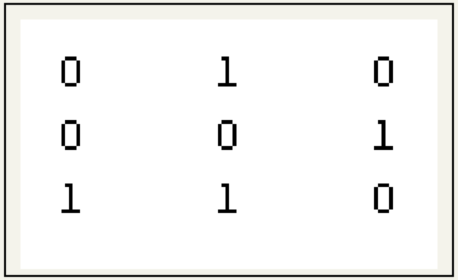
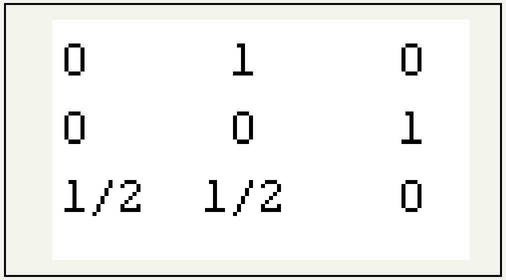
Random Walks

Paolo Ferragina
Dipartimento di Informatica
Università di Pisa

Definitions



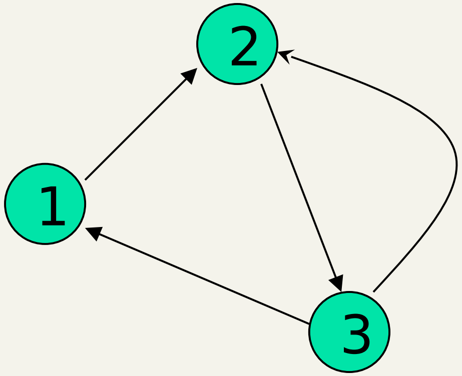
$$\begin{matrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{matrix}$$

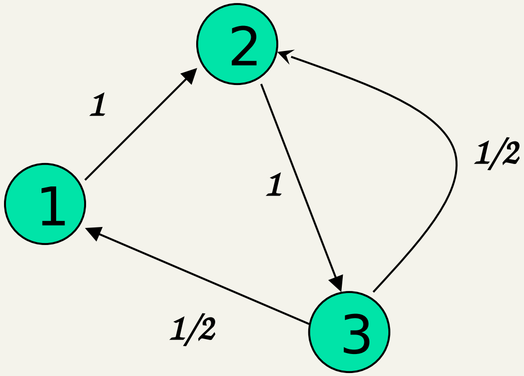


$$\begin{matrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1/2 & 1/2 & 0 \end{matrix}$$

