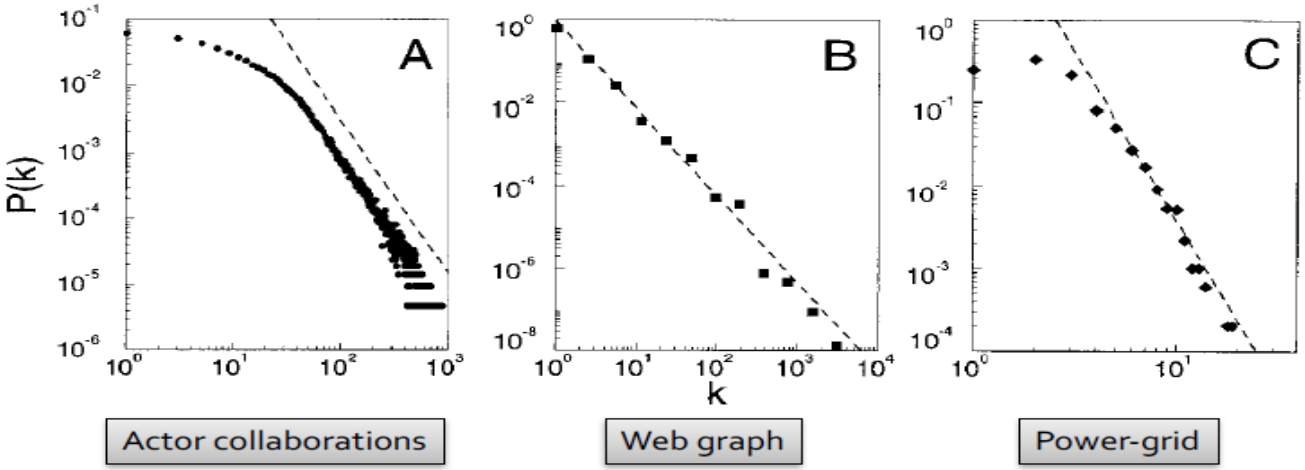


Esempi

- **The World Wide Web** [Broder et al., 2000]



Esempi



[Barabasi-Albert, 1999]

Esponente è normalmente $2<\alpha<3$

- **Web graph:**
  - $\alpha_{in} = 2.1$, $\alpha_{out} = 2.4$ [Broder et al. 00]
- **Actor-collaborations:**
  - $\alpha = 2.3$ [Barabasi-Albert 00]
- **Citations to papers:**
  - $\alpha \approx 3$ [Redner 98]
- **Online social networks:**
  - $\alpha \approx 2$ [Leskovec et al. 07]

Perchè le funzioni Power Law sono importanti?

- Descrivono squilibri estremi ("pochi hanno molto, molti hanno poco").

- Esempi
  - reti sociali (es. numero di follower su Instagram, visualizzazioni di video TikTok)

  - biologia (La scala metabolica nei mammiferi , cioè la relazione tra il tasso metabolico di un organismo e la sua massa corporea)

  - Reti economiche (capitalizzazione delle cripovalute vs. numero di holder)

Perchè le funzioni Power Law sono importanti?

- **Zipf's Law**

La Legge di Zipf , strettamente correlata (e a volte usata in modo intercambiabile) alle distribuzioni power law,

mette in relazione la frequenza, *f* (cioè il conteggio), di qualcosa con il suo rango r:

Il rango r = 1 indica l'elemento più frequente, r = 2 il secondo più frequente, e così via.

Un fenomeno soddisfa la Legge di Zipf quando

$$f \approx br^{-\alpha}$$

per qualche coppia di costanti α e b.

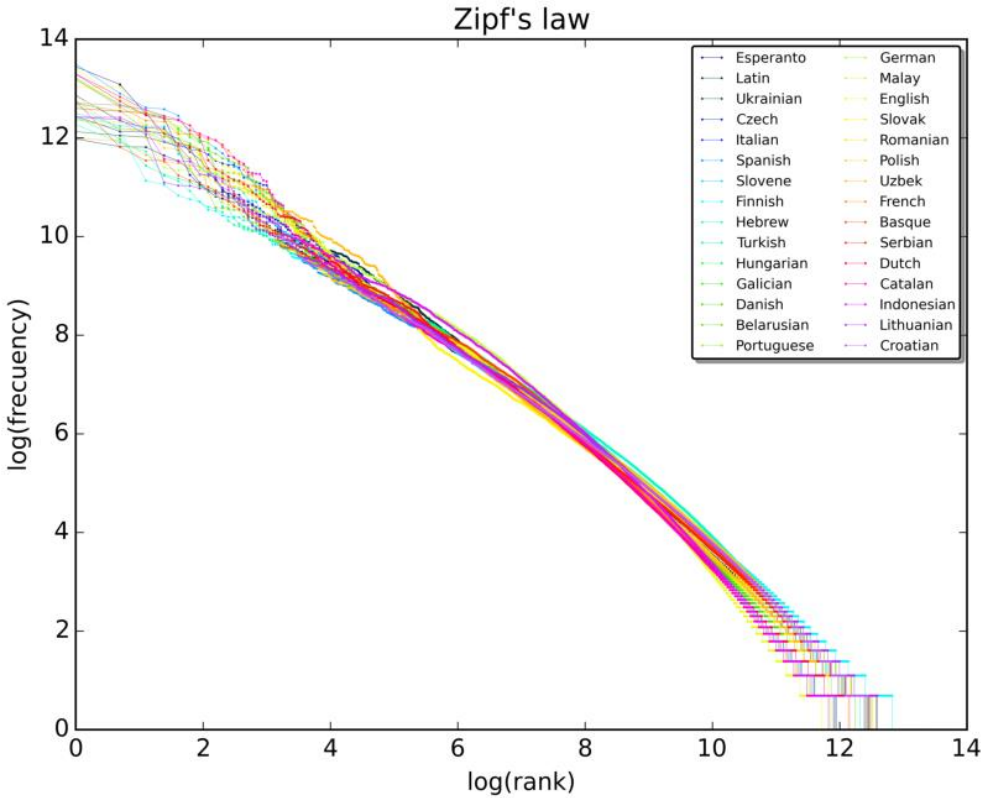
Perchè le funzioni Power Law sono importanti?

