Web-search

Reti e informazione

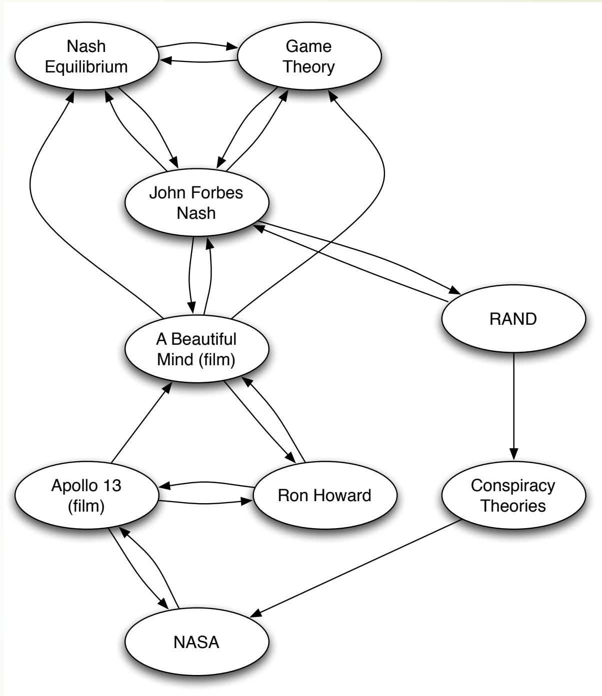
- Il materiale descritto in queste lezioni costituisce i Capitoli 13 e 14 del testo.
- Nella serie di lezioni corrispondenti a questi lucidi ci interessiamo al problema di individuare all'interno di una rete ed estrarre da essa informazioni rilevanti ad una data richiesta
- e lo faremo nel caso in cui la rete sia il Web
  - che è una rete diversa dalle reti fino ad ora considerate in questo corso
    - che sono costituite da individui collegati da relazioni
    - o da dispositivi di calcolo collegati mediante dispositivi fisici (reti wireless)
  - perché è una rete costituita da documenti
  - collegati mediante hyperlink
- Naturalmente, utilizzeremo ancora il grafo come modello di rete
  - e, questa volta, lo strumento sarà l'algebra lineare

Il World Wide Web

- È un'applicazione progettata per condividere informazione su Internet
  - fra il 1989 e il 1991, da Tim Berners-Lee
- il cui progetto è basato su una architettura client-server
  - documenti sono resi disponibili, sotto forma di pagine web, memorizzandoli in zone pubblicamente accessibili di taluni computer (server)
  - e l'accesso a tali pagine avviene mediante un'applicazione (browser) eseguita dal client e che accede agli spazi pubblici dei server
- Il principio logico fondazionale del web è l'**ipertesto**
  - nel quale l'informazione è organizzata in una struttura di rete
  - ossia, i documenti che costituiscono il web sono nodi di un grafo diretto
  - si tratta, perciò di una organizzazione non lineare
- L'**ipertesto** è una implementazione del concetto di *memoria associativa*
  - che ha una lunga storia...

Memoria associativa

- Esistono, in effetti, numerosi precursori dell'ipertesto
  - le citazioni – quando nei libri o negli articoli si citano altri libri o articoli che contengono, ad esempio, informazioni supplementari
  - i crossing reference – i riferimenti incrociati, che permettono di collegare argomenti diversi mediante relazioni di vario tipo
    - come nell'esempio in figura, nel quale, si riesce a trovare una catena di collegamenti che mettono in relazione la teoria dei giochi con l'Apollo 13
    - catena di collegamenti che ricorda un po' il flusso di coscienza, ossia il modo in cui avvengono le associazioni mentali
  - le mappe concettuali – che molti studenti sperimentano nella preparazione di tesine per vari livelli di esami...



MEMEX

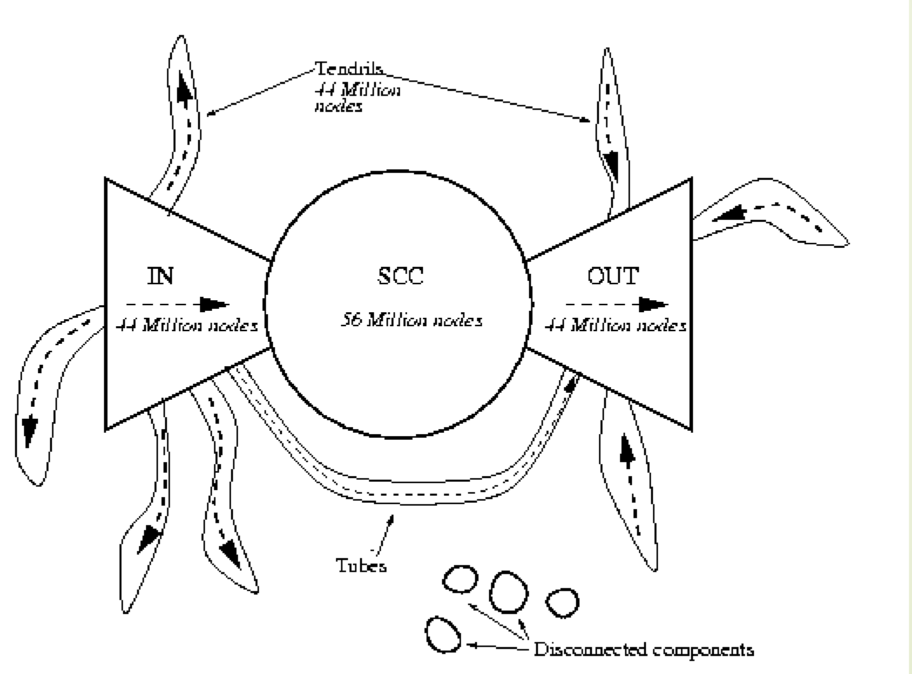
- Vannevar Bush, in un articolo del 1945 ("*As we may think*") provò a immaginare in che modo la nascente tecnologia dell'informazione poteva modificare il modo
  - in cui le informazioni vengono memorizzate
  - e con il quale si può accedere ad esse
- Egli osservò che, mentre i metodi tradizionali per rappresentare l'informazione (in libri, ma anche in biblioteche) sono sostanzialmente lineari
  - un esempio eclatante è il dizionario enciclopedico
  - una collezione di dati elencati *sequenzialmente*
- il modo in cui l'essere umano pensa e memorizza (e accede a) le sue esperienze è *associativo*
  - inizia a pensare a una cosa, che gliene fa venire in mente una seconda, e così via
- Bush immaginò anche un prototipo: MEMEX
  - una versione digitalizzata della conoscenza umana nella quale i dati sono collegati da link associativi
- E, in effetti, Bush è stato citato esplicitamente da Tim Berners Lee nel suo progetto del Web

Dal Web al Web 2.0

- Inizialmente, il Web era costituito da un insieme di pagine al cui interno erano (possibilmente) presenti alcuni **hyperlink** ciascuno dei quali collegava quella pagina ad un'altra pagina
  - sono appunto gli **hyperlink** che implementano l'idea di *memoria associativa*!
- Con il tempo, il Web si è evoluto
  - intorno all'ossatura costituito dalle pagine "statiche",
    - con un contenuto e hyperlink che portano in altre pagine
  - sono state progettate anche pagine che permettevano di *eseguire azioni*
    - sottomettere una richiesta, effettuare un pagamento...
  - i link contenuti in queste pagine permettono di concludere le azioni
  - e sono detti ***transazionali***, per distinguerli dai link ***navigazionali***
- E altre evoluzioni sono avvenute, con lo sviluppo di strumenti
  - per la creazione e il mantenimento di contenuti condivisi (Wikipedia)
  - per il mantenimento di dati personali on-line (web-mail)
  - per la connessione di individui, invece che di documenti (social networks)
- evoluzioni così rilevanti che si è parlato di avvento del Web 2.0

