AALBORG UNIVERSITY

Link Analysis – PageRank & HITS

Peter Dolog
dolog@cs.aau.dk
http://people.cs.aau.dk/~dolog

Based (in part) on Stanford slides by Christopher Manning & Pandu Nayak and on the 'Introduction to Information Retrieval ' book (Chap. 21) by Christopher Manning, Prabhakar Raghavan & Hinrich Schütze. Also, some slides are provided by Bo Thiesson, Manfred Jaeger, Thomas D. Nielsen, AAU.

Outline

Link-based techniques for ranking

- Motivation
- Anchor text
- Structural measures
  - PageRank (query-independent)
  - HITS (query-dependent)

Today's lecture – hypertext and links (for ranking)

We look beyond the *content* of documents

- We begin to look at the hyperlinks between them

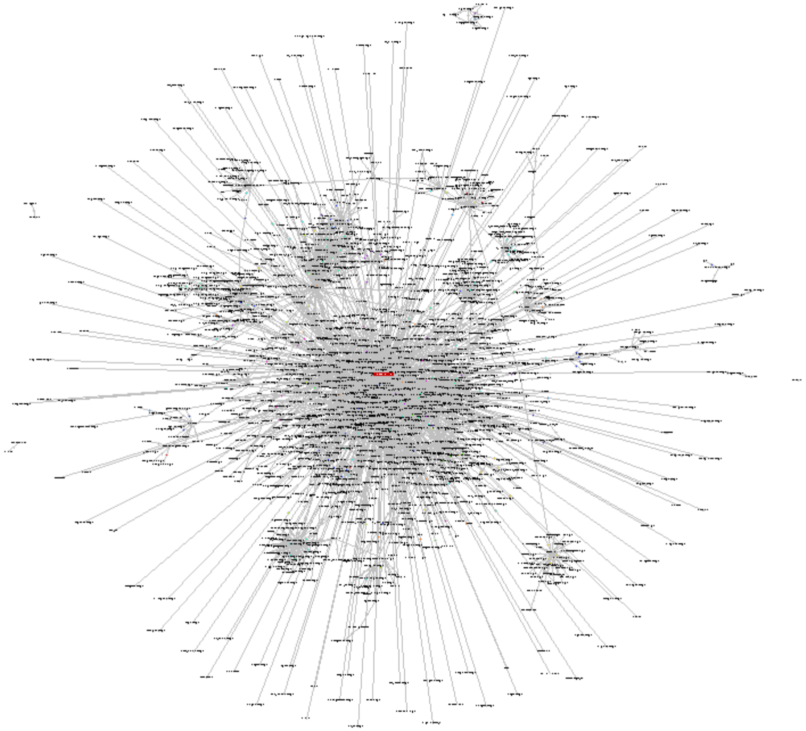
Links are everywhere

- The Web (today)
- Email
- Social networks
- Phone call logs
- Usage (connectivity) logs
- …

Web Structure – Social Network

- **A** node **for each** person
- **A** (directed) edge **for each (directed)** relationship

ENRON social network



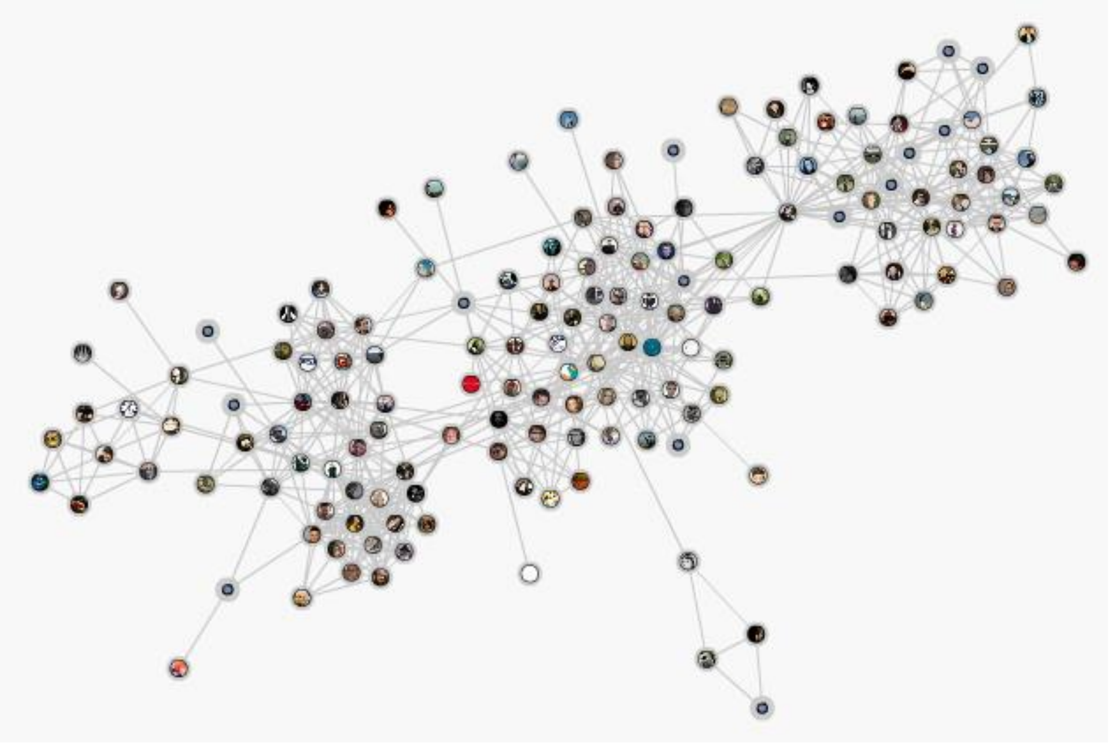
Nodes: ENRON employees

(Directed) link: "sent email to"

Web Structure – Social Network

- A node for each person
- A (directed) edge for each (directed) relationship

Facebook – Social graph



Nodes: Facebook profile & friends

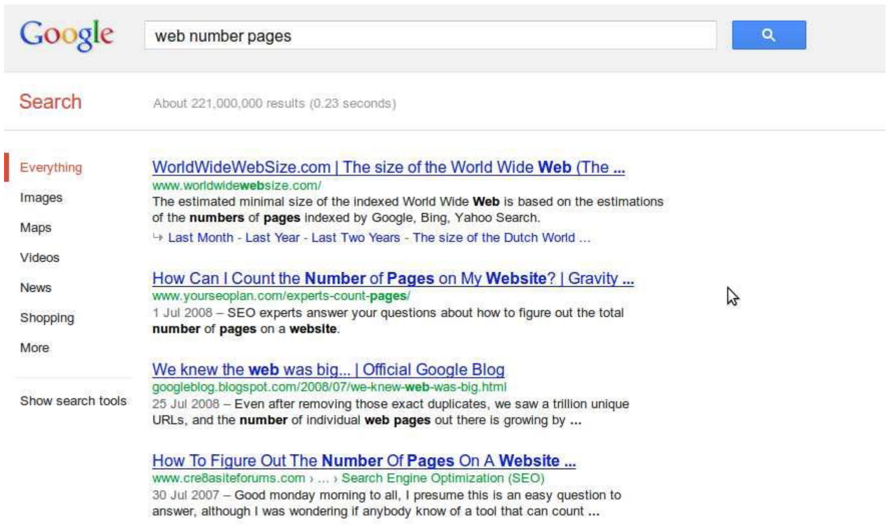
(Directed) link: "friend of"

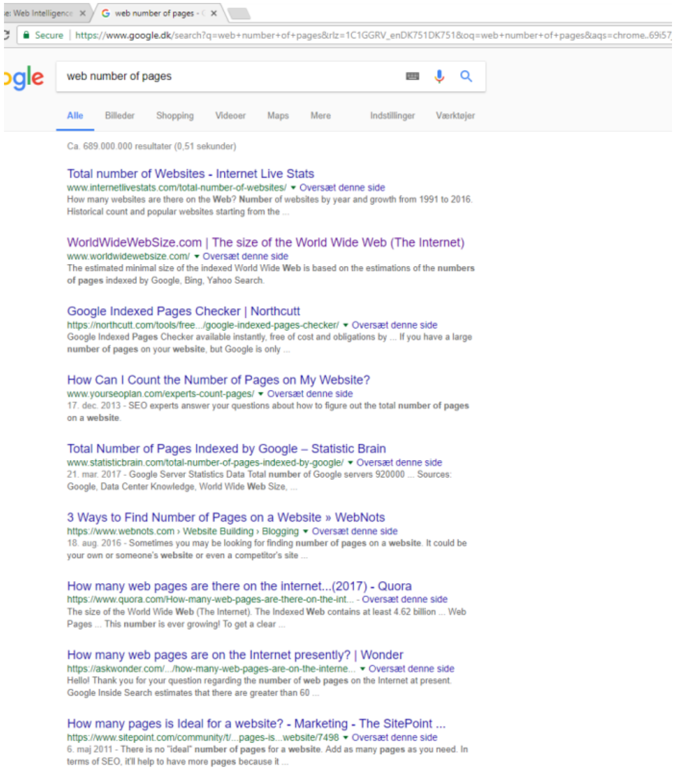
Ranking motivation

Google search for "web number pages":