In un qualsiasi libro, sia *w* una parola al suo interno. Se calcoliamo

la sua frequenza f_w (cioè il numero di occorrenze nel testo) e

il suo rango r_w (cioè se la parola è la prima, la seconda, ... parola più comune nel libro),

allora scopriamo che $f_w \propto 1/r_w$ (o equivalentemente, log $f_w \approx -\log r_w + c$)."



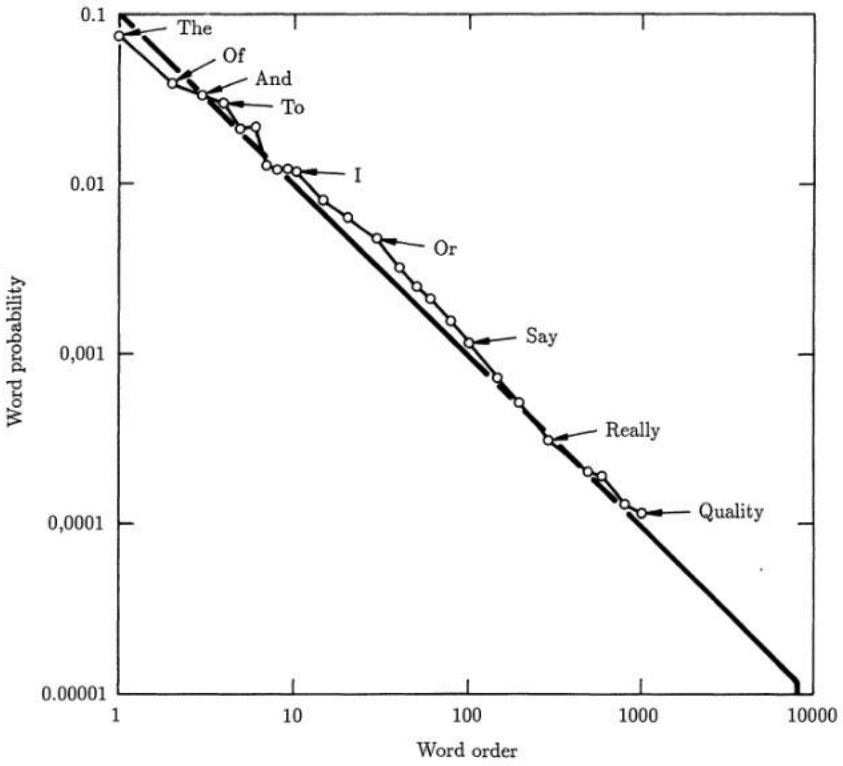
Perchè le funzioni Power Law sono importanti?

In un qualsiasi libro, sia *w* una parola al suo interno. Se calcoliamo

la sua frequenza f_w (cioè il numero di occorrenze nel testo) e

il suo rango r_w (cioè se la parola è la prima, la seconda, ... parola più comune nel libro),

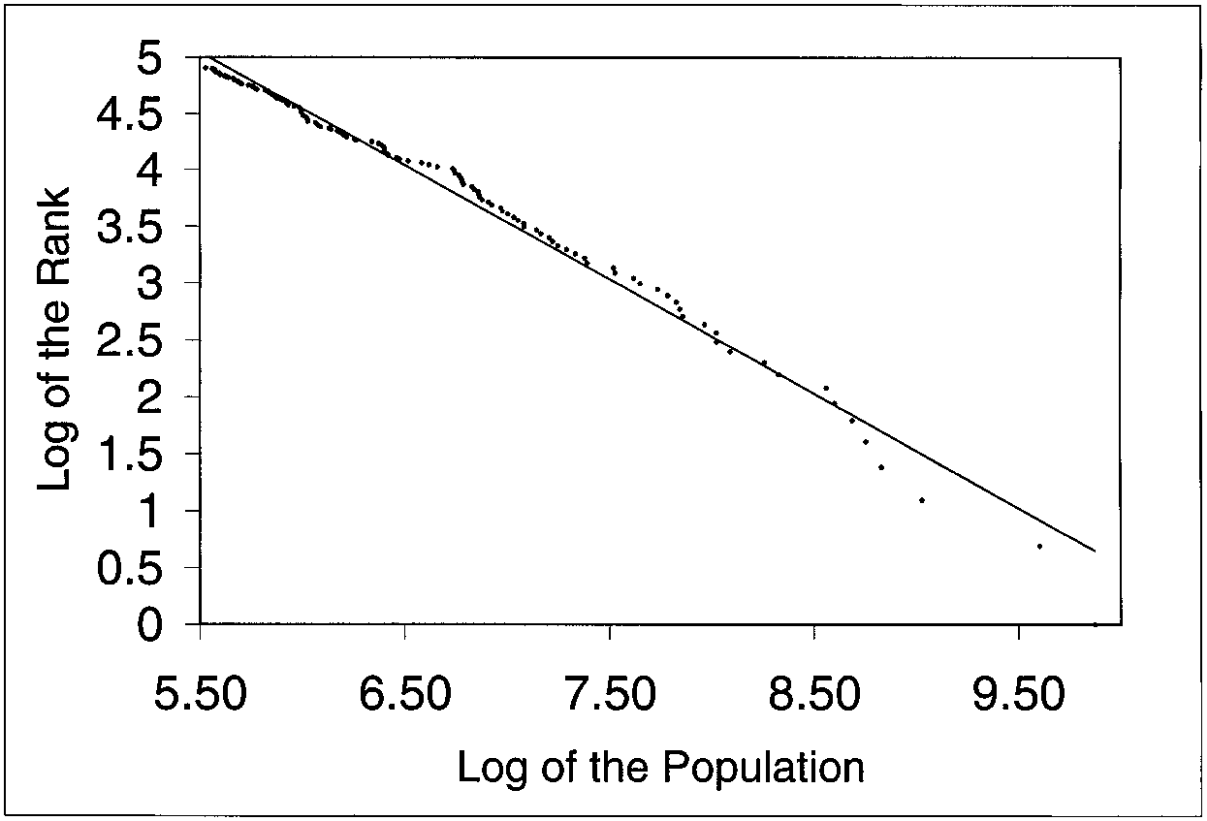
allora scopriamo che $f_w \propto 1/r_w$ (o equivalentemente, log $f_w \approx -\log r_w + c$)."



