

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

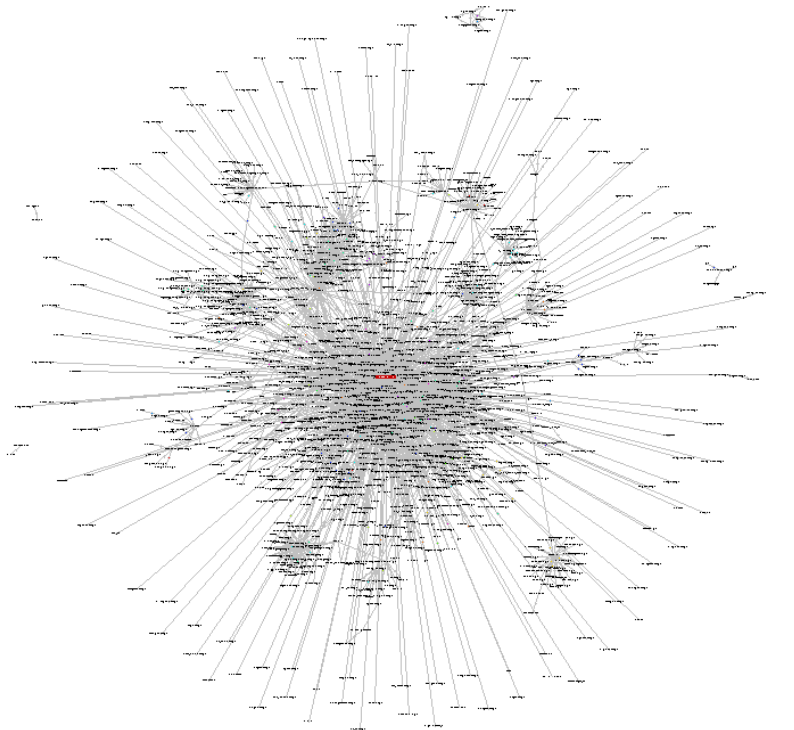
Links are everywhere

- 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



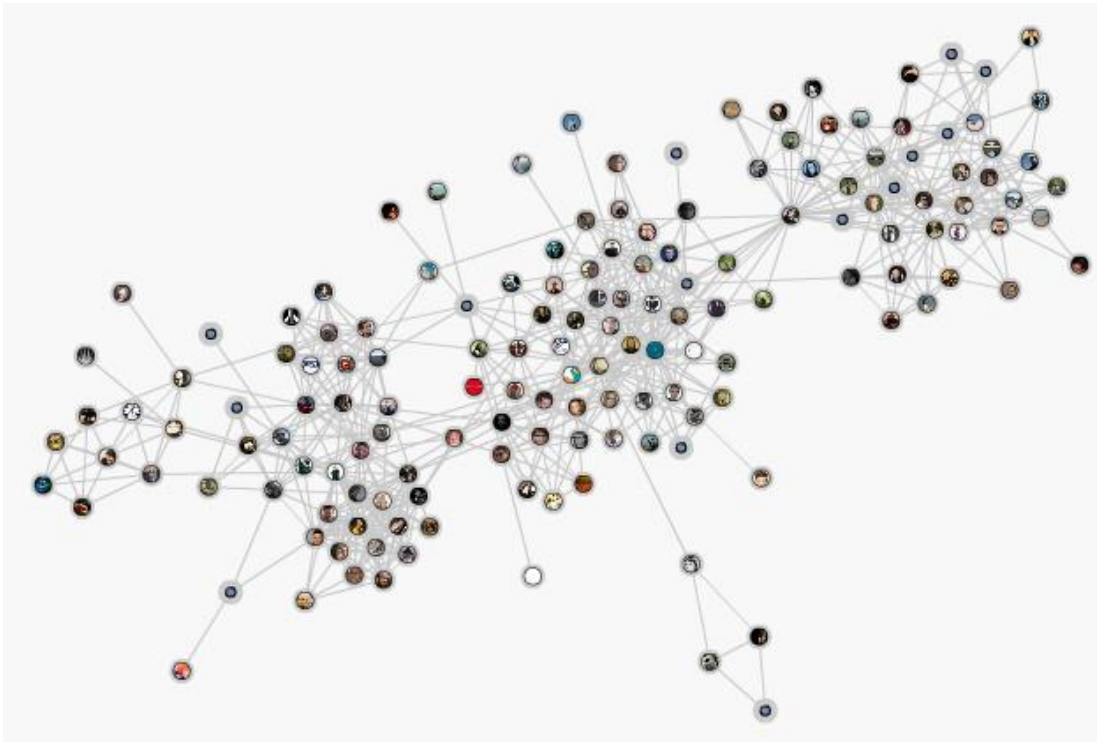
Nodes: ENRON employees

(Directed) link: “sent email to”

