Power Laws

In molte situazioni 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

Ad esempio, la popolarità all'interno di un social network 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

Lo stesso si potrebbe dire di libri, film o quasi tutto ciò che dipende dalle scelte del pubblico.

Ci poniamo le seguenti domande:

- Come possiamo quantificare questi squilibri?
- Perché sorgono?
- Sono in qualche modo intrinseci all'idea di popolarità?

Popolarità sul WEB

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

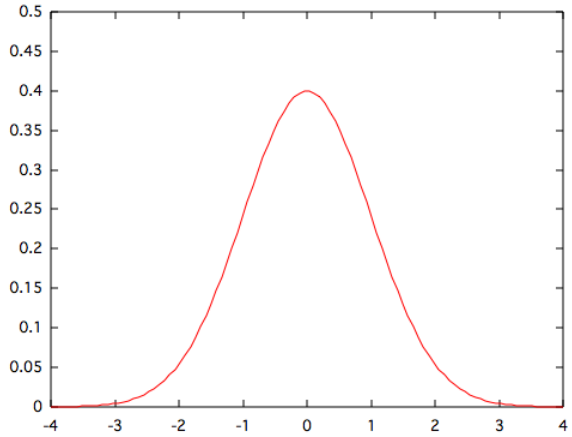
Possiamo prendere il *numero di in-link di una pagina Web* come misura della popolarità della pagina stessa.

A questo scopo, facciamo un'istantanea del Web completo e contiamo il numero di in-link per ogni pagina.

A questo punto possiamo valutare la distribuzione della popolarità sulle pagine Web come segue: In funzione del numero k di in-link, *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). Questa è usata ampiamente in probabilità e statistica ed è caratterizzato da due quantità: un valore medio e una deviazione standard attorno a tale valore-



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

La distribuzione normale è onnipresente nelle scienze naturali, per le quali si applica il teorema del limite centrale.

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.

Questo implica che la distribuzione della media del campione è nota anche se non si conosce la distribuzione della popolazione da cui è tratto il campione.

Ad esempio. 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.

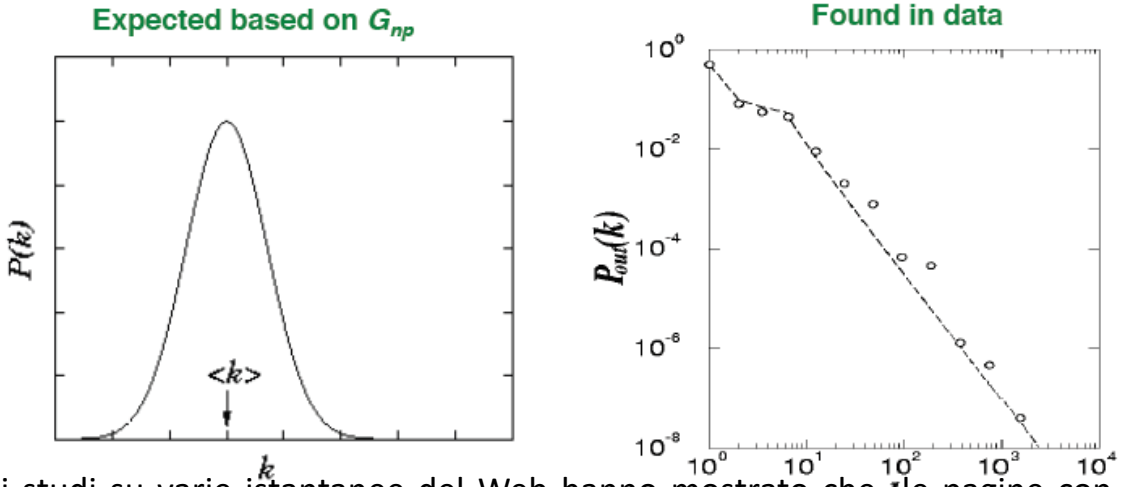
Quindi la distribuzione dei valori misurati dovrebbe essere approssimativamente normale

La distribuzione 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.

Questo implicherebbe che il numero di collegamenti in entrata a una determinata pagina è la somma di molte quantità casuali indipendenti e viene normalmente distribuito.Ne consegue che il numero di pagine con k in-link dovrebbe diminuire esponenzialmente in k.

Diverse misurazioni hanno però dimostrato che ciò è sbagliato.



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 aspetteremmo con una distribuzione normale.