La struttura del Web

- Anche il Web ''navigazionale'' può essere modellato come un grafo- in particolare, come un grafo diretto
  - i cui nodi sono le pagine
  - ed esiste un arco diretto (u, v) se la pagina u contiene un hyperlink alla pagina v
    - in questo caso diciamo che ***u punta a v***
  - osserviamo che i nodi non hanno coscienza degli archi entranti: una pagina Web non ha modo di conoscere le pagine che la puntano!
- In [Broder et al., 2000] viene studiata la struttura del grafo del web
  - ogni nodo del grafo studiato è una SCC del grafo del Web
  - ed è stato rilevato che esso aveva una struttura bow-tie contenente una (unica) SCC gigante



Ricerca e classificazione

- La ricerca di documenti è un problema antico
- ma, tradizionalmente, era richiesta in ambienti specifici
  - ossia, cercatore e cercato erano dello stesso ambiente
  - ad esempio, avvocati che cercavano “precedenti” sentenze di tribunali
  - o medici che cercavano pazienti con le stesse sintomatologie dei pazienti che si trovavano a dover trattare
  - e ciascuna categoria cercava i documenti di interesse nelle collezioni di documenti su quell’argomento
- e, tipicamente, la difficoltà della ricerca era causata dalla scarsità di documenti disponibili
- L’avvento del Web ha cambiato sostanzialmente la questione:
  - il Web è un’immensa collezione di documenti appartenenti alle categorie più disparate
  - e chiunque cerca nel Web: lo scienziato, l’avvocato, il medico, il letterato
- e la difficoltà è ora la sovrabbondanza di documenti disponibili

Ricerca non specializzata

- Tradizionalmente, la ricerca di documenti avveniva in ambienti specifici
  - cercatore e cercato erano dello stesso ambiente
  - così che le parole chiave erano in numero limitato
  - e ciascuna di esse aveva un significato ben definito, in riferimento al settore nel quale avveniva la ricerca
- L'avvento del Web ha cambiato sostanzialmente la questione:
  - il Web è un'immensa collezione di documenti appartenenti alle categorie più disparate
  - e chiunque cerca nel Web: lo scienziato, l'avvocato, il medico, il letterato
  - così che *il numero di parole chiave da utilizzare per effettuare una ricerca sul web è potenzialmente illimitato*
  - e <u>la stessa parola può avere tanti diversi significati</u>, dipendentemente dal settore al quale la si riferisce
  - ad esempio , la parola "albero"
    - che è, certamente, un membro del regno vegetale
    - o un albero genealogico
    - o anche un grafo privo di cicli…

La necessità di sfoltire le alternative

- Nella ricerca tradizionale la fonte di difficoltà era la scarsità di documenti disponibili
- Ma, con l'avvento del Web la questione si è sostanzialmente ribaltata:
  - il Web contiene un'infinità di documenti collegati, in qualche modo, alle parole chiave utilizzate per eseguire una ricerca
  - e per un individuo non sapere dove cercare un documento o sapere di doverlo cercare in una collezione di qualche milione di documenti... è all'incirca la stessa cosa!
- Perciò, affinché uno strumento di web search sia effettivamente efficace, è necessario che disponga di mezzi che gli permettano di "sfoltire" l'immensa mole di pagine individuate in prima battuta come risposta ad una richiesta
  - eliminando le pagine ritenute meno rilevanti
- così da proporre al cercatore una lista di pagine relativamente breve
- e, soprattutto, nella quale le pagine siano elencate in ordine di (presunta) rilevanza non decrescente
  - ossia, individuando un **ranking** delle pagine più rilevanti

Link analysis

- Le caratteristiche interne a una pagina non consentono, da sole, di capire quanto una pagina sia rilevante per una ricerca
- Gli hyperlink sono strumento che può essere utile a questo scopo
  - per una ragione ben nota!
- ESEMPIO: stiamo preparando la nostra tesi di laurea
  - il relatore ci indica un articolo di rassegna sull'argomento sul quale verterà la tesi, e ci dice che, per approfondire, possiamo attingere alla bibliografia di quell'articolo
    - ossia, agli articoli da esso citati
  - allora, andiamo fiduciosi a guardare la bibliografia e ci accorgiamo che consta di 574 riferimenti
    - ARGH!
  - Mica possiamo leggerli tutti quanti!
    - Ci piacerebbe laurearci prima di raggiungere l'età pensionabile....
  - Come facciamo a capire quali sono quelli più rilevanti?!
  - Beh, se c'è in bibliografia un articolo che è citato dalla maggioranza degli altri, esso sarà, probabilmente, un articolo molto rilevante per il nostro argomento!

Link analysis

- Generalizziamo dall'esempio:
- dopo aver individuato un insieme di pagine significative rispetto a una richiesta
  - i riferimenti nella bibliografia dell'articolo di rassegna
- consideriamo maggiormente rilevanti le pagine che sono molto puntate da altre pagine nell'insieme
  - i riferimenti che sono molto referenziati dagli altri
- Ossia, è la struttura stessa del grafo del Web
  - o meglio, della porzione del grafo che ha qualche attinenza con la richiesta
- a fornire indicazioni sulla rilevanza di una pagina
- **La rilevanza di una pagina è calcolata tramite analisi dei link che la coinvolgono**

HITS: Hubs and Authorities

- Dopo aver individuato un insieme di pagine rilevanti per una richiesta
- consideriamo maggiormente rilevanti le pagine che sono molto puntate da altre pagine nell'insieme
- Ma una pagina potrebbe essere puntata da tante pagine poco rilevanti ai fini della richiesta
  - ad esempio, sappiamo che per finalità di marketing possono essere acquistati puntatori ad una data pagina, così da far aumentare il suo contatore di *in-link*
- Allora, è necessario considerare la seguente questione: quando una pagina è autorevole nel suo conferire rilevanza ad una pagina alla quale punta?
- Il metodo Hyperlink-Induced Topic Search (*HITS*; conosciuta anche come Hubs and Authorities) proposto da Kleinberg è un modo per considerare la questione
- Osserviamo, intanto, che ad ogni pagina possiamo associare due indici
  - un indice di **autorità** – che esprime la **rilevanza** della pagina ai fini della ricerca
  - un indice di **hub** (= centro, fulcro) - che esprime l'**autorevolezza** *della pagina a conferire autorità* alle pagine alle quali punta

HITS: Hubs and Authorities

- Potremmo pensare, ad esempio, che il valore di **autorità** di una pagina i, che indichiamo con a_i, sia il numero di pagine nell'insieme di pagine attinenti alla ricerca che puntano ad essa
- ossia, $a_i = |\{ j: j \text{ è attinente alla ricerca e } j \rightarrow i\}|$
  - dove con "$j \rightarrow i$" indichiamo che la pagina j punta alla pagina i
  - ovvero, che la pagina j contiene un hyperlink alla pagina i
- così che, se indichiamo con M la matrice di adiacenza del sottografo del Web attinente alla ricerca – e assumiamo che esso contenga n pagine
  - ossia, per $1 \leq i,j \leq n$, $M[i, j] = \begin{cases} 1 & \text{se } i \rightarrow j \\ 0 & \text{altrimenti} \end{cases}$
- allora, $a_i = \sum_{1 \leq j \leq n} M[j, i]$
- Analogamente, il valore di **hub** di una pagina i, h_i, potrebbe essere il numero di pagine nell'insieme di pagine attinenti alla ricerca alle quali essa punta
- ossia, $h_i = |\{ j: j \text{ è attinente alla ricerca e } i \rightarrow j\}|$
- e, quindi, $h_i = \sum_{1 \leq j \leq n} M[i, j]$

