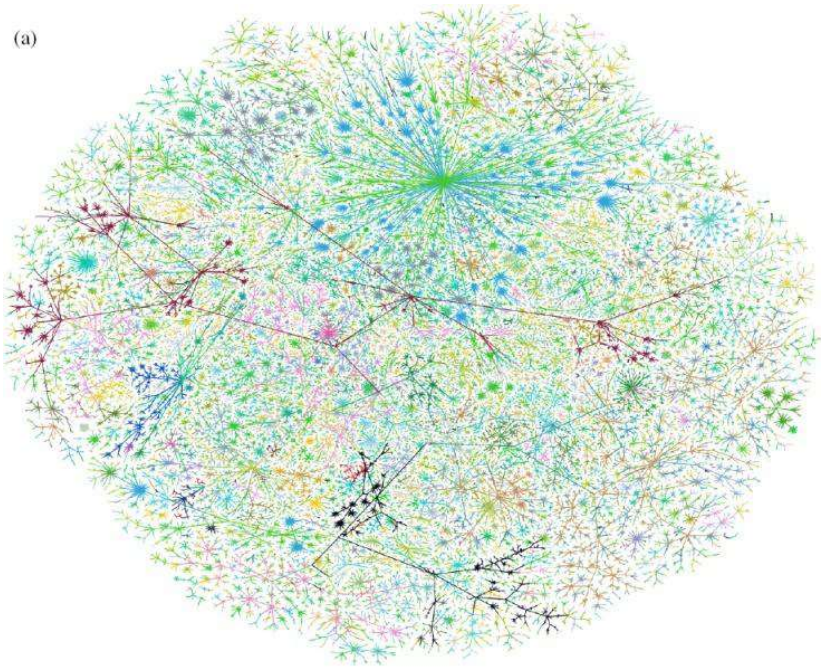
Reti Sociali



Link Analysis e Ricerca sul Web

Il problema della ricerca sul web

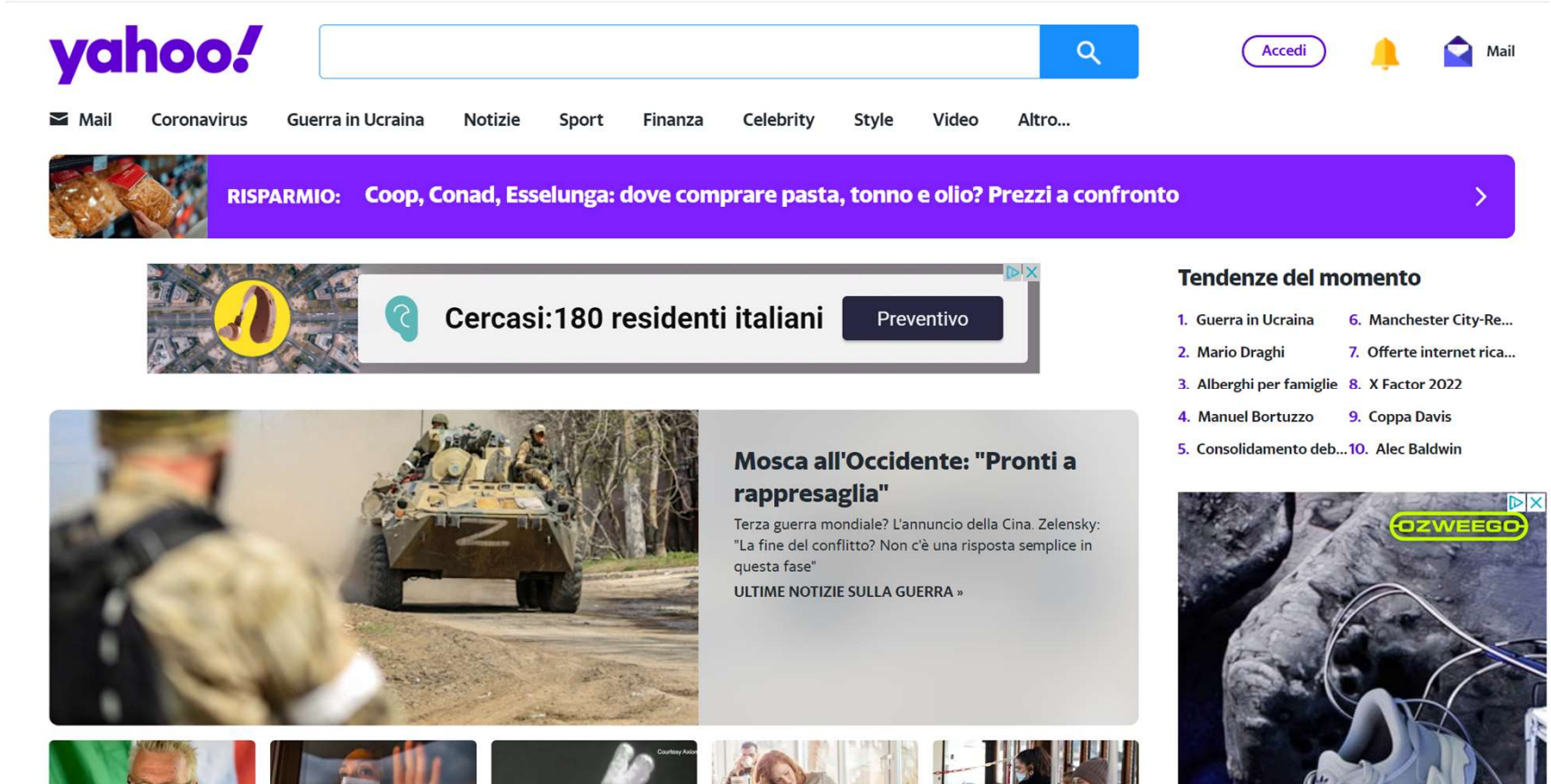
- Il Web è soprattutto una grande rete di informazioni
  - La sua utilità dipende fortemente da quanto efficacemente possiamo trovare le informazioni
  - Il motore di ricerca è un elemento essenziale del Web

- Sin dagli esordi del Web ci si è posti il problema di come organizzarlo in modo da rendere più efficace la ricerca

Come organizzare il WEB?

Primo tentativo: Web Directories

- A cura dell'uomo
- Pagine gestite da un comitato di redazione che forniscono link a contenuti rilevanti per ciascun argomento
- Yahoo!



Come organizzare il WEB?

Primo tentativo: Web Directories

- a cura dell'uomo
- DMOZ
  - Chiuso

As of Mar 17, 2017, dmoz.org is no longer available.

The editors have set up a static mirror here.

If you are interested in staying in touch with the DMOZ community,

please visit www.resource-zone.com.

Thank you all, especially the editors, for your interest and dedication to this project.

Welcome!

This site includes information formerly made available via DMOZ.

Visit resource-zone to stay in touch with the community.

#OrganizeTheWeb

Arts
Movies, Television, Music...

Business
Jobs, Real Estate, Investing...

Computers
Internet, Software, Hardware...

Games
Video Games, RPGs, Gambling...

Health
Fitness, Medicine, Alternative...

Home
Family, Consumers, Cooking...

News
Media, Newspapers, Weather...

Recreation
Travel, Food, Outdoors, Humor...

Reference
Maps, Education, Libraries...

Regional
US, Canada, UK, Europe...

Science
Biology, Psychology, Physics...

Shopping
Clothing, Food, Gifts...

Society
People, Religion, Issues...

Sports
Baseball, Soccer, Basketball...

Kids & Teens Directory
Arts, School Time, Teen Life...

World
Deutsch, Français, 日本語, Italiano, Español, Русский, ...

91,929 Editors | 1,031,722 Categories | 3,861,210 Sites | 90 Languages

Come organizzare il Web?

Secondo tentativo: Web search

- Information Retrieval
  - Tecniche sviluppate a partire dagli anni 60 per cercare in archivi di documenti strutturati utilizzando delle parole chiave
    - Articoli di giornale, articoli scientifici, brevetti, sentenze, leggi, ecc.
  - Fino agli anni 80 esistevano persone specializzate nel fare ricerche bibliografiche
    - Può risultare efficace se si devono trovare documenti in un insieme piccolo ed affidabile (dove chi scrive usa un lessico tecnico)
  - Le parole chiave sono poco espressive e ci sono i sinonimi e dei significati multipli assegnati ad una parola
- Ma: il Web è grande, pieno di documenti inaffidabili, cose scritte da chiunque, spam, etc.

Come organizzare il Web?

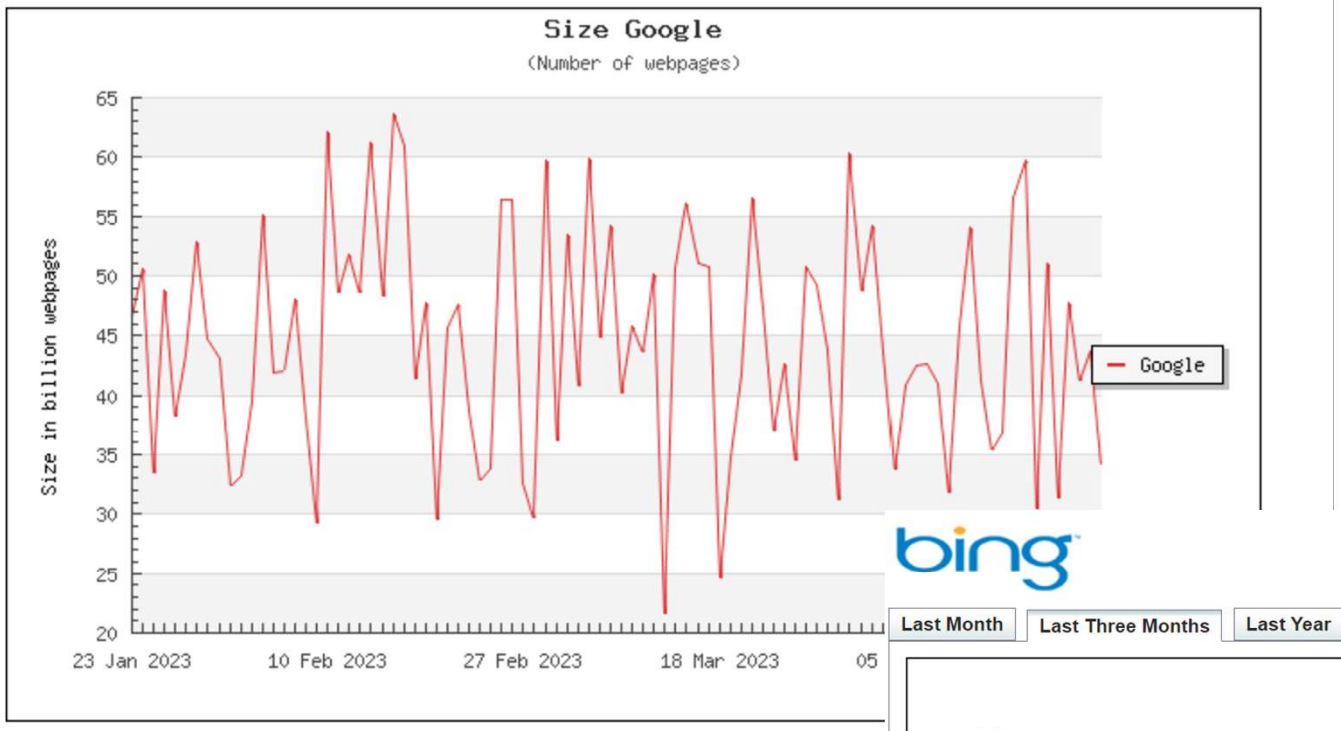
- Le tecniche di information retrieval si sono rivelate inefficaci per il Web
  - Non scalano alle dimensioni del Web
  - Le pagine Web sono poco strutturate e scritte utilizzando stili differenti
  - L'utenza è molto eterogenea e questo acuisce il problema dei significati multipli
  - Il Web è dinamico ed il suo contenuto e la sua struttura è in continua evoluzione

Taglia del Search Index

Google

The size of the World Wide Web:
Estimated size of Google's index

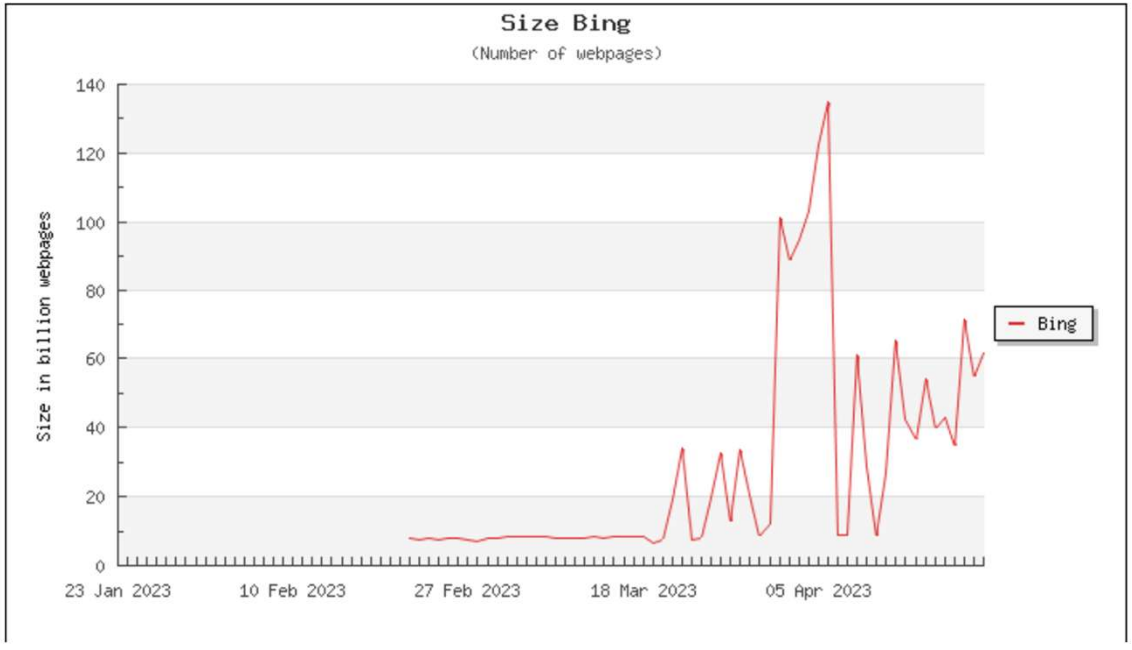
Last Month | Last Three Months | Last Year | Last Two Years | Last Five Years | Last Ten Years



bing

The size of the World Wide Web:
Estimated size of Bing index

Last Month | Last Three Months | Last Year | Last Two Years | Last Five Years | Last Ten Years



http://www.worldwidewebsize.com/

Come organizzare il Web?