2017

2015



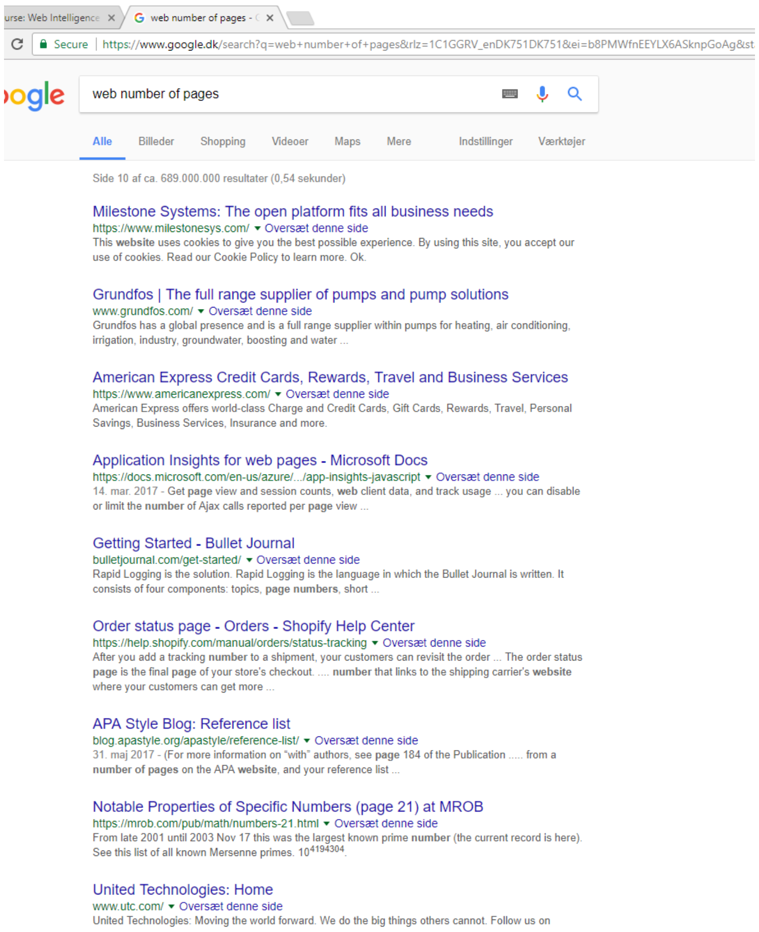


- 221.000.000 hits in 2015, 689.000.000 in 2017
- Top 4 results contains useful and trustworthy information

Ranking Motivation cntd.

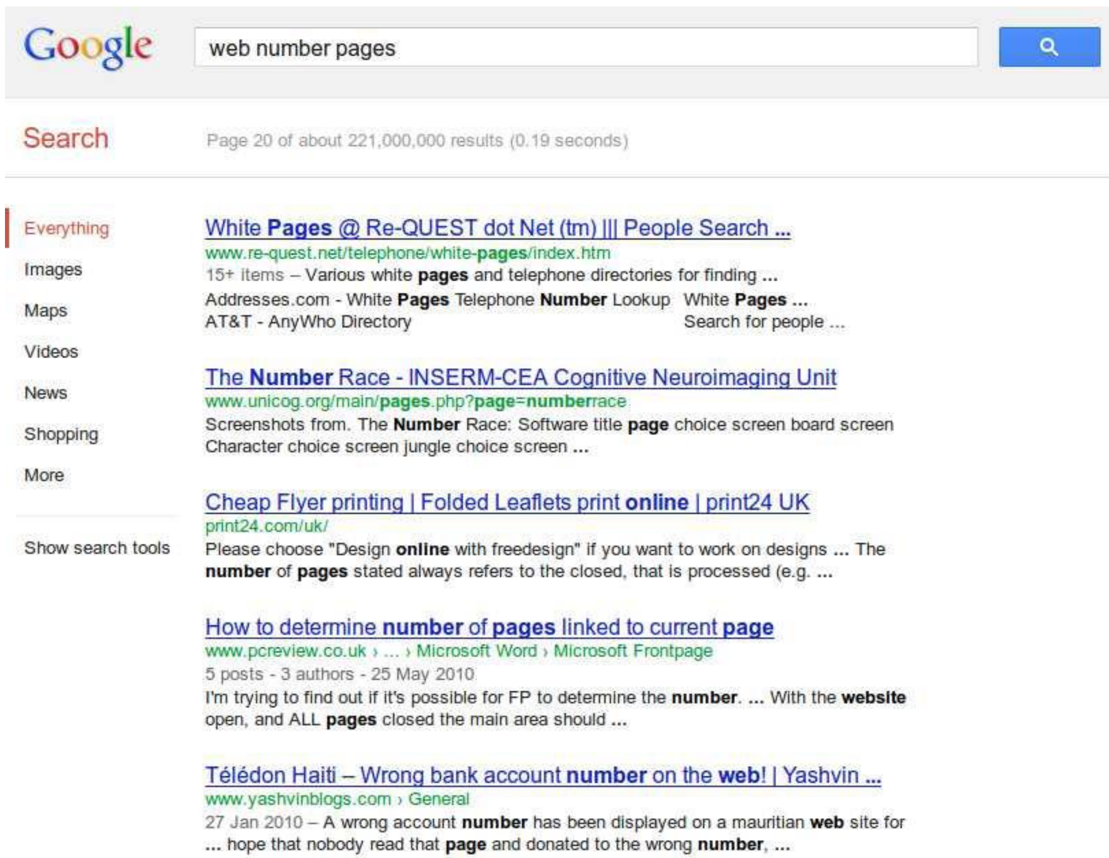
Look at 10th result page
(top 90-100)

Not so useful anylonger



Ranking motivation (cont.)

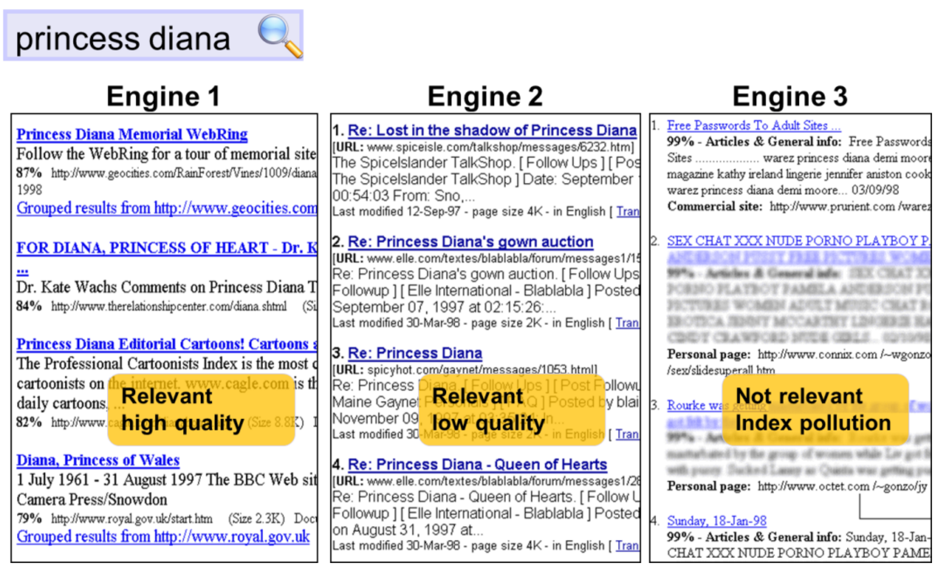
And results 200-205 cnt. less useful:



Ranking

Goal: order the answers to a query in decreasing order of value

- Just show top k (≈10) results (at a time)
- User is not overwhelmed



Some Ranking Criteria

- Content-based techniques (e.g. vector space model) – query-dependent
- Ad-hoc factors (anti-porn heuristics, location on page, publication/location data, length …) – mostly query-independent
- Human annotations
- Structure-based techniques (this lecture)
  - PageRank – query-independent
  - HITS – query-dependent

Ranking criteria defines a **ranking score** that measures how well a document and query "match".

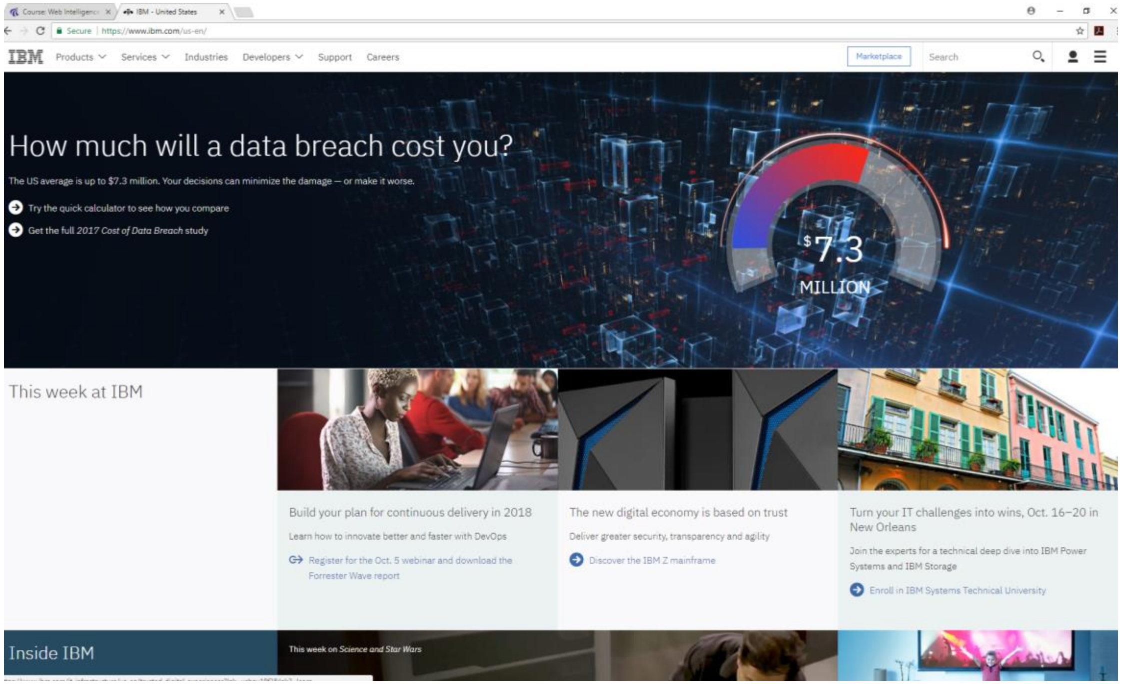
- E.g., ranking score $\in [0,1]$

IBM.com is first hit for search "IBM" – How?