HITS: Hubs and Authorities

- In realtà, possiamo raffinare l'idea appena illustrata osservando che
  - il valore di autorità di una pagina è tanto più elevato quanto più autorevoli sono le pagine che la puntano
  - e, simmetricamente, il valore di hub di una pagina è tanto più elevato quanto più rilevanti sono le pagine a cui essa punta
- ossia, se, in qualche modo ci venisse suggerito un valore di hub iniziale $\mathbf{h}_i^{(0)}$ per ciascuna pagina i, allora,

$$\mathbf{a}_i^{(1)} = \sum_{1 \leq j \leq n} M[j, i]\ \mathbf{h}_j^{(0)}$$

- ma, allora, essendo variato il valore di autorità di ciascuna pagina, potremmo ricalcolare i valori di hub per ottenere una valutazione più raffinata:

$$\mathbf{h}_i^{(1)} = \sum_{1 \leq j \leq n} M[i, j]\ \mathbf{a}_j^{(1)}$$

- E, a questo punto, i nuovi valori di hub ci permetterebbero di raffinare i valori di authorities, e così via
- ottenendo un procedimento iterativo:

$$\mathbf{a}_i^{(k+1)} = \sum_{1 \leq j \leq n} M[j, i]\ \mathbf{h}_j^{(k)} \qquad \text{e} \qquad \mathbf{h}_i^{(k+1)} = \sum_{1 \leq j \leq n} M[i, j]\ \mathbf{a}_j^{(k+1)}$$

HITS: Hubs and Authorities

- Se, in qualche modo, ci venisse suggerito un valore di hub iniziale $h_i^{(0)}$ per ciascuna pagina i, allora,

$$a_i^{(1)} = \sum_{1 \leq j \leq n} M[j, i]\ h_j^{(0)}$$

- ma, in genere, all'inizio non si hanno informazioni circa il valore di hub delle pagine che si stanno considerando
- In questo caso, si può assumere che tali valori siano uguali per tutte le pagine, ossia: $\mathbf{h_i^{(0)} = 1}$ **per ogni** $\mathbf{1 \leq i \leq n}$

- Le forme $a_i^{(k+1)} = \sum_{1 \leq j \leq n} M[j, i]\ h_j^{(k)}$ e $h_i^{(k+1)} = \sum_{1 \leq j \leq n} M[i, j]\ a_j^{(k+1)}$ possono essere scritte in forma compatta, ossia:

$$\mathbf{a^{(k+1)} = M^T h^{(k)}} \quad \text{e} \quad \mathbf{h^{(k)} = M\, a^{(k)}}$$

  - dove $a^{(k+1)}$, $a^{(k)}$ e $h^{(k)}$ sono vettori colonna – ad esempio, $a^{(k)} = (a_1^{(k)}, a_2^{(k)}, \dots, a_n^{(k)})$
  - M^T è la matrice trasposta di M (ossia, $M^T[i, j] = M[j, i]$)
  - $M^T h^{(k)}$ e $M\, a^{(k)}$ indicano il consueto prodotto righe per colonne fra matrice e vettore

HITS: Hubs and Authorities

- Abbiamo, quindi, introdotto un metodo iterativo che calcola per raffinamenti successivi il valore di hub e di authority di ciascuna pagina:

$$a^{(k+1)} = M^T h^{(k)} \quad e \quad h^{(k)} = \sum_{1 \le j \le n} M\, a^{(k)}$$

  - partendo dal vettore hub iniziale $h_i^{(0)} = 1$ per ogni $1 \le i \le n$
- L'obiettivo è, naturalmente, ottenere ad ogni iterazione valori che descrivano meglio (rispetto all'iterazione precedente) la rilevanza di una pagina ai fini della ricerca
- ma, perché questo accada, è necessario che vi sia una sorta di convergenza verso certi valori limite che, in tal caso, sarebbero i "veri" valori di hub e di authority delle pagine
  - il procedimento non sarebbe di alcuna utilità se, ad esempio, esistessero due pagine i e ℓ tali che $a_i^{(k)} > a_\ell^{(k)}$, e $a_i^{(k+1)} < a_\ell^{(k+1)}$, e $a_i^{(k+2)} > a_\ell^{(k+2)}$, ...
- Osserviamo subito che convergenza vera e propria non si può avere: infatti le espressioni che permettono di calcolare i vettori $a^{(k)}$ e $h^{(k)}$ sono somme di termini non negativi
  - e, quindi, tendono a crescere con il numero di iterazioni
- Dunque, possiamo parlare di convergenza solo in presenza di opportuna **normalizzazione**

HITS: Hubs and Authorities

- In effetti, vale il seguente

  Teorema. Esistono un valore $c \in \mathbb{R}^+$ e un vettore $z \in \mathbb{R}^n$ non nullo tali che comunque si scelga un vettore $\mathrm{h}^{(0)}$ a coordinate positive,

$$\lim_{k \to \infty} \frac{\mathrm{h}^{(k)}}{c^k} = z \quad \text{e} \quad \lim_{k \to \infty} \frac{\mathrm{a}^{(k)}}{c^k} = z$$

- Prima di dimostrare il teorema, richiamiamo qualche nozione di algebra lineare:
- data una matrice quadrata $n \times n$ A, un vettore (complesso) n-dimensionale x e un numero $\lambda \in \mathbb{C}$ non nullo sono, rispettivamente, un ***autovettore*** e un ***autovalore*** per A se

$$\mathbf{Ax = \lambda\, x}$$

- Vediamo, adesso, un paio di strumenti concernenti la relazione fra la struttura di A e quella degli insiemi degli autovettori e degli autovalori di A che giocheranno un ruolo fondamentale ai fini della dimostrazione della convergenza del metodo HITS
  - ossia, del Teorema enunciato all'inizio di questa pagina

HITS: Hubs and Authorities

- **Teorema.** Se A è una matrice reale e simmetrica n × n, allora A ha n autovettori e n autovalori reali e, inoltre, gli autovettori formano una base ortonormale per $\mathbb{R}^n$

  - una **base** per $\mathbb{R}^n$ è un insieme di n vettori $z_1, z_2, \ldots, z_n$ tali che:

    ogni altro vettore x in $\mathbb{R}^n$ può essere espresso come combinazione lineare di $z_1, z_2, \ldots, z_n$, ossia,

    $$\forall\, x \in \mathbb{R}^n : x = \sum_{1 \leq i \leq n} p_i\, z_i$$

  - **ortonormale**: detto $z_i = (z_{i1}, z_{i2}, \ldots, z_{in})$, si ha che

    per ogni i = 1, … , n : $z_i \cdot z_i = z_{i1}^2 + z_{i2}^2 + \ldots + z_{in}^2 = 1$ (normalità)

    e

    per ogni $i \neq \ell$: $z_i \cdot z_\ell = z_{i1} z_{\ell 1} + z_{i2} z_{\ell 2} + \ldots + z_{in} z_{\ell n} = 0$ (ortogonalità)

HITS: Hubs and Authorities

- Una matrice A n × n simmetrica e reale è *semidefinita positiva* se

  $$\forall\, \mathbf{x} \in \mathbb{R}^n : \mathbf{x}^T \mathbf{A}\, \mathbf{x} \geq \mathbf{0}$$

  - NB: con x si indica il vettore colonna e x^T è il vettore riga trasposto di x.
  - NB: x^T A x è un numero reale

- **Teorema**. Se A è una matrice semidefinita positiva allora:
  1) gli autovalori di A sono non negativi
  2) se A è non nulla allora almeno un autovalore di A è strettamente positivo