Web Structure – Social Network

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

Facebook – Social graph

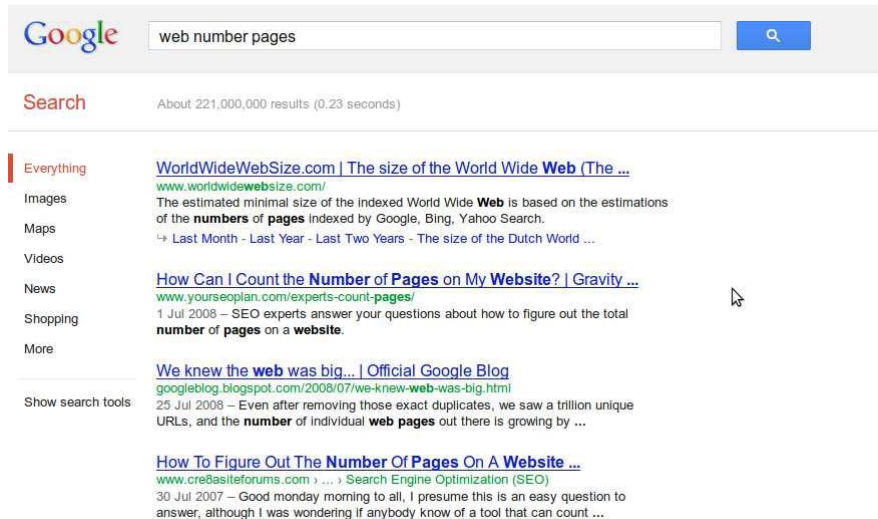


**Nodes: Facebook profile
& friends**

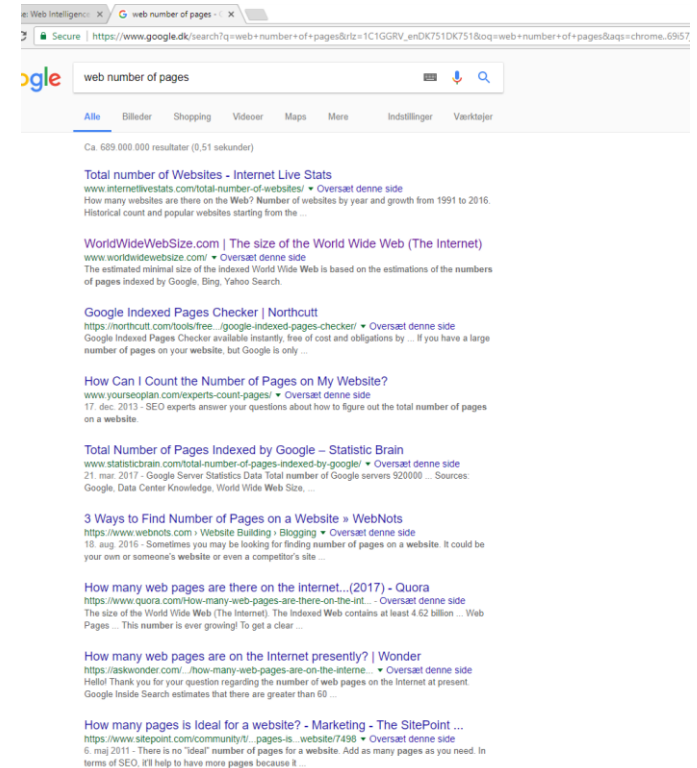
**(Directed) link: “friend
of”**

Ranking motivation

Google search for “web number pages”:
2015



2017

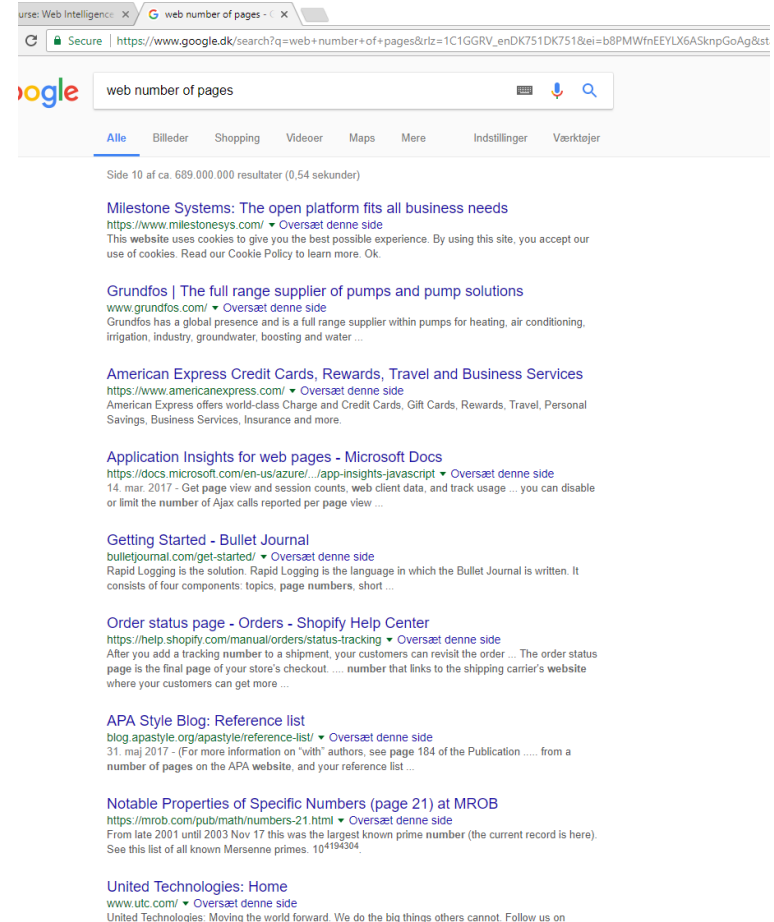


- 221.000.000 hits in 2015, 689.000.000 in 2017
- Top 4 results contains useful and trustworthy information

Ranking Motivation cntd.

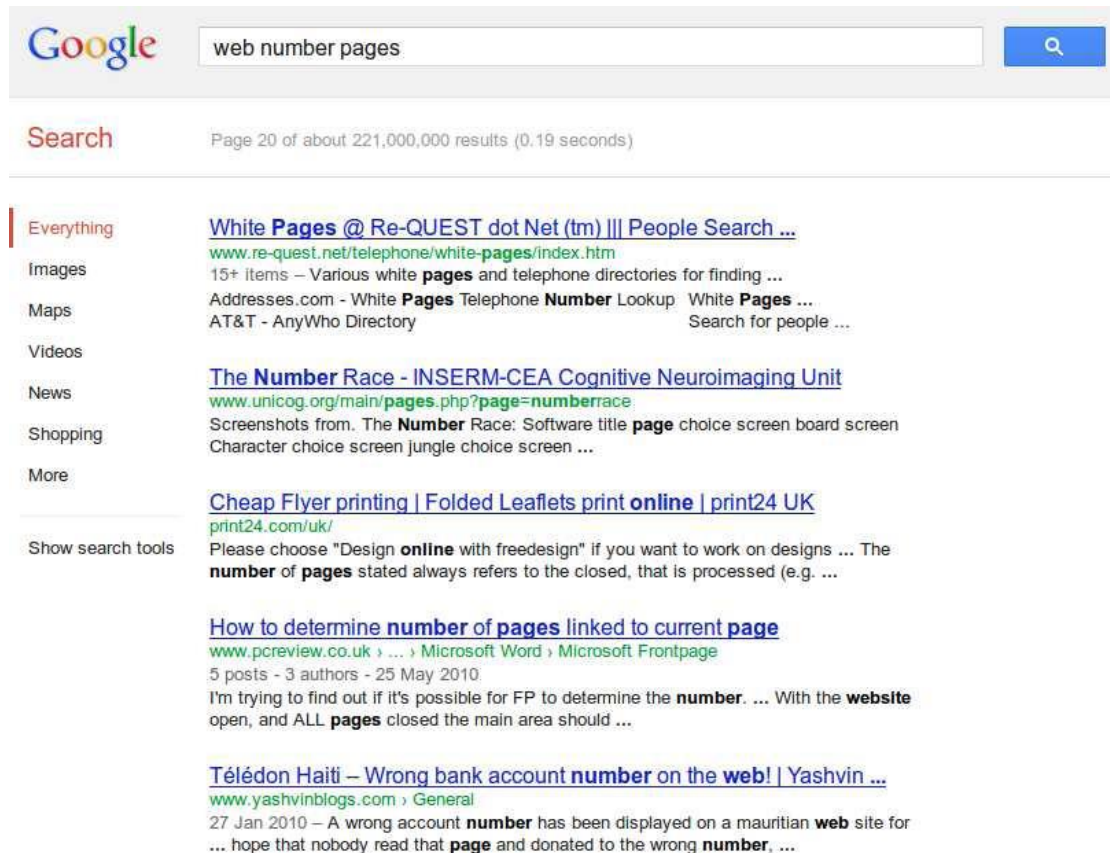
Look at 10th result page
(top 90-100)