IBM.com mostly graphical – not much mentioning of IBM in text-content

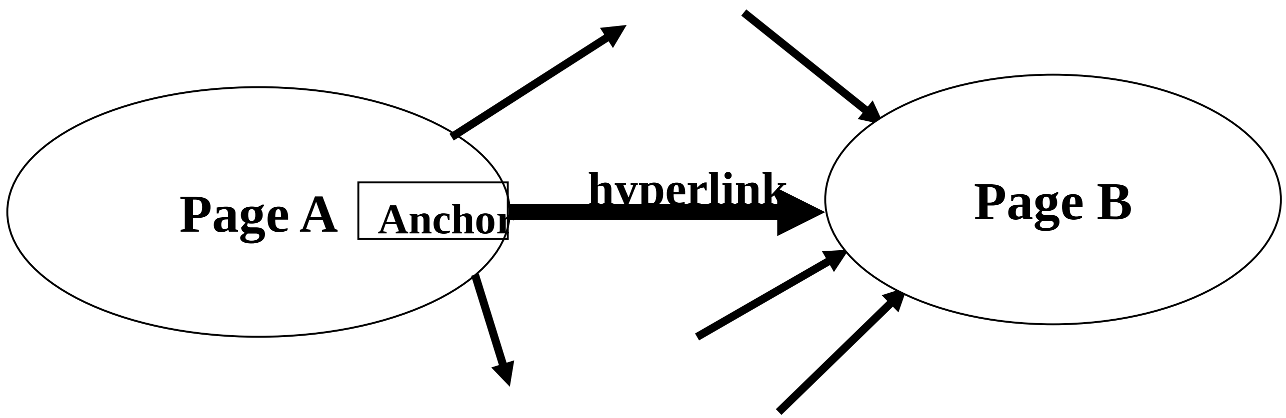
Why not:

- **IBM's copyright page – has high term frequency for "IBM"**
- **Spam page – with high term frequency for "IBM"**



The Web as a Directed Graph

- A node for each page
- A directed edge for each hyperlink



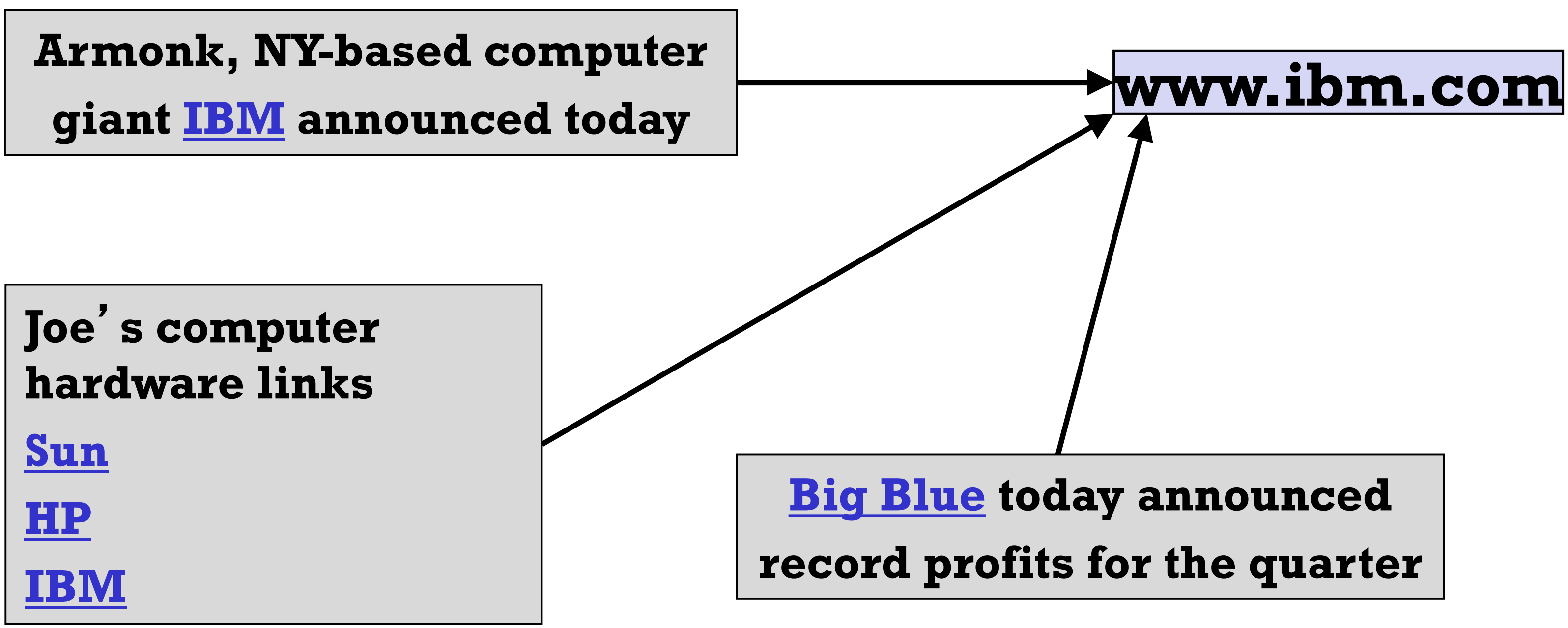
Hypothesis 1: A hyperlink between pages is a conferral of authority (quality signal)

Hypothesis 2: The text in the anchor of the hyperlink on page A describes the target page B

Hypothesis 2: The text in the anchor of the hyperlink on page A describes the target page B

Indexing anchor text

When indexing a document *D*, include (with some weight) anchor text from links pointing to *D*.



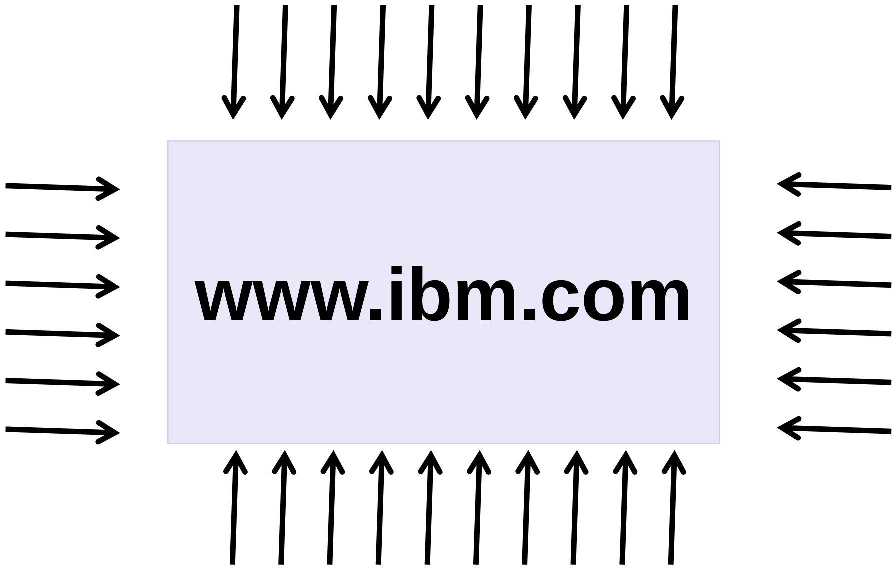
Use anchor text for indexing!

- Anchor is often a useful descriptor of page that it links to
  - (e.g., …the computer giant IBM today announces…)
- Anchor might be different from the text on the page → broader view of the page that better reflect "the crowds" multiple views
  - (e.g., …Big Blue today announced…)
  - Can be exploited by spammers (link spam) to generate anchors with text dishonesting some web sites
  - Weight anchor according to authority of referring page!
- Some undescriptive anchors: here, click, this…
  - (e.g., Information about IBM can be found here.)
  - Extended Anchor Text (i.e., neighbourhood of a link anchor may help!