Adjacency matrix A

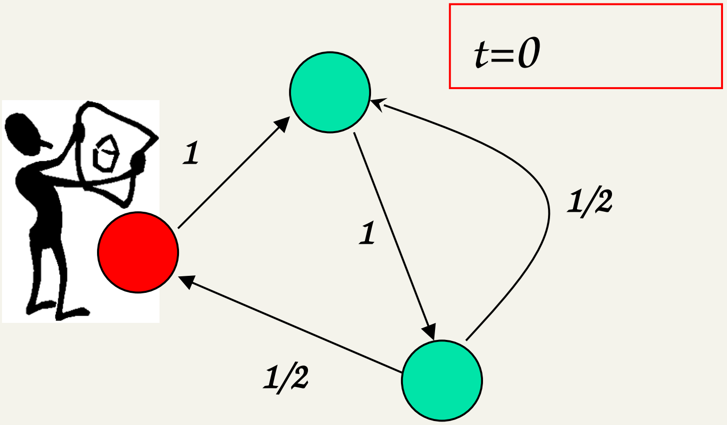
Transition matrix P





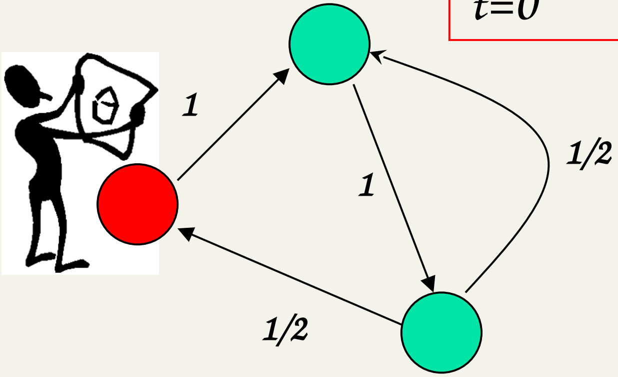
Any edge weigthing is possible

What is a random walk

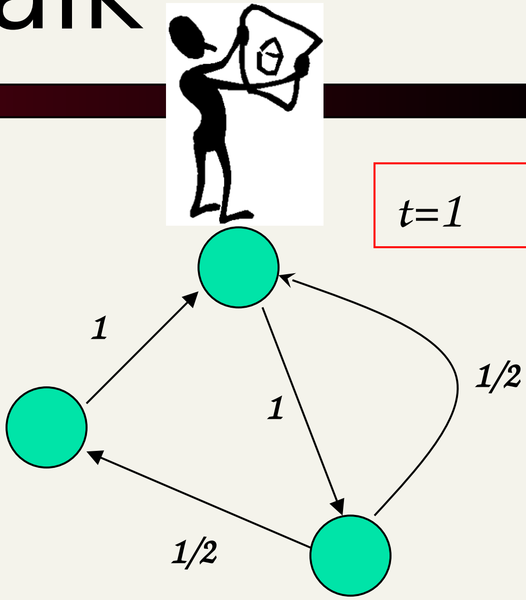


What is a random walk

t=0

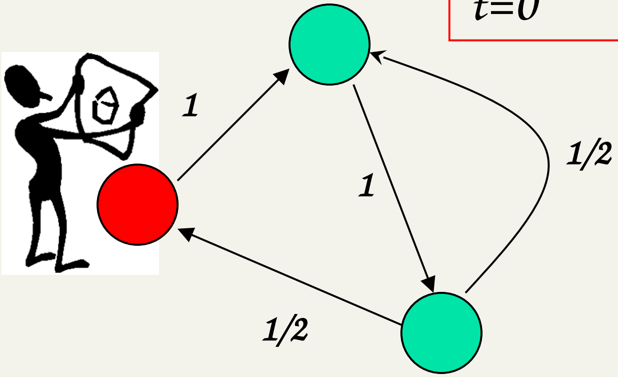


t=1

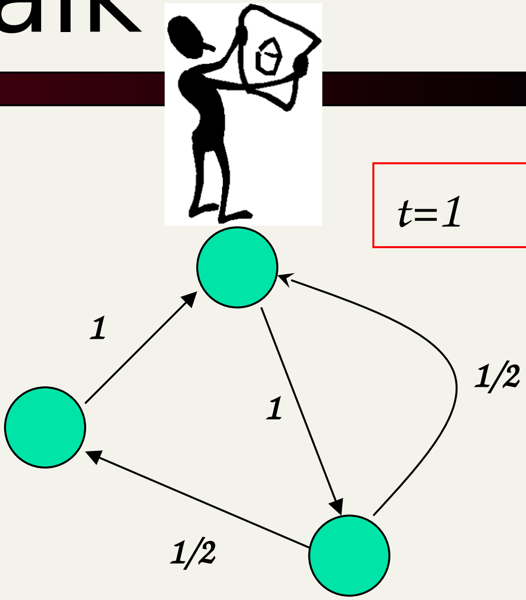


What is a random walk

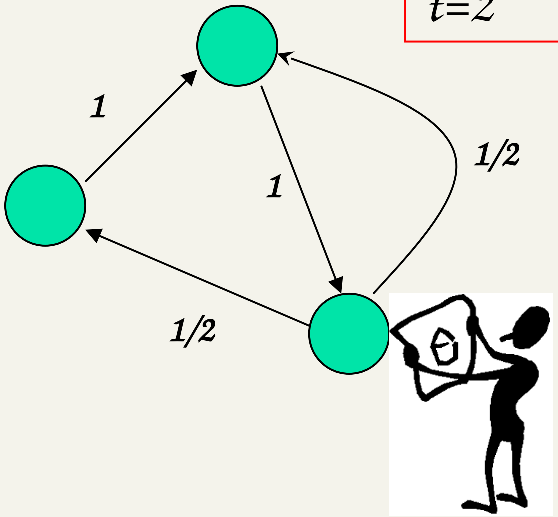
t=0



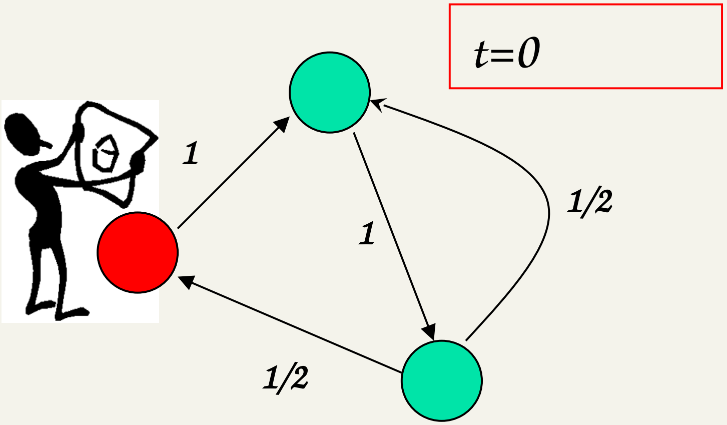
t=1

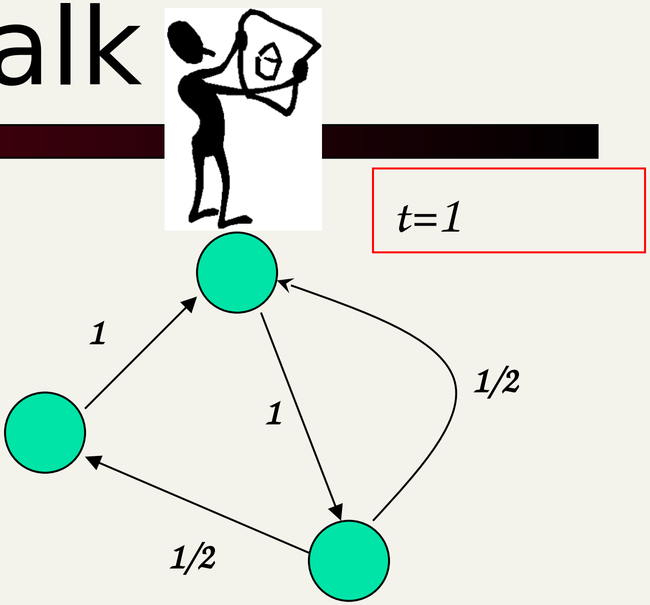


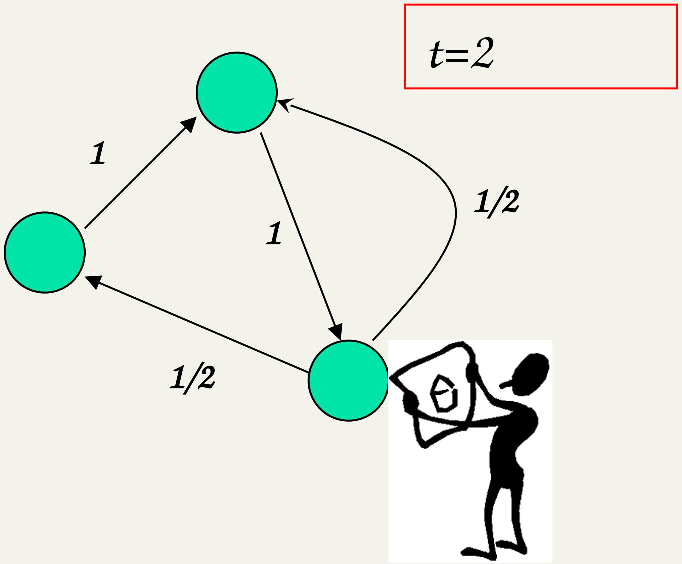
t=2

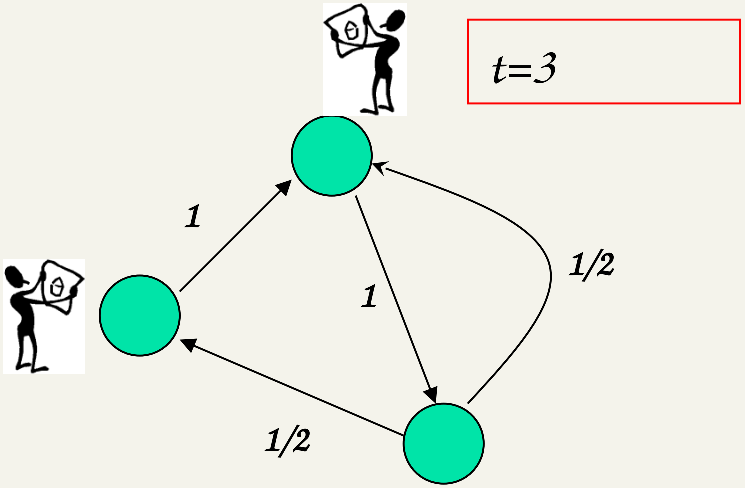


What is a random walk









Probability Distributions

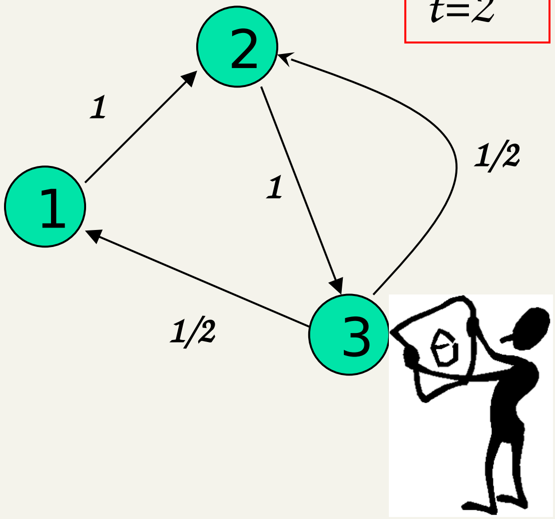
- $x_t(i)$ = probability that surfer is at node *i* at time *t*
- $x_{t+1}(i) = \sum_j$(Probability of being at node j)*Pr(j->i)

$$= \sum_j x_t(j)*P(j,i) = x_t * P$$

$$\underset{x_t}{\boxed{0 \quad 0 \quad 1}} \quad \boxed{\begin{matrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1/2 & 1/2 & 0 \end{matrix}} = \underset{x_{t+1}}{\boxed{½ \ ½ \ 0}}$$

Transition matrix P

t=2



Probability Distributions

Recall that:

- $x_t(i)$ = probability that surfer is at node *i* at time *t*
- $x_{t+1}(i) = \sum_j$(Probability of being at node j)*Pr(j->i)