Not so useful anymore



Ranking motivation (cont.)

And results 200-205 cnt. less useful:



The screenshot shows a Google search interface with the query 'web number pages'. The results are displayed on page 20 of approximately 221,000,000 results, with a search time of 0.19 seconds. The left sidebar contains navigation links for 'Everything', 'Images', 'Maps', 'Videos', 'News', 'Shopping', and 'More', along with a 'Show search tools' option. The main content area lists several search results, each with a title, URL, and a brief description. The results are as follows:

- White Pages @ Re-QUEST dot Net (tm) ||| People Search ...**
www.re-quest.net/telephone/white-pages/index.htm
 15+ items – Various white **pages** and telephone directories for finding ...
 Addresses.com - White **Pages** Telephone **Number** Lookup White **Pages** ...
 AT&T - AnyWho Directory Search for people ...
- The Number Race - INSERM-CEA Cognitive Neuroimaging Unit**
www.unicog.org/main/pages.php?page=numberrace
 Screenshots from. The **Number** Race: Software title **page** choice screen board screen
 Character choice screen jungle choice screen ...
- Cheap Flyer printing | Folded Leaflets print online | print24 UK**
print24.com/uk/
 Please choose "Design **online** with freedesign" if you want to work on designs ... The
number of pages stated always refers to the closed, that is processed (e.g. ...
- How to determine number of pages linked to current page**
www.pcreview.co.uk › ... › Microsoft Word › Microsoft Frontpage
 5 posts - 3 authors - 25 May 2010.
 I'm trying to find out if it's possible for FP to determine the **number**. ... With the **website**
 open, and ALL **pages** closed the main area should ...
- Télédon Haiti – Wrong bank account number on the web! | Yashvin ...**
www.yashvinblogs.com › General
 27 Jan 2010 – A wrong account **number** has been displayed on a mauritian **web** site for
 ... hope that nobody read that **page** and donated to the wrong **number**, ...

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

princess diana

Engine 1

Princess Diana Memorial WebRing
Follow the WebRing for a tour of memorial site
87% <http://www.geocities.com/RainForest/Vines/1009/diana1998>
[Grouped results from http://www.geocities.com/RainForest/Vines/1009/diana1998](http://www.geocities.com/RainForest/Vines/1009/diana1998)

FOR DIANA, PRINCESS OF HEART - Dr. K
84% <http://www.therelationshipcenter.com/diana.shtml> (Size 8.8K)

Princess Diana Editorial Cartoons! Cartoons!
The Professional Cartoonists Index is the most c
cartoonists on <http://www.cartoonist.com> is th
daily cartoons. **Relevant high quality**
82% <http://www.cartoonist.com> (Size 8.8K)

Diana, Princess of Wales
1 July 1961 - 31 August 1997 The BBC Web sit
Camera Press/Snowdon
79% <http://www.royal.gov.uk/start.htm> (Size 2.3K) Doc
[Grouped results from http://www.royal.gov.uk/start.htm](http://www.royal.gov.uk/start.htm)

Engine 2

1. Re: Lost in the shadow of Princess Diana
[URL: www.spiceisle.com/talkshop/messages/6232.htm]
The Spicelander TalkShop. [Follow Ups] [Post
The Spicelander TalkShop] Date: September
00:54:03 From: Sno,...
Last modified 12-Sep-97 - page size 4K - in English [Tran

2. Re: Princess Diana's gown auction
[URL: www.elle.com/textes/blablaba/forum/messages1/14]
Re: Princess Diana's gown auction. [Follow Ups
Followup] [Elle International - Blablaba] Posted
September 07, 1997 at 02:15:26...
Last modified 30-Mar-98 - page size 2K - in English [Tran

3. Re: Princess Diana
[URL: spicyhot.com/gaynet/messages/1053.html]
Re: Princess Diana [Follow Ups] [Post Followu
Maine Gayne [Follow Ups] [Post Followu
November 09, 1997 at 02:15:26...
Last modified 30-Mar-98 - page size 2K - in English [Tran

4. Re: Princess Diana - Queen of Hearts
[URL: www.elle.com/textes/blablaba/forum/messages1/28]
Re: Princess Diana - Queen of Hearts. [Follow U
Followup] [Elle International - Blablaba] Posted
on August 31, 1997 at...
Last modified 30-Mar-98 - page size 4K - in English [Tran

Engine 3

1. Free Passwords To Adult Sites...
99% - Articles & General info: Free Password
Sites warez princess diana demi moore
magazine kathy ireland lingerie jennifer aniston cook
warez princess diana demi moore... 03/09/98
Commercial site: <http://www.prirent.com/warez>

2. SEX CHAT XXX NUDE PORNO PLAYBOY P
[URL: www.connix.com/~wgonzo/ses/slidesuperal.htm]
99% - Articles & General info: SEX CHAT XXX
PORNO PLAYBOY PAMELA ANDERSON P
FECTRESS WOMEN ADULT MUSIC CHAT B
BROTICA JERRY MCCARTHY LINGERIE SA
CRISTY CRAWFORD MISS GILLS... 03/09/98
Personal page: <http://www.connix.com/~wgonzo/ses/slidesuperal.htm>

3. Rourke was...
99% - Articles & General info: Rourke was...
countdown by the group of women while Lin got it
with prom. Richard Lacey as Quade was getting pr
Personal page: <http://www.octet.com/~gonzo/jy>

4. Sunday, 18-Jan-98
99% - Articles & General info: Sunday, 18-Jan-
CHAT XXX NUDE PORNO PLAYBOY PAME

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”.