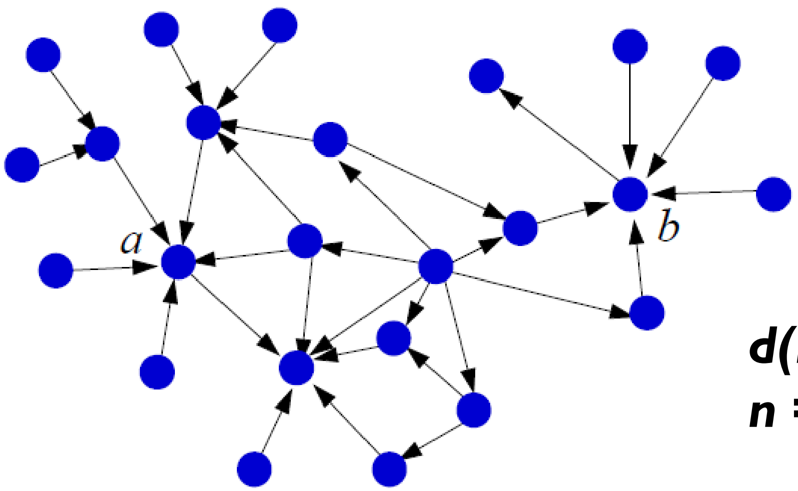
Hypothesis 1: A hyperlink between pages is a conferral of authority (quality signal)

Wisdom of the crowd!

A million links send a strong signal of prestige



Prestige as a simple local measure



d(i) **= in-degree (# incoming links)**
n **= # nodes**

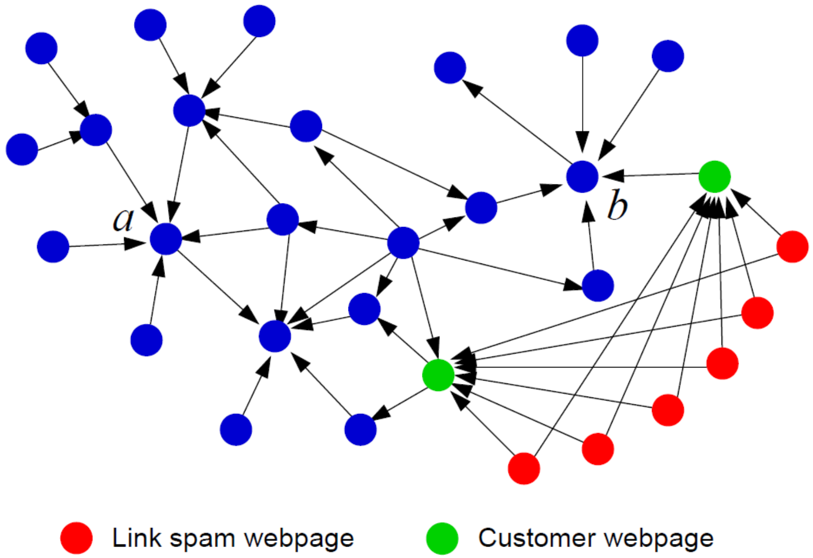
Prestige

$$P(i) = d(i) \,/\, (n-1)$$

Example: P(a) = P(b) = 5/23

Prestige (cont.)

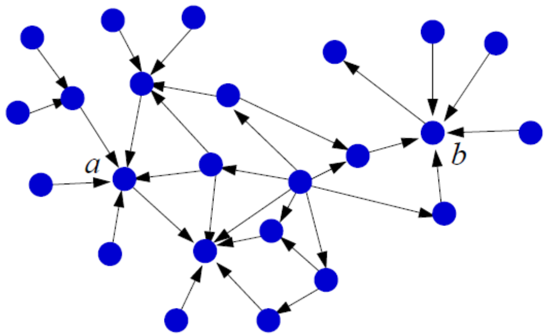
Local measures can be easily manipulated by link spamming:



On the other hand, more difficult to manipulate global measures (though still possible)

A global prestige measure

Rank prestige:



$$P(j) = \sum_{i:i \to j} P(i)$$

Recursion: Prestige of page depends on

- Its in-degree, and
- The prestige of pages linking to it
- This does not directly defines prestige – only a mutual relationship between values

Can we find a prestige measure that satisfies these relationships?

PageRank

Definition of PageRank
(Brin and Page '98)

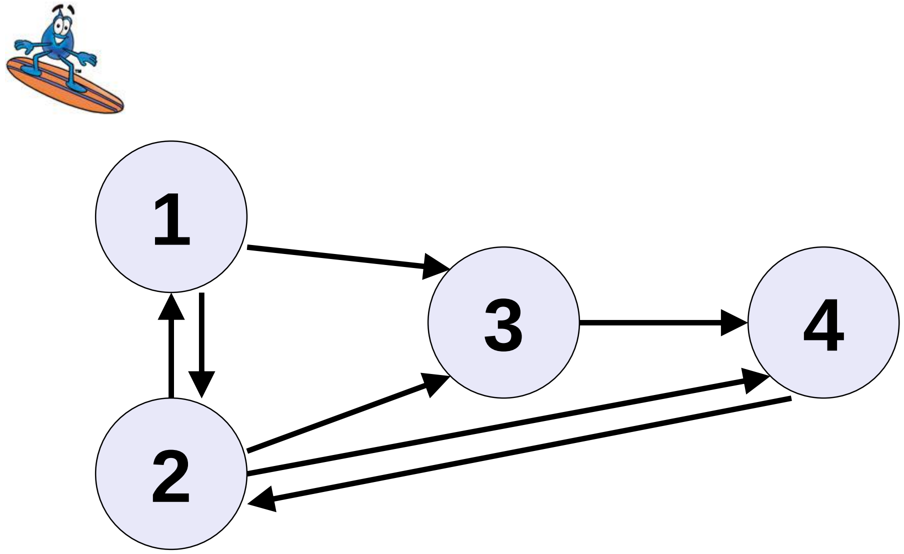
Consider the following infinite random walk (surf):

- Initially the surfer is at a random page
- At each step, the surfer proceeds
  - to a randomly chosen web page with probability α
  - to a randomly chosen successor of the current page with probability $1 - \alpha$

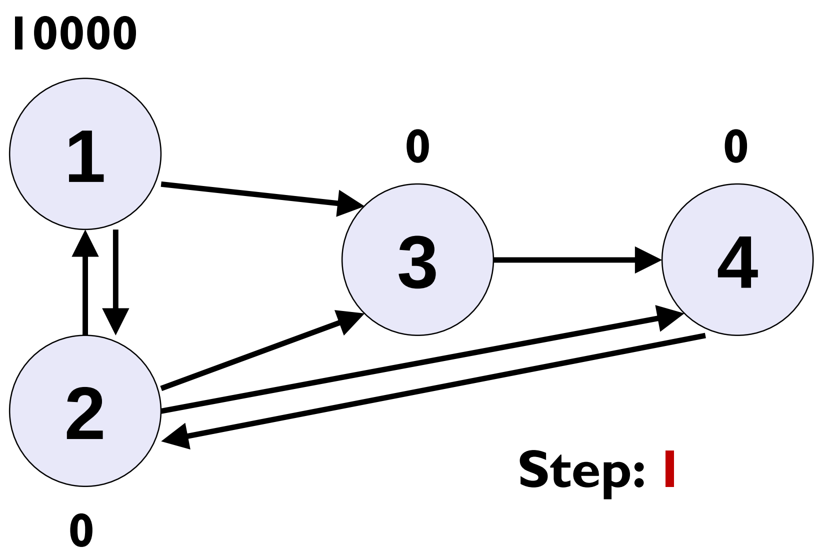
The PageRank of a page p is the fraction of steps the surfer spends at p in the limit. (A score between 0 and 1)

Notice: Using link structure only!