HITS: Hubs and Authorities

- **Teorema**. Comunque si scelga un vettore $\mathrm{h}^{(0)}$ a coordinate positive, esistono un valore $c \in \mathbb{R}^+$ e un vettore $z \in \mathbb{R}^n$ non nullo tali che $\lim_{k \to \infty} \frac{\mathrm{h}^{(k)}}{c^k} = z$
- Vedremo a breve il ruolo giocato dal **Teorema** e dal **Teorema** nella dimostrazione del **Teorema** sopra...
- Intanto, osserviamo che
  - $(M\,M^T)$ è una matrice reale e simmetrica
    - infatti: $(M\,M^T)_{ij} = \sum_{1 \le k \le n} M_{ik} M_{kj} = (M\,M^T)_{ji}$
  - allora, $(M\,M^T)$ ha una base ortonormale di autovettori per $\mathbb{R}^n$ e tutti gli autovalori reali
  - $(M\,M^T)$ è semidefinita positiva
    - infatti: comunque scelto $x \in \mathbb{R}^n$ si ha $x^T (M\,M^T)\, x = (x^T M)\,(M^T x)$ e dunque,

      ponendo $(M^T x) = y = (y_1, y_2, \ldots, y_n)$

      si ottiene $x^T (M\,M^T)\, x = y^T y = \sum_{1 \le k \le n} y_k^2 \ge 0$
  - allora, poiché $(M\,M^T)$ è non nulla, il suo autovalore massimo è strettamente positivo

HITS: Hubs and Authorities

- **Teorema**. Comunque si scelga un vettore $h^{(0)}$ a coordinate positive, esistono un valore c ∈ $\mathbb{R}^+$ e un vettore z ∈ $\mathbb{R}^n$ non nullo tali che $\lim_{k \to \infty} \frac{h^{(k)}}{c^k} = z$

  (Teorema e dimostrazione analoghi valgono per $a^{(k)}$)
- Dimostrazione.
  - Cominciamo scrivendo in modo diverso le formule per calcolare $h^{(k)}$ e $a^{(k)}$:

    $h^{(1)} = M\, a^{(1)} = M\, M^T\, h^{(0)}$ poiché $a^{(k+1)} = M^T\, h^{(k)}$ per ogni k,

    $h^{(2)} = M\, a^{(2)} = M\, M^T\, h^{(1)} = (M\, M^T)\, (M\, M^T)\, h^{(0)} = (M\, M^T)^2\, h^{(0)}$

    $h^{(3)} = M\, a^{(3)} = M\, M^T\, h^{(2)} = (M\, M^T)\, (M\, M^T)^2\, h^{(0)} = (M\, M^T)^3\, h^{(0)}$ ...

  - e così via, induttivamente, possiamo esprimere $h^{(k)}$ nelle due forme

    $h^{(k)} = M\, M^T\, h^{(k-1)}$ e $h^{(k)} = (M\, M^T)^k\, h^{(0)}$

HITS: Hubs and Authorities

- **Teorema**. Comunque si scelga un vettore $\mathrm{h}^{(0)}$ a coordinate positive, esistono un valore c ∈ $\mathbb{R}^+$ e un vettore z ∈ $\mathbb{R}^n$ non nullo tali che $\lim_{k \to \infty} \frac{h^{(k)}}{c^k} = z$
- Dimostrazione. Scegliamo un vettore $\mathrm{h}^{(0)}$ a coordinate positive
  - Abbiamo mostrato che $h^{(k)} = (M\, M^T)^k\, h^{(0)}$
  - e $(M\, M^T)$ è una matrice simmetrica, semidefinita positiva e non nulla
    - siano $z_1, z_2, \ldots, z_n$ i suoi autovettori reali – **base** ortonormale per $\mathbb{R}^n$
    - e sia, per i = 1, … , n, c_i l'autovalore **reale e non negativo** corrispondente a z_i ossia $(MM^T)z_i=c_iz_i$
    - senza perdita di generalità, assumiamo che $c_1 \geq c_2 \geq \ldots \geq c_n$ ricordiamo che $\mathbf{c_1 > 0}$
  - **allora**, possiamo esprimere $h^{(0)}$ come combinazione lineare degli autovettori:

$$h^{(0)} = \sum_{1 \leq i \leq n} q_i\, z_i$$

  - e, dunque, $h^{(k)} = (M\, M^T)^k\, h^{(0)} = (M\, M^T)^k \sum_{1 \leq i \leq n} q_i\, z_i$

$$= (M\, M^T)^{k-1} \sum_{1 \leq i \leq n} q_i\, (M\, M^T)\, z_i$$

$$= (M\, M^T)^{k-1} \sum_{1 \leq i \leq n} q_i\, c_i\, z_i$$ **per definizione di autovettore**

$$= (M\, M^T)^{k-2} \sum_{1 \leq i \leq n} q_i\, c_i\, (M\, M^T)\, z_i = (M\, M^T)^{k-2} \sum_{1 \leq i \leq n} q_i\, c_i^2\, z_i$$

$$= \ldots = \sum_{1 \leq i \leq n} q_i\, c_i^k\, z_i$$

HITS: Hubs and Authorities

- **Teorema**. Comunque si scelga un vettore $\mathrm{h}^{(0)}$ a coordinate positive, esistono un valore $c \in \mathbb{R}^+$ e un vettore $z \in \mathbb{R}^n$ non nullo tali che $\lim_{k \to \infty} \frac{h^{(k)}}{c^k} = z$
- Dimostrazione.
  - Abbiamo mostrato che $\mathbf{h^{(k)}} = (M\, M^T)^k\, h^{(0)} = \sum_{1 \leq i \leq n} q_i\, c_i^k\, z_i$ **(*)**
    - dove $z_1, z_2, \dots, z_n$ sono gli autovettori di $(M\, M^T)$ – base ortonormale per $\mathbb{R}^n$
    - e, per $i = 1, \dots, n$, c_i è l'autovalore corrispondente a z_i, con $c_1 \geq c_2 \geq \dots \geq c_n \geq 0$
  - Ora, dividiamo ambo i membri in (*) per c_1^k e otteniamo

$$\frac{h^{(k)}}{c_1^k} = \sum_{1 \leq i \leq n} q_i \left(\frac{c_i}{c_1}\right)^k z_i$$

  - Sia $\ell \leq n$ tale che $c_1 = c_2 = \dots = c_\ell > c_{\ell+1} \geq \dots \geq c_n$
  - allora, $\lim_{k \to \infty} \frac{h^{(k)}}{c_1^k} = \lim_{k \to \infty} \left[q_1\, z_1 + \dots + q_\ell\, z_\ell + q_{\ell+1} \left(\frac{c_{\ell+1}}{c_1}\right)^k z_{\ell+1} \dots + q_n \left(\frac{c_n}{c_1}\right)^k z_n \right]$

$$= q_1\, z_1 + q_2\, z_2 + \dots + q_\ell\, z_\ell$$

  - **perché** $c_i \in \mathbb{R}^+ \cup \{0\}$ **per ogni** $i = 1, \dots, n$ **e, per ogni** $i = \ell + 1, \dots, n$, $0 \leq \frac{c_i}{c_1} < 1$