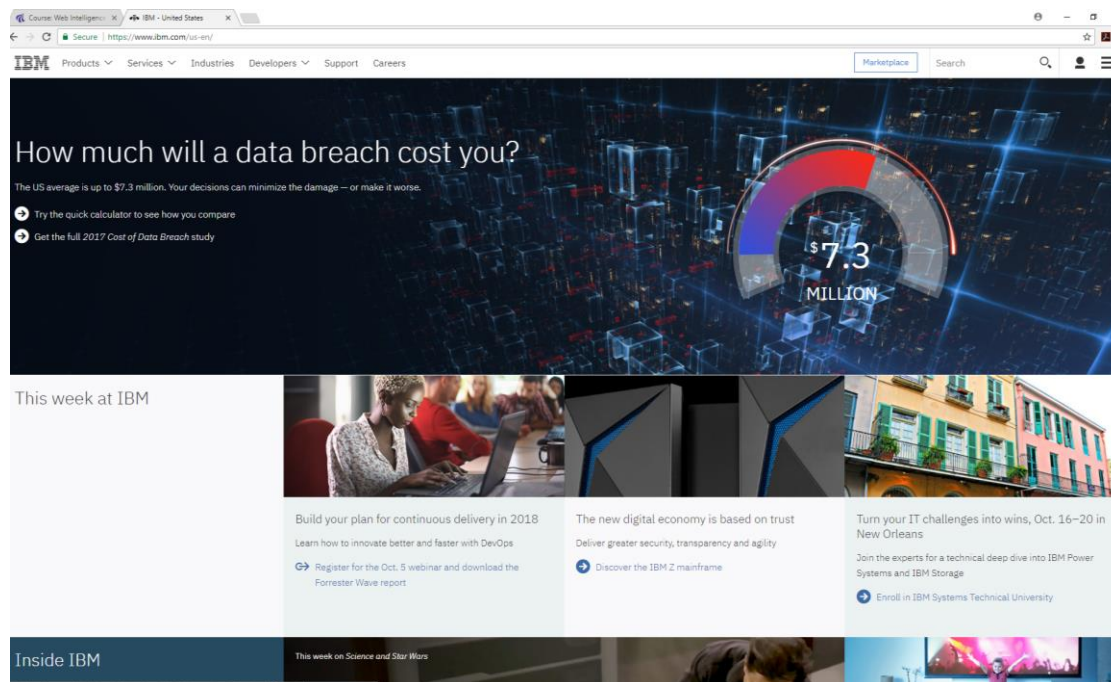
- E.g., ranking score $\in [0,1]$

IBM.com is first hit for search “IBM” – How?

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

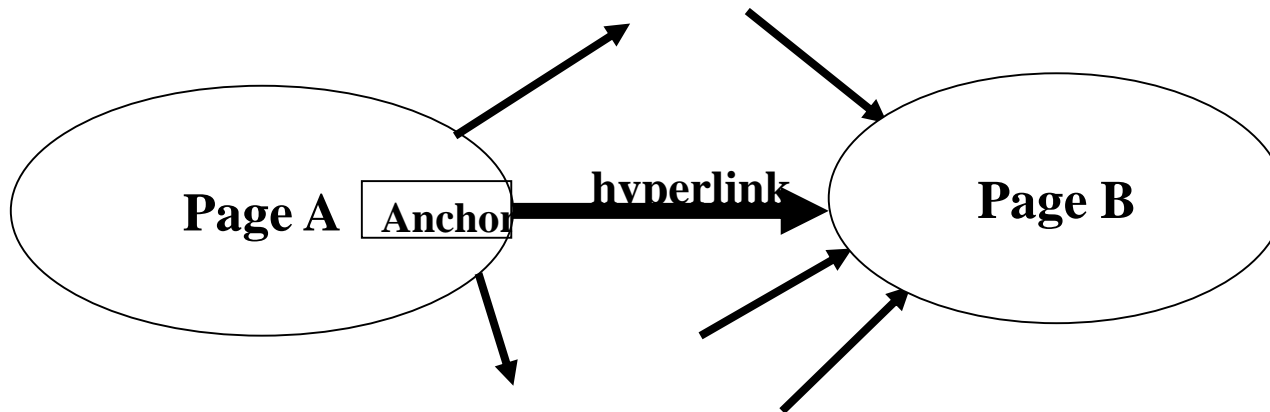
Why not:

- **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



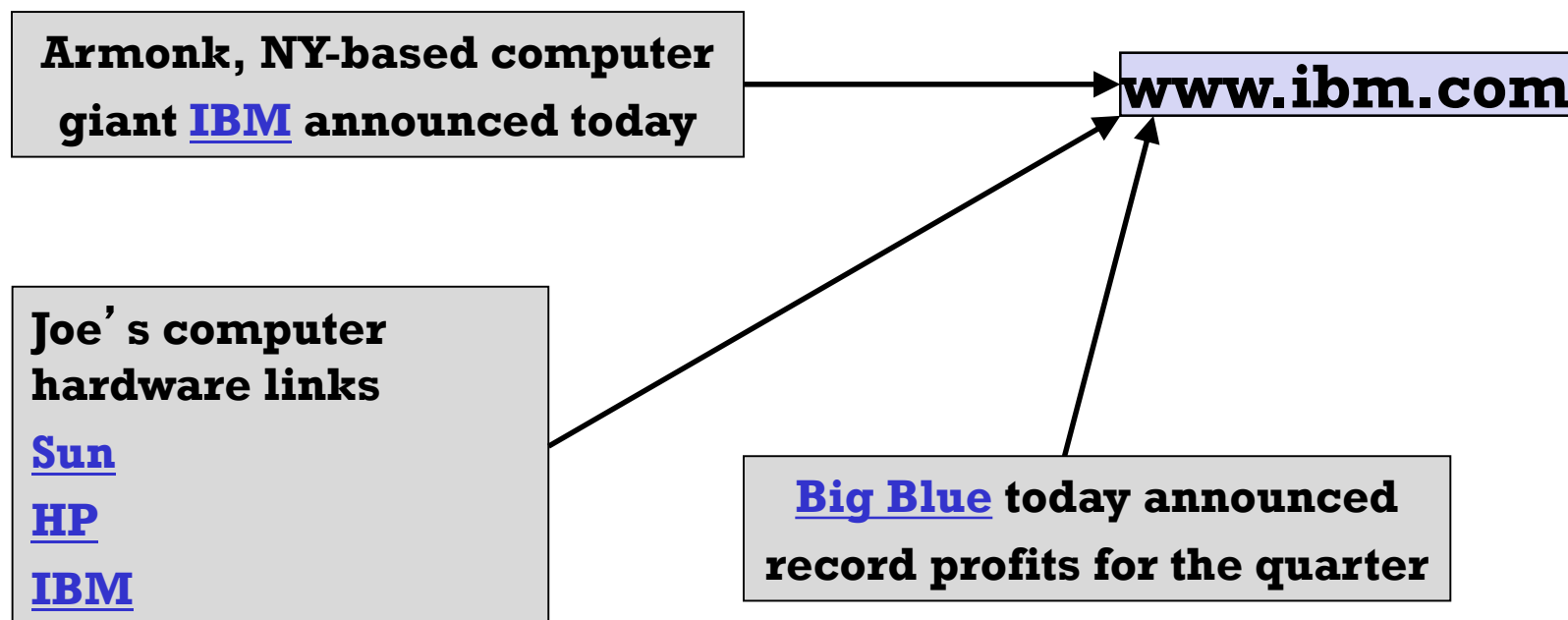
Hypothesis 1: A hyperlink between pages is a conferral of authority (quality signal)

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 .



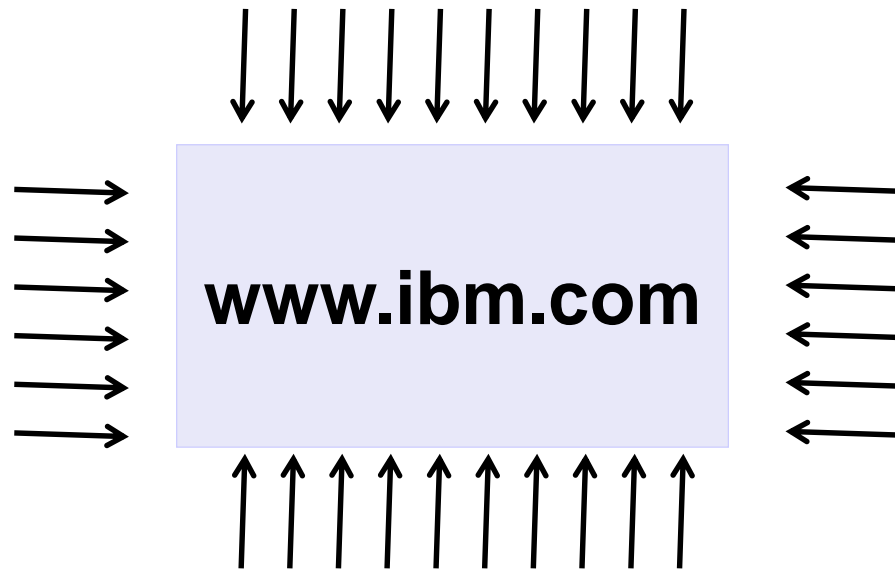
Use anchor text for indexing!

- 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!)

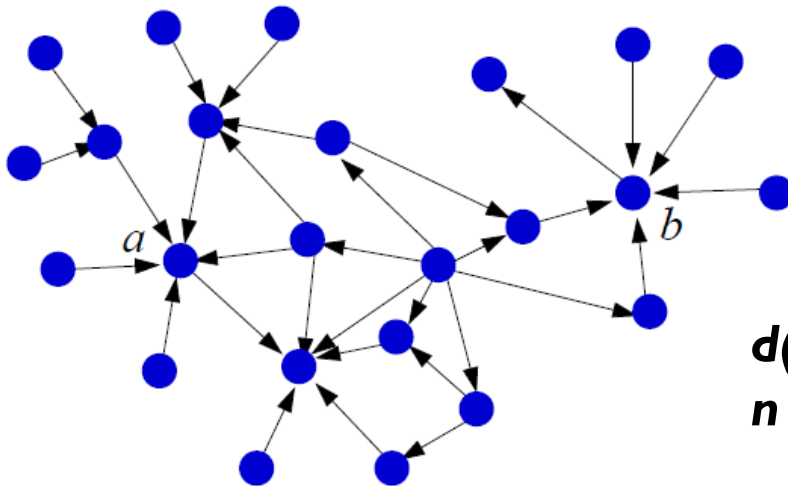
Hypothesis 1: A hyperlink between pages is a conferral of authority (quality signal)

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

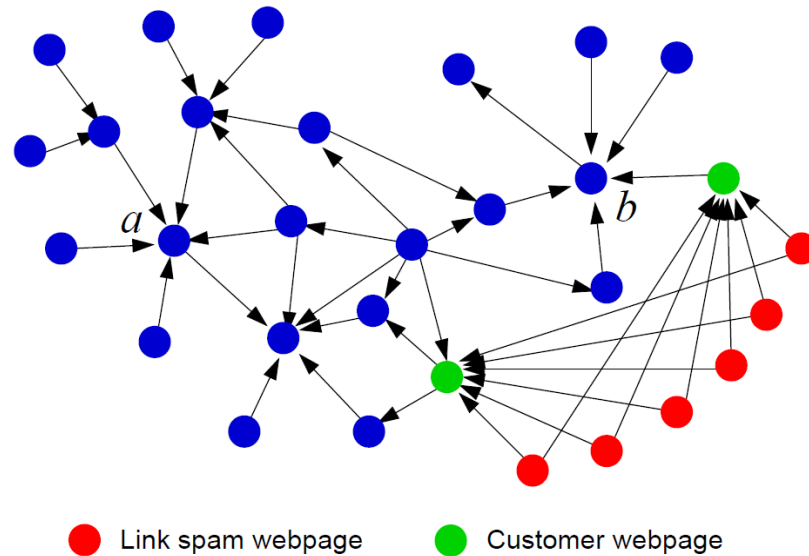
Prestige

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

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

Prestige (cont.)