Perchè le funzioni Power Law sono importanti?

Allo stesso modo, se consideriamo una città *t* ed indichiamo con con f_t la sua popolazione e con r_t il suo rango nell'ordinamento delle città per popolazione risulta

$$f_t \approx 1/r_t$$



Log Size versus Log Rank of the 135 largest U. S. Metropolitan Areas in 1991
Source: Statistical Abstract of the United States [1993].

Perchè le funzioni Power Law sono così diffuse?

Le idee tratte dall'analisi delle cascate di informazione e dagli effetti di rete forniscono la base per un meccanismo molto naturale per generare funzioni Power Law

- **Le distribuzioni normali derivano dalla media di una serie di decisioni casuali indipendenti**
  - Teorema del Limite Centrale

- **Le distribuzioni Power law derivano dal *feedback* introdotto da decisioni correlate su una popolazione**
  - È un problema ancora aperto fornire un **modello** pienamente soddisfacente di distribuzioni power law a partire da modelli semplici di decisioni individuali

Meccanismi di Formazione

Preferential Attachment (attaccamento preferenziale):
Nuovi nodi si collegano a nodi già popolari
(es. citazioni accademiche).

Effetto "Rich-Get-Richer":
Feedback positivo: la popolarità attuale aumenta la crescita futura.

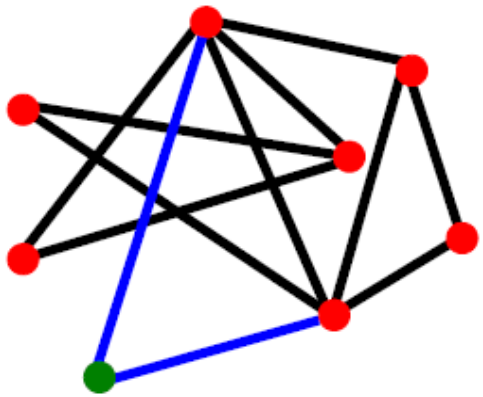
Modello Rich-Get-Richer

- Modello basato su alcune conseguenze osservabili dei processi in presenza di *cascate di informazioni*

- Assume che le persone hanno la tendenza a copiare **con maggiore probabilità**:
  - **le decisioni delle persone che agiscono prima di loro**
  - **le decisioni degli individui popolari**

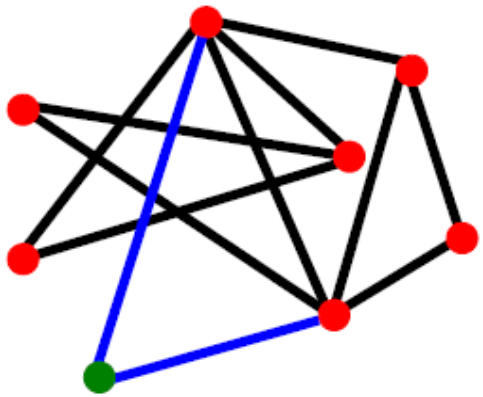
Preferential Attachment

- Nodi arrivano in ordine *1,..,N*
- Al passo *j*
  - sia d_i il grado del **nodo *i*** (per ogni i=1,…,j-1)
  - **Il nuovo nodo *j*** arriva e crea *m* link
  - Probabilità che *j* crei un link ad *i* è $P(j \rightarrow i)=\frac{d_i}{\sum d_v}$



Preferential Attachment

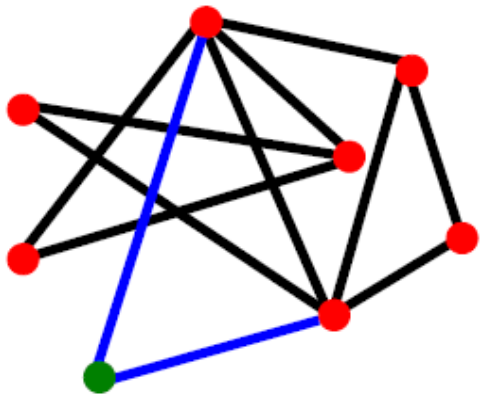
- Nodi arrivano in ordine *1,..,N*
- Al passo *j*
  - sia d_i il grado del **nodo *i*** (per ogni i=1,…,j-1)
  - **Il nuovo nodo *j*** arriva e crea *m* link
  - Probabilità che *j* crei un link ad *i* è $P(j \rightarrow i)=\frac{d_i}{\sum d_v}$



➔ **Nuovi nodi hanno probabilità maggiore di scegliere nodo con grado già alto**

Preferential Attachment

- Nodi arrivano in ordine *1,..,N*
- Al passo *j*
  - sia d_i il grado del **nodo *i*** (per ogni i=1,…,j-1)
  - **Il nuovo nodo *j*** arriva e crea *m* link
  - Probabilità che *j* crei un link ad *i* è $P(j \rightarrow i)=\frac{d_i}{\sum d_v}$



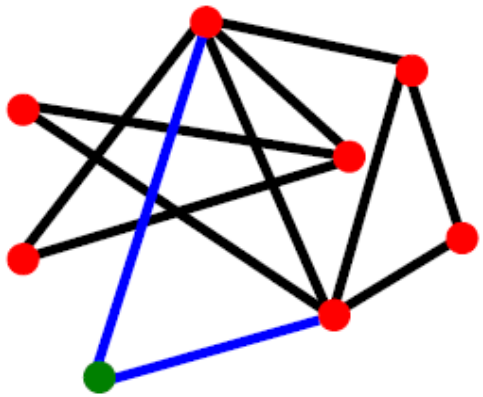
➔ **Nuovi nodi hanno probabilità maggiore di scegliere nodo con grado già alto**

Es.