  $= \sum_j x_t(j)*P(j,i) = x_t * P$

We can write:

- $x_{t+1} = x_t * P = (x_{t-1} * P) * P = (x_{t-2} * P) * P * P = \ldots = x_0 P^{t+1}$

- What happens when the surfer keeps walking for a long time? **Called Stationary distribution**

Stationary Distribution

- The stationary distribution at a node is related to the amount of time a random walker spends visiting that node.
- It is when the distribution does not change:

$x_{T+1} = x_T \rightarrow x_T P = 1 * x_T$ *(left eigenvector, with eigenvalue 1)*

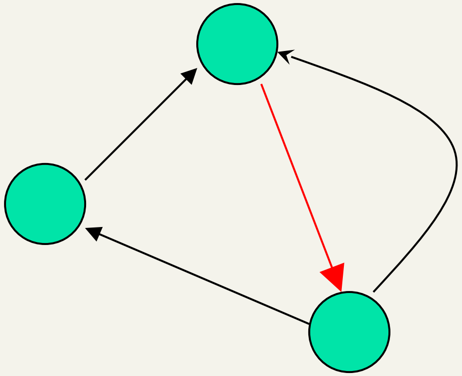
- For "well-behaved" graphs this does not depend on the start distribution x_0

Interesting questions

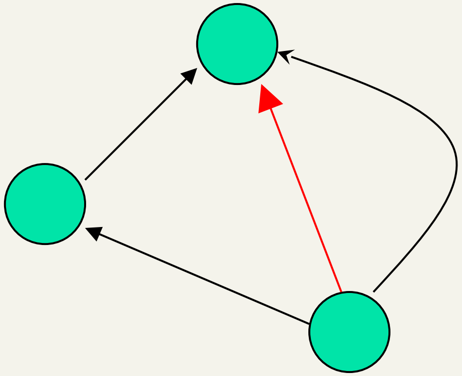
- Does a stationary distribution always exist? Is it unique?
  - Yes, if the graph is “well-behaved”, namely the markov chain is irreducible and aperiodic.
- How fast will the random surfer approach this stationary distribution?
  - Mixing Time!

Well behaved graphs

- Irreducible: There is a path from every node to every other node (→ it is an SCC).



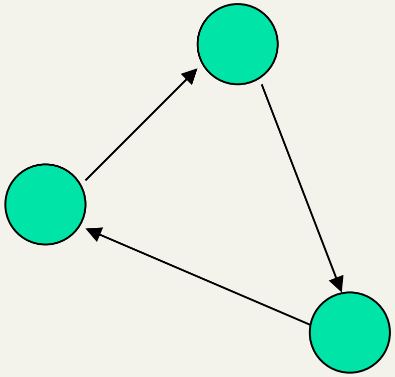
Irreducible



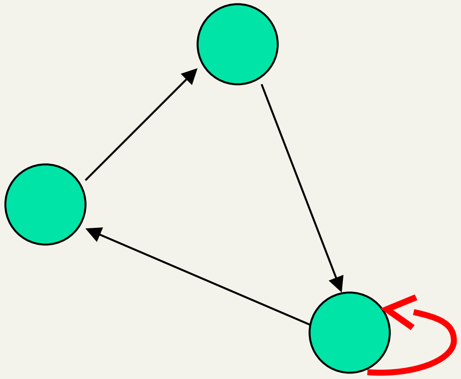
Not irreducible

Well behaved graphs

- Aperiodic: The GCD of all cycle lengths is *1.* The GCD is also called period.