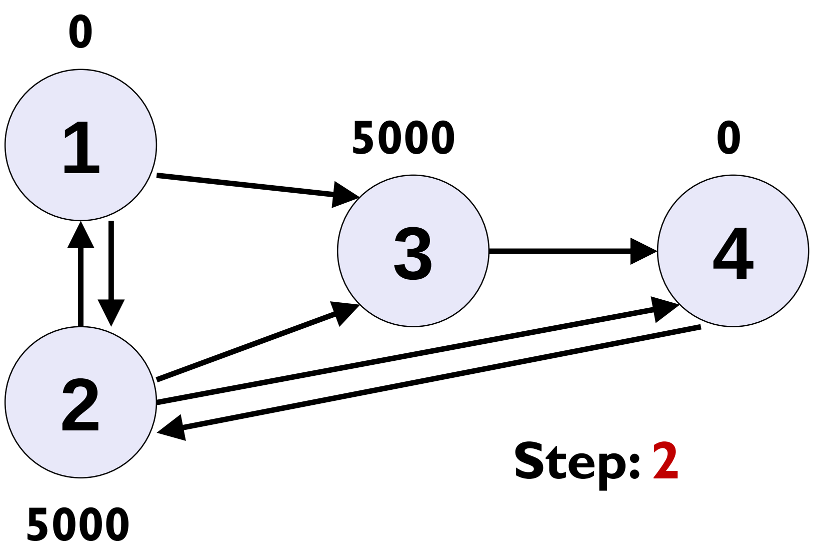
The Random Web Surfer



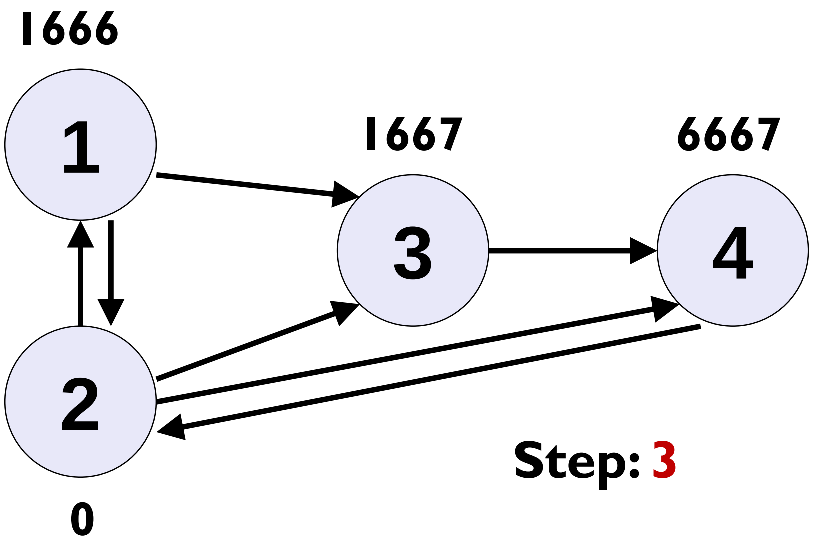
10.000 Random Surfers



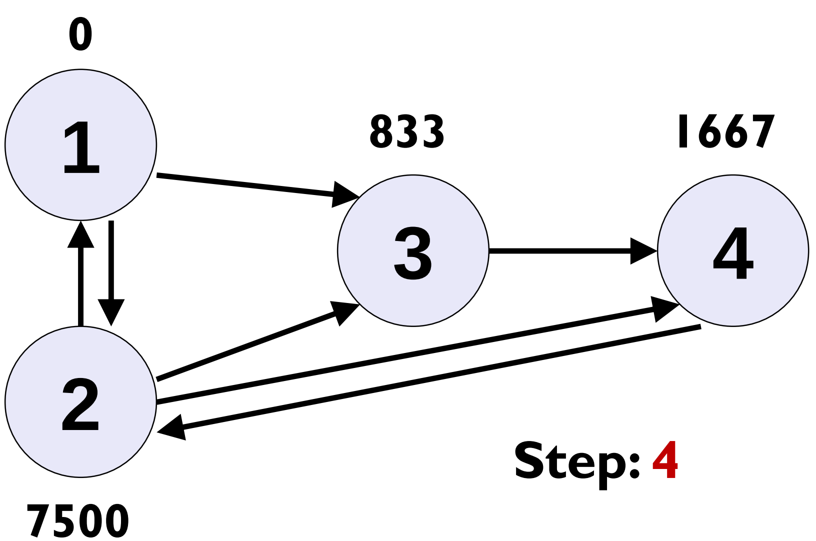
10.000 Random Surfers



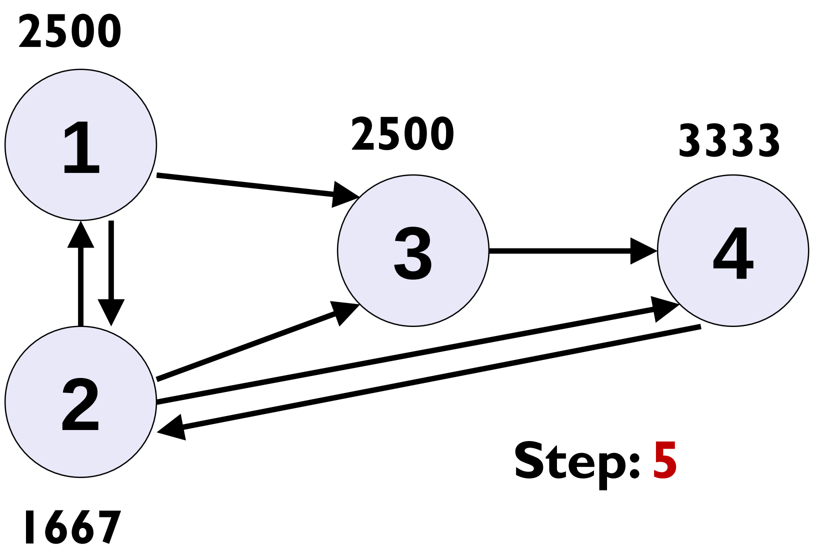
10.000 Random Surfers



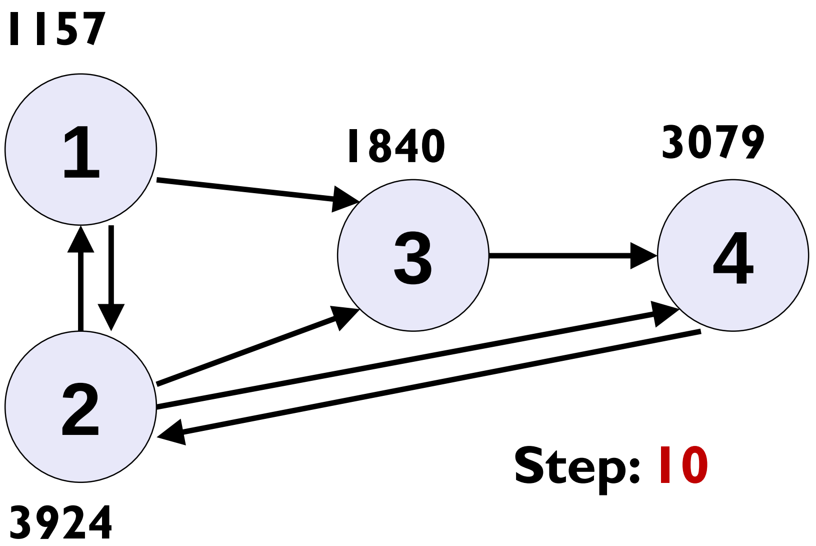
10.000 Random Surfers



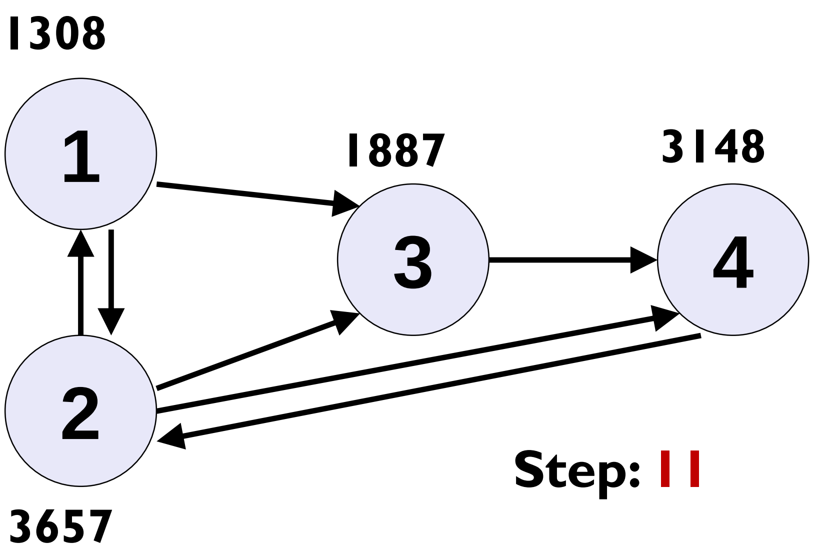
10.000 Random Surfers



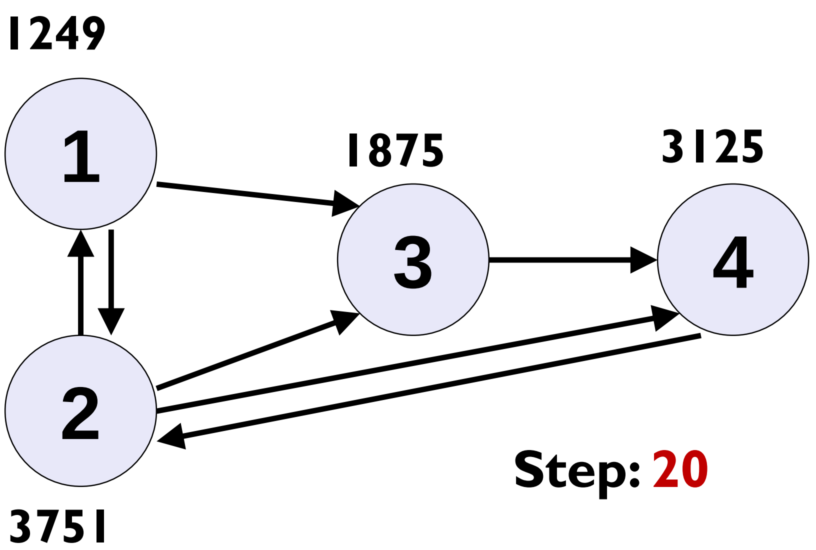
10.000 Random Surfers



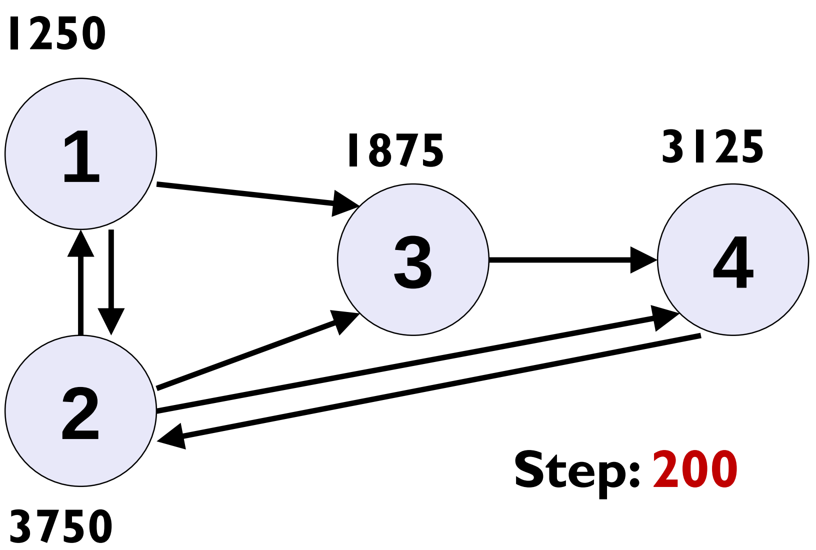
10.000 Random Surfers



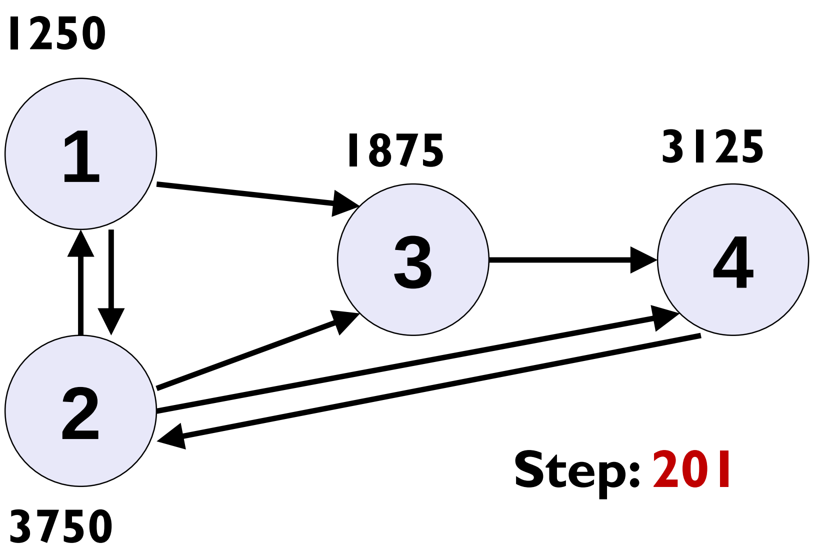
10.000 Random Surfers



10.000 Random Surfers



10.000 Random Surfers

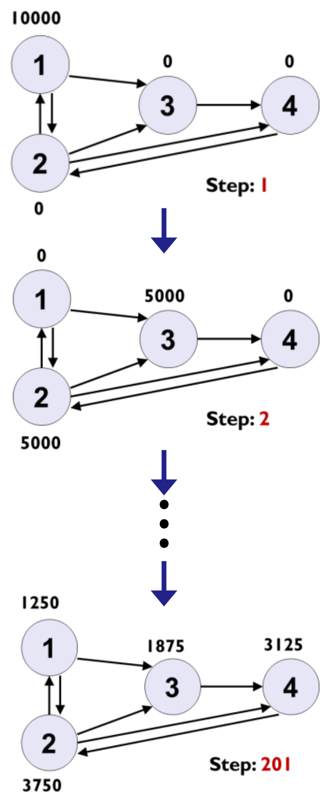


Surfers visit some nodes more often than others.

These are those with many in-links from other frequently visited pages.

Pages visited more often has higher prestige (rank)

PageRank Intuition



Imagine users doing a random walk on web pages:

- Start on a **random** page
- At each step, continue to next page along one of the links on the current page, **equiprobably**

PageRank idea: Run to convergence; the rank (prestige) of a web-page is then proportional to:

- the **proportion** of random web-surfers that will be visiting the page at a given point in time
- = the **probability** that a random web-surfer is at this page at any point in time

$$(P(1), P(2), P(3), P(4)) \;=\; (0.1250, 0.3750, 0.1875, 0.3125)$$

PageRank – The Markov Chain Model
(Markov chains are abstractions of random walks)

The random surfer is described by

- An initial (step 0) probability distribution over all (n) web-pages

$$\boldsymbol{q}^{(0)} = \left(q_1^{(0)}, \dots, q_n^{(0)}\right)$$

- A transition probability matrix

$$\boldsymbol{P} = \begin{pmatrix} P_{1,1} & \cdots & P_{1,j} & \cdots & P_{1,n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ P_{i,1} & \cdots & P_{i,j} & \cdots & P_{i,n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ P_{n,1} & \cdots & P_{n,j} & \cdots & P_{n,n} \end{pmatrix}$$