HITS: Hubs and Authorities

- **Teorema**. Comunque si scelga un vettore $\mathrm{h}^{(0)}$ a coordinate positive, esistono un valore c ∈ $\mathbb{R}^+$ e un vettore z ∈ $\mathbb{R}^n$ non nullo tali che $\lim_{k \to \infty} \frac{h^{(k)}}{c^k} = z$
- Dimostrazione.
  - Abbiamo mostrato che $\mathbf{h}^{(k)} = (M\ M^T)^k\ h^{(0)} = \sum_{1 \leq i \leq n} q_i\, c_i^k\, z_i$ **(*)**
    - dove $z_1, z_2, \ldots, z_n$ sono gli autovettori di $(M\ M^T)$ – base ortonormale per $\mathbb{R}^n$
    - e, per i = 1, … , n, c_i è l'autovalore corrispondente a z_i, con $c_1 \geq c_2 \geq \ldots \geq c_n$
  - e che, se $\ell \leq n$ è tale che $c_1 = c_2 = \ldots = c_\ell > c_{\ell+1} \geq \ldots \geq c_n$,
  - allora, $\lim_{k \to \infty} \frac{h^{(k)}}{c_1{}^k} = q_1 z_1 + q_2 z_2 + \ldots + q_\ell z_\ell$
  - resta da mostrare che $q_1 z_1 + q_2 z_2 + \ldots + q_\ell z_\ell$ è un vettore non nullo
    - ossia ha almeno una componente non nulla
  - e lo dimostriamo solo nel caso particolare $\ell = 1$, ossia, $c_1 > c_2$

HITS: Hubs and Authorities

- **Teorema**. Comunque si scelga un vettore $\mathrm{h}^{(0)}$ a coordinate positive, esistono un valore c ∈ $\mathbb{R}^+$ e un vettore z ∈ $\mathbb{R}^n$ non nullo tali che $\lim_{k \to \infty} \frac{h^{(k)}}{c^k} = z$
- Dimostrazione.
  - se $c_1 > c_2$: allora, $\lim_{k \to \infty} \frac{h^{(k)}}{{c_1}^k} = \lim_{k \to \infty} \frac{(MM^T)^k h^{(0)}}{{c_1}^k} = q_1 z_1$
  - per completare la dimostrazione del teorema è sufficiente mostrare che, *comunque si scelga un vettore $h^{(0)}$ a coordinate positive*, $q_1 \neq 0$
    - infatti, z_1 è un vettore reale perché elemento di una base per $\mathbb{R}^n$
    - z_1 è un vettore non nullo perché la base è ortonormale: $z_1 \cdot z_1 = z_{11}^2 + z_{12}^2 + \ldots + z_{1n}^2 = 1$
  - innanzi tutto, osserviamo che $q_1 = \mathrm{h}^{(0)} \cdot z_1$
    - infatti, $\mathrm{h}^{(0)} \cdot z_1 = (\sum_{1 \leq i \leq n} q_i z_i) \cdot z_1 = \sum_{1 \leq i \leq n} q_i \, (z_i \cdot z_1) = q_1 \, (z_1 \cdot z_1) = q_1$
    - allora, $q_1 \neq 0$ se e solo se $\mathrm{h}^{(0)} \cdot z_1 \neq 0$
  - allora dobbiamo dimostrare, che, *comunque si scelga* $\mathrm{h}^{(0)}$ *a coordinate positive*, $\mathrm{h}^{(0)} \cdot z_1 \neq 0$

HITS: Hubs and Authorities

- **Teorema**. Comunque si scelga un vettore $\mathrm{h}^{(0)}$ a coordinate positive, esistono un valore c ∈ $\mathbb{R}^+$ e un vettore z ∈ $\mathbb{R}^n$ non nullo tali che $\lim_{k \to \infty} \frac{h^{(k)}}{c^k} = z$
- Dimostrazione. Se $c_1 > c_2$: allora, $\lim_{k \to \infty} \frac{h^{(k)}}{c_1{}^k} = q_1 z_1$ e dobbiamo dimostrare, che, <u>*comunque si scelga*</u> $\mathrm{h}^{(0)}$ *a coordinate positive*, $\mathrm{h}^{(0)} \cdot z_1 \neq 0$
  - a) iniziamo dimostrando che <u>*esiste*</u> un vettore y a coordinate positive tale che $y \cdot z_1 \neq 0$
    - ricordiamo che $z_1 = (z_{11}, z_{12}, \dots, z_{1n})$ nel sistema di riferimento iniziale
    - supponiamo per assurdo che per ogni x ∈ $\mathbb{R}^n$ a coordinate positive sia $x \cdot z_1 = 0$:
    - sia y ∈ $\mathbb{R}^n$ tale che $y = (y_1, y_2, \dots, y_n)$ con $y_i > 0$ per ogni i = 1, … , n
    - e sia, per ogni ℓ = 1, … ,n, $y^{(\ell)}$ il vettore tale che $y^{(\ell)}{}_i = y_i$ per ogni $i \neq \ell$, e $y^{(\ell)}{}_\ell = y_\ell + 1$
    - allora, per ogni ℓ = 1, … ,n, $\begin{cases} y \cdot z_1 = 0 \\ y^{(\ell)} \cdot z_1 = 0 \end{cases}$, ossia,

      esempio:
      $y^{(2)} = (y_1, y_2+1, \dots, y_n)$

$$\begin{cases} y_1 z_{11} + \dots + y_\ell z_{1\ell} + \dots + y_n z_{1n} = 0 \\ y_1 z_{11} + \dots + (y_\ell + 1) z_{1\ell} + \dots + y_n z_{1n} = 0 \end{cases}$$

      - da cui, sottraendo la prima equazione dalla seconda, $z_{1\ell} = 0$
    - ossia, per ogni ℓ = 1, … ,n, $z_{1\ell} = 0$ – un assurdo perché $z_1 \cdot z_1 = 1$

HITS: Hubs and Authorities

- **Teorema**. Comunque si scelga un vettore $\mathrm{h}^{(0)}$ a coordinate positive, esistono un valore $c \in \mathbb{R}^+$ e un vettore $z \in \mathbb{R}^n$ non nullo tali che $\lim_{k \to \infty} \frac{\mathrm{h}^{(k)}}{c^k} = z$
- Dimostrazione. Se $c_1 > c_2$: allora, $\lim_{k \to \infty} \frac{\mathrm{h}^{(k)}}{c_1{}^k} = \lim_{k \to \infty} \frac{(MM^T)^k \mathrm{h}^{(0)}}{c_1{}^k} = q_1 z_1$ e dobbiamo dimostrare, che, *comunque si scelga* $\mathrm{h}^{(0)}$ *a coordinate positive*, $\mathrm{h}^{(0)} \cdot z_1 \neq 0$
  - a) esiste un vettore y a coordinate positive tale che $y \cdot z_1 \neq 0$
  - b) mostriamo che per ogni vettore x a coordinate positive vale che $x \cdot z_1 \neq 0$
    - sia $y \in \mathbb{R}^n$ a coordinate positive un vettore tale che $y \cdot z_1 \neq 0$ (y esiste per a))
    - esprimiamo y nel sistema $z_1, \ldots, z_n$: $y = p_1 z_1 + p_2 z_2 + \ldots + p_n z_n$
    - allora, come abbiamo visto, l'espressione $\frac{(M\,M^T)^k y}{c_1{}^k}$ converge a $p_1 z_1$
    - e poiché l'espressione $\frac{(M\,M^T)^k y}{c_1{}^k}$ contiene solo valori non negativi allora le coordinate di $p_1 z_1$ sono tutte non negative
    - e, poiché $p_1 = y \cdot z_1 \neq 0$, allora $p_1 z_1$ ha almeno una coordinata strettamente positiva