- **Citazioni:** Le nuove citazioni di una pubblicazione sono proporzionali al numero di citazioni che già ha
  - *se molte persone citano un documento, allora deve essere buono, e quindi dovrei citarlo anch'io*

Preferential Attachment

- Nodi arrivano in ordine *1,..,N*
- Al passo *j*
  - sia d_i il grado del **nodo *i*** (per ogni i=1,…,j-1)
  - **Il nuovo nodo *j*** arriva e crea *m* link
  - Probabilità che *j* crei un link ad *i* è $P(j \rightarrow i)=\frac{d_i}{\sum d_v}$



➔ **Nuovi nodi hanno probabilità maggiore di scegliere nodo con grado già alto**

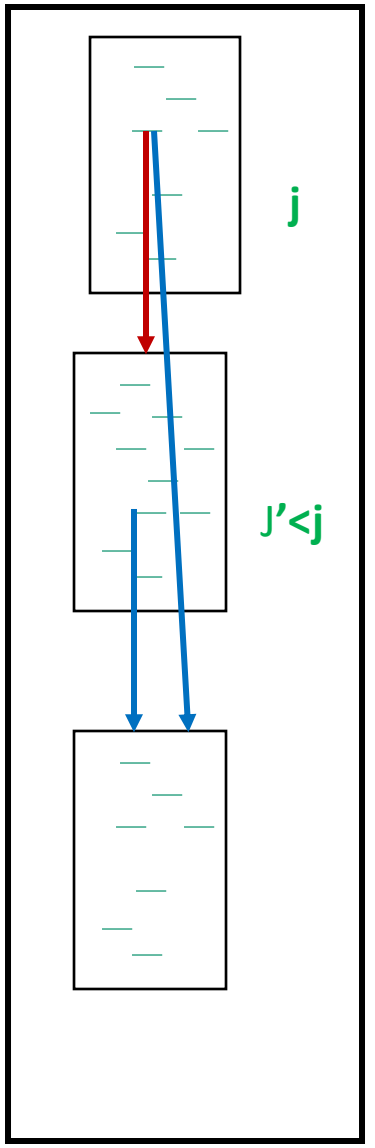
Es.

- **Citazioni:** Le nuove citazioni di una pubblicazione sono proporzionali al numero di citazioni che già ha
  - *se molte persone citano un documento, allora deve essere buono, e quindi dovrei citarlo anch'io*
- **Sociologia:** Effetto Matteo [Merton, 1968], "Per chi avrà, più sarà dato e avranno in abbondanza. A chiunque non ha, anche quello che ha sarà preso''(parabola dei talenti)
  - Eminenti scienziati ottengono spesso più credito di un ricercatore relativamente sconosciuto, anche se il loro lavoro è simile

Esempio- Creazione di una pagina WEB

- Per semplicità, supponiamo che ogni pagina crei **un solo** link in uscita

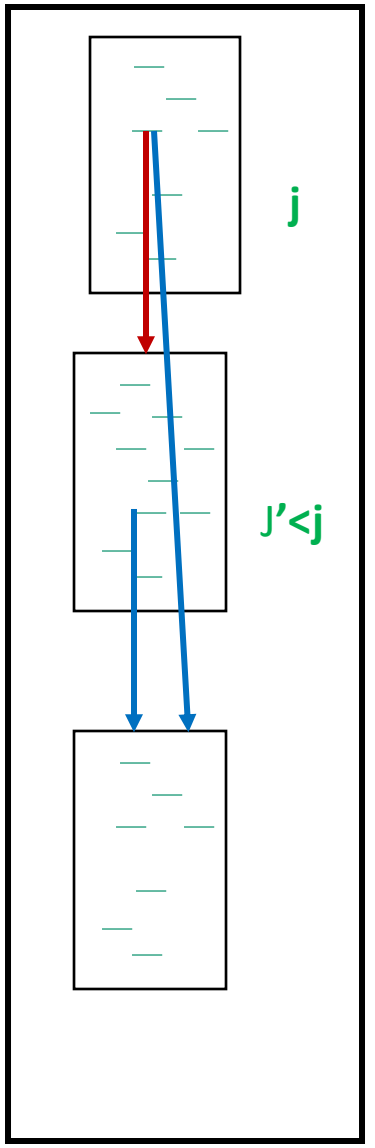
- Le pagine sono create in ordine e denotate 1, 2,. . , N



Esempio- Creazione di una pagina WEB

- Per semplicità, supponiamo che ogni pagina crei **un solo** link in uscita

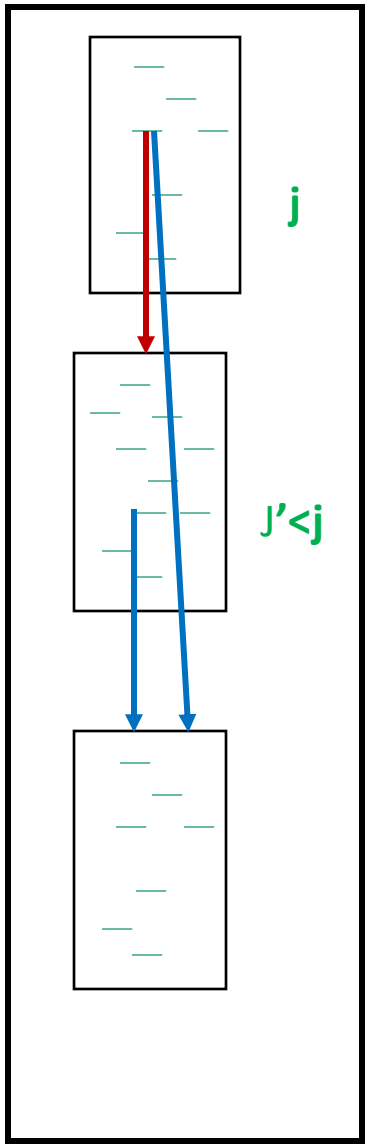
- Le pagine sono create in ordine e denotate 1, 2,. . , N
- Quando viene creata, la pagina j produce un collegamento a una pagina Web precedente in base alla seguente regola probabilistica
  - **Con probabilità p,** la pagina j sceglie una pagina in modo uniforme a caso tra tutte le pagine precedenti e crea un collegamento a questa pagina