Il problema principale nel Web

- Non è trovare pochi elementi in un (grande) insieme
- Ma è filtrare l'enorme mole di dati trovati per individuare quelli più rilevanti

- Abbiamo bisogno di una buona tecnica di classificazione (**ranking**) delle pagine Web

Web search: 2 sfide

1) Il web contiene molte sorgenti di informazioni.
   Di chi «fidarsi»?
   - Intuizione: pagine affidabili possono puntarsi l'un l'altra

2) Quale è la risposta «migliore» in una ricerca con la parola «libro»?
   - Può non esserci una sola risposta giusta
   - Intuizione: le pagine che contengono libro potrebbero puntare a molti libri

Come organizzare il web?

Terzo tentativo (l'era Google): usare il grafo del web

- Spostare l'attenzione dalla rilevanza alla *autorevolezza*
- Non è solo importante che la pagina sia rilevante (nel senso che è attinente a quello che cercavo), ma che sia anche importante sul web

Abbiamo bisogno di una buona tecnica di classificazione (**ranking**) delle pagine Web

Il ranking delle pagine Web

- I moderni motori di ricerca non hanno difficoltà a trovare nel Web ed indicizzare migliaia di pagine che soddisfano una query
- L'utente medio controlla soltanto le prime 5–10 pagine proposte dal motore di ricerca
  - Se non trova quello che cerca prova con una nuova query o abbandona e sceglie un altro motore di ricerca
- Per il motore di ricerca è fondamentale classificare le pagine rispetto alla loro "**rilevanza**" rispetto alla query
  - Sulla prima pagina di output vengono visualizzate le pagine più rilevanti
  - Massimizza la probabilità di soddisfare la richiesta dell'utente

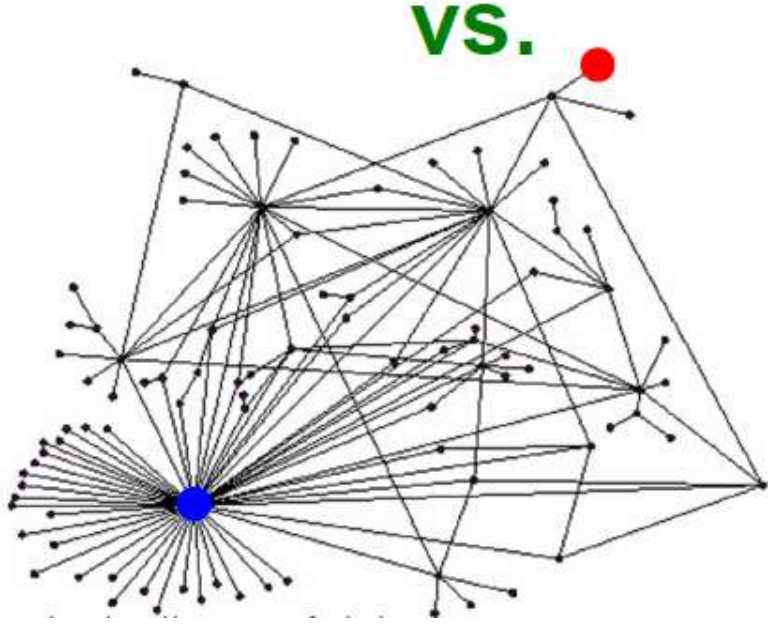
Ranking e Centralità

- L'idea è usare la struttura del grafo del web per classificare le pagine
  - Le misure di centralità definiscono quanto un nodo del grafo è importante rispetto ad una certa proprietà
- Per fare il ranking delle pagine Web possiamo utilizzare una misura di centralità che misura la rilevanza di un nodo

Link Analysis

- Non tutti i nodi hanno lo stesso valore
  - Possono essere pagine che vogliono criticare
  - Possono essere pubblicità a pagamento
  - Possono essere pagine fuori contesto
- I link agiscono come un endorsement (raccomandazione):
  - Quando una pagina *p* link a *q* essa fa un endorse (raccomanda) non solo q ma anche il contenuto di *q*

Quale è il modo più semplice per misurare l'importanza di una pagina web?

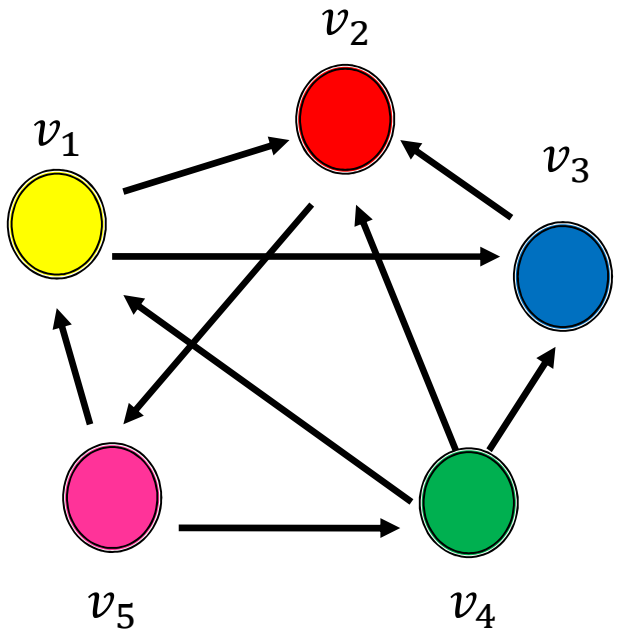


Link Analysis

- Possiamo considerare i link del grafo come voti sulla rilevanza di una pagina
  - Se io referenzio una pagina sto implicando che la reputo rilevante
- Dall'esame dei link del grafo possiamo recuperare informazioni sulla rilevanza delle pagine

Classifica per popolarità …

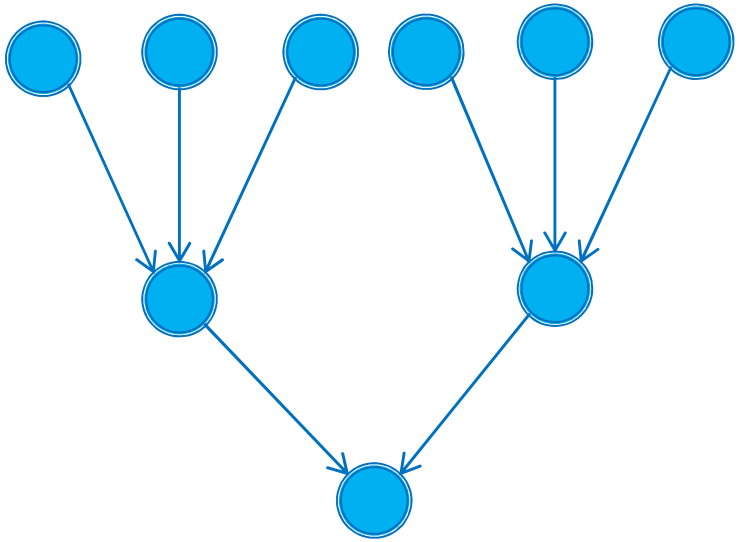
- Assumiamo che nel complesso una pagina che ha più link in ingresso è più rilevante di una pagina che ne ha meno



1. **Pagina rossa**
2. **Pagina gialla**
3. **Pagina blu**
4. **Pagina viola**
5. **Pagina verde**

… e per importanza

- Non è importante solo quanti link puntano ad una pagina, ma anche quanto importanti sono le pagine che vi puntano
  - Good authority sono puntate da good authority
    - Definizione ricorsiva di importanza



Algoritmi di Link Analysis

- Gli algoritmi di link analysis più utilizzati sono
  - HITS (Hubs and Authorities)
    - Hyperlink-Induced Topic Search
  - PageRank
  - versioni di Pagerank customizzate su un particolare argomento

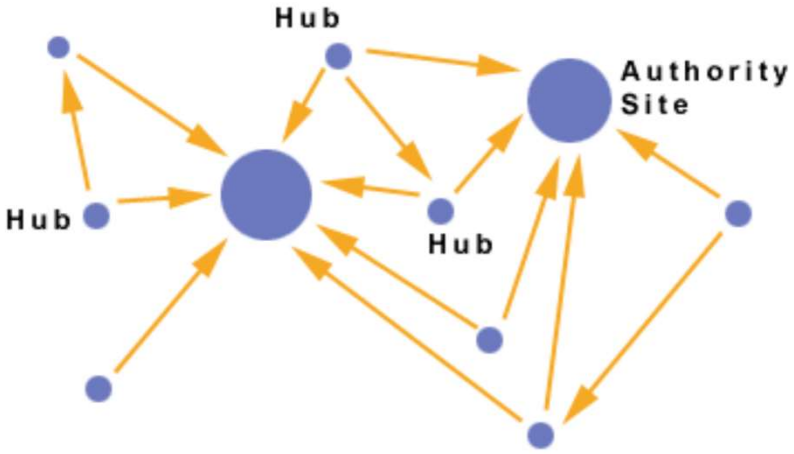
HITS (Hubs and Authorities)

L’algoritmo HITS

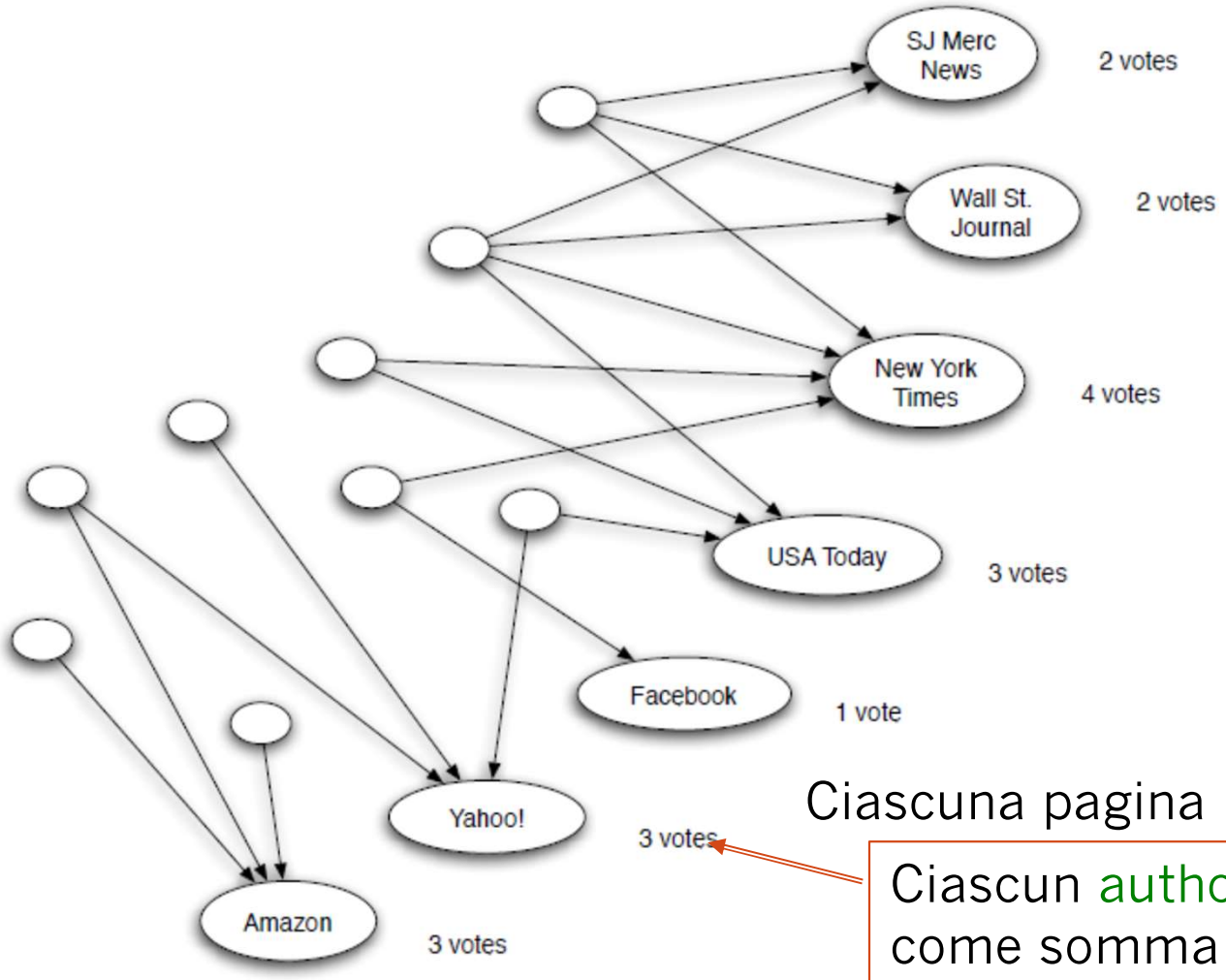
- Usa gli hyperlink presenti nelle pagine web per classificarle, cioè per assegnare loro un rank
  - dovuto a Kleinberg e risale al 1998
  - a quel tempo Kleinberg lavorava in IBM Almaden
  - IBM non ne face niente
- Si basa su una doppia identità da assegnare alle pagine
  - Hub
  - Authority