HITS: Hubs and Authorities

- **Teorema**. Comunque si scelga un vettore $\mathrm{h}^{(0)}$ a coordinate positive, esistono un valore $c \in \mathbb{R}^+$ e un vettore $z \in \mathbb{R}^n$ non nullo tali che $\lim_{k \to \infty} \frac{\mathrm{h}^{(k)}}{c^k} = z$
- Dimostrazione. Se $c_1 > c_2$: allora, $\lim_{k \to \infty} \frac{\mathrm{h}^{(k)}}{c_1{}^k} = q_1\, z_1$ e dobbiamo dimostrare, che, *comunque si scelga* $\mathrm{h}^{(0)}$ *a coordinate positive*, $\mathrm{h}^{(0)} \cdot z_1 \neq 0$
  - a) esiste un vettore y a coordinate positive tale che $y \cdot z_1 \neq 0$
  - b) mostriamo che per ogni vettore x a coordinate positive vale che $x \cdot z_1 \neq 0$
    - sia $y \in \mathbb{R}^n$ a coordinate positive tale che $y \cdot z_1 \neq 0$: con $y = p_1 z_1 + p_2 z_2 + \ldots + p_n z_n$,
      - $\mathbf{p_1 \neq 0}$ (perché $y \cdot z_1 \neq 0$)
      - e il vettore $\mathbf{p_1\, z_1}$ ha almeno una coordinata strettamente positiva e tutte le altre coordinate non negative
    - sia $x = (x_1, x_2, \ldots , x_n)$ un qualunque vettore in $\mathbb{R}^n$ <u>a coordinate positive</u> tale che $x \neq y$
    - allora nell'espressione $\mathbf{p_1\, z_1} \cdot x = p_1 z_{11} x_1 + p_1 z_{12} x_2 + \ldots + p_1 z_{1n} x_n$
      tutti gli addendi sono non negativi e almeno uno di essi è positivo
    - e, quindi, $x \cdot z_1 = z_1 \cdot x = \frac{1}{p_1} \left(p_1 z_{11} x_1 + p_1 z_{12} x_2 + \ldots + p_1 z_{1n} x_n\right) \neq 0$ QED

HITS: Hubs and Authorities

- **Teorema**. Comunque si scelga un vettore $\mathrm{h}^{(0)}$ a coordinate positive, esistono un valore c ∈ $\mathbb{R}^+$ e un vettore z ∈ $\mathbb{R}^n$ non nullo tali che $\lim_{k \to \infty} \frac{h^{(k)}}{c^k} = z$
- **Osservazione 1**: abbiamo dimostrato che z è non nullo nel solo caso $c_1 > c_2$: tecniche analoghe permettono di dimostrarlo nel caso generale
- **Osservazione 2**: come abbiamo visto, nel caso $c_1 > c_2$,
  - situazione che si presenta quasi sempre
- qualunque sia il vettore iniziale $\mathrm{h}^{(0)}$ a coordinate positive, $\frac{h^{(k)}}{{c_1}^k}$ converge al vettore $q_1 z_1$, dove $q_1 = \mathrm{h}^{(0)} \cdot z_1$
- questo significa che a partire da qualunque valutazione dei valori di hub iniziali, se $c_1 > c_2$, il metodo HITS converge sempre ad un vettore parallelo all'autovettore z_1
- ossia, la valutazione dei valori di hub delle pagine individuata da HITS è sempre la stessa da qualunque valutazione dei valori di hub iniziali si parta
  - ed è individuata dai valori relativi delle coordinate di z_1
- e, quindi dipende dalla sola matrice M

HITS: considerazioni

- HITS valuta ciascuna pagina rispetto a due ruoli diversi:
  - come authority – la sua rilevanza ai fini della ricerca in atto
  - come hub – la sua attitudine ad assegnare rilevanza alle pagine alle quali punta
- e, per ciascuna pagina, ciascuno dei due ruoli è valutato utilizzando un diverso insieme di link che coinvolgono quella pagina:
  - i link entranti nella pagina concorrono alla valutazione come authority
  - i link uscenti dalla pagina concorrono alla valutazione come hub
- Questo significa che una pagina può conferire rilevanza, ai fini di una ricerca, ad un'altra pagina pur essendo essa poco rilevante per quella ricerca
- e, dunque, il metodo HITS ben si presta a modellare situazioni nelle quali le pagine sono naturalmente partizionate in due sottoinsiemi "semanticamente" distinti
- Come avviene, ad esempio, in ricerche di prodotti da acquistare
  - laddove i venditori o i marchi non si puntano l'un l'altro (sono concorrenti!)
  - piuttosto, taluni siti (ad esempio, eBay) puntano ai venditori – o i venditori puntano ai marchi

PageRank

- In altre situazioni, invece, assumere un tale partizionamento delle pagine non è ragionevole
  - ad esempio, quando cerchiamo un articolo di ricerca: articoli "importanti" citano articoli "importanti" e sono citati da articoli "importanti"
  - sono le pagine rilevanti a decretare direttamente la rilevanza delle pagine alle quali puntano
- Il PageRank ben modella queste situazioni
  - prende il nome da Larry Page, uno dei fondatori di Google (di cui è marchio)
- Si tratta ancora di un metodo iterativo, basato, però, sull'analisi dei soli link entranti in una pagina
  - ossia, i soli link entranti in una pagina concorrono a determinarne il rank
- Esso assume che nella (porzione di) rete (attinente alla ricerca in atto) sia presente una unità di fluido
  - inizialmente distribuita equamente fra tutti i nodi
  - ossia, se il numero di nodi è n, ciascuno di essi possiede inizialmente $\frac{1}{n}$ di fluido
- e che poi, iterativamente, ciascun nodo distribuisca equamente il fluido fra i suoi vicini – e, alla fine, una pagina sarà tanto più rilevante quanto maggiore è la quantità di fluido in suo possesso

PageRank

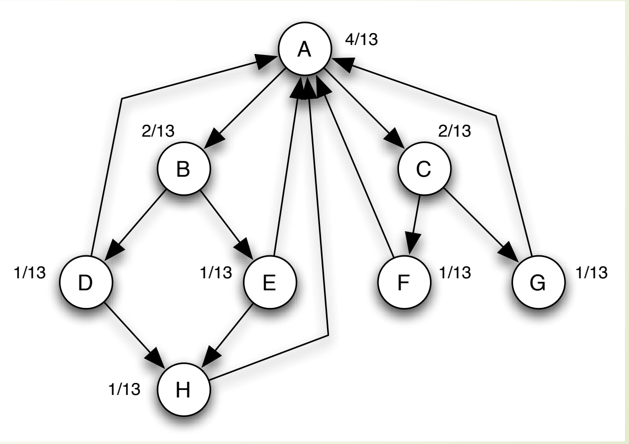
- Esso assume che nella rete sia presente una unità di fluido
- inizialmente, ciascuno degli n nodi possiede $\frac{1}{n}$ di fluido
  - ossia, per ogni i = 1, … , n, $\mathbf{f}_i^{(0)} = \frac{1}{n}$
- poi, ad ogni iterazione, ciascun nodo distribuisce equamente il fluido fra i suoi vicini
  - ossia, distribuisce il fluido in suo possesso e, contestualmente, riceve quello dei vicini
- Formalmente:
  - per ogni j = 1, … , n, indichiamo con ω_j il numero di archi uscenti dalla pagina j
  - ossia, $\boldsymbol{\omega}_j = |\{ i \in [n]: j \rightarrow i \}|$

    dove, al solito con "j → i" indichiamo che la pagina j contiene un link alla pagina i
  - allora, per ogni i = 1, … , n, $\mathbf{f}_i^{(k+1)} = \sum_{1 \leq j \leq n:\ j \rightarrow i} \frac{\mathbf{f}_j^{(k)}}{\omega_j}$
- Analogamente a quanto abbiamo fatto con HITS, indichiamo con $f^{(k)}$ il vettore $(f_1^{(k)}, f_2^{(k)}, \dots, f_n^{(k)})$