Periodicity is 3



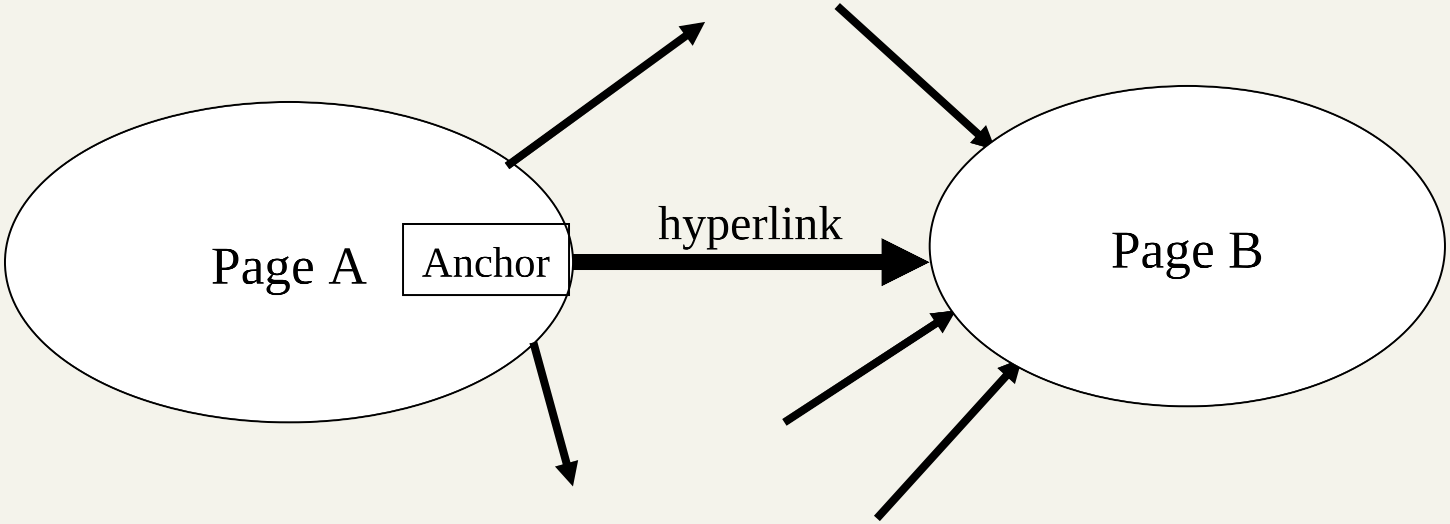
Aperiodic

Ranking

Link-based Ranking

(2° generation)

The Web as a Directed Graph

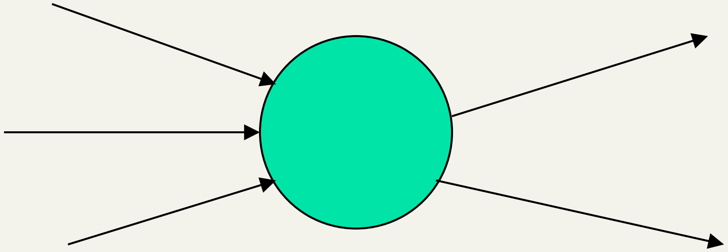


Assumption 1: A hyperlink between pages denotes author perceived relevance (quality signal)

Assumption 2: The text in the anchor of the hyperlink describes the target page (textual context)

Query-independent ordering

- First generation: using link counts as simple measures of popularity.
    - Undirected popularity:
        - Each page gets a score given by the number of in-links plus the number of out-links (es. 3+2=5).
    - Directed popularity:
        - Score of a page = number of its in-links (es. 3).

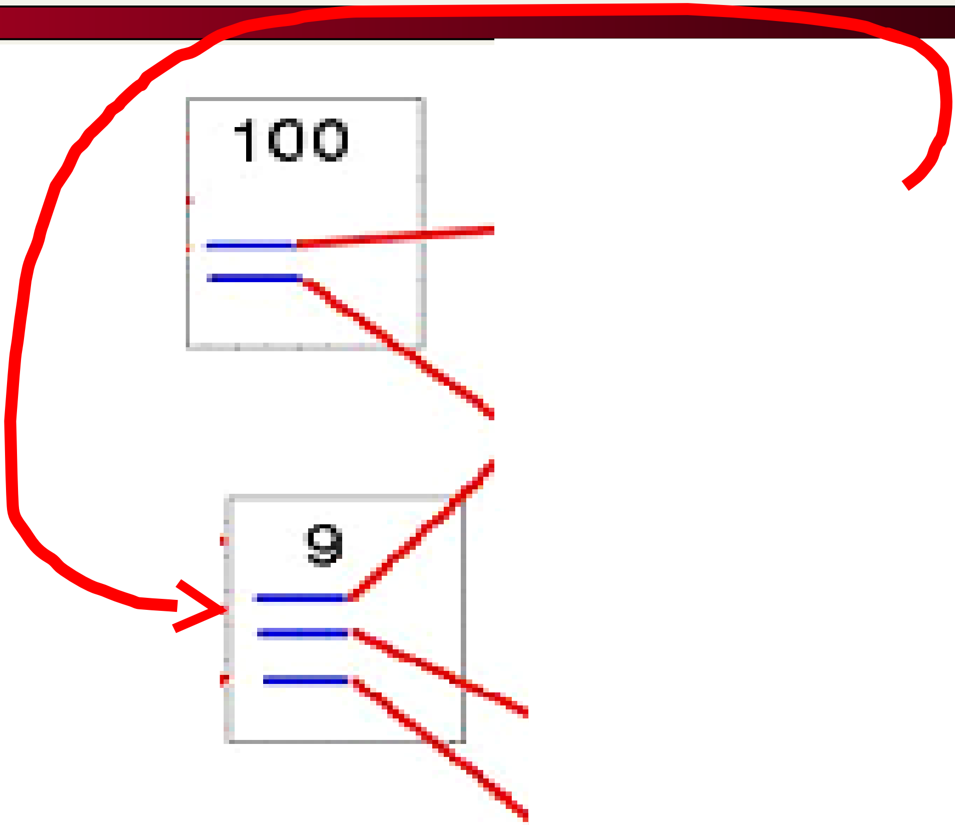


Easy to SPAM

Second generation: **PageRank**

- *Each link has its own importance!!*
- *PageRank* is
  - independent of the query
  - many interpretations…