Hub e Authority

- Authority sono pagine contenenti informazioni utili
  - Home page di quotidiani
  - Home page di corsi di laurea
  - Home page di case automobilistiche
- Hub sono pagine che contengono link ad authority
  - Liste di quotidiani
  - Liste di corsi di laurea
  - Liste di case automobilistiche



Conto degli in-links: Authority

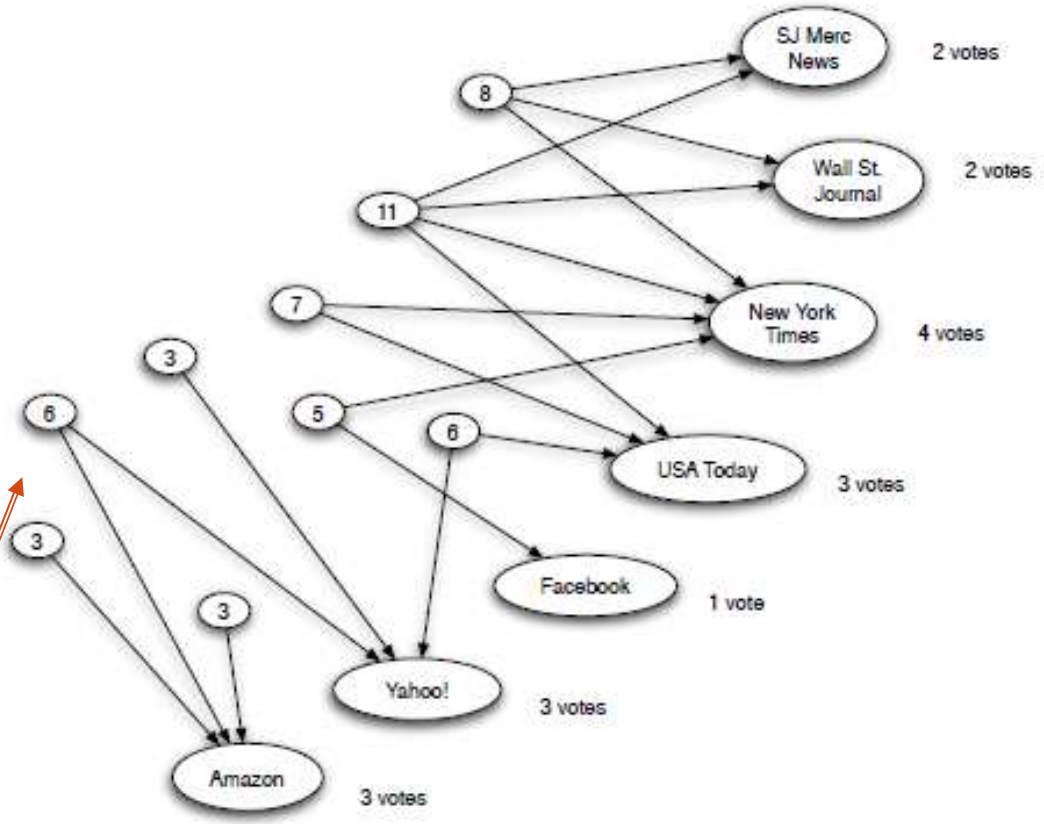


Ciascuna pagina parte con un hub score =1

Ciascun authority colleziona i suoi voti come somma degli hub score che fanno riferimenti ad essa

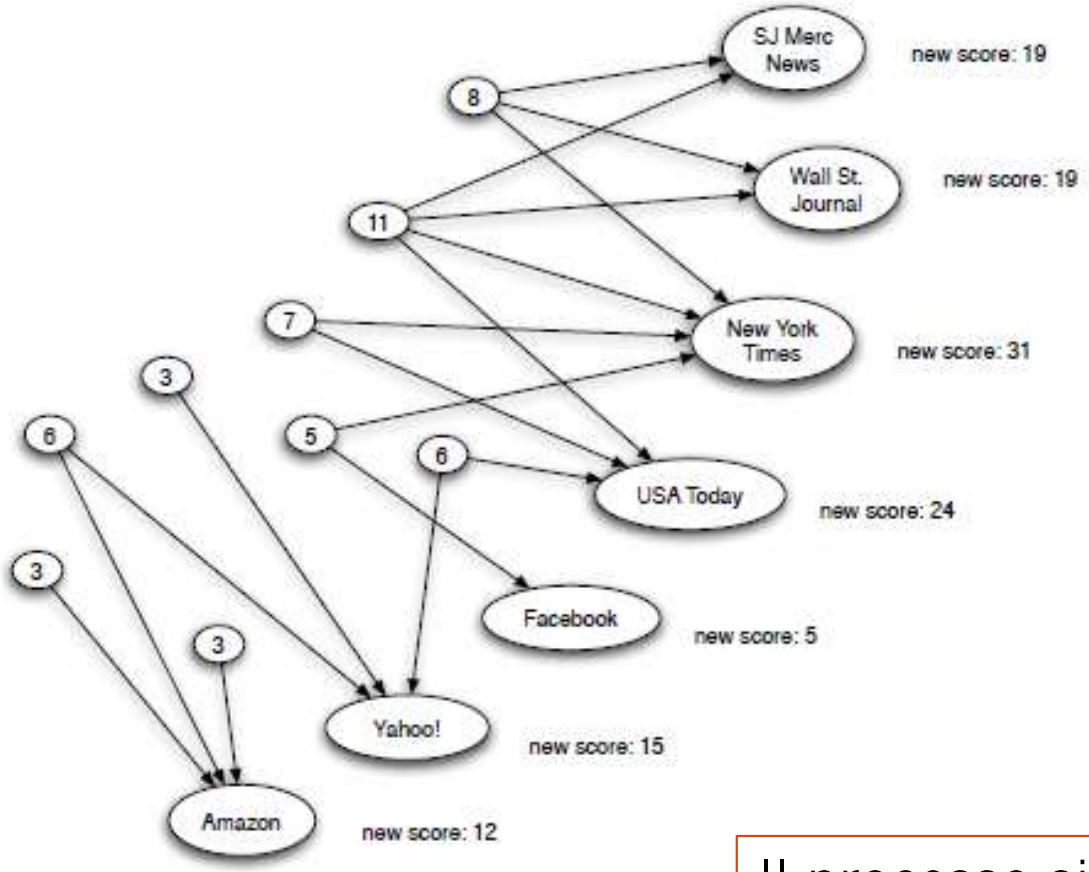
NB: questo è un esempio idealizzato; nella realtà il grafo non è bipartito e ciascuna pagina ha sia un hub score che un authority score

Expert quality: Hub



Ciascun hub colleziona i suoi voti sommando gli authority scores a cui si riferisce

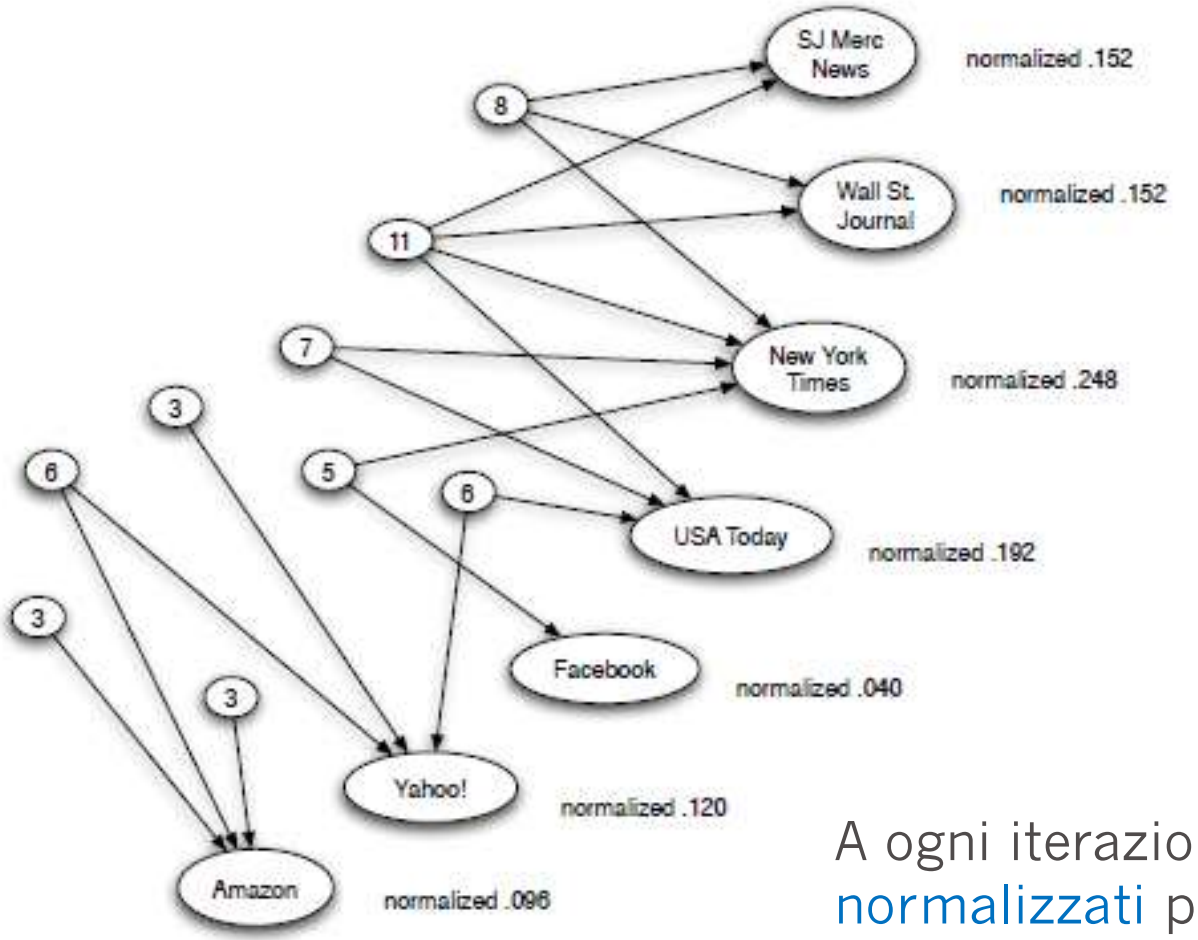
Re-weighting



Il processo si itera:

- le authorithies collezionano i nuovi hub score
- gli hub collezionano i nuovi authorities score

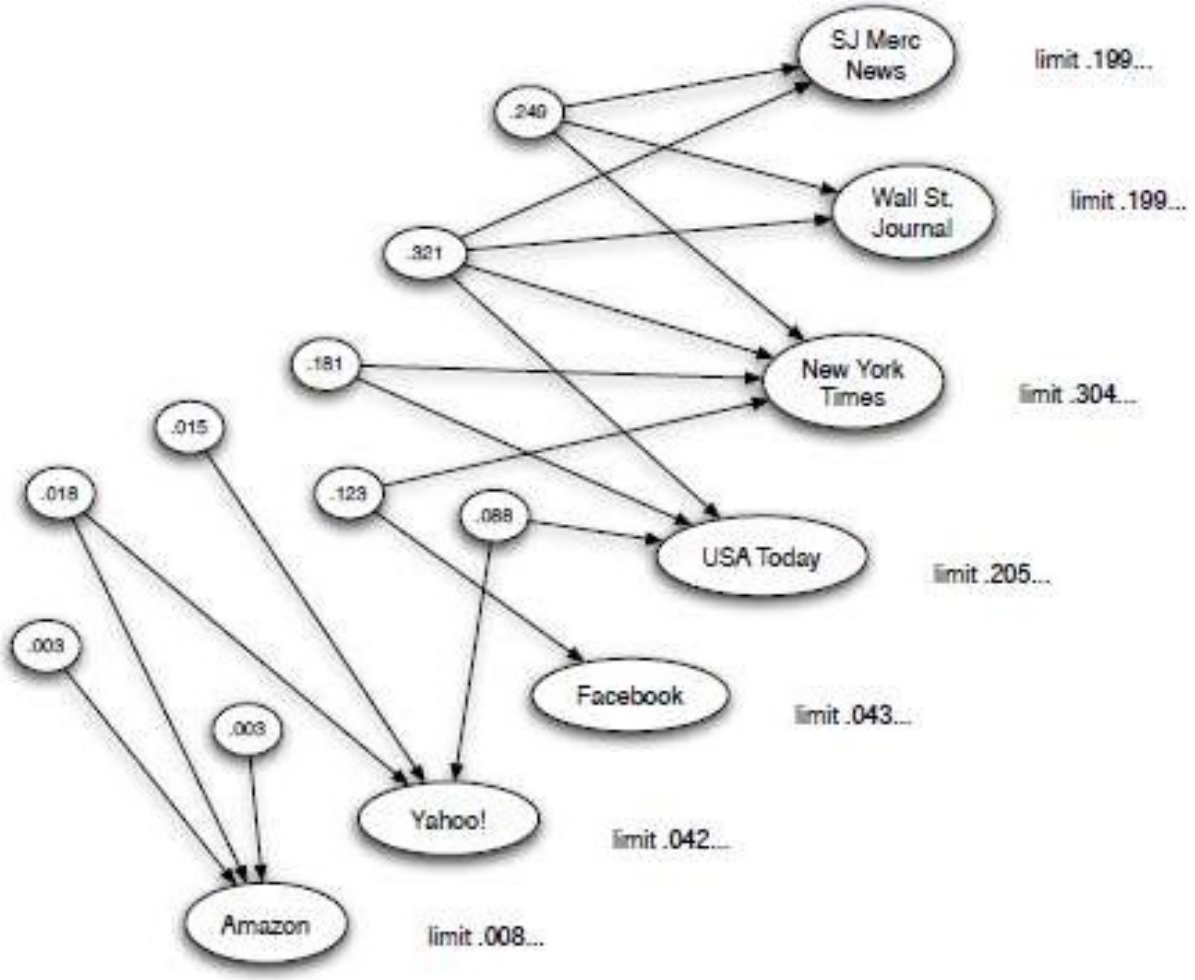
Hub e Authority



A ogni iterazione gli score vengono normalizzati per evitare di gestire numeri troppo alti,