Esempio- Creazione di una pagina WEB

- Per semplicità, supponiamo che ogni pagina crei **un solo** link in uscita

- Le pagine sono create in ordine e denotate 1, 2,. . , N
- Quando viene creata, la pagina j produce un collegamento a una pagina Web precedente in base alla seguente regola probabilistica
  - **Con probabilità p,** la pagina j sceglie una pagina in modo uniforme a caso tra tutte le pagine precedenti e crea un collegamento a questa pagina
  - **Con probabilità 1 - p,** la pagina j sceglie una pagina in modo uniforme a caso tra tutte le pagine precedenti e ne copia il link (cioè crea un collegamento alla pagina a cui essa rimanda)



Rich-get-Richer implica Power Law

- Il punto chiave del modello è che l'autore della pagina j con probabilità (1-p) copia la decisione dell'autore della pagina i

- Vedremo che, se N è molto grande, la frazione di pagine con k in-link sarà distribuita approssimativamente come

$$P(d_i = k) \propto k^{-\alpha(p)}$$ crescente nel valore di *p*

➔ **power law con α=$\alpha(p)$**

se *p* diminuisce (si copia più spesso), allora

l'esponente diminuisce,

(risulta più probabile l'esistenza di pagine molto popolari)

- Questo meccanismo di copiatura è un'implementazione di una dinamica Rich-get-Richer

Rich-get-Richer implica Power Law

Poniamo d_i = numero di pagine che puntano alla pagina *i*

Arriva la pagina t+1

Rich-get-Richer implica Power Law

Poniamo d_i = numero di pagine che puntano alla pagina *i*

Arriva la pagina t+1

Quando si copia la decisione di una pagina casuale precedente, la probabilità di collegarsi a qualche pagina è proporzionale al numero totale di pagine che attualmente linkano a tale pagina

Prob{t+1 decide di copiare il link e sceglie i}

$$=(1\text{-}p)\, d_i \frac{1}{\textit{numero di pagine esistenti}} = (1-p)\, \frac{d_i}{t}$$

Rich-get-Richer implica Power Law

Poniamo d_i = numero di pagine che puntano alla pagina *i*

Arriva la pagina t+1

Quando si copia la decisione di una pagina casuale precedente, la probabilità di collegarsi a qualche pagina è proporzionale al numero totale di pagine che attualmente linkano a tale pagina

Prob{t+1 decide di copiare il link e sceglie i}

$$=(1\text{-}p)\, d_i \frac{1}{\textit{numero di pagine esistenti}} = (1-p)\,\frac{d_i}{t}$$

In totale

$$\textit{Prob}\{t+1 \rightarrow i\} = p\frac{1}{t} + (1-p)\,\frac{d_i}{t}$$

Quindi, la probabilità che la pagina *i* aumenti la sua popolarità è proporzionale alla popolarità attuale di *i*

Approssimazione dei gradi

- **Poniamo** $d_j(t)$ funzione continua che rappresenta il grado di *j* al tempo *t*
- Nodo t arriva al tempo t

- Abbiamo $d_t(t)=0$ (nodo t è ultimo arrivato)
- $P(t+1 \rightarrow i)=p/t + (1-p)\, d_i(t)/t$

Incremento del grado

$$d_i(t+1) - d_i(t) = p\frac{1}{t} + (1-p)\frac{d_i(t)}{t}$$

Quando *t* cresce

- $\underbrace{d_i(t+1) - d_i(t)}_{} = p\frac{1}{t} + (1-p)\frac{d_i(t)}{t}$
- $\frac{\mathrm{d}d_i(t)}{\mathrm{d}t} = p\frac{1}{t} + (1-p)\frac{d_i(t)}{t} = \frac{p+qd_i(t)}{t}$ ← $q = (1-p)$

Quando *t* cresce

- $\underbrace{d_i(t+1) - d_i(t)}_{} = p\frac{1}{t} + (1-p)\frac{d_i(t)}{t}$
- $\frac{\mathrm{d}d_i(t)}{\mathrm{d}t} = p\frac{1}{t} + (1-p)\frac{d_i(t)}{t} = \frac{p+qd_i(t)}{t}$ ← $q = (1-p)$
- $\frac{1}{p+qd_i(t)}\mathrm{d}d_i(t) = \frac{1}{t}\mathrm{d}t$ ← **Dividiamo per** $p + q\, d_i(t)$

Quando *t* cresce

- $\underbrace{d_i(t+1) - d_i(t)}_{} = p\frac{1}{t} + (1-p)\frac{d_i(t)}{t}$
- $\frac{\mathrm{d}d_i(t)}{\mathrm{d}t} = p\frac{1}{t} + (1-p)\frac{d_i(t)}{t} = \frac{p+qd_i(t)}{t}$ ← $q = (1-p)$
- $\frac{1}{p+qd_i(t)}\mathrm{d}d_i(t) = \frac{1}{t}\mathrm{d}t$ ← Dividiamo per $p + q\, d_i(t)$
- $\int \frac{1}{p+qd_i(t)}\mathrm{d}d_i(t) = \int \frac{1}{t}\mathrm{d}t$ ← Integriamo