The (classic) PageRank

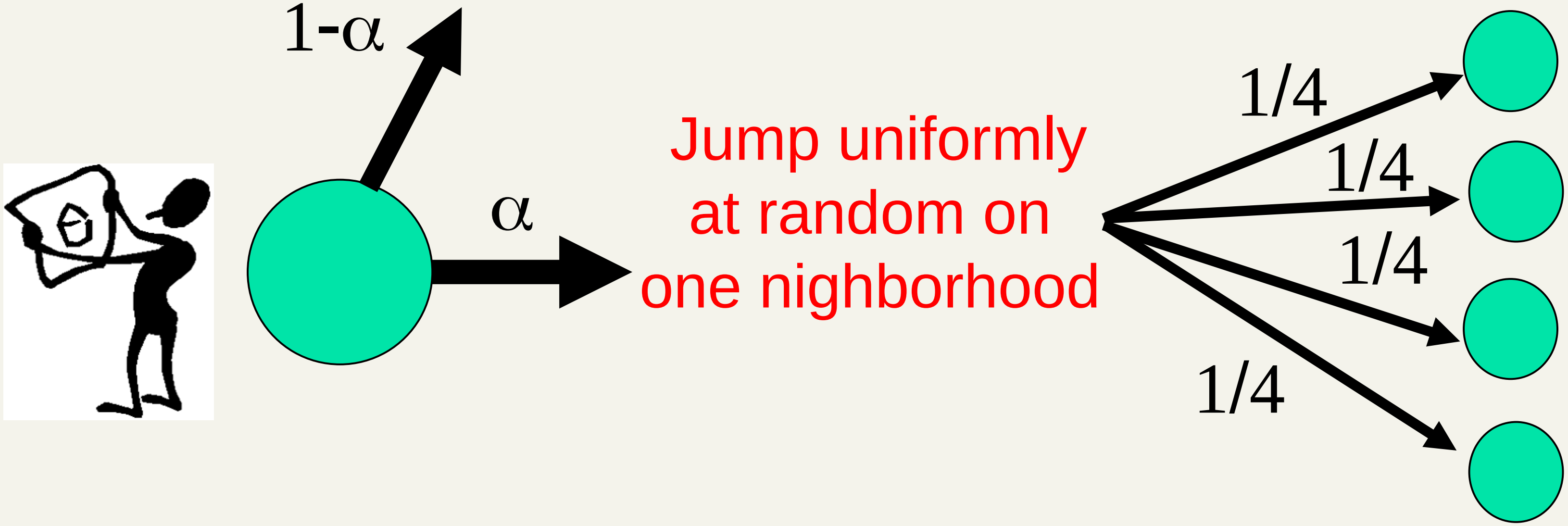


If we make these values flow,
Do they stabilize?

- Various interpretations: linear system of equations with billion variables and billion constraints
- Random walks

PageRank, as a Random Walk on the Web Graph

Jump uniformly at random at any page (node) in the Web



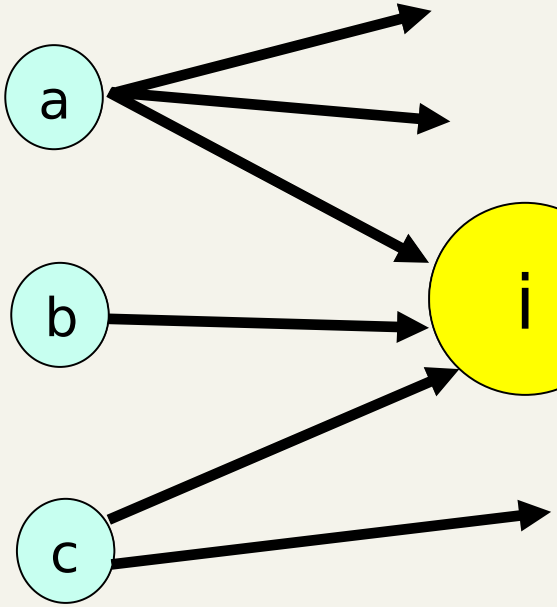
Jump uniformly at random on one nighborhood

PageRank of a node is the «frequency of visit» that node by assuming an infinite random walk

A «measure of centrality» of a node in a (directed) graph

PageRank, as a Linear System of Equations

$$r(i) = \alpha \cdot \sum_{j \in B(i)} \frac{r(j)}{\#out(j)} + (1-\alpha) \cdot \frac{1}{N}$$



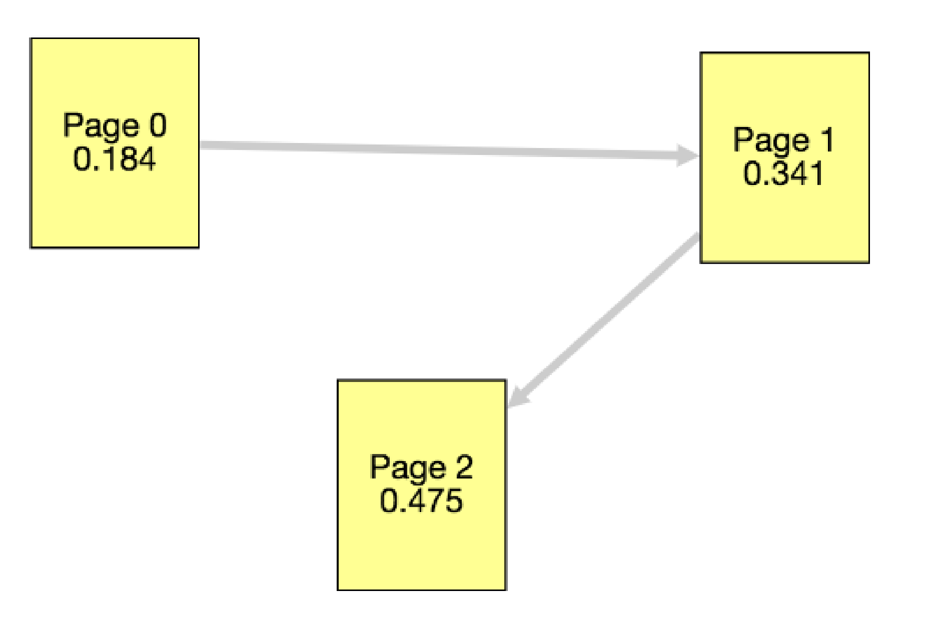
α = 0.85

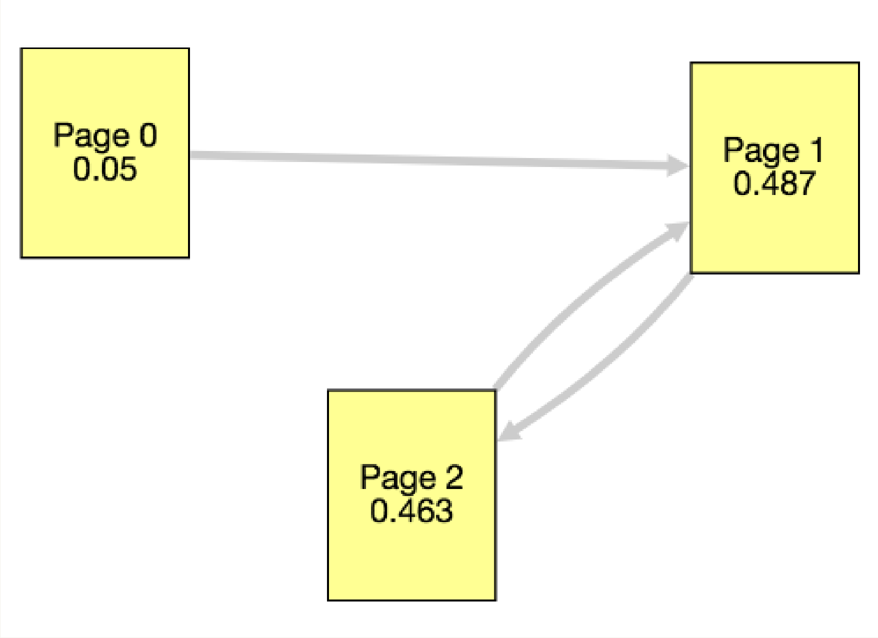
N = # nodes in graph

r(i) = α (r(a) / 3 +
r(b) / 1 +
r(c) / 2)
+ (1- α) / N

It is «related» to the eigenvalues of the matrix describing the linear system of equations