- ciascun hub score viene diviso per la somma di tutti gli hub score
- ciascun authority score viene diviso per la somma di tutti gli authority score

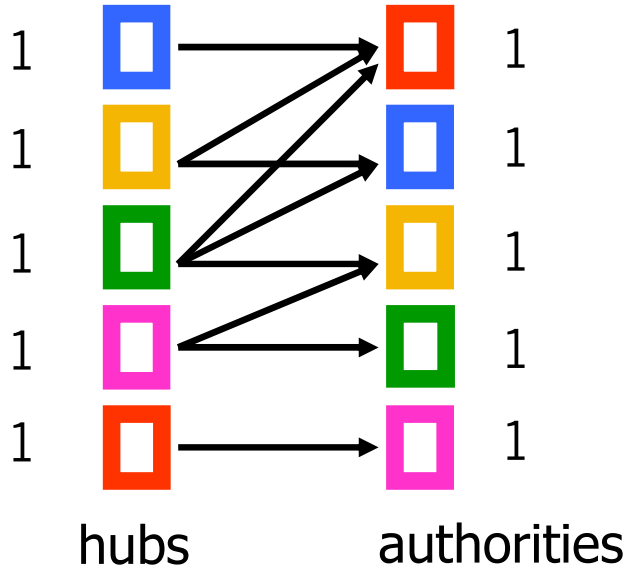
Hub e Authority



Il processo di conteggio va avanti fino a che converge (cioè, la differenza tra i vecchi ed i nuovi score diventa piccola)

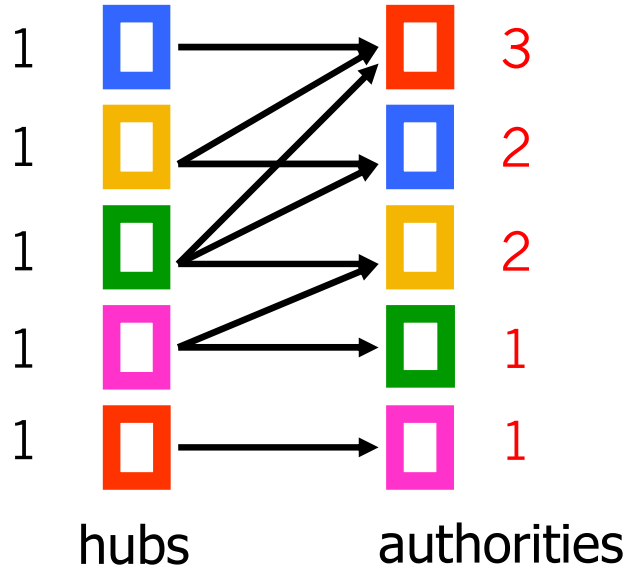
Esempio

Inizializzazione



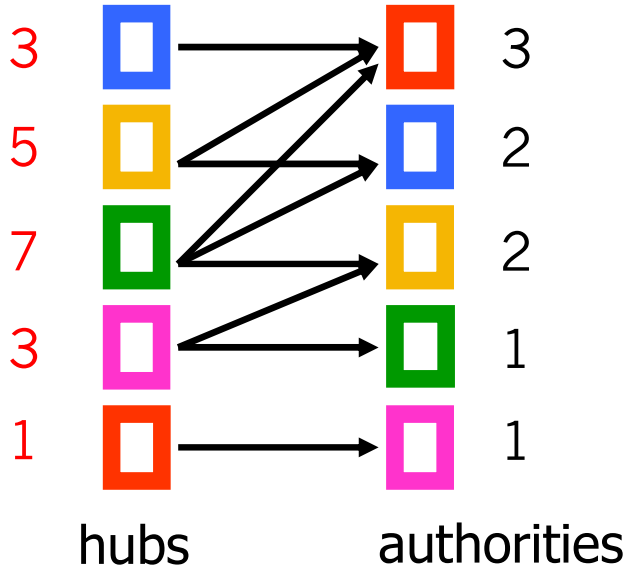
Esempio

Step 1: calcolo dei pesi authority



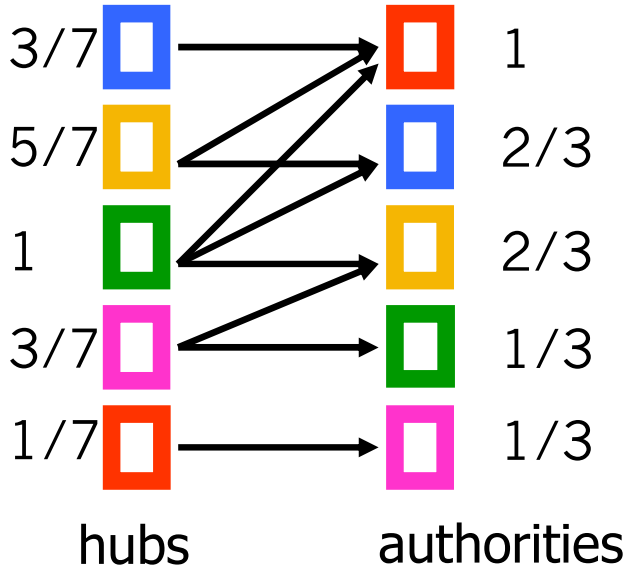
Esempio

Step 1: calcolo dei pesi hub



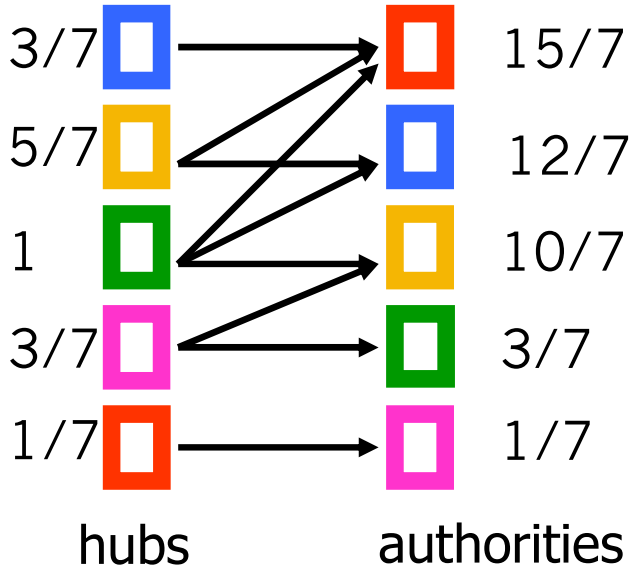
Esempio

Step 1: Normalizzazione (attraverso il max)



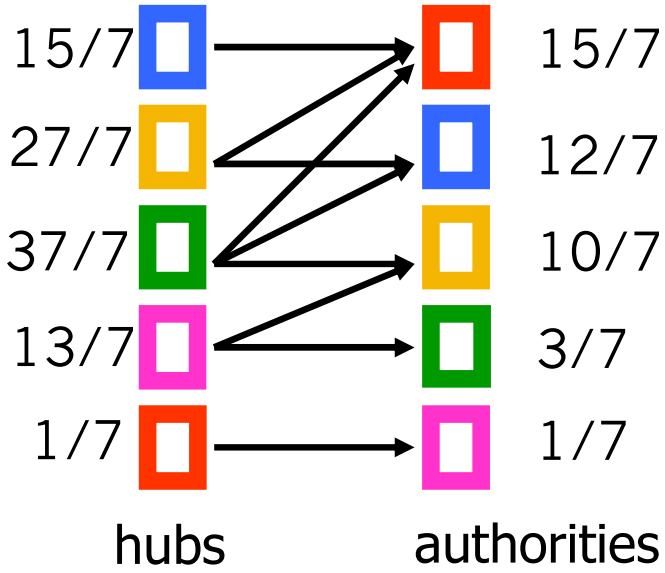
Esempio

Step 2: calcolo dei pesi authority



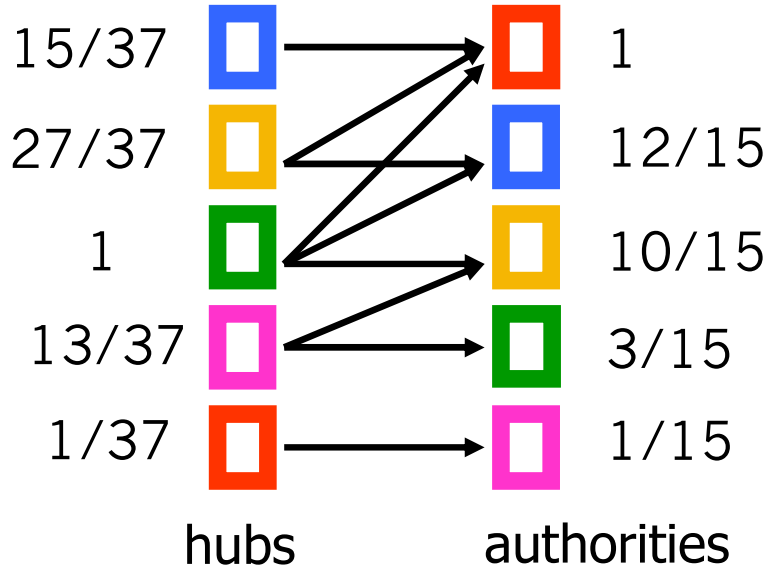
Esempio

Step 2: calcolo dei pesi hub



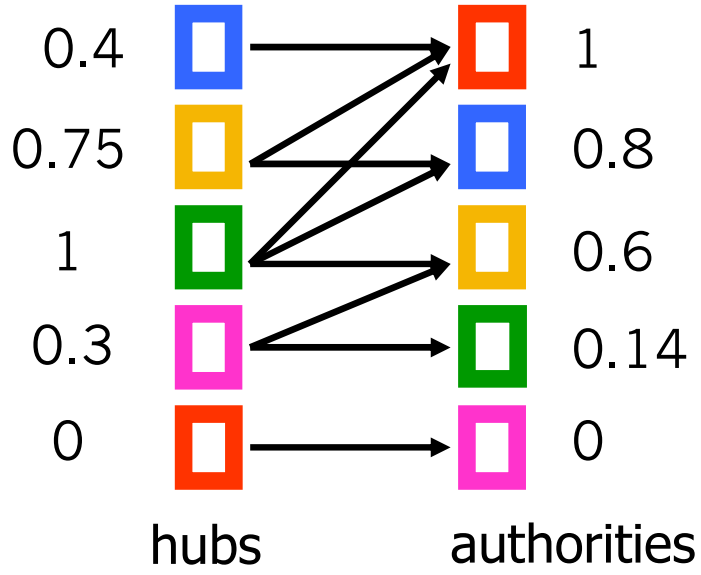
Esempio

Step 2: Normalizzazione



Esempio

Convergenza



Definizione mutuamente ricorsiva

- Un buon hub ha link a molte buone authority
- Una buona authority è linkata da molti buoni hub
- Formalmente:
  - ad ogni pagina i si assegnano due pesi: hub score h_i ed authority score a_i
  - Inizialmente tali valori vengono impostati ad 1
  - Ad ogni iterazione dell'algoritmo vengono effettuati due passi
    1. Viene valutato l' authority score che è definito come la somma dei hub score delle pagine che puntano a quella pagina authority

$$a_i = \sum_{j:j\to i} h_j$$

    2. Viene valutato l' hub score che è definito come la somma dei authority score delle pagine authority puntate da quell'hub

$$h_i = \sum_{k:i\to k} a_k$$

Definizione mutuamente ricorsiva

$$h_i = \sum_{k:i\to k} a_k \qquad\qquad a_i = \sum_{j:j\to i} h_j$$

- Se consideriamo la matrice delle adiacenze A del grafo
  - Riga $\boldsymbol{i}$ –sima di A contiene un 1 in corrispondenza di un arco uscente da $\boldsymbol{i}$
  - Colonna $\boldsymbol{i}$ –sima di A contiene un 1 in corrispondenza di un arco entrante in $\boldsymbol{i}$
    - Ma la colonna i-sima di A non è altro che la riga i-sima della trasposta di A^T

- Questo significa che nella prima iterazione
  - nel passo 1 a_i è il prodotto scalare della $\boldsymbol{i}$ –sima riga di A^T e del vettore di tutte gli hub score h_i settati inizialmente, cioè vettorialmente

$$\underline{a}^1 = A^T \underline{h}^0$$

  - nel passo 2 della prima iterazione h_i è il prodotto scalare della $\boldsymbol{i}$ –sima riga di A e del vettore di tutte gli hub score a_i settati al passo 1, cioè vettorialmente

$$\underline{h}^1 = A \quad \underline{a}^1$$

Definizione mutuamente ricorsiva

$$\underline{a}^1 = A^T \underline{h}^0 \qquad \underline{h}^1 = A \ \ \underline{a}^1$$