Quando *t* cresce

- $\underbrace{d_i(t+1) - d_i(t)}_{} = p\frac{1}{t} + (1-p)\frac{d_i(t)}{t}$
- $\frac{\mathrm{d}d_i(t)}{\mathrm{d}t} = p\frac{1}{t} + (1-p)\frac{d_i(t)}{t} = \frac{p+qd_i(t)}{t}$ ← $q = (1-p)$
- $\frac{1}{p+qd_i(t)}\mathrm{d}d_i(t) = \frac{1}{t}\mathrm{d}t$ ← Dividiamo per $p + q\,d_i(t)$
- $\int \frac{1}{p+qd_i(t)}\mathrm{d}d_i(t) = \int \frac{1}{t}\mathrm{d}t$ ← Integriamo
- $\frac{1}{q}\ln\left(p + qd_i(t)\right) = \ln t + c$

Quando *t* cresce

- $\underbrace{d_i(t+1) - d_i(t)}_{} = p\frac{1}{t} + (1-p)\frac{d_i(t)}{t}$
- $\frac{\mathrm{d}d_i(t)}{\mathrm{d}t} = p\frac{1}{t} + (1-p)\frac{d_i(t)}{t} = \frac{p+qd_i(t)}{t}$ ← $q = (1-p)$
- $\frac{1}{p+qd_i(t)}\mathrm{d}d_i(t) = \frac{1}{t}\mathrm{d}t$ ← Dividiamo per $p + q\, d_i(t)$
- $\int \frac{1}{p+qd_i(t)}\mathrm{d}d_i(t) = \int \frac{1}{t}\mathrm{d}t$ ← Integriamo
- $\frac{1}{q}\ln\left(p + qd_i(t)\right) = \ln t + c$ ← Esponenziamo e poniamo $A = e^c$
- $p + qd_i(t) = e^{qc}\, t^q \Rightarrow \boldsymbol{d_i(t) = \frac{1}{q}\left((At)^q - p\right)}$ **A=?**

Quando *t* cresce

$$\boldsymbol{d_i(t) = \frac{1}{q}((At)^q - p)}$$

Sapendo che d_i(i)=0, abbiamo

$$d_i(i) = \frac{1}{q}((Ai)^q - p) = 0$$

Quindi Ai^q=p, cioè

$$A^q = p(1/i^q)$$

Sostituendo, otteniamo

$$\boldsymbol{d_i(t) = \frac{p}{q}\left(\left(\frac{t}{i}\right)^q - 1\right)}$$

Power Law?

Abbiamo visto come cresce $\boldsymbol{d_i}$(t)

$$d_i(t) = \frac{p}{q}\left(\left(\frac{t}{i}\right)^q - 1\right)$$

Per un dato k, e un tempo t,

quale è la frazione dei nodi che hanno almeno k in-link al tempo t?

Poiché $\boldsymbol{d_i}$(t) approssima il numero di in-link del nodo t, consideriamo la disuguaglianza

$$d_i(t) = \frac{p}{q}\left(\left(\frac{t}{i}\right)^q - 1\right) \geq k$$

Power Law?

$$d_i(t) = \frac{p}{q}\left(\left(\frac{t}{i}\right)^q - 1\right) \geq k$$

➔

$$(t/i)^q \geq k(q/p) + 1 \quad \Leftrightarrow \quad (t/i) \geq [k(q/p) + 1]^{-q}$$

Quindi

$$i \leq t\left[\frac{q}{p}k + 1\right]^{-1/q}$$

Dividendo per il numero t di valori che i può assumere

$$\frac{t\left[\frac{q}{p}k+1\right]^{-1/q}}{t} = \left[\frac{q}{p}k + 1\right]^{-1/q}$$

otteniamo **la frazione di tutte le funzioni d_i(t) che soddisfano d_i(t) ≥ k**

Essendo p e q costanti, abbiamo che essa **è proporzionale a**

$$k^{-1/q}$$

Perchè Rich-Get-Reacher?

Quando si copia la decisione di una pagina casuale precedente, la probabilità di collegarsi a qualche pagina è proporzionale al numero totale di pagine che attualmente linkano a tale pagina

➔la probabilità che la pagina i aumenti la sua popolarità è proporzionale alla popolarità attuale di i

Questo fenomeno è anche noto come **preferential attachment**:

i link sono "preferenzialmente" a pagine che hanno già grande popolarità

Fornisce una ragione per cui la popolarità dovrebbe mostrare dinamiche del tipo rich-get-richer:

più qualcuno è conosciuto,

più è probabile che se ne senta il nome, e quindi

più è probabile che si finisca per conoscerlo

Non predicibilità dell'effetto Rich-Get-Richer

- Una volta che un elemento è ben consolidato, è probabile che le dinamiche di popolarità **Rich-Get-Richer** lo spingano ancora più in alto

- **L'ascesa alla popolarità** di qualsiasi oggetto di attenzione popolare è una cosa relativamente fragile

  - Le dinamiche della popolarità suggeriscono che **gli effetti casuali all'inizio del processo** giocano un ruolo fondamentale

  - Ripetendo esperimenti più volte risulta che
    - la distribuzione della popolarità è del tipo **Rich-Get-Richer** (quasi sempre)
    - gli articoli più popolari NON sono sempre gli stessi

Un esperimento interessante -- 1

- Salgankik, Dodds, and Watts hanno eseguito un esperimento che testimonia la fragilità del fenomeno rich-get-richer
  - Hanno creato un sito per il download di musica, popolato da 48 canzoni ignote
  - Ai visitatori è stato presentato un elenco di brani con il numero di download di ogni canzone ed è stata data l'opportunità di ascoltarli
  - Alla fine di una sessione, il visitatore poteva scaricare copie delle canzoni preferite