Hands-on test





http://bit.ly/2iwHH3e

http://faculty.chemeketa.edu/ascholer/cs160/WebApps/PageRank/

Pagerank: use in Search Engines

- Preprocessing:
  - Given graph, build P
  - Compute **r = [1/N, .., 1/N] * P^t** for t=0, 1, …
  - r[i] is the pagerank of page *i*

We are interested in the relative order

- Query processing:
  - Retrieve pages containing query terms
  - Rank them by their Pagerank

The final order is query-independent

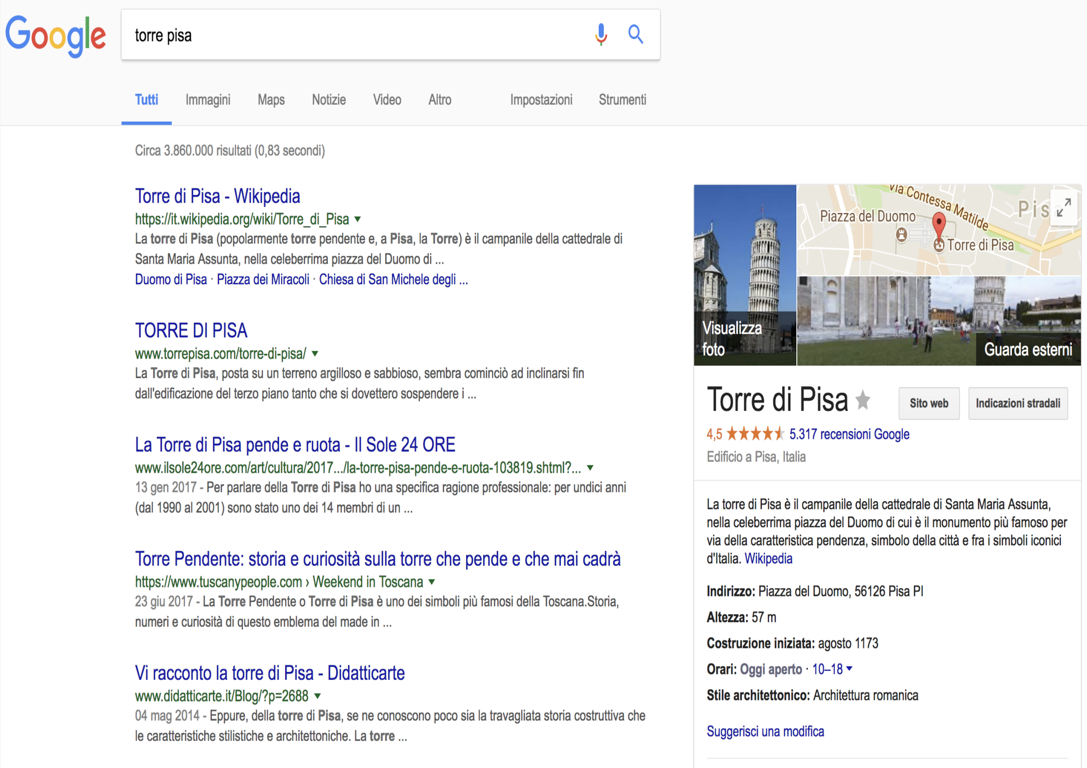
Nowadays

Relevance is a not well defined mathematical concept, which is actually not even depending on the single user because its needs may change over time too

For every page we compute a series of ***features***:

- TF-IDF of tokens
- PageRank
- Their proximity in the page
- Their occurrence in URL
- Their occurrence in the title
- …

there are > 200 «features»…



Circa 3.860.000 risultati (0,83 secondi)

Torre di Pisa - Wikipedia
https://it.wikipedia.org/wiki/Torre_di_Pisa ▾
La **torre** di **Pisa** (popolarmente **torre** pendente e, a **Pisa**, la **Torre**) è il campanile della cattedrale di Santa Maria Assunta, nella celeberrima piazza del Duomo di ...
Duomo di Pisa · Piazza dei Miracoli · Chiesa di San Michele degli ...

TORRE DI PISA
www.torrepisa.com/torre-di-pisa/ ▾
La **Torre** di **Pisa**, posta su un terreno argilloso e sabbioso, sembra cominciò ad inclinarsi fin dall'edificazione del terzo piano tanto che si dovettero sospendere i ...

La Torre di Pisa pende e ruota - Il Sole 24 ORE
www.ilsole24ore.com/art/cultura/2017.../la-torre-pisa-pende-e-ruota-103819.shtml?... ▾
13 gen 2017 - Per parlare della **Torre** di **Pisa** ho una specifica ragione professionale: per undici anni (dal 1990 al 2001) sono stato uno dei 14 membri di un ...

Torre Pendente: storia e curiosità sulla torre che pende e che mai cadrà
https://www.tuscanypeople.com › Weekend in Toscana ▾
23 giu 2017 - La **Torre** Pendente o **Torre** di **Pisa** è uno dei simboli più famosi della Toscana.Storia, numeri e curiosità di questo emblema del made in ...

Vi racconto la torre di Pisa - Didatticarte
www.didatticarte.it/Blog/?p=2688 ▾
04 mag 2014 - Eppure, della **torre** di **Pisa**, se ne conoscono poco sia la travagliata storia costruttiva che le caratteristiche stilistiche e architettoniche. La **torre** ...

Torre di Pisa ★
4,5 ★★★★★ 5.317 recensioni Google
Edificio a Pisa, Italia

La torre di Pisa è il campanile della cattedrale di Santa Maria Assunta, nella celeberrima piazza del Duomo di cui è il monumento più famoso per via della caratteristica pendenza, simbolo della città e fra i simboli iconici d'Italia. Wikipedia