Worth noticing:

- $P_{i,j} = 0$, if no link from i to j
  Happens many, many times!!!
- Each row sums to 1

$P_{i,j}$ = probability of moving from page $\boldsymbol{i}$ to page $\boldsymbol{j}$

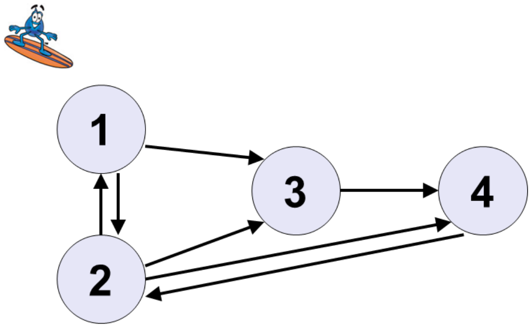
$$= \frac{1}{out\!-\!degree(i)}$$

Example (from before)

Initial probability distribution

(Here, surfer starts on page 1, but could have started anywhere!)

$$\boldsymbol{q}^{(0)} = \begin{pmatrix} 1 & 0 & 0 & 0 \end{pmatrix}$$



Transition probability matrix

$$\boldsymbol{P} = \begin{pmatrix} 0 & 1/2 & 1/2 & 0 \\ 1/3 & 0 & 1/3 & 1/3 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

Markov Chain Transitions

$P_{i,j} = 0$ when $i \not\to j$

The way we compute transitions:

$$q_j^{(t)} = \sum_{i=1}^{n} q_i^{(t-1)} P_{i,j} = \sum_{i:i \to j}^{n} q_i^{(t-1)} P_{i,j}$$

computed for all pages $j \in \{1, \dots n\}$

Matrix notation, the way we talk about it:

$$\boldsymbol{q}^{(t)} = \boldsymbol{q}^{(t-1)} \boldsymbol{P}$$

$$\left(q_1^{(t)} \dots q_j^{(t)} \dots q_n^{(t)}\right) = \left(q_1^{(t-1)} \dots q_j^{(t-1)} \dots q_n^{(t-1)}\right) \begin{pmatrix} P_{1,1} & \cdots & P_{1,j} & \cdots & P_{1,n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ P_{i,1} & \cdots & P_{i,j} & \cdots & P_{i,n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ P_{n,1} & \cdots & P_{n,j} & \cdots & P_{n,n} \end{pmatrix}$$

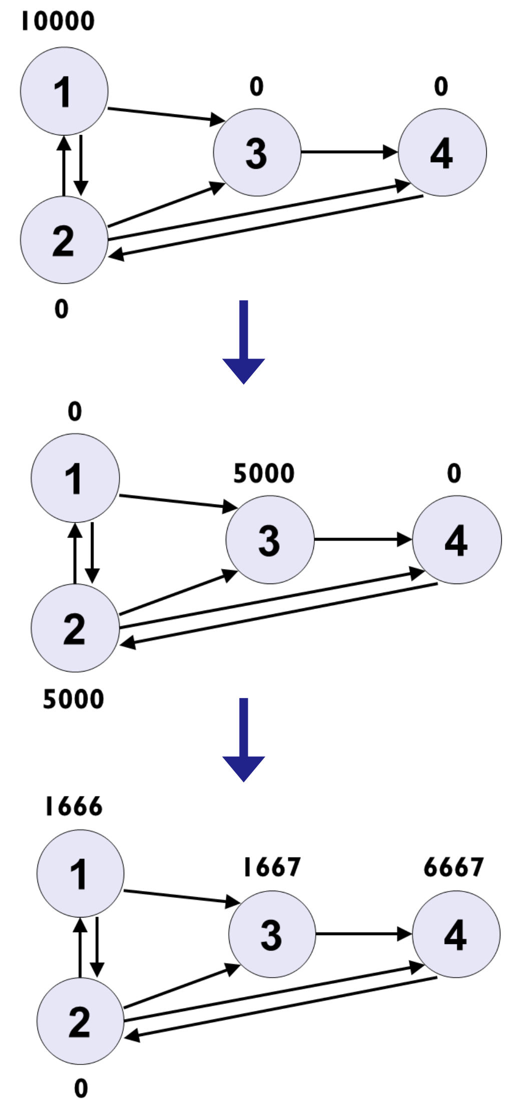
Example (cont.)

$$q^{(1)} = q^{(0)} P$$

$$\begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{3} & 0 & \frac{1}{3} & \frac{1}{3} \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

$$q^{(2)} = q^{(1)} P$$

$$\begin{pmatrix} \frac{1}{6} & 0 & \frac{1}{6} & \frac{4}{6} \end{pmatrix} = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix} \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{3} & 0 & \frac{1}{3} & \frac{1}{3} \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$



…and for the Bayesian network folks!