Un esperimento interessante -- 2

- Durante l'esperimento sono state usate 8 copie "parallele" del server
- Ogni copia parte con le stesse configurazioni iniziali (stesse canzoni e 0 download per ogni canzone)
- Visitatori
  - sono assegnati in modo casuale a una copia del server
  - sono inconsapevoli delle copie parallele
  - Possono ascoltare i brani e vedere i download di ogni brano
  - alla fine possono scaricare i brani preferiti
- **Risultato: nelle diverse copie parallele la popolarità delle canzoni variava molto**

  anche se le canzoni migliori non sono mai finite in fondo e le canzoni peggiori non sono mai finite in cima

Un esperimento interessante -- 2

- **Un nono server**

  è stato creato senza conteggi di download

  (➔ no feedback ➔ no dinamica rich-get-richer)

  La popolarità delle canzoni è cambiata significativamente

  - non ha mostrato un andamento power-law
  - Meno variazioni in popolarità tra brani

Feedback

L'esperimento illustra come

- Il successo di un libro, film, celebrità o sito Web è fortemente influenzato da questi tipi di effetti di feedback
- Quindi può essere intrinsecamente imprevedibile

Studi su feedback in molti ambiti, specialmente nelle recensioni

Recensioni iniziali positive ➔ recensioni migliori

Recensioni iniziali negative ➔ recensioni peggiori

Effetto degli algoritmi di ricerca sulla popolarità

Come Google (e i motori di ricerca) aiutano a scoprire cose "nascoste»

Google scansiona e cataloga **tutte le pagine web**, anche quelle con zero visite.

Quando un utente digita una query, Google non cerca solo i siti più popolari, ma quelli **più pertinenti** al significato della query

- **Query *particolari***

  Esempio: Immaginiamo di cercare *"come coltivare un raro fiore del deserto"*.

  Senza Google, dovremmo sfogliare libri di giardinaggio o chiedere a esperti

  Con Google, anche se pochissime persone cercano questo fiore, indica subito un blog di un appassionato o un forum specializzato.

- **Navigazione vs. Ricerca**

  Navigazione tradizionale: Se usano solo menu e categorie (es. su un sito di e-commerce), si vedono solo i prodotti più venduti o in homepage.

  Problema: I prodotti di nicchia restano invisibili.

  Ricerca con Google: Digitando parole precise, trovi anche oggetti "strani", perché il motore di ricerca scandaglia tutto il web, non solo le cose famose.

Effetto degli algoritmi di ricerca sulla popolarità

Contrastare le dinamiche Rich-get-Richer

- Senza motori di ricerca, solo i siti più famosi otterrebbero visite (es. Amazon o Wikipedia). I piccoli siti morirebbero.
- **Come Google rompe questo ciclo:**
  - Un blog sconosciuto ma *super pertinente* su *"come riparare orologi a cucù"* può apparire in prima pagina se la query è specifica.
  - Senza Google, solo i grandi siti generalisti vincerebbero, anche se hanno contenuti meno utili per quella ricerca.

Esempio:

- **Caso A (senza ricerca):** Vuoi un *"libro sugli squali preistorici"*. In libreria vedi solo i bestseller (es. *"Il grande squalo bianco"*). Quello di nicchia non lo trovi.
- **Caso B (con Google):** Cerchi *"Megalodon libri rari PDF"* e trovi un articolo di un'università o un ebook gratuito di un paleontologo oscuro.

Perché è importante?

Democratizza l'informazione: Piccoli creatori/negozi hanno una chance.

Evasione dalla "bolla": Gli **algoritmi dei social** (es. TikTok) mostrano solo cose già popolari. I motori di ricerca no.

Effetto degli algoritmi di ricerca sulla popolarità

Ricapitolando

- Le persone usano motori di ricerca come Google per trovare pagine
  - Google usa misure di popolarità per classificare le pagine Web e le pagine con un alto ranking sono le alternative preferite per il collegamento
    - Questo tipo di feedback rafforza le dinamiche rich-get-richer, producendo ancora più ineguaglianza nella popolarità
  - Gli utenti digitano una gamma molto ampia di query in Google
    - ottenendo risultati su query relativamente oscure, gli utenti vengono portati a pagine che probabilmente non avrebbero mai scoperto attraverso la sola navigazione
    - Gli strumenti di ricerca utilizzati in questo stile consentono alle persone di trovare più facilmente oggetti impopolari
    - Questo tipo di feedback contrasta dinamiche rich-get-richer

The Long Tail

- La distribuzione della popolarità può avere importanti conseguenze economiche, in particolare nel settore dei media

- La maggior parte delle vendite di media company (con un enorme inventario) è generata da
  - pochi articoli che sono enormemente popolari, o
  - molti articoli che sono individualmente meno popolari ?

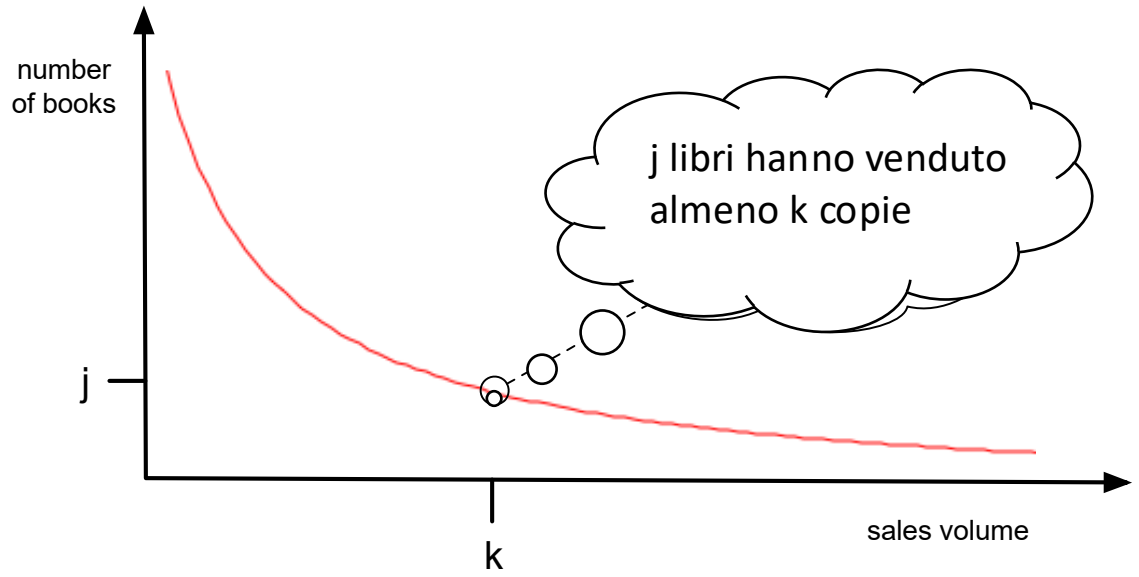
- La distribuzione basata su Internet e altri fattori stanno rendendo la seconda alternativa dominante