Indirizzo: Piazza del Duomo, 56126 Pisa PI

Altezza: 57 m

Costruzione iniziata: agosto 1173

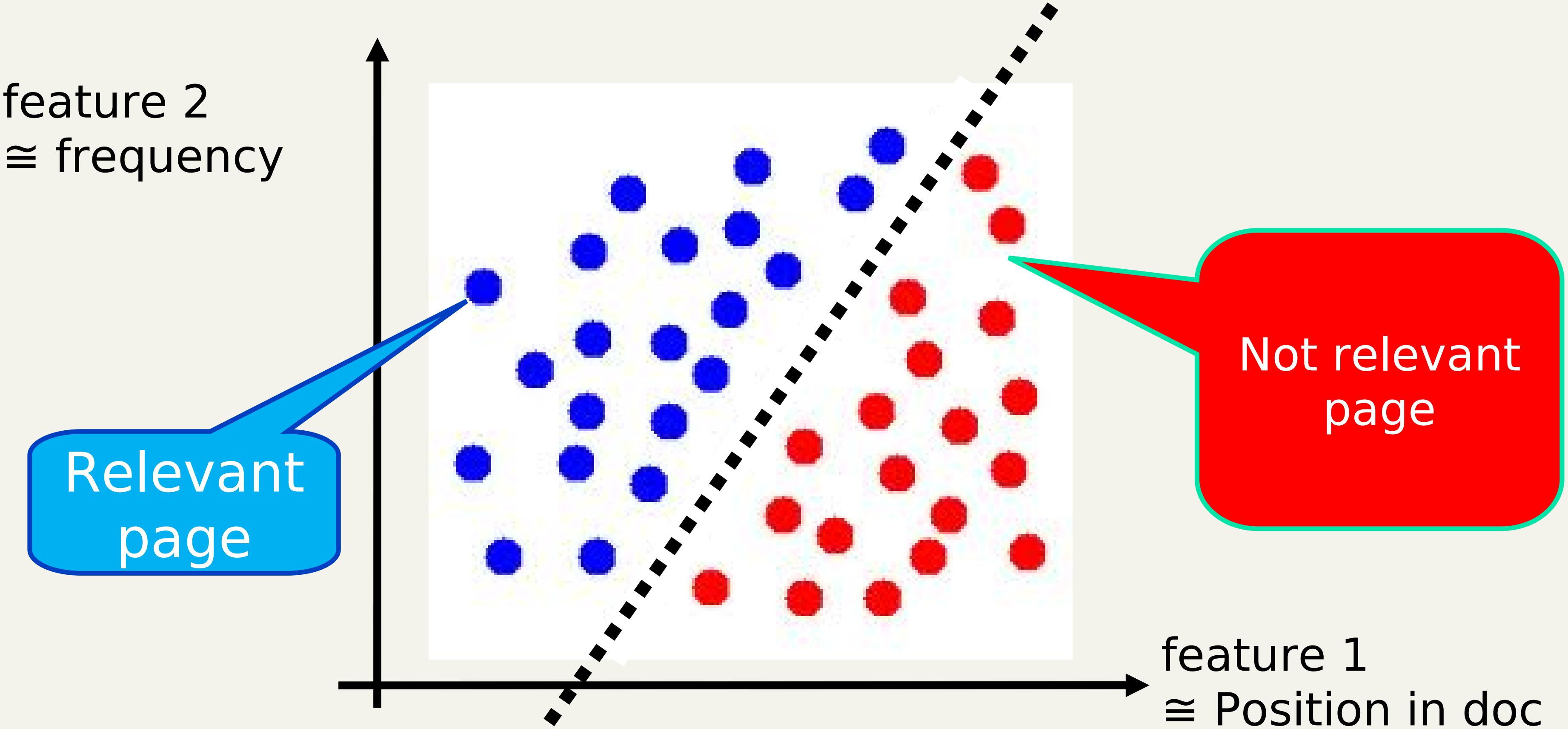
Orari: Oggi aperto · 10–18 ▾

Stile architettonico: Architettura romanica

Suggerisci una modifica

Computing the Ranking

Strong use of **AI e Machine Learning**

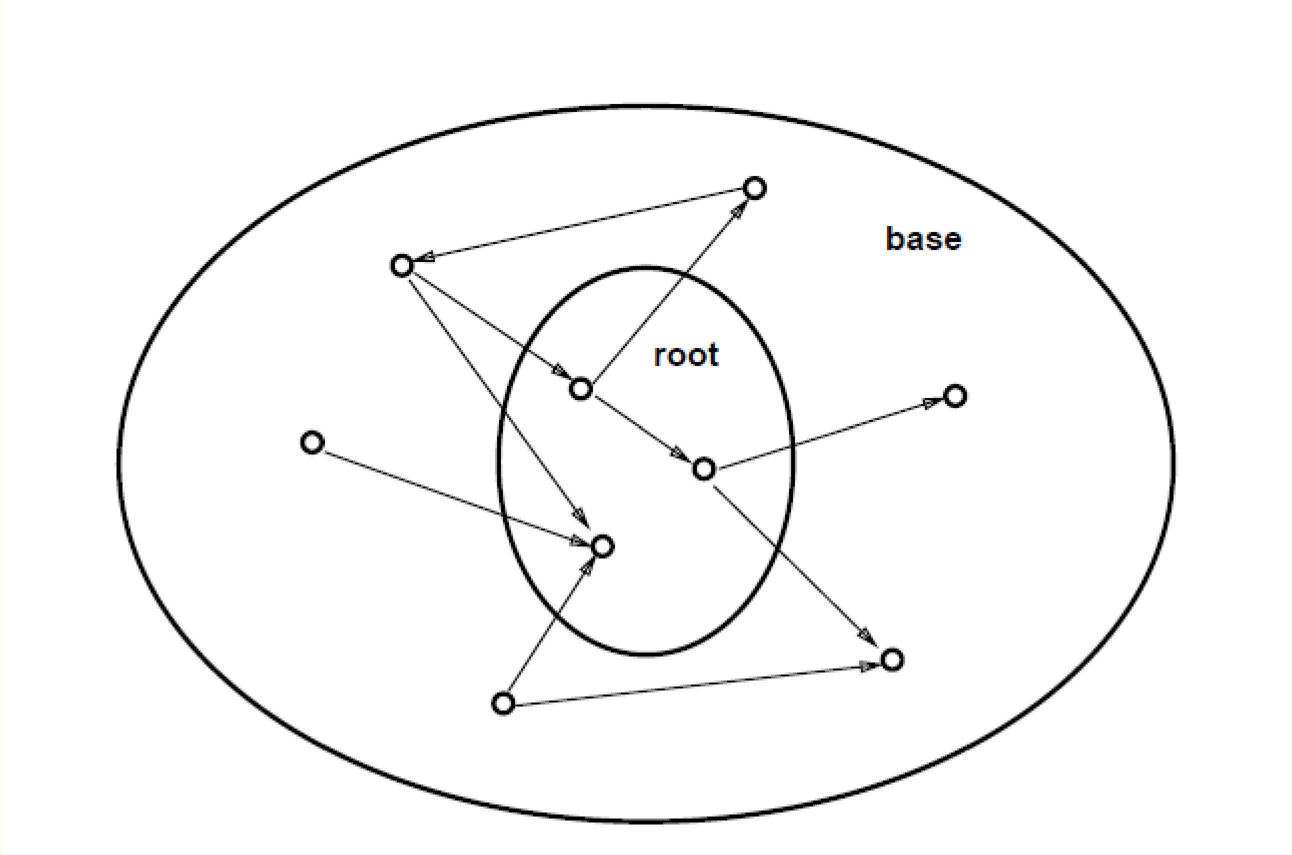


Personalized Pagerank

- Bias the random jump substituting the **uniform** jump to **all** nodes with the jump to **one** specific node (second term is (1-α) only for that node, the others are 0)
- … or uniform jump to **some** set **S** of preferred nodes (second term is (1-α)/|S| only for that set of nodes, the others are 0)
- Possibly not a uniform jump (change 1/#out(j) with the proper weight of the edge (j,i))

$$r(i) = \alpha \cdot \sum_{j \in B(i)} \frac{r(j)}{\#out(j)} + (1 - \alpha) \cdot \frac{1}{N}$$