- Questo significa che nella t-sima iterazione
  - nel passo 1

$$\underline{a}^t = A^T \underline{h}^{t-1}$$

  - nel passo 2

$$\underline{h}^t = A \ \ \underline{a}^t$$

  - quindi abbiamo

$$\underline{h}^t = A \ \ \underline{a}^t = AA^T \underline{h}^{t-1}$$

$$\underline{a}^t = A^T \underline{h}^{t-1} = A^T A \underline{a}^{t-1}$$

HITS e gli autovettori

- L'algoritmo HITS è un metodo per il calcolo degli autovettori
  - In termini vettoriali alla *t*-sima iterazione dell'algoritmo abbiamo (A matrice di adiacenza del grafo)

$$\underline{a}^t = A^T A \;\; \underline{a}^{t-1} \qquad\qquad \underline{h}^t = A \;\; A^T \underline{h}^{t-1}$$

  - iterazioni ripetute convergeranno agli autovettori

Definition: Se vale $\boldsymbol{R} \cdot \boldsymbol{x} = \boldsymbol{\lambda} \cdot \boldsymbol{x}$
dove $\boldsymbol{\lambda}$ è uno scalare, $\boldsymbol{x}$ è un vettore, $\boldsymbol{R}$ è una matrice, allora $\boldsymbol{x}$ **è un autovettore**, e $\boldsymbol{\lambda}$ **è un autovalore**

- il vettore dei pesi authority $\underline{a}$ è l'autovettore di $A^T A$
- il vettore dei pesi hub $\underline{h}$ è l'autovettore di $A \;\; A^T$

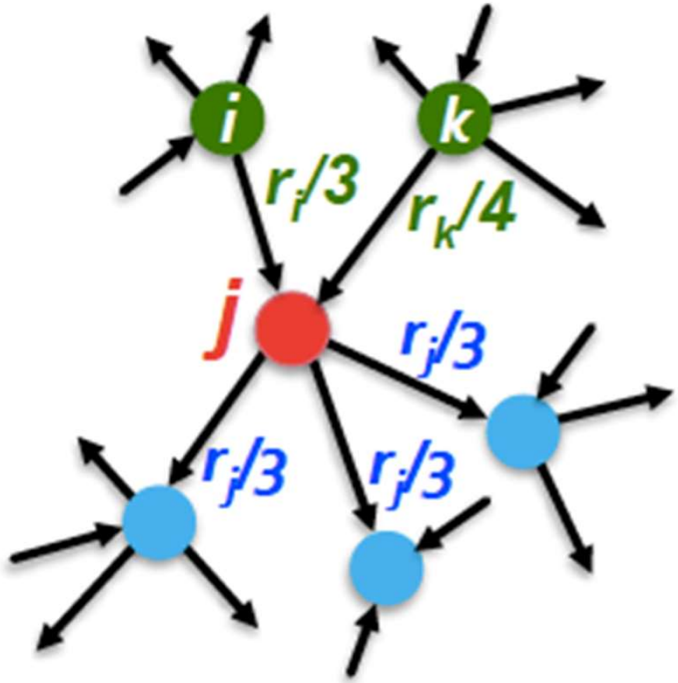
PageRank

Page rank

- Link usati come voti
  - Una pagina è tanto più importante quanti più link puntano ad essa
- I link entranti sono tutti uguali?
  - È naturale pensare che link entranti provenienti da pagine importanti contano di più
  - ... ma questo è un problema ricorsivo

Page rank: flow model

- Un "voto" da una pagina importante vale di più
  - Il voto derivante da ciascun link è proporzionale all'importanza della sua pagina sorgente
  - Se la pagina **i** ha importanza r_i e d_i link uscenti, allora ciascun link uscente da **i** da un voto r_i/d_i
  - L'importanza della pagina **j**, cioè r_j è la somma dei voti dei suoi link entranti



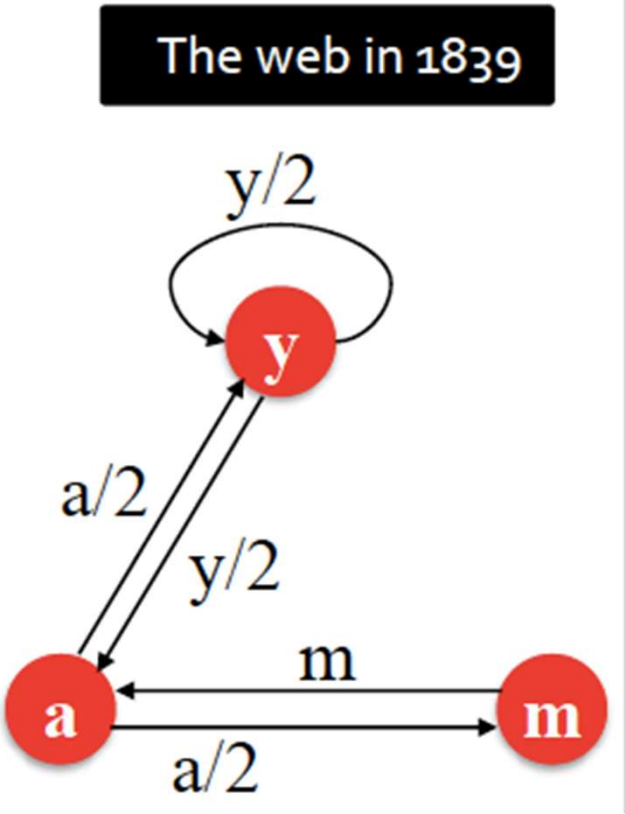
$$r_j = r_i/3 + r_k/4$$

Page rank: flow model

- Definiamo "rank" r_j del nodo j

$$r_j = \sum_{i \to j} \frac{r_i}{\mathrm{d_i}}$$

d_i = out-degree del nodo i



Equazioni del flusso:

$$r_y = r_y/2 + r_a/2$$
$$r_a = r_y/2 + r_m$$
$$r_m = r_a/2$$

Page rank: come calcolarlo?

Dato un grafo con **n** nodi, dove i nodi sono le pagine e gli archi sono gli hyperlink

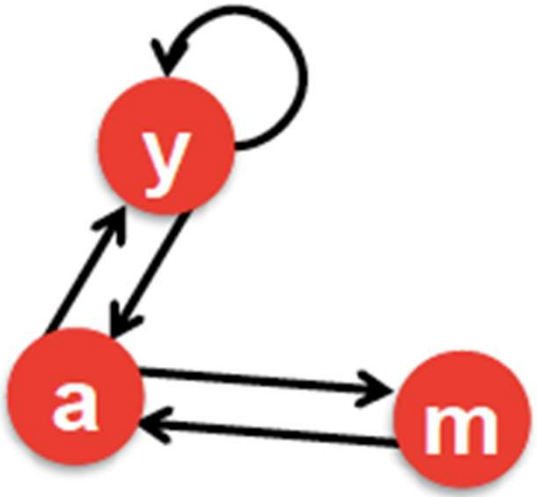
- Assegnamo a tutti i nodi lo stesso iniziale PageRank $= 1/n$
  - Set $r_j = 1/n$ per ciascun nodo j
- Scegliamo un numero di step **k**
- Aggiorniamo per ciascuno step $t \leq$ **k** i PageRank di ciascun nodo

$$r_j^{(t+1)} = \sum_{i \to j} \frac{r_i^{(t)}}{d_i}$$

definizione ricorsiva

Page rank: come calcolarlo?

Esempio:



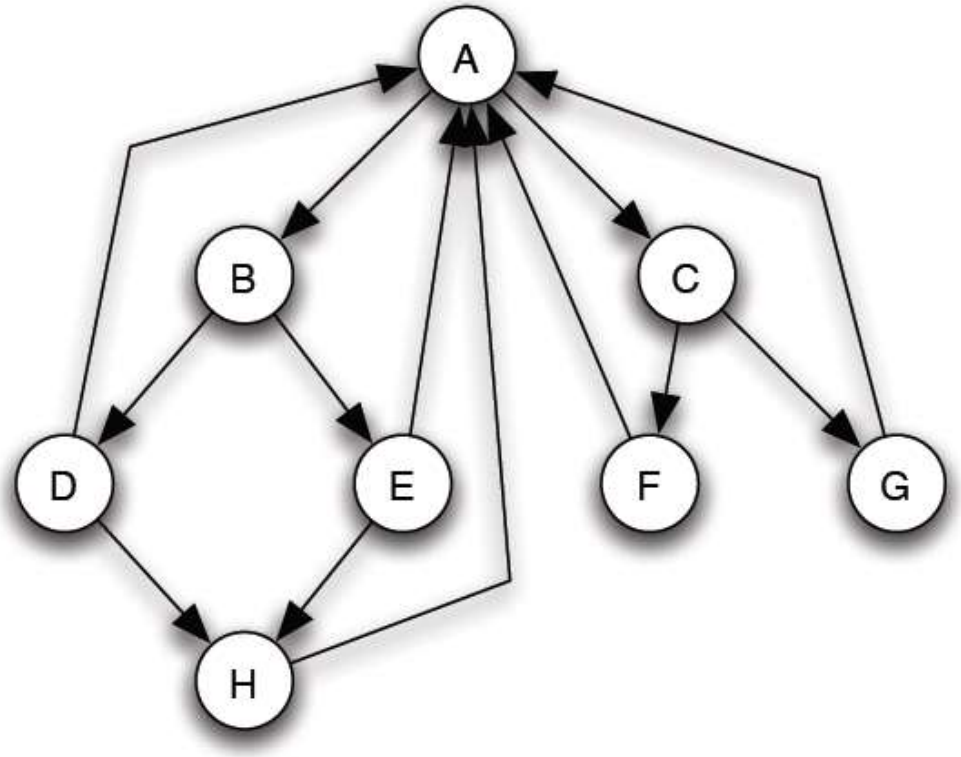
$$r_j^{(t+1)} = \sum_{i \to j} \frac{r_i^{(t)}}{d_i}$$

$$\begin{bmatrix} r_y \\ r_a \\ r_m \end{bmatrix} = \begin{matrix} 1/3 & 1/3 & 5/12 & 9/24 & & 6/15 \\ 1/3 & 3/6 & 1/3 & 11/24 & \ldots & 6/15 \\ 1/3 & 1/6 & 3/12 & 1/6 & & 3/15 \end{matrix}$$

Iterazione	0	1	2	3	.....

PageRank

Inizialmente, tutti I nodi hanno PageRank 1/8



Step	A	B	C	D	E	F	G	H
1	1/2	1/16	1/16	1/16	1/16	1/16	1/16	1/8
2	3/16	1/4	1/4	1/32	1/32	1/32	1/32	1/16

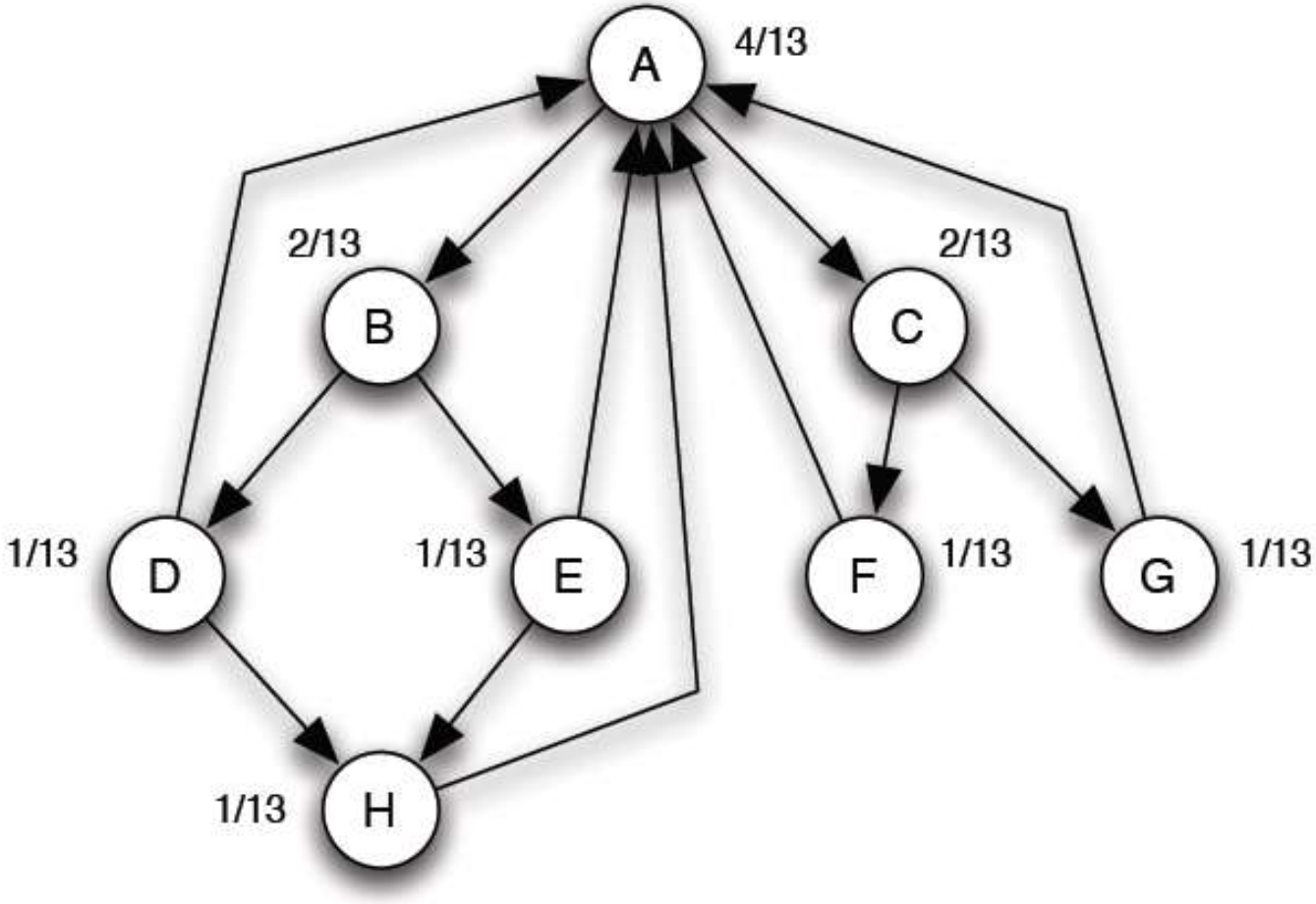
✓ similmente ad un “fluido” che circola in una rete

✓ Il PageRank totale nella rete rimane costante (non c’è necessità di normalizzazione); esso è uguale a 1

Page rank: converge?