Local measures can be easily manipulated by link spamming:

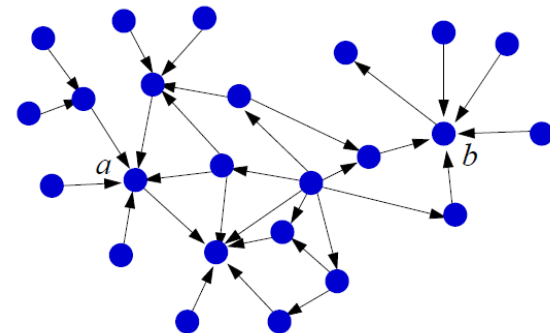


On the other hand, more difficult to manipulate global measures (though still possible)

A global prestige measure

Rank prestige:

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



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)

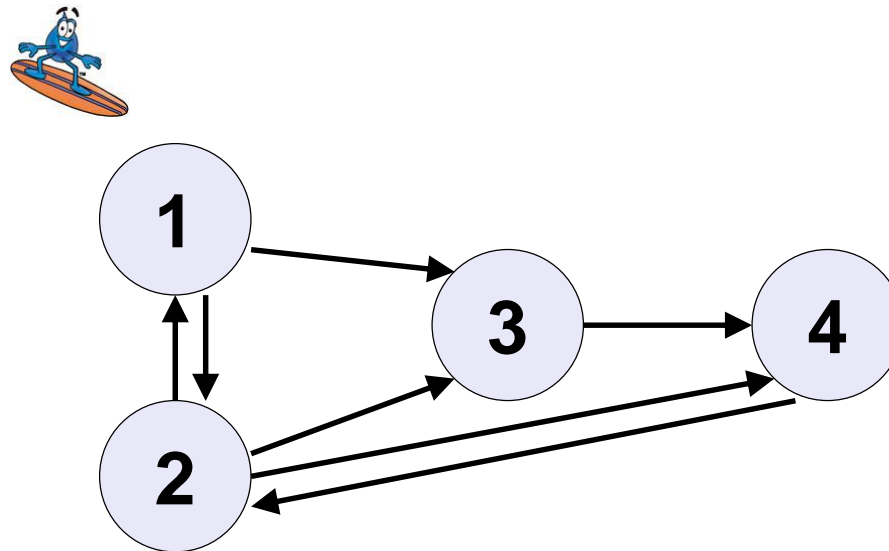
Consider the following infinite random walk (surf):

- 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 $1 - \alpha$

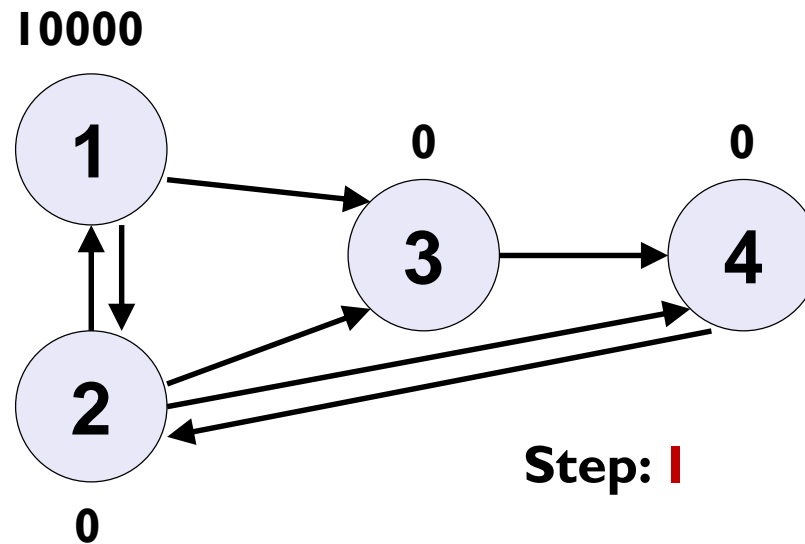
The PageRank of a page p is the fraction of steps the surfer spends at p in the limit. (A score between 0 and 1)

Notice: Using link structure only!