HITS: Hypertext Induced Topic Search

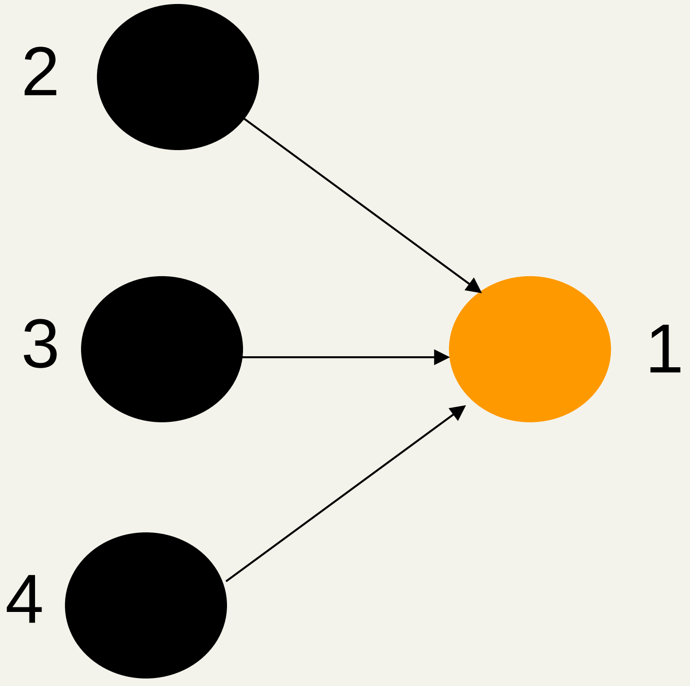


Calculating HITS

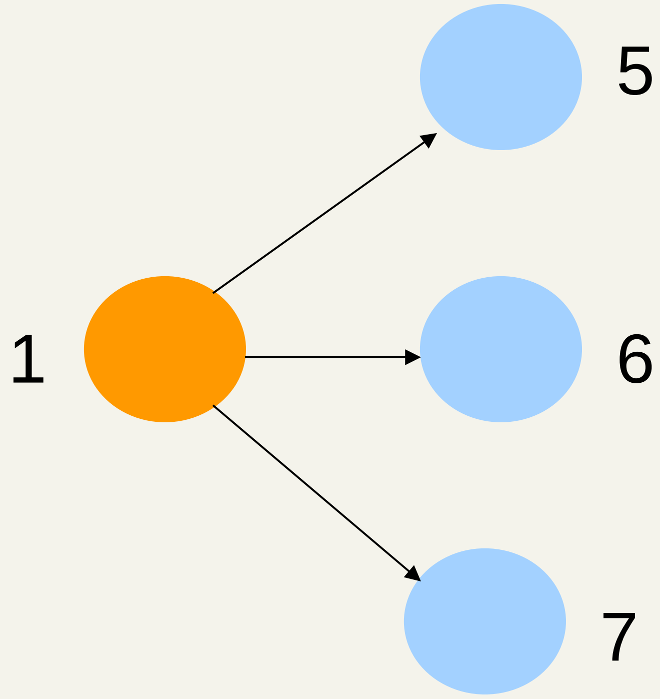
- *It is query-dependent*

- Produces two scores per page:
  - **Authority score**: a *good authority* page for a topic is *pointed* to by many good hubs for that topic.
  - **Hub score:** A *good hub* page for a topic *points* to many authoritative pages for that topic.

Authority and Hub scores



a(1) = h(2) + h(3) + h(4)



h(1) = a(5) + a(6) + a(7)

HITS: Link Analysis Computation

$$\left.\begin{aligned} a &= A^T h \\ h &= Aa \end{aligned}\right\} \Rightarrow \begin{aligned} a &= A^T A a \\ h &= A A^T h \end{aligned}$$

Where

a: Vector of Authority’s scores

h: Vector of Hub’s scores.

A: Adjacency matrix in which $a_{i,j} = 1$ if i→j

Symmetric matrices

Thus, h is an eigenvector of AA^t

a is an eigenvector of A^tA

Weighting links

Weight more if the query occurs in the neighborhood of the link (e.g. anchor text).

$$h(x) \leftarrow \sum_{x \to y} a(y)$$

$$a(x) \leftarrow \sum_{y \to x} h(y)$$

⇒

$$h(x) = \sum_{x \to y} w(x,y) \cdot a(y)$$

$$a(x) = \sum_{y \to x} w(x,y) \cdot h(y)$$

Summarization via Random Walks

Paolo Ferragina

UNIVERSITÀ DI PISA

The key simple idea

Rank (and select) sentences by saliency score of their constituting words ***w*** , computed as:

- **TF-IDF** for *weight(w)*

$$saliency(S_i) = \sum_{w \in S_i} \frac{weight(w)}{|S_i|}$$

- **Centrality over proper graphs:** PageRank, HITS, or other measures

TextRank

- The key issue is how the GRAPH is built
  - **Nodes =** terms or sentences
  - **Edges =** similarity relation between nodes

$$Similarity(S_i, S_j) = \frac{|S_i \cap S_j|}{\log |S_i| + \log |S_j|}$$

  - Use **PageRank** over weighted graph (directed by S’s position) and compute the score of the nodes

Lexical PageRank (LexRank)

- The main difference with TextRank resides in the way they compute **edge weights:**
  - **Cosine similarity** via Tf-Idf between sentences, so it is not pure content overlap (binary)
  - Edges are **pruned** if weight < threshold
- Scoring of **nodes** via **weigthed HITS** to ensure a mutual reinforcement between words and sentences

1) Do exist more sophisticate construction of graphs
2) What about multi-topic documents?