In particolare, la frazione di pagine Web con k in-link è approssimativamente proporzionale a $\boldsymbol{1/k^2}$

Power Laws

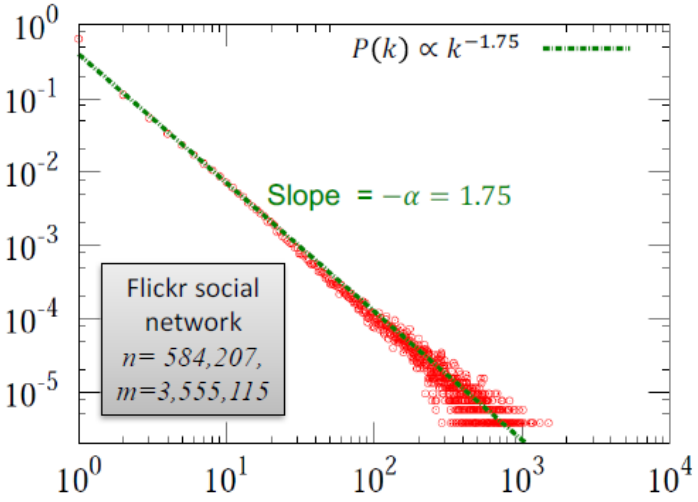
Una funzione che diminuisce come k^{-c} è detta **power law.**

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?



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

Ponendo $f(k) = bk^{-\alpha}$ **otteniamo** $\log f(k) = \log b - \alpha \log k$.

Quindi una distribuzione power law su un grafico di tipo log-log, si presenta come una *retta*: α è la pendenza della linea e *log b* è l'intersezione con l'asse y

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; quindi vale il Teorema del Limite Centrale
- Le distribuzioni Power law derivano dal feedback introdotto da decisioni correlate su una popolazione.

Modello Rich-Get-Richer

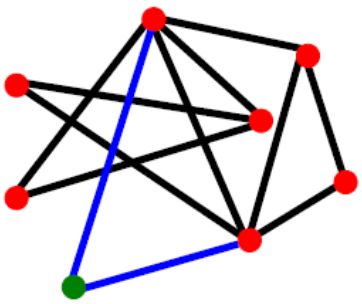
È un modello basato su alcune conseguenze osservabili dei processi in presenza di cascate di informazioni

Tale modello assume che le persone hanno la tendenza a copiare con maggiore probabilità sia le decisioni delle persone che agiscono prima di loro, sia le decisioni degli individui popolari. Come esempio consideriamo lo schema di preferential attachment per la formazione dei link in una rete.

Preferential Attachment

Assumiamo che:

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



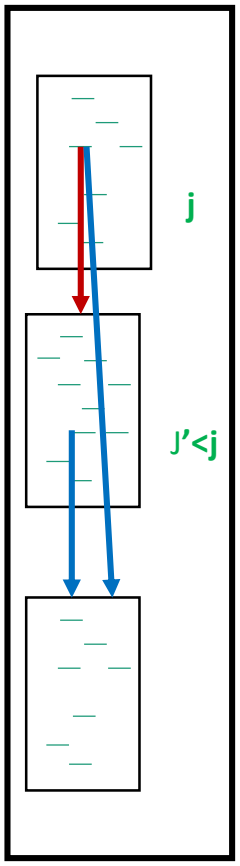
Questo schema implica che i nodi hanno una probabilità maggiore di scegliere un nodo con grado già alto.

Ad esempio, 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.*

Come altro esempio, consideriamo l' effetto Matteo [Merton, 1968], studiato in sociologia, *"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

Ritornando al preferential attachment, per semplicità, supponiamo che ogni pagina crei **un solo** link in uscita. Quindi,

- 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 crea un collegamento alla pagina che cui essa rimanda



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

Si mostrache 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*

Si ottiene così una funzione power law con $\alpha=\alpha(p)$. Se p diminuisce (si copia più spesso), allora l'esponente diminuisce e risulta più probabile l'esistenza di pagine molto popolari.

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

Rich-get-Richer implica Power Law

Poniamo d_i pari al numero di pagine che puntano alla pagina *i.*

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

$$Prob\{t+1 \text{ sceglie } i\} = (1-p)\, d_i \frac{1}{\text{numero di pagine esistenti}} = (1-p)\, \frac{d_i}{t}$$

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

In totale

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

Approssimazione dei gradi

Sia d_j(t) una funzione continua che rappresenta il grado di *j* al tempo *t.*

Poiché il nodo j arriva al tempo j, abbiamo d_j(t)=0 (nodo j è ultimo arrivato).

La probabilità che il nodo t+1 scelga i come nodo a cui puntare è

$$P(t+1 \rightarrow i)=p/t + (1-p)\, d_i(t)/t$$

L'incremento del grado possiamo quindi scriverlo come

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

Al crescere di t avremo

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

Quindi

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

Ricordando che d_i(i)=0, otteniamo

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

Da cui

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

Abbiamo visto come cresce $\boldsymbol{d_i}$(t). Ci chiediamo ora, per un dato k e un tempo t,

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

Poiché d_i(t) approssima il numero di in-link del nodo t, valutiamo quale è la frazione di tutte le funzioni d_i(t) che soddisfano d_i(t) ≥ k?