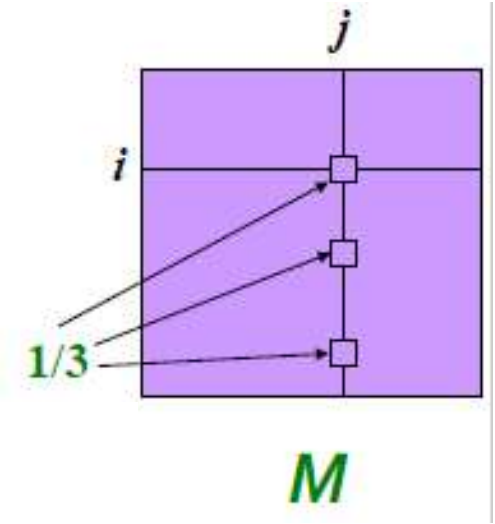
- Fatta eccezione per alcuni casi speciali, i valori dei PageRank di tutti i nodi convergono a valori limite per **k** che tende all'infinito
- Dal punto di vista pratico:
  - Possiamo fermare il processo quando il valore aggiornato del rank rimane uguale a quello prima dell'aggiornamento.

PageRank: equilibrio



PageRank: formulazione matriciale

- Matrice di adiacenza stocastica **M:**
  - d_j = out-degree della pagina j
  - se $j \rightarrow i$ allora $M_{ij} = 1/d_j$
  - ciascuna colonna di M (archi uscenti) somma ad 1



- Vettore dei rank $\boldsymbol{r}$:
  - Ha un entry per pagina
  - $\boldsymbol{r_i}$ è l'importanza della pagina $\boldsymbol{i}$
  - $\Sigma_i \boldsymbol{r_i} = \mathbf{1}$

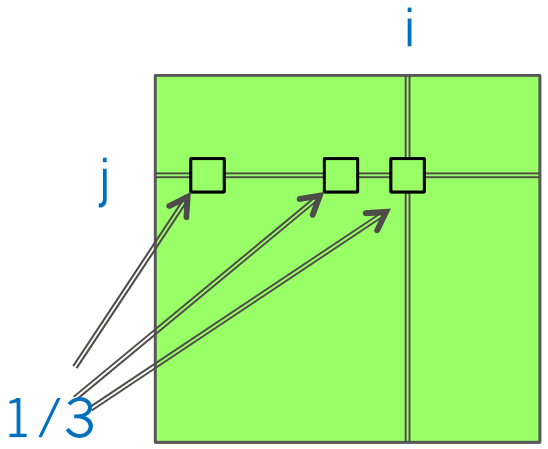
- Considerando che la riga j-sima di M considera tutti gli archi entranti in j, abbiamo che l'equazione del flusso può essere scritta

$$\boldsymbol{r} = \boldsymbol{M} \cdot \boldsymbol{r}$$

$$r_j = \sum_{i \rightarrow j} \frac{r_i}{\mathrm{d_i}}$$

PageRank: formulazione matriciale

- Matrice di adiacenza stocastica **M:**
  - d_j = out-degree della pagina j
  - se j → i allora $M_{ji} = 1/d_j$
  - ciascuna colonna di M somma ad 1



- Vettore dei rank $\boldsymbol{r}$:
  - Ha un entry per pagina
  - $\boldsymbol{r_i}$ è l'importanza della pagina $\boldsymbol{i}$
  - $\Sigma_i \boldsymbol{r_i} = \mathbf{1}$

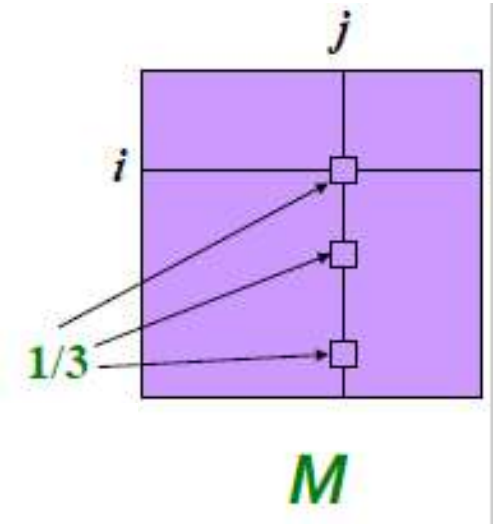
- Considerando che
  - la colonna j-sima di M considera tutti gli archi entranti in j, e che quindi la riga j-sima di $\boldsymbol{M}^{\top}$ considera tutti gli archi entranti in j abbiamo che l'equazione del flusso può essere scritta

$$\boldsymbol{r} = \boldsymbol{M}^{\top} \cdot \boldsymbol{r}$$

$$r_j = \sum_{i \to j} \frac{r_i}{\mathrm{d_i}}$$

PageRank: formulazione matriciale

- Matrice di adiacenza stocastica **M:**
  - d_i = out-degree della pagina i
  - se i → j allora $M_{ij} = 1/d_i$
  - ciascuna colonna di M somma ad 1



- Vettore dei rank $\boldsymbol{r}$:
  - Ha un entry per pagina
  - $\boldsymbol{r_i}$ è l'importanza della pagina $\boldsymbol{i}$
  - $\Sigma_i \boldsymbol{r_i} = \mathbf{1}$

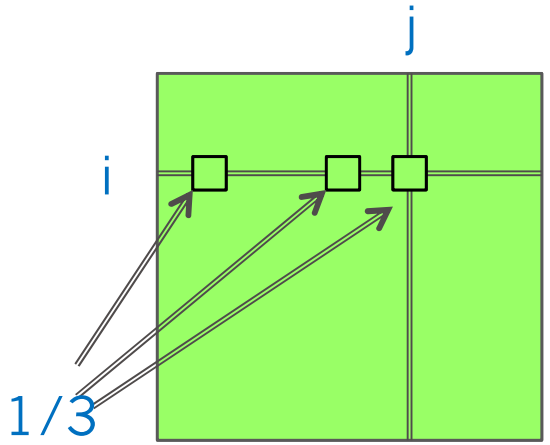
- Considerando che
  - la colonna j-sima di M considera tutti gli archi entranti in j, e che quindi la riga j-sima di $\boldsymbol{M}^{\mathsf{T}}$ considera tutti gli archi entranti in j abbiamo che l'equazione del flusso può essere scritta

$$\boldsymbol{r}^{(t+1)} = \boldsymbol{M}^{\mathsf{T}} \cdot \boldsymbol{r}^{(t)}$$

$$r_j^{(t+1)} = \sum_{i \to j} \frac{r_i^{(t)}}{d_i}$$

PageRank: formulazione matriciale

- Matrice di adiacenza stocastica **M:**
  - d_i = out-degree della pagina i
  - se i → j allora $M_{ij} = 1/d_i$
  - ciascuna riga di M somma ad 1



- Vettore dei rank $\boldsymbol{r}$:
  - Ha un entry per pagina
  - $\boldsymbol{r_i}$ è l'importanza della pagina $\boldsymbol{i}$
  - $\Sigma_i \boldsymbol{r_i} = \mathbf{1}$

- Considerando che
  - la colonna j-sima di M considera tutti gli archi entranti in j, e che quindi la riga j-sima di $\boldsymbol{M}^{\mathsf{T}}$ considera tutti gli archi entranti in j abbiamo che l'equazione del flusso può essere scritta

$$\boldsymbol{r}^{(t+1)} = \boldsymbol{M}^{\mathsf{T}} \cdot \boldsymbol{r}^{(t)}$$

$$r_j^{(t+1)} = \sum_{i \to j} \frac{r_i^{(t)}}{d_i}$$

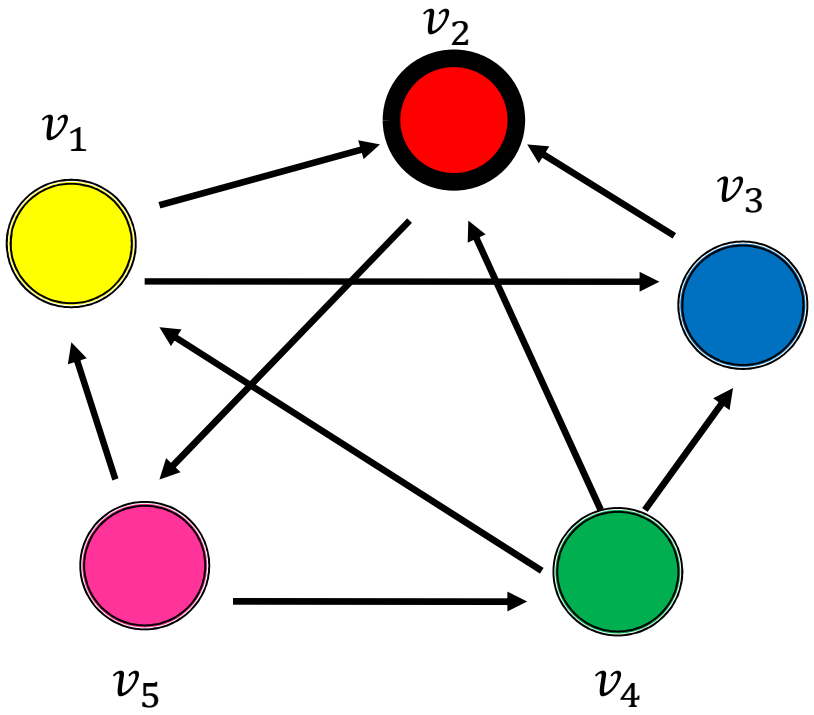
Interpretazione attraverso i random walk

- Immaginiamo un navigatore del web che a caso sceglie un cammino nella rete
  - Al tempo **t**, il navigatore è sulla pagina ***i***
  - Al tempo **t+1**, il navigatore segue uniformemente a caso un link che esce da ***i***, finendo in ***j***
  - Il processo si ripete continuamente