PageRank

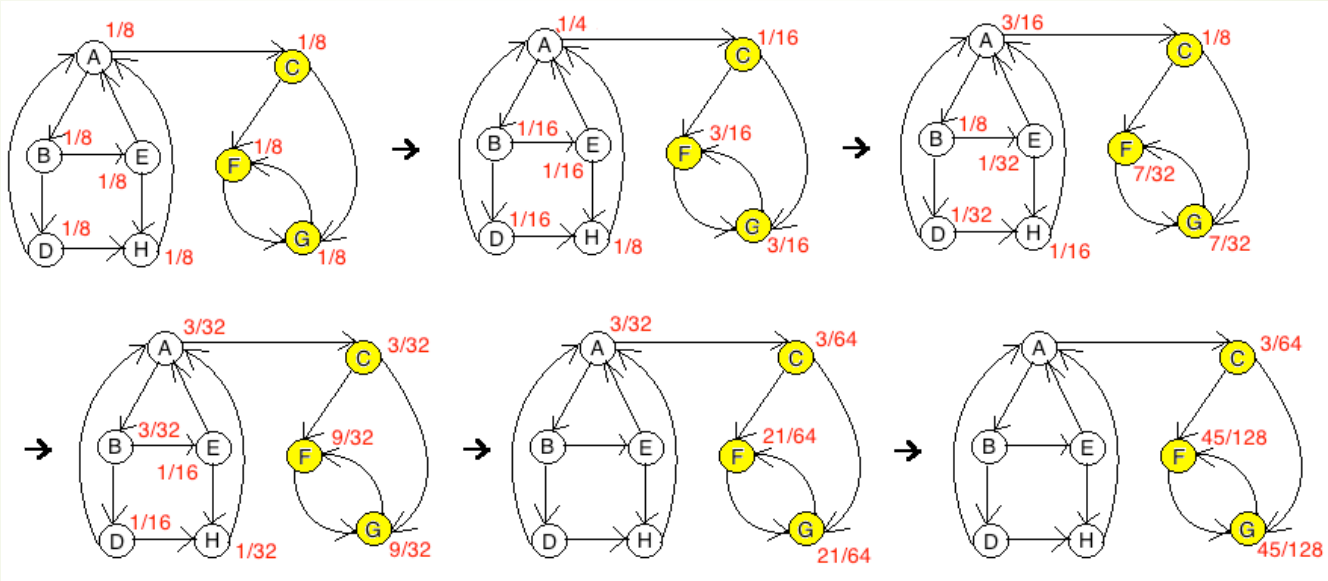
- **Teorema**. Se il grafo delle pagine attinenti alla ricerca è fortemente connesso allora esiste ed è unico il $\lim_{k \to \infty} f^{(k)}$
- Osserviamo che, ad ogni iterazione k, la quantità di fluido totale presente nel grafo è sempre pari ad 1
  - al passo k+1, ciascun nodo redistribuisce il fluido in suo possesso al passo k
  - senza generarne di nuovo
- Allora, possiamo pensare al vettore limite degli $f^{(k)}$, chiamiamolo f*, come ad un vettore che esprime una sorta di **configurazione di equilibrio** del fluido
  - ossia, redistribuendo il fluido a partire da f* il vettore non varia:

    per ogni i = 1, ... , n, $\mathbf{f_i^* = \sum_{1 \le j \le n:\ j \to i} \frac{f_j^*}{\omega_j}}$
- In figura è mostrato il vettore limite di un grafo:
  - a partire da $f_i^{(0)}$ = 1/8 per i = A,B,C,D,E,F,G,H
  - si raggiungono i valori in figura (provare!)
  - che sono valori di equilibrio (verificare!)



PageRank

- **Teorema**. Se il grafo delle pagine attinenti alla ricerca è fortemente connesso allora esiste ed è unico il $\lim_{k\to\infty} f^{(k)}$
- Ma, se il grafo non è fortemente connesso, le cose cambiano:



- il flusso tende ad accumularsi nei nodi gialli:
  - inizialmente, i tre nodi (C, F, G) possiedono 3/8 < 1/2 del flusso totale
  - nell'ultima iterazione mostrata ne possiedono 96/128 = 3/4

PageRank scalato

- Se il grafo non è fortemente connesso, il fluido tende a concentrarsi nelle regioni del grafo dalle quali non si può uscire
  - e, da ciò che conosciamo della struttura del Web, sappiamo che di tali regioni esso ne contiene
- Allora, è necessario modificare il procedimento iterativo che aggiorna il fluido associato ai nodi:
  - ad esempio, non distribuendo tutto il *fluido* (d'ora in avanti, il ***rank***) contenuto in ciascun nodo equamente lungo gli archi uscenti da quel nodo
  - ma "mettendone un po' da parte" per evitare che esso si concentri nei "vicoli ciechi" del grafo
- Nello Scaled PageRank viene fissato un parametro $s \in [0,1]$ e, ad ogni iterazione,
  - una frazione pari a s del rank contenuto in ciascun nodo è distribuito equamente lungo gli archi uscenti da quel nodo
  - e la parte rimanente, ossia, una frazione pari a $(1 – s)$ del rank contenuto in ciascun nodo, è distribuita uniformemente fra tutti i nodi del grafo
  - e, poiché il rank totale presente nella rete è 1, allora, ad ogni iterazione, ciascun nodo riceve una quantità di rank almeno pari a $\frac{(1-s)}{n}$

PageRank scalato

- Nello *Scaled* PageRank viene fissato un parametro s ∈ [0,1] e
  - detto $r_i^{(k)}$ il rank posseduto dal nodo i all'iterazione k
- $\mathbf{r_i^{(k+1)} = \left(\sum_{1 \le j \le n:\ j \to i} s \frac{r_j^{(k)}}{\omega_j} \right) + \frac{1-s}{n}}$
- La matrice N che descrive l'evoluzione del rank è tale che, per 1 ≤ i,j ≤ n,

$$\mathbf{N[i,j]} = \begin{cases} \frac{s}{\omega_j} + \frac{1-s}{n} & \textbf{se } j \to i \\ \frac{1-s}{n} & \textbf{altrimenti} \end{cases}$$

  - infatti, detto $r^{(k)}$ il vettore $(r_1^{(k)}, r_2^{(k)}, \ldots, r_n^{(k)})$, l'elemento i-esimo di $\mathbf{N}\, \boldsymbol{r}^{(k)}$ è

$$\sum_{1 \le j \le n} N[i,j]\, r_j^{(k)} = \sum_{1 \le j \le n:\, j \to i} \left[\frac{s}{\omega_j} + \frac{1-s}{n} \right] r_j^{(k)} + \sum_{1 \le j \le n:\, j \text{ non punta a } i} \frac{1-s}{n} r_j^{(k)}$$

$$= \sum_{1 \le j \le n:\, j \to i} \frac{s}{\omega_j} r_j^{(k)} + \sum_{1 \le j \le n} \frac{1-s}{n} r_j^{(k)}$$

$$= \sum_{1 \le j \le n:\, j \to i} \frac{s}{\omega_j} r_j^{(k)} + \frac{1-s}{n} \sum_{1 \le j \le n} r_j^{(k)} = \sum_{1 \le j \le n:\, j \to i} \frac{s}{\omega_j} r_j^{(k)} + \frac{1-s}{n}$$

- e, dunque, $\mathbf{r^{(k+1)} = N\, r^{(k)}}$

PageRank scalato

- Analogamente al caso del PageRank non scalato, anche in quello scalato, ad ogni iterazione k, la quantità di rank totale presente nel grafo è sempre pari ad 1
  - al passo k, ciascun nodo redistribuisce il fluido in suo possesso al passo k
  - senza generarne di nuovo
  - e questa caratteristica è stata usata nell'ultima uguaglianza nella slide precedente
- Allora, possiamo pensare al vettore limite degli $\mathbf{r}^{(k)}$, chiamiamolo r*, come ad un vettore che esprime una sorta di **configurazione di equilibrio** del rank
- ossia, redistribuendo il rank a partire da r* il vettore non varia: $\mathbf{r}^* = \mathbf{N}\ \mathbf{r}^*$
- Ossia, **il vettore limite dovrebbe essere un autovettore della matrice N**
  - il cui corrispondente autovalore è 1
  - e un autovettore ad elementi non negativi
  - e un autovettore la somma dei cui elementi è pari a 1
  - e, magari, se fosse anche l'unico autovettore con queste proprietà...
- Ma chi ci dice che N ha una coppia autovettore-autovalore con tutte queste proprietà1?!