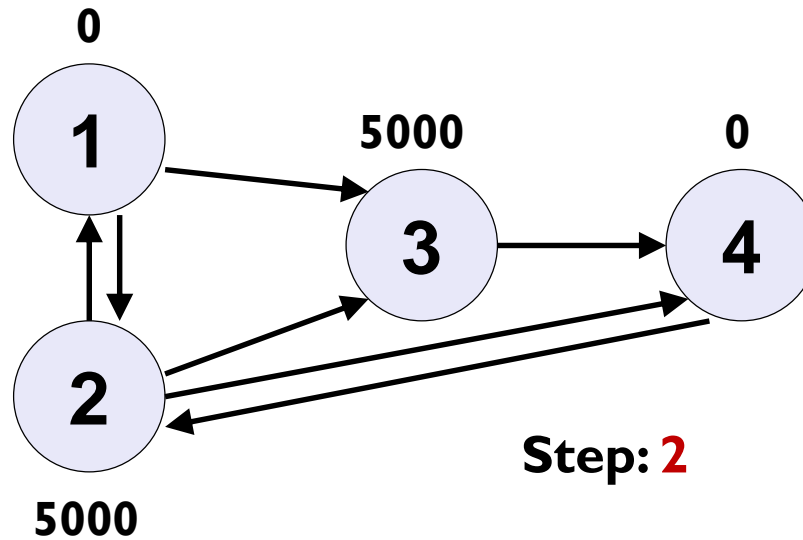
The Random Web Surfer



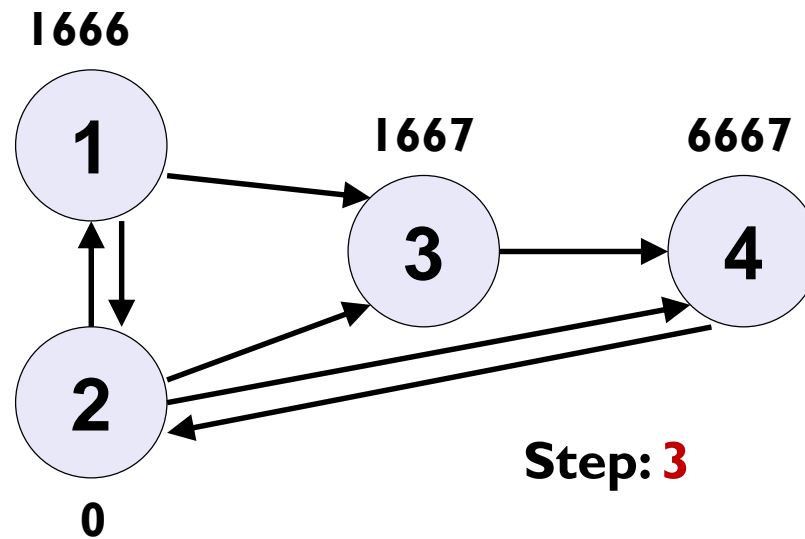
10.000 Random Surfers



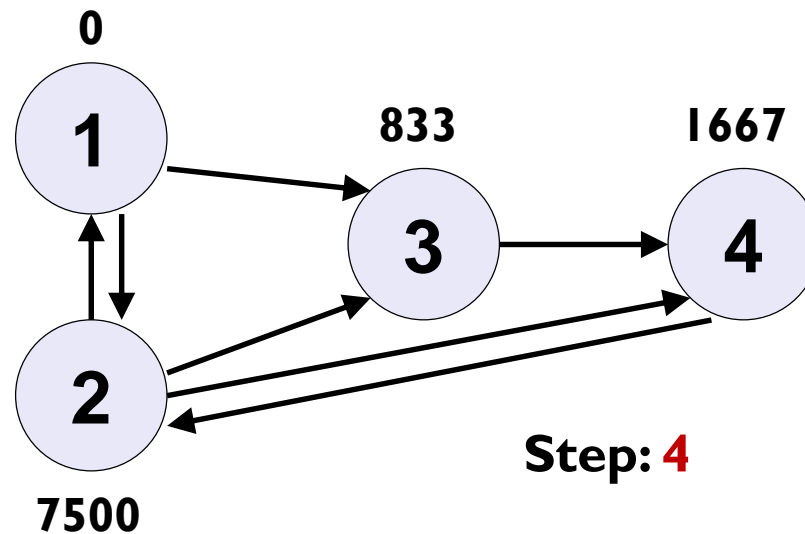
10.000 Random Surfers



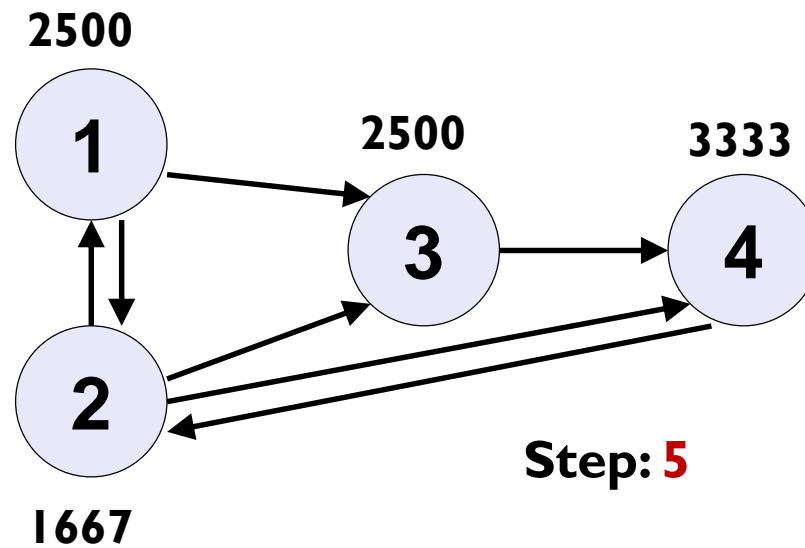
10.000 Random Surfers



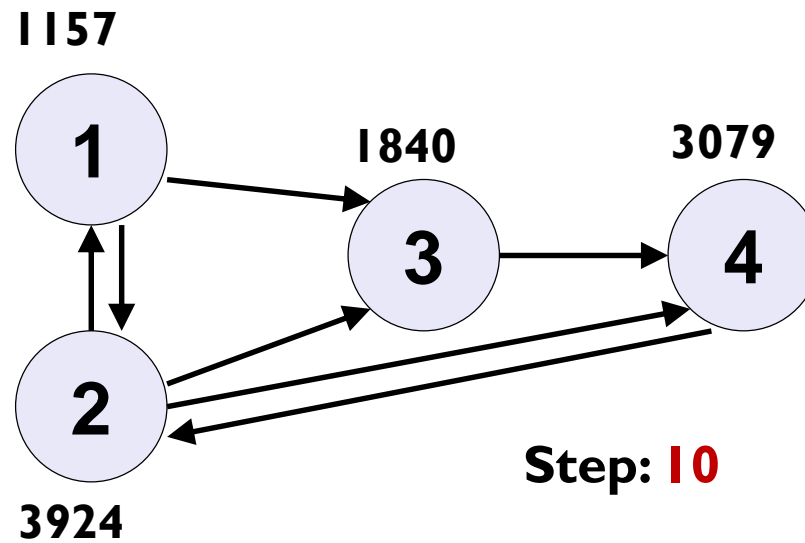
10.000 Random Surfers



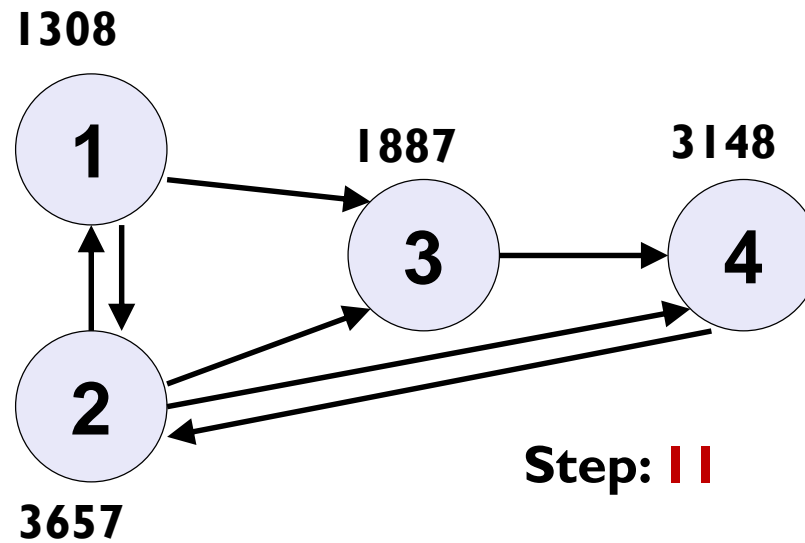
10.000 Random Surfers



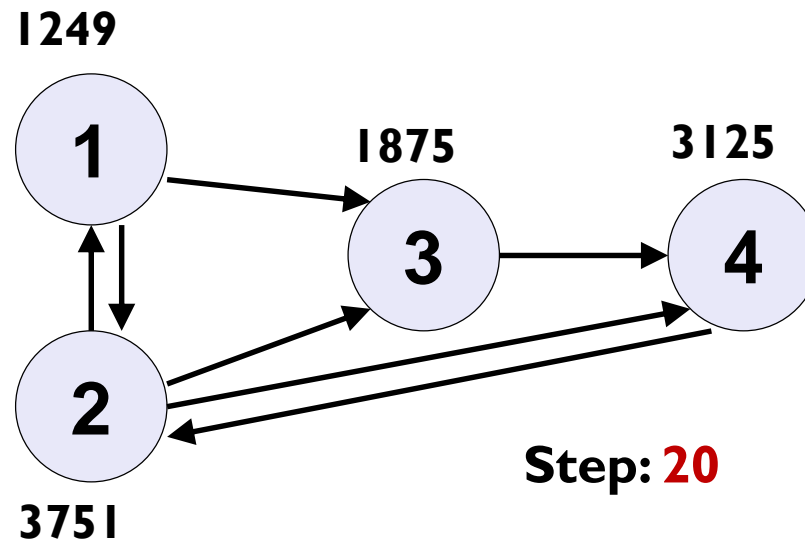
10.000 Random Surfers



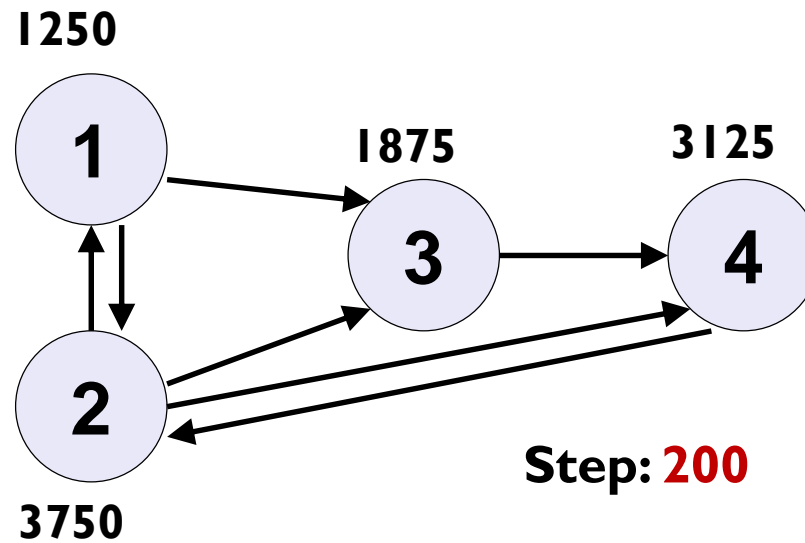
10.000 Random Surfers