$$r_j^{(t+1)} = \sum_{i \to j} \frac{r_i^{(t)}}{\mathrm{d_i}}$$

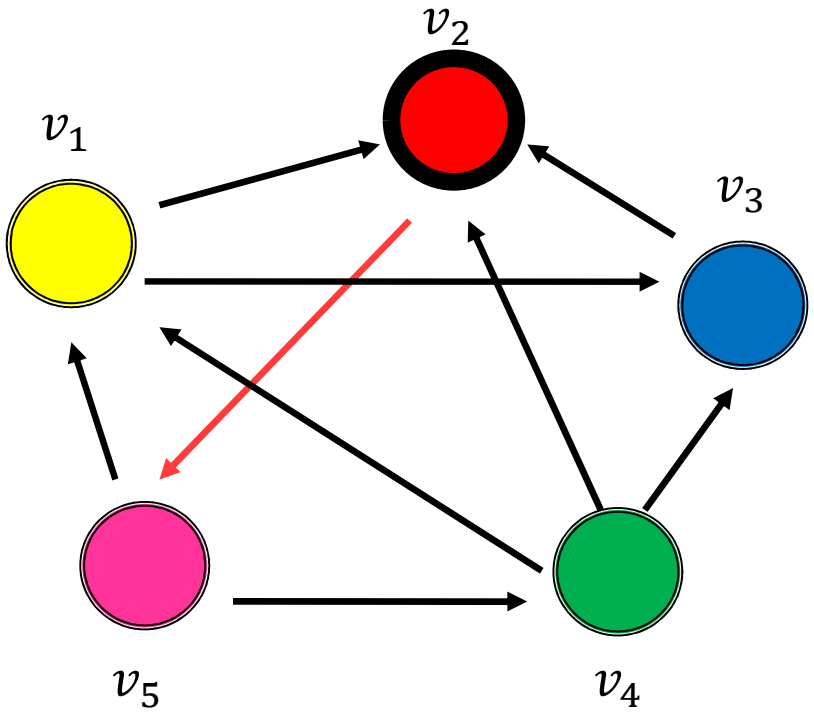
Esempio

- Step 0



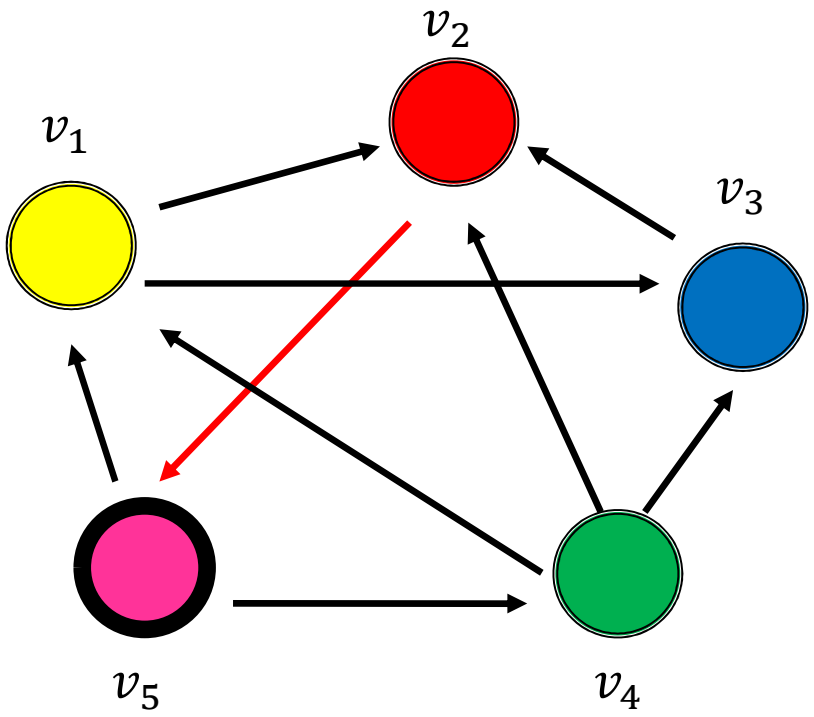
Esempio

- Step 0



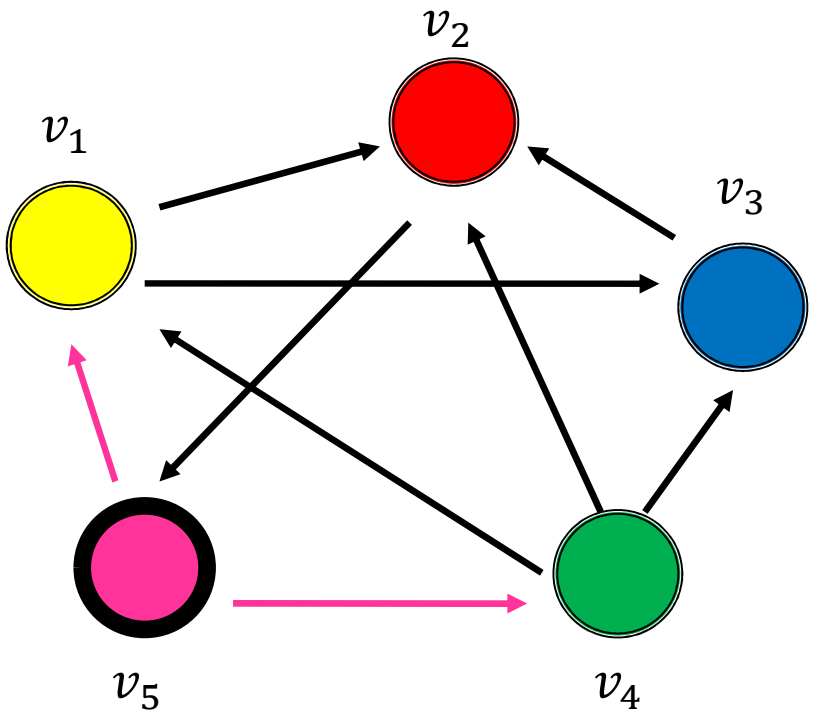
Esempio

- Step 1



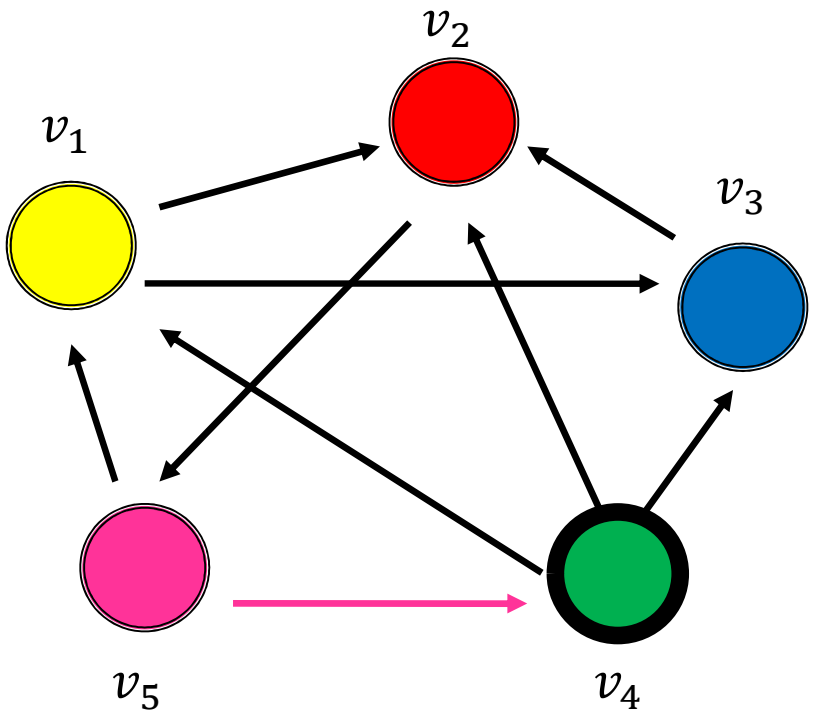
Esempio

- Step 1



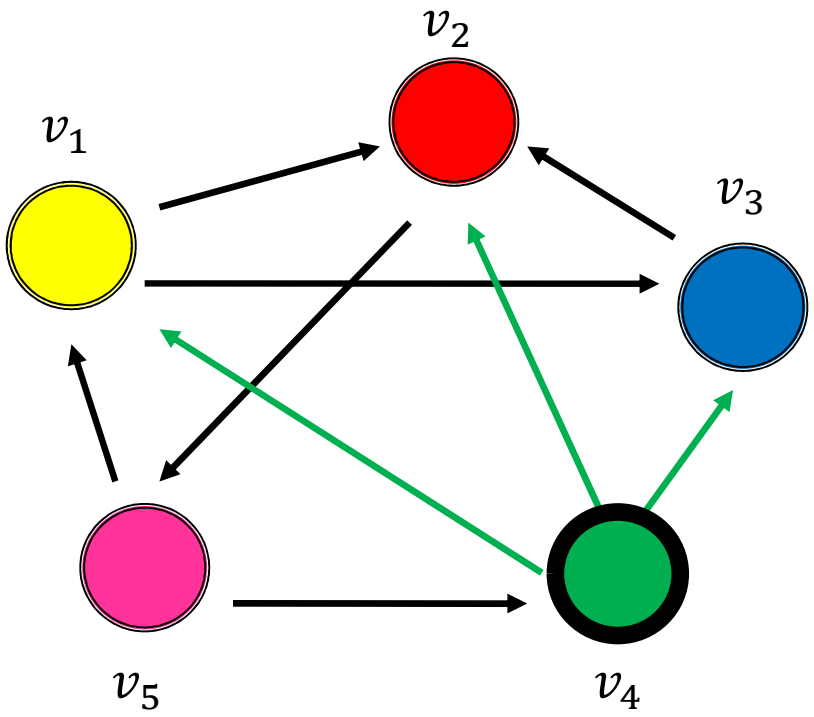
Esempio

- Step 2



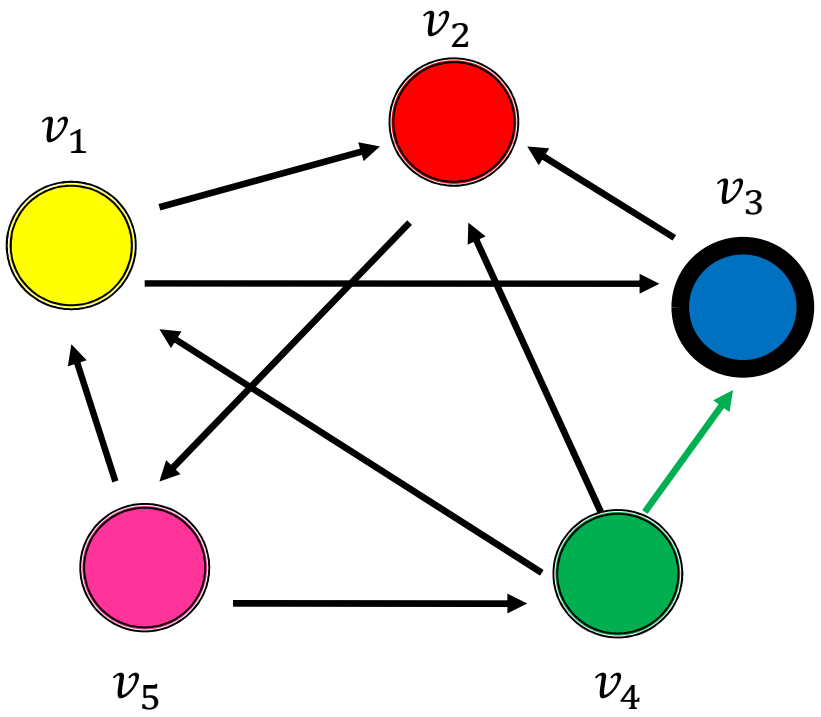
Esempio

- Step 2



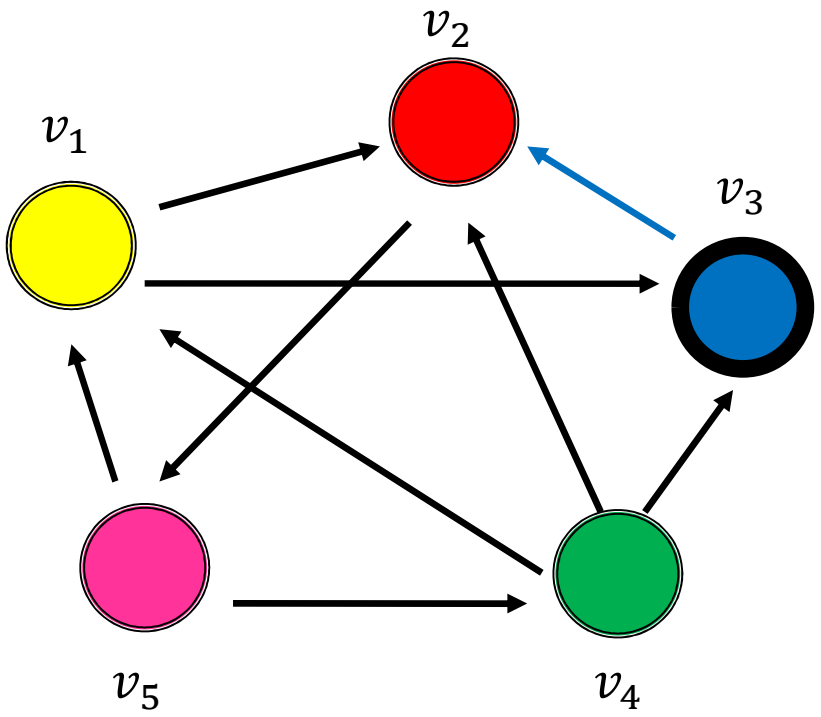
Esempio

- Step 3



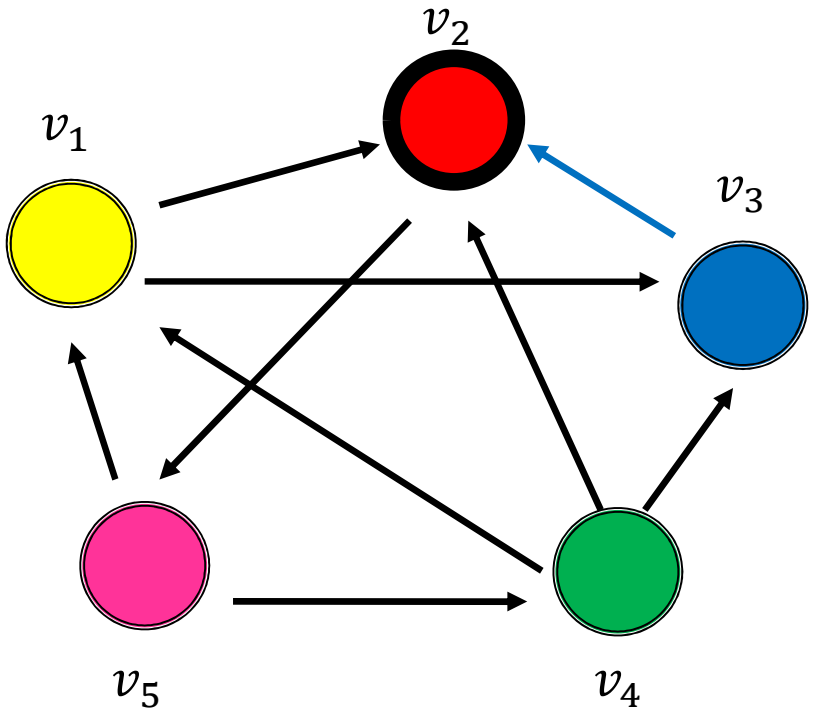
Esempio

- Step 3



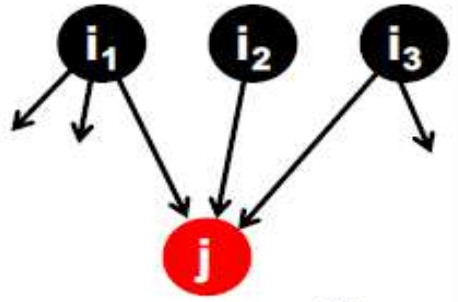
Esempio

- Step 4...