Poniamo quindi

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

Otteniamo

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

Dividendo per il numero t di valori di i, otteniamo la frazione

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

di tutte le funzioni d_i(t) che soddisfano d_i(t) ≥ k.

Essendo p e q costanti, abbiamo che tale frazione è proporzionale a

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

Perchè Rich-Get-Reacher?

Abbiamo visto che, 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 e, di conseguenza, la probabilità che la pagina i aumenti la sua popolarità è proporzionale alla popolarità attuale di i.

Questo fenomeno, noto come **preferential attachment**, 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. Tuttavia, 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 giochino un ruolo fondamentale. Infatti, ripetendo esperimenti più volte risulta che la distribuzione della popolarità è del tipo Rich-Get-Richer (quasi sempre), ma gli articoli più popolari NON sono sempre gli stessi

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

Durante l'esperimento sono state usate 8 copie "parallele" del server; ogni copia partiva con le stesse configurazioni iniziali (stesse canzoni e 0 download per ogni canzone). I visitatori sono stati assegnati in modo casuale a una copia del server:

- Erano inconsapevoli delle copie parallele;
- potevano ascoltare i brani e vedere i download di ogni brano
- alla fine potevano 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)

Esisteva però anche un nono server che era stato creato senza conteggi di download.

Quindi senza feedback e, di conseguenza, senza dinamica rich-get-richer.

In questo caso la popolarità delle canzoni è cambiata significativamente, non ha mostrato un andamento power-law e vi sono state meno variazioni in popolarità tra brani.

L'esperimento illustra come il successo di un libro, film, celebrità o sito Web è fortemente influenzato da questi tipi di effetti di **feedback** e può quindi essere intrinsecamente imprevedibile.

The Long Tail

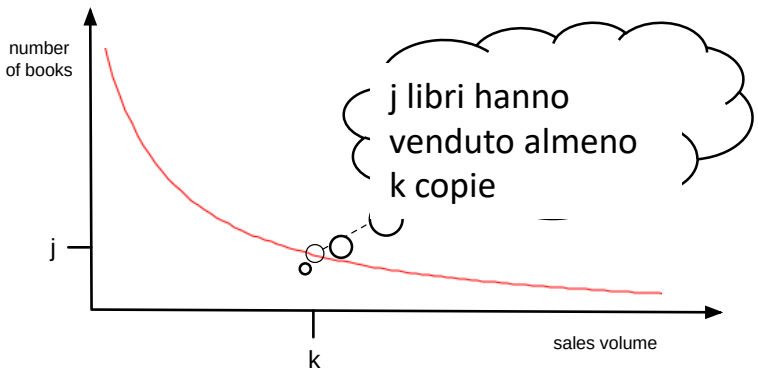
La distribuzione della popolarità può avere importanti conseguenze economiche, in particolare nel settore dei media

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.

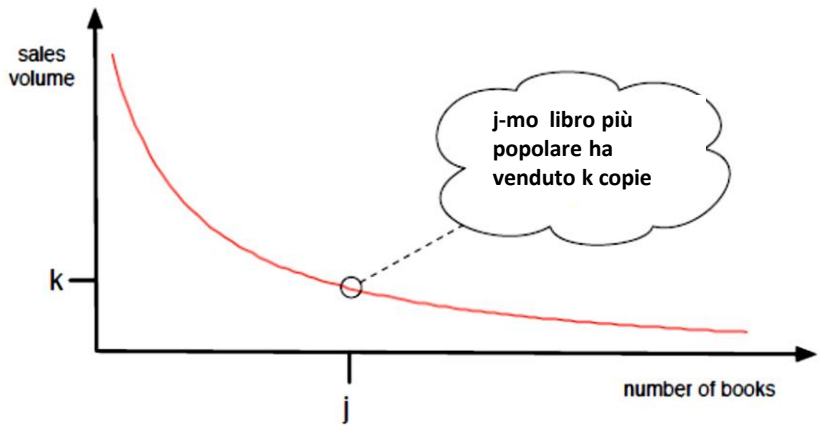
Es. Aziende come Amazon hanno enormi inventari senza restrizioni di negozi fisici ed il loro volume di vendite consiste in *un'enorme quantità di prodotti, ciascuno venduto in quantità molto piccola*

Consideriamo la seguente domanda. In funzione di k, quanti oggetti hanno popolarità $\geq k$?



Abbiamo ancora una distribuzione power law. Quindi, quando k aumenta il numero di prodotti con popolarità $\geq k$ diventa sempre più piccolo. Ma quanti sono?

Scambiamo gli assi ed ordiniamo i prodotti in base alla "classifica vendite«. Osserviamo quindi 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, come parti importanti delle loro strategie aziendali, i sistemi di raccomandazione:

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