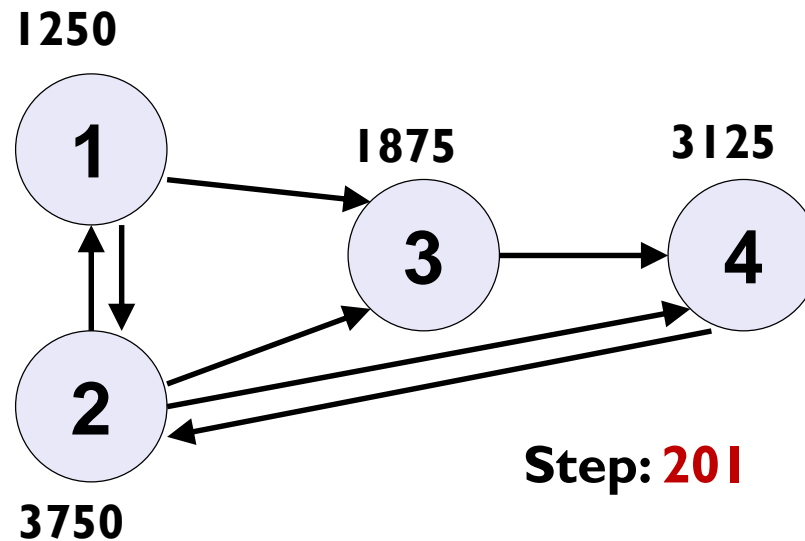
10.000 Random Surfers



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)

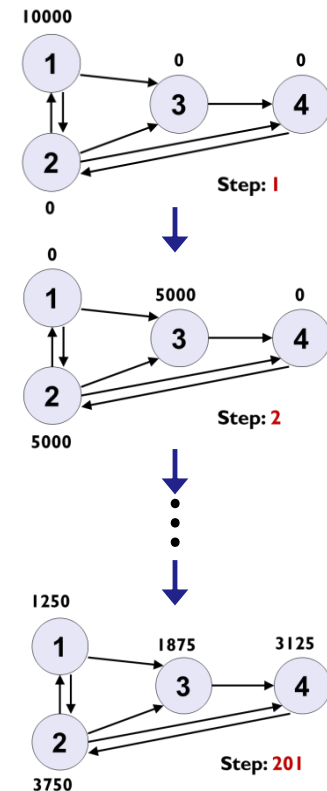
PageRank Intuition

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

$$\mathbf{q}^{(0)} = (q_1^{(0)}, \dots, q_n^{(0)})$$

- A **transition probability matrix**

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

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 i to page j

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

Example (from before)