PageRank scalato

- **Teorema di Perron** (versione semplificata per i nostri scopi) .
  Se A è una matrice reale $n \times n$ a elementi positivi, allora
  - A ha un autovalore $c \in \mathbb{R}^+$ tale che, $c > |c'|$ per ogni altro autovalore c' di A
  - l'autovettore x di A corrispondente a c è unico ed è a elementi reali e positivi la cui somma è 1.

- La matrice N soddisfa le ipotesi del teorema di Perron
  - e, dunque, ha una coppia autovettore-autovalore con *quasi* tutte le proprietà che ci occorrono

- In effetti, N ha tutte le proprietà che ci occorrono tranne una:
  - se potessimo esser certi che c = 1 potremmo concludere che $\mathbf{r}^* = \mathbf{N}\, \mathbf{r}^*$

PageRank scalato

- Una matrice quadrata si dice stocastica se i suoi elementi sono non negativi e
  - la somma degli elementi su ciascuna riga è 1 (stocastica per righe)
  - oppure la somma degli elementi su ciascuna colonna è 1 (stocastica per colonne)
- **Teorema**. Se A è una matrice stocastica allora A ha un autovalore λ tale che $|\lambda|=1$ e λ è l'autovalore di modulo massimo di A

- La matrice N è stocastica per colonne: poiché $\sum_{1 \leq i \leq n:\, j \to i} \frac{1}{\omega_j} = 1$, allora

$$\sum_{1 \leq i \leq n} N[i,j] = \sum_{1 \leq i \leq n:\, j \to i} \left[\frac{s}{\omega_j} + \frac{1-s}{n}\right] + \sum_{1 \leq i \leq n:\, j \text{ non punta a } i} \frac{1-s}{n} = \sum_{1 \leq i \leq n:\, j \to i} \frac{s}{\omega_j} + \sum_{1 \leq i \leq n} \frac{1-s}{n} = 1$$

- allora:
  - per il **teorema di Perron** N ha un autovalore $c \in \mathbb{R}^+$ tale che $c > |c'|$ per ogni altro autovalore c' di N e un unico autovettore x corrispondente a c a elementi reali e positivi la cui somma è 1
  - e per il **teorema sulle matrici stocastiche** $c = 1$
- perciò, **N x = x** e **r* = x**

PageRank e random walks

- Abbiamo introdotto il PageRank descrivendolo come una sorta di fluido che circola nella rete
  - che viene iterativamente redistribuito fra i nodi
  - senza che ne venga alterata la quantità
- Il PageRank ha anche un'altra interpretazione, legata al concetto di random walk in un grafo
  - inizialmente scegliamo uniformemente a caso un nodo u del grafo
  - e, da u, scegliamo (uniformemente) a caso un arco uscente da quel nodo...
  - ... che ci condurrà ad un altro nodo dal quale sceglieremo (uniformemente) a caso un arco uscente...
  - ... e, continuando in tal modo, percorreremo un *percorso aleatorio* nel grafo
- per individuare la relazione fra PageRank e random walk, calcoliamo la probabilità di trovarsi in un qualsiasi nodo del grafo dopo un random walk di k passi

PageRank e random walks

- Per eseguire un random walk in un grafo diretto G = (V,A)
  - inizialmente scegliamo uniformemente a caso un nodo u del grafo
  - e, da u, scegliamo (uniformemente) a caso un arco uscente da quel nodo...
  - ... che ci condurrà ad un altro nodo del quale sceglieremo (uniformemente) a caso un arco uscente... e così via
- per ogni i ∈ V, indichiamo con w_i^k la variabile aleatoria il cui valore è
  - 1 se al passo k del random walk ci troviamo nel nodo i
  - 0 altrimenti
- allora, per ogni i ∈ V, $P(w_i^0 = 1) = \frac{1}{n} = f_i^{(0)}$
- e, per ogni i ∈ V, detto ω_j il numero di archi uscenti dal nodo j,

$$P(w_i^1 = 1) = \sum_{j \in V:\ (j,i) \in A} \frac{1}{\omega_j} P(w_j^0 = 1) = \sum_{j \in V:\ (j,i) \in A} \frac{f_j^{(0)}}{\omega_j} = f_i^{(1)}$$

- e, induttivamente, per ogni k > 0 e per ogni i ∈ V,

$$P(w_i^k = 1) = \sum_{j \in V:\ (j,i) \in A} \frac{1}{\omega_j} P(w_j^{k-1} = 1) = \sum_{j \in V:\ (j,i) \in A} \frac{f_j^{(k-1)}}{\omega_j} = f_i^{(k)}$$

Scaled PageRank e random walks

- Cosideriamo ora un random walk scalato in un grafo diretto G = (V,A) eseguito in accordo alle seguenti regole: sia $s \in [0,1]$
  - inizialmente scegliamo uniformemente a caso un nodo u del grafo
  - e, da u,
    - con probabilità s scegliamo (uniformemente) a caso un arco uscente da quel nodo,
    - e con probabilità (1 – s) scegliamo uniformemente a caso un altro nodo del grafo ...
  - ... che ci condurrà ad un altro nodo nel quale ripeteremo il procedimento... e così via
- per ogni $i \in V$, indichiamo con sw_i^k la variabile aleatoria il cui valore è
  - 1 se al passo k del nuovo random walk scalato ci troviamo nel nodo i
  - 0 altrimenti
- allora, per ogni $i \in V$, $P(\mathrm{sw}_i^0 = 1) = \frac{1}{n} = r_i^{(0)}$
- e, per ogni $k > 0$ e per ogni $i \in V$, è semplice verificare che

$$P(\mathrm{sw}_i^k = 1) = \sum_{j \in V:\, (j,i) \in A} \frac{s}{\omega_j}\, P(\mathrm{sw}_j^{k-1} = 1) + \frac{1-s}{n} = r_i^{(k)}$$

Modern Web search

- L'enorme crescita del Web
  - in dimensioni – numero di pagine delle quali è costituito
  - in contenuti – numero di argomenti trattati
- ha richiesto la revisione delle tradizionali tecniche di link analysis per la web search
- Da tecniche di link analysis pura si è passati
  - a tecniche che combinano la link analysis all'analisi del contenuto testuale
    - come, ad esempio, usando gli anchor text - le parole (spesso visualizzate in azzurro) che costituiscono un hyperlink descrivendone sinteticamente il contenuto
  - o che considerano la frequenza con la quale una pagina individuata da una search viene effettivamente aperta
- e, poiché esistono frequentemente interessi di vario tipo ad apparire nelle prime posizioni di una search
  - spesso interessi commerciali, ma anche ragioni di prestigio
- gli stili di scrittura delle pagine web tengono conto delle tecniche di ranking usate dai motori di ricerca
  - con l'obiettivo di ottenere rank elevato nelle ricerche

Modern Web search

- Per l'interesse ad ottenere rank elevato nelle search
- è nata una vera e propria industria relativa alla produzione di pagine web che tiene in debito conto le tecniche di ranking usate dai motori di ricerca
  - così alterando di fatto quello che sarebbe il rank "naturale" di una pagina
  - e inducendo i progettisti dei motori di ricerca a rivedere gli algoritmi utilizzati per aggirare le tecniche utilizzate dai web-designer
  - che rivedono le loro tecniche
  - inducendo a rivedere gli algoritmi di web search…
  - …
- Ecco perché gli algoritmi di ranking utilizzati vengono tenuti segreti!