Interpretazione attraverso i random walk

- Immaginiamo un navigatore del web che a caso sceglie un cammino nella rete
  - Al tempo **t**, il navigatore è sulla pagina ***i***
  - Al tempo **t+1**, il navigatore segue uniformemente a caso un link che esce da ***i***, finendo in ***j***
  - Il processo si ripete continuamente
- Sia
  - **p(t)** = vettore la cui ***i***-sima componente è la probabilità che il navigatore sia alla pagina ***i*** al tempo **t**
  - **p(t)** è la distribuzione di probabilità su tutte le pagine



$$r_j^{(t+1)} = \sum_{i \to j} \frac{r_i^{(t)}}{d_i}$$

Random walk

- Quale è la probabilità p_i^t di essere al nodo i dopo t step?

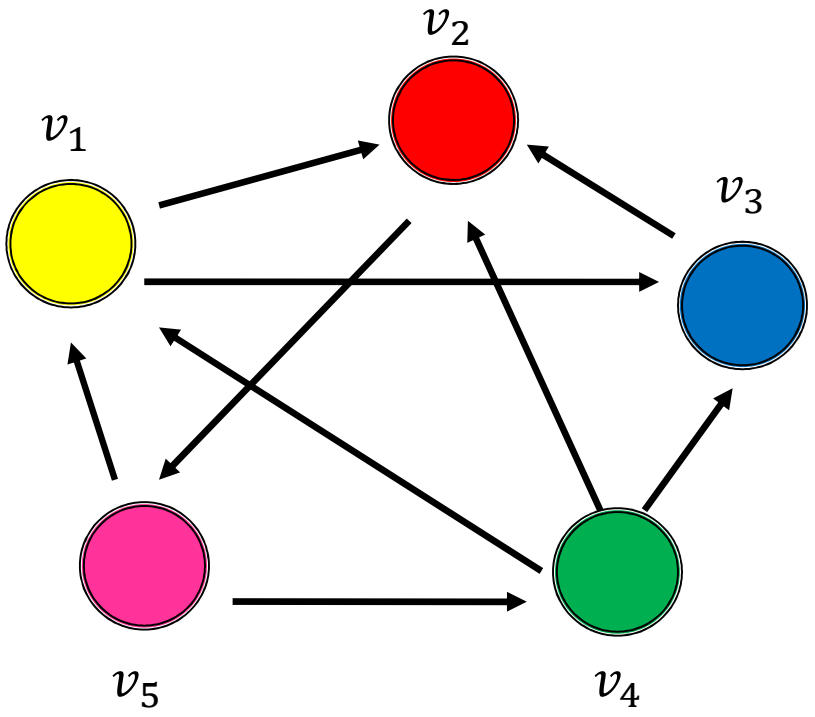
$$p_1^0 = \frac{1}{5} \qquad p_1^t = \frac{1}{3}p_4^{t-1} + \frac{1}{2}p_5^{t-1}$$

$$p_2^0 = \frac{1}{5} \qquad p_2^t = \frac{1}{2}p_1^{t-1} + p_3^{t-1} + \frac{1}{3}p_4^{t-1}$$

$$p_3^0 = \frac{1}{5} \qquad p_3^t = \frac{1}{2}p_1^{t-1} + \frac{1}{3}p_4^{t-1}$$

$$p_4^0 = \frac{1}{5} \qquad p_4^t = \frac{1}{2}p_5^{t-1}$$

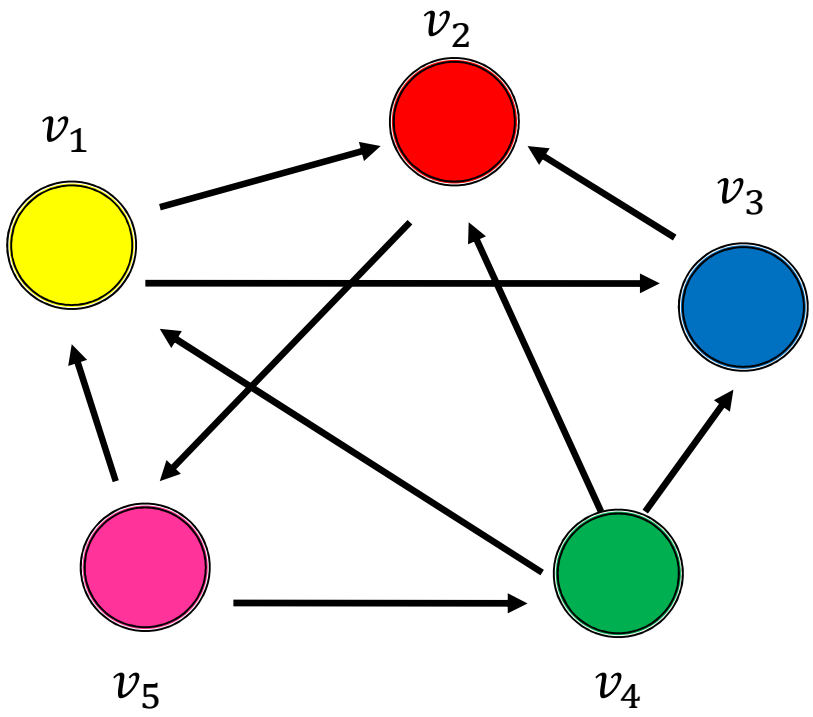
$$p_5^0 = \frac{1}{5} \qquad p_5^t = p_2^{t-1}$$



elementi della matrice M

Esempio

$$M = \begin{bmatrix} 0 & 1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 1/3 & 1/3 & 1/3 & 0 & 0 \\ 1/2 & 0 & 0 & 1/2 & 0 \end{bmatrix}$$



$$p_1^t = \frac{1}{3}p_4^{t-1} + \frac{1}{2}p_5^{t-1}$$

$$p_2^t = \frac{1}{2}p_1^{t-1} + p_3^{t-1} + \frac{1}{3}p_4^{t-1}$$

$$p_3^t = \frac{1}{2}p_1^{t-1} + \frac{1}{3}p_4^{t-1}$$

$$p_4^t = \frac{1}{2}p_5^{t-1}$$

$$p_5^t = p_2^{t-1}$$

M la si può vedere come **la matrice di transizione di una catena di Markov** con i nodi del grafo che fanno da stati della catena

Interpretazione attraverso i random walk

- Dove è il navigatore al tempo **t+1** ?
  - Segue un link scelto uniformemente a caso
    - $\mathbf{p(t+1) = M^T p(t)}$
- Supponiamo che il navigatore raggiunge uno stato

$$\mathbf{p(t+1) = M^T p(t) = p(t)}$$

allora **p(t)** è la distribuzione stazionaria π del random walk con matrice di transizione **M**

- π_i = la frazione di volte che visitiamo lo stato i quando $t \rightarrow \infty$
- Markov Chain Theory: Il random walk converge ad un' unica distribuzione stationaria indipendente dal vettore iniziale se il grafo è fortemente connesso e non bipartito

Interpretazione attraverso i random walk

Quindi

- considerando che per la distribuzione stazionaria π del random walk con matrice di transizione $\boldsymbol{M}$ vale

$$\pi = \boldsymbol{M}^{\top} \pi$$

- e ricordando che per il vettore dei rank $\boldsymbol{r}$ valeva

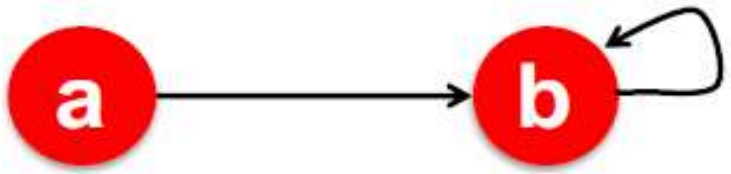
$$\boldsymbol{r} = \boldsymbol{M}^{\top} \cdot \boldsymbol{r}$$

- abbiamo

$\boldsymbol{r}$ è la distribuzione stazionaria del random walk

Problemi del PageRank

- “ Spider trap ” problem: i link in uscita rimangono all’interno di un gruppo; ci sono dei loop



$$r_j^{(t+1)} = \sum_{i \to j} \frac{r_i^{(t)}}{d_i}$$

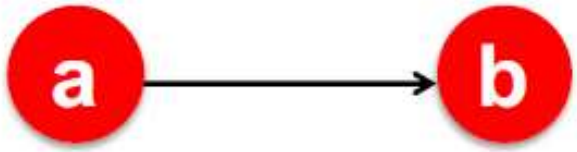
Esempio:

Step:	0,	1,	2,	3…
r_a	1/2	0	0	0
r_b	1/2	1	1	1

Arrivati in b non si esce più, il processo cicla all’infinito; il flusso viene assorbito

Problemi del PageRank

- “ Dead end ” problem: alcune pagine non hanno link uscenti



$$r_j^{(t+1)} = \sum_{i \to j} \frac{r_i^{(t)}}{d_i}$$

Esempio:

Step:	0,	1,	2,	3…
r_a	1/2	0	0	0
r_b	1/2	1/2	0	0

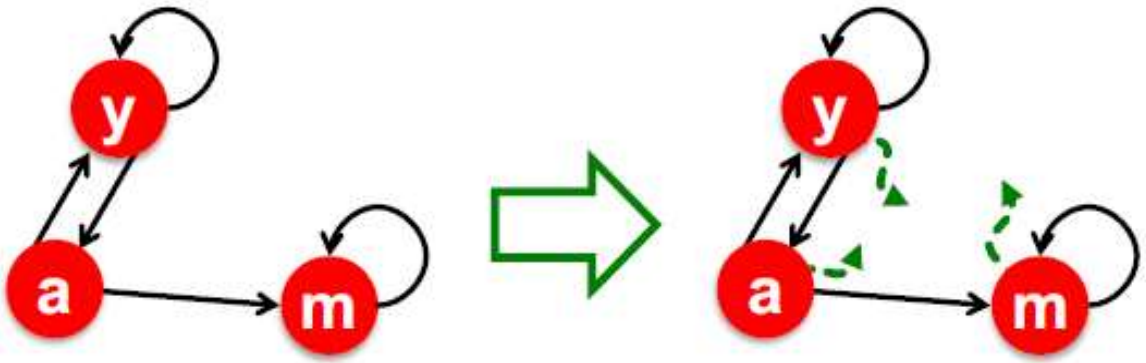
Arrivati in b, il processo finisce; il flusso si blocca

Soluzione: Teleport

Spider trap: soluzione proposta da Google

- In ciascun time step, il navigatore ha due opzioni:
  - con probabilità β segue un out-link a caso
  - con probabilità 1- β salta ad una pagina a caso (teleport link)
  - valori comuni per β sono nel range da 0.8 a 0.9

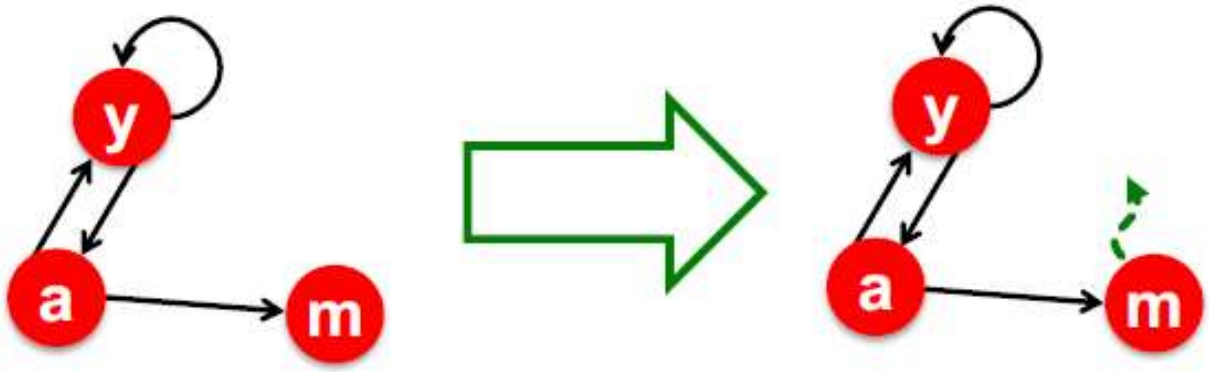
- Quando il navigatore si ritrova in uno spider trap entro pochi step ne sarà fuori



Soluzione: Teleport

Dead end:

- Arrivato in una pagina dead-end segui con probabilità 1 un teleport link scelto a caso



Equazione finale per il PageRank

Google's solution [Brin-Page, '98]

In ciascuno step, il navigatore sceglie a caso tra due opzioni:

- con probabilità β, segue un link a caso
- con probabilità 1-β, salta ad una pagina a caso

PageRank equation

$$r_j = \beta \sum_{i \to j} \frac{1}{d_i} r_i + (1 - \beta) \frac{1}{n}$$

P. Berkhin, *A Survey on PageRank Computing*, Internet Mathematics, 2005

PageRank e HITS

- PageRank ed HITS sono due soluzioni allo stesso problema:
  - Nel PageRank l'importanza di una pagina i dipende dai link entranti in i
  - In HITS, essa dipende dal valore dei link uscenti da i

Storia di PageRank

- Google si avvantaggiò molto nei primi giorni di applicazione dell'algoritmo
  - Esso dava un modo per valutare una pagina
    - Utile per creare un ordine tra le pagine del web
  - Dopo, divenne chiaro che l' 'anchor text' (testo presente in una pagina a cui è associato un link – testo su cui si clicca per seguire un link) era ancora più importante per effettuare il ranking
  - Ancora, i link spam potevano essere un arma che alterava il valore del rank
- Una grande quantità di ricerca è stata fatta
  - Analisi numeriche
  - Un enorme numero di varianti di PageRank
  - Le compagnie sono reticenti sulle tecniche attualmente in uso