PageRank is just a (first order) Markov chain represented by a Bayesian network as follows

$$q^0 \xrightarrow{P} q^1 \xrightarrow{P} q^2 \xrightarrow{P} q^3 \xrightarrow{P} \dots$$

PageRank – Stationary Distribution

$\boldsymbol{q}$ is stationary, if

$$\boldsymbol{q} = \boldsymbol{q}\boldsymbol{P}$$

Example

$$\begin{pmatrix} \frac{1}{8} & \frac{3}{8} & \frac{3}{16} & \frac{5}{16} \end{pmatrix} = \begin{pmatrix} \frac{1}{8} & \frac{3}{8} & \frac{3}{16} & \frac{5}{16} \end{pmatrix} \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{3} & 0 & \frac{1}{3} & \frac{1}{3} \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

Under *some conditions* (more detail on next slide)

- For any $\boldsymbol{q}^{(0)}$
- a Markov chain has a *unique* stationary distribution $\boldsymbol{q}^*$

$$\lim_{t \to \infty} \boldsymbol{q}^{(t)} = \boldsymbol{q}^*$$

"Eigen-Stuff"

$\boldsymbol{q}$ with $\boldsymbol{q} = \boldsymbol{q}\boldsymbol{P}$ is also called an Eigen-vector of $\boldsymbol{P}$ with Eigen-value 1.
In fact, $\boldsymbol{q}$ is the principal Eigen-vector, because $\boldsymbol{P}$ is a transision probalility matrix.

Many ways to find the principal Eigen-vector in practice:

- SVD
- Power iteration – PageRank algorithm

Stationarity conditions

The PageRank iterations $\boldsymbol{q}^{(t)} = \boldsymbol{q}^{(t-1)}\boldsymbol{P}$ must be representable by an **ergodic** Markov chain.

A Markov chain is ergodic if it is

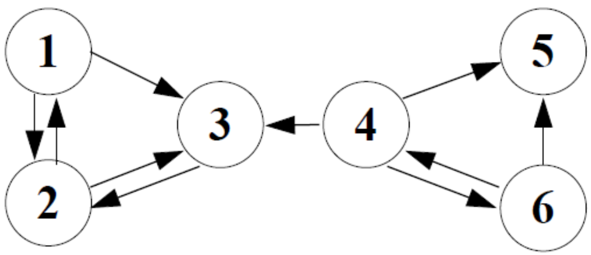
- *Irreducible*: "every node is reachable from every other node"
- *Aperiodic*: "nodes are not partitioned into sets such that all transitions occur cyclically from one set to another

Sufficient condition:

- A Markov chain is ergodic, if there is a strictly positive probability to pass from any state to any other state in one step.

We will be using this one

Complication: Dangling Pages



Ups! – A dangling page
⇒ ~~irreducible~~

"Transition matrix":

$$P = \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/3 & 0 & 1/3 & 1/3 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 1/2 & 0 \end{pmatrix}$$

Problem: Not a proper transition matrix, because dangling pages (Ex.: page **5**) have no defined transitions

Q: What do we do now – our theory breaks down!

Teleporting

At a dangling node, jump to a random web page.

At any non-dangling node, with probability α (e.g.,10%), jump to a random web page.

- With remaining probability (1-α) (e.g., 90%), go out on a random link.
- α - a parameter. Should mirror our belief that a random web surfer would leave a page in another way than following a link

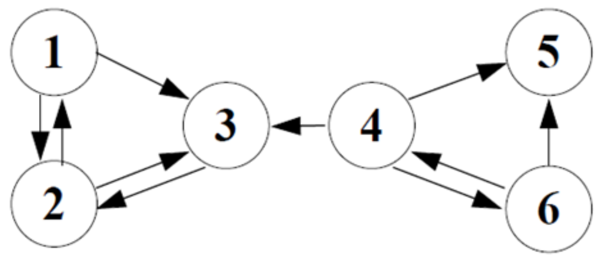
Result of teleporting:

- Now we cannot get stuck locally.
- there is a strictly positive probability to pass from any state to any other state in one step $\Rightarrow$ guaranteed ergodicity
- The power iteration converges to a unique $\boldsymbol{p}^*$ $\leftarrow$PageRank

Teleporting …in math

New transition matrix: $P_{PageRank} = (1-\alpha)P + \alpha U$

Example:



$$P_{pageRank} = (1-\alpha)\begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/3 & 0 & 1/3 & 1/3 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 1/2 & 0 \end{pmatrix} + \alpha \begin{pmatrix} 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \end{pmatrix}$$

Definition of PageRank

(Brin and Page '98)

Consider the following infinite random walk (surf):

- Initially the surfer is at a random page
- At each step, the surfer proceeds
  - to a randomly chosen web page with probability α
  - to a randomly chosen successor of the current page with probability $1 - \alpha$

The PageRank of a page p is the fraction of steps the surfer spends at p in the limit. (A score between 0 and 1)

Notice: Using link structure only!

PageRank -- Summary

- The page rank of webpage *i* is

$$n \cdot q_i^*$$

where q_i^* is the (unique) limit distribution of the Markov chain defined by $\boldsymbol{P}_{PageRank}$

- We use Power Iterations to approximate the PageRank by iterating

$$\boldsymbol{q}^{(t)} = \boldsymbol{q}^{(t-1)}\boldsymbol{P}_{PageRank}$$

until $\boldsymbol{q}^{(t)}$ does not change very much

- Used as one of the ranking criteria in Google, Bing,…

Scalability issues

Large matrix – 50bil x 50 bil

With teleport fix dense

Hard to compute in main memory

Fixes:

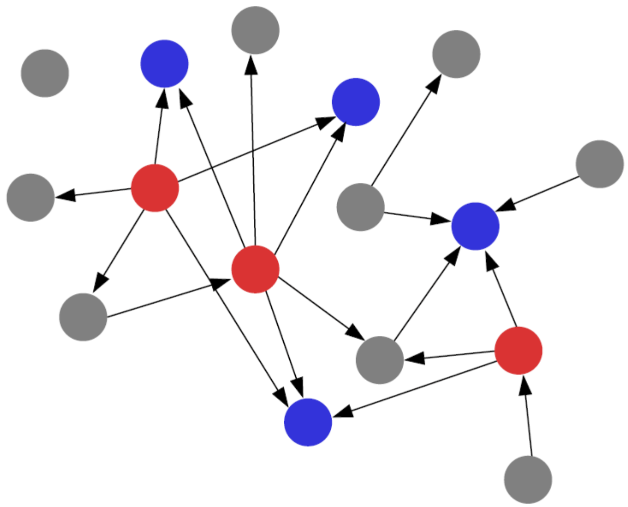
- Compression
- I/O operations and swaps
- Splitting computation on dangling and non dangling parts
- Looking just at some values from each column/raw
- Only computing few (say 10?) iterations to get order right but not necessarily the rank values
- …

More in:

Amy N. Langville and Carl D. Meyer: Deeper Inside PageRank. Internet Mathematics Vol. 1, No. 3: 335-380. http://www.internetmathematicsjournal.com/article/1388-deeper-inside-pagerank

HITS
(Hyperlink-Induced Topic Search)

Authorities and Hubs



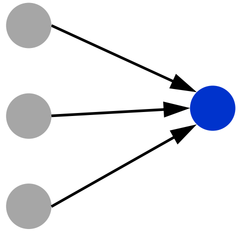
Authorities: Web pages linked to by many other pages.
Example: important company homepages

Hubs: Web pages pointing to many (relevant) pages.
Example: Business listings (yellow pages)

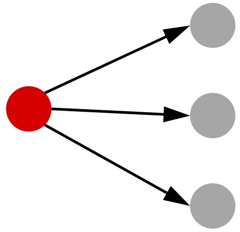
HITS – The Goal

Given a query, find

- Good sources of content (authorities)

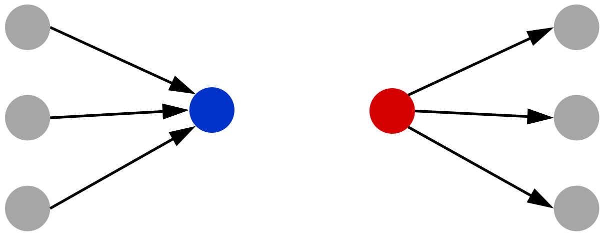


- Good sources of links (hubs)

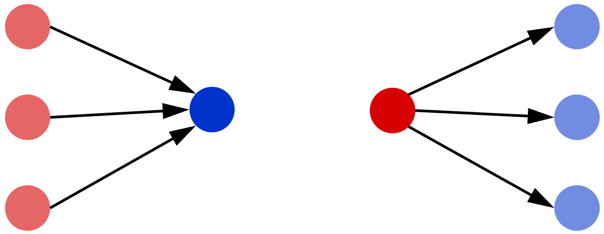


Intuition

Authority **comes from** in-edges.
Being a good hub **comes from** out-edges.



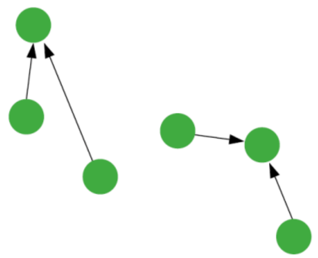
Better Authority **comes from** in-edges from good hubs.
Being a better hub **comes from** out-edges to good authorities.



HITS Algorithm

Step 1 (of 3)

Retrieve top *t* webpages for query (mostly content-based):

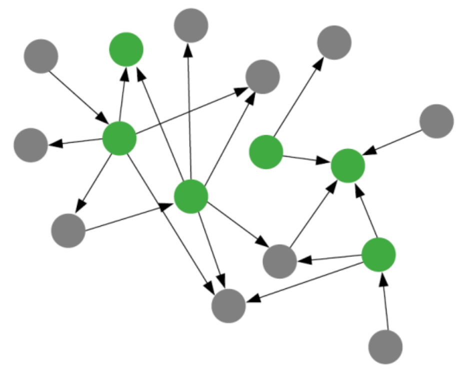


Result: the **root** set

HITS Algorithm (cont.)