  Es. Aziende come Amazon hanno
  - enormi inventari senza restrizioni di negozi fisici ed
  - il loro volume di vendite consiste in *un'enorme quantità di prodotti, ciascuno venduto in quantità molto piccola*

The Long Tail

Consideriamo la seguente domanda

In funzione di k, quanti oggetti hanno popolarità $\geq k$



Funzione Power Law?

- A random variable X attains certain values, say for some constants β, α

$$\Pr[X = k] = \frac{\beta}{k^{\alpha}}, \text{ for } k = 1, 2, \ldots, n$$

- Then

$$\Pr[X \geq k] = \sum_{x=k}^{n} \Pr[X = x] = \sum_{x=k}^{n} \frac{\beta}{x^{\alpha}} \approx \beta \int_{k}^{n} \frac{1}{x^{\alpha}} dx$$

$$= \beta \left[-\frac{x^{1-\alpha}}{1-\alpha} \right]_{k}^{n} = \beta \frac{k^{1-\alpha}}{1-\alpha} - \beta \frac{n^{1-\alpha}}{1-\alpha}$$

- Note that β is a normalizing coefficient

$$1 = \sum_{k=1}^{n} \frac{\beta}{k^{\alpha}} \approx \beta \int_{1}^{n} \frac{1}{x^{\alpha}} dx$$

- A non-negative random variable X is said to have a *power law* distribution if

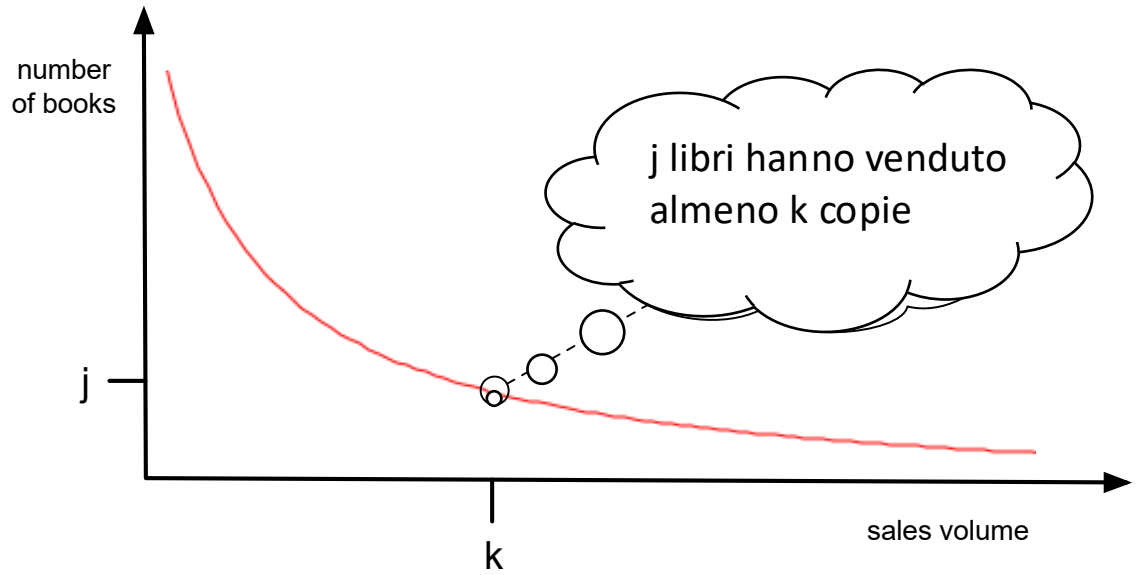
$$Pr[X \geq x] \sim cx^{-\alpha}$$

for constants $c > 0$ and $\alpha > 0$.

The Long Tail

Consideriamo la seguente domanda

In funzione di k, quanti oggetti hanno popolarità $\geq k$



Abbiamo ancora una distribuzione power law

➔

Quando k aumenta il numero di prodotti con popolarità $\geq k$ diventa sempre più piccolo

Quanti sono?

The Long Tail

- Quando guardiamo articoli sempre **meno popolari**, quali volumi di vendita vediamo?

- Scambiamo gli assi
  - Ordiniamo i prodotti in base alla "*classifica vendite*"
  - Osserviamo la popolarità dei libri mentre passiamo a livelli di vendita sempre più piccoli
  - L'area sotto la coda destra della curva è il volume delle vendite dovuto a *prodotti di nicchia*

Recommendation Systems

- **Per guadagnare** da un gigantesco inventario di prodotti di nicchia, un'azienda deve rendere i propri clienti consapevoli di **questi prodotti**

- Aziende come Amazon e Netfix hanno adottato **sistemi di raccomandazione** come parti importanti delle loro strategie aziendali
- strumenti di ricerca progettati per esporre le persone a elementi che potrebbero non essere generalmente popolari, ma che corrispondono agli interessi degli utenti come dedotti dalla loro cronologia degli acquisti precedenti

- Sistemi di raccomandazione: contrastare rich-get-richer effect