Step 2 (of 3)

Add all neighbors of the pages in the root set:



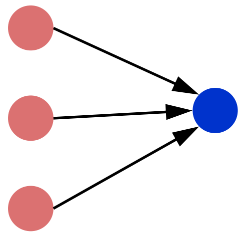
Result: the **base** set

HITS Algorithm (cont.)

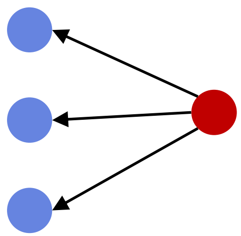
Step 3 (of 3)

Iteratively update (normalized) **authority** and **hub** **scores** for all pages in the base set – *until convergence*:

$$a(j) = \sum_{i:i \to j} h(i)$$



$$h(j) = \sum_{i:j \to i} a(i)$$



Notice: Each page both have an authority and a hub score

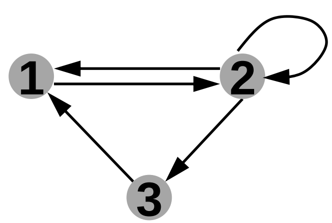
Math…

$n \times n$ <u>link-matrix</u> $\boldsymbol{L}$:

- each of the n pages in the base set has a row and column in the matrix.
- Entry $L_{ij} = 1$ if page i links to page j, else = 0.

$n \times n$ <u>*transposed* link-matrix</u> $\boldsymbol{L}^T$:

- Entry $L^T_{ij} = 1$ if page i is linked to from page j, else = 0.



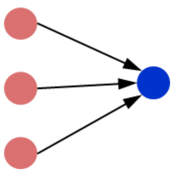
$$\begin{array}{ccc} & & j \\ & \boldsymbol{L} & \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ i & \begin{array}{c} 1 \\ 2 \\ 3 \end{array} & \begin{bmatrix} 0 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & 0 & 0 \end{bmatrix} \end{array} \qquad \begin{array}{ccc} & & j \\ & \boldsymbol{L}^T & \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ i & \begin{array}{c} 1 \\ 2 \\ 3 \end{array} & \begin{bmatrix} 0 & 1 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 0 \end{bmatrix} \end{array}$$

HITS Algorithm (cont.)

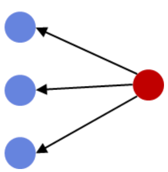
Step 3 (of 3) – the way we compute it

Iteratively update (normalized) **authority** and **hub scores** for all pages in the base set – *until convergence*:

$$a(j) = \sum_{i:i \rightarrow j} h(i)$$



$$h(j) = \sum_{i:j \rightarrow i} a(i)$$



Matrix notation – the way we talk about it

$$\boldsymbol{a} = \boldsymbol{L}^T \boldsymbol{h}$$

$$\boldsymbol{h} = \boldsymbol{L}\boldsymbol{a}$$

Eigen-stuff

HITS - Step 3 (of 3)

$$\boldsymbol{a} = \boldsymbol{L}^T \boldsymbol{h}$$

$$\boldsymbol{h} = \boldsymbol{L}\,\boldsymbol{a}$$

Substituting $\Rightarrow$

$$\boldsymbol{a} = \boldsymbol{L}^T \boldsymbol{L}\,\boldsymbol{a}$$

$$\boldsymbol{h} = \boldsymbol{L}\boldsymbol{L}^T \boldsymbol{h}$$

What does this look like?

- $\boldsymbol{a}$ is an eigenvector of $\boldsymbol{L}^T \boldsymbol{L}$
- $\boldsymbol{h}$ is an eigenvector of $\boldsymbol{L}\boldsymbol{L}^T$

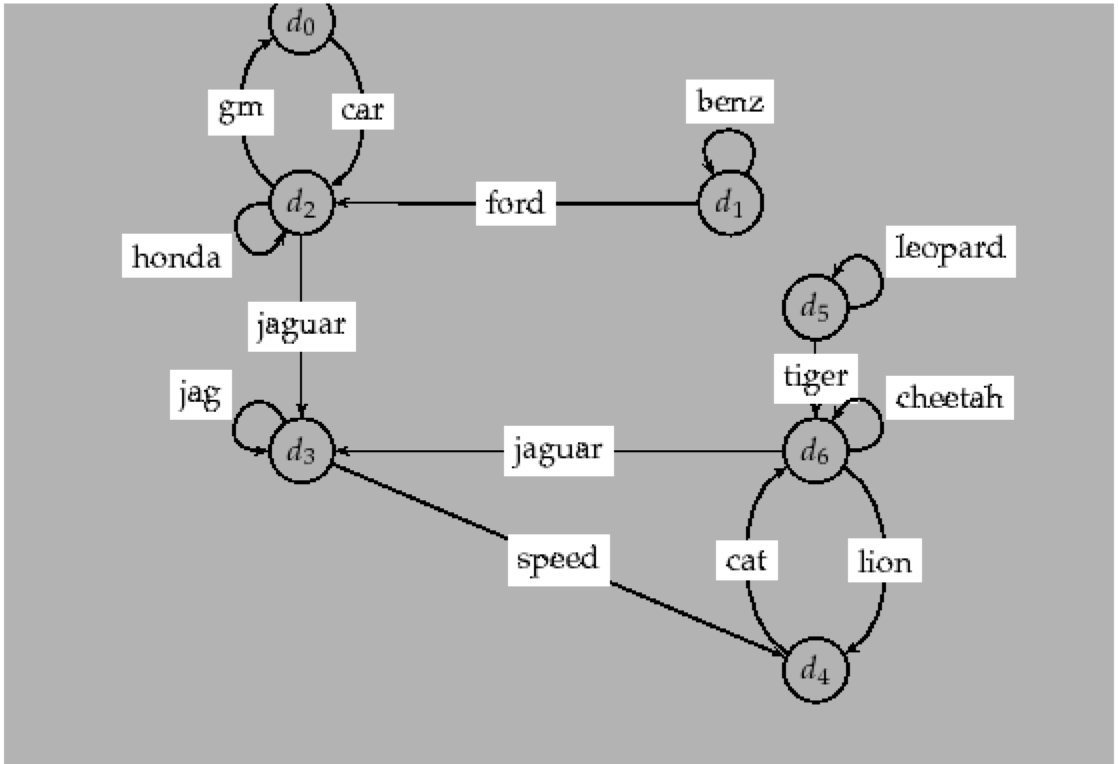
The HITS algorithm is a particular, known algorithm for computing eigenvectors: the *power iteration* method.

Guaranteed to converge

HITS -- Recommendations

- 200 or so pages in the root set is sufficient
- 5 iterations are sufficient to get good results
- Matrix operations are not necessary, just additive updates as the web graph is quite sparse (average 10 links per page)
- Experiments has shown that HITS works to certain extent multi-lingualy

PageRank vs. HITS -- Example



► **Figure 21.1** A small web graph. Arcs are annotated with the word that occurs in the anchor text of the corresponding link.

http://nlp.stanford.edu/IR-book/html/htmledition/the-pagerank-computation-1.html

PageRank vs. HITS – Example (cont.)

$\boldsymbol{P}_{PageRank}$ from PageRank:

0.02	0.02	0.88	0.02	0.02	0.02	0.02
0.02	0.45	0.45	0.02	0.02	0.02	0.02
0.31	0.02	0.31	0.31	0.02	0.02	0.02
0.02	0.02	0.02	0.45	0.45	0.02	0.02
0.02	0.02	0.02	0.02	0.02	0.02	0.88
0.02	0.02	0.02	0.02	0.02	0.45	0.45
0.02	0.02	0.02	0.31	0.31	0.02	0.31

$\boldsymbol{L}$ from HITS:

0	0	1	0	0	0	0
0	1	1	0	0	0	0
1	0	1	2	0	0	0
0	0	0	1	1	0	0
0	0	0	0	0	0	1
0	0	0	0	0	1	1
0	0	0	2	1	0	1

PageRank scores:

$$\vec{x} = (0.05 \quad 0.04 \quad 0.11 \quad 0.25 \quad 0.21 \quad 0.04 \quad 0.31)$$

Hub & Authority scores:

$$\vec{h} = (0.03 \quad 0.04 \quad 0.33 \quad 0.18 \quad 0.04 \quad 0.04 \quad 0.35)$$

$$\vec{a} = (0.10 \quad 0.01 \quad 0.12 \quad 0.47 \quad 0.16 \quad 0.01 \quad 0.13)$$

http://nlp.stanford.edu/IR-book/html/htmledition/hubs-and-authorities-1.html
http://nlp.stanford.edu/IR-book/html/htmledition/the-pagerank-computation-1.html

PageRank vs. HITS

PageRank	HITS
Computation: • Expensive • Once for all documents and queries (offline)	Computation: • Expensive • Requires computation for each query (online)
Query-independent • requires combination with query-dependent criteria	Query-dependent
Hard to spam	Relatively easy to spam • partly local
	Quality depends on quality of start set Gives hubs as well as authorities

Outline

Link-based techniques for ranking

- Motivation
- Anchor text
- Structural measures
  - PageRank (query-independent)
  - HITS (query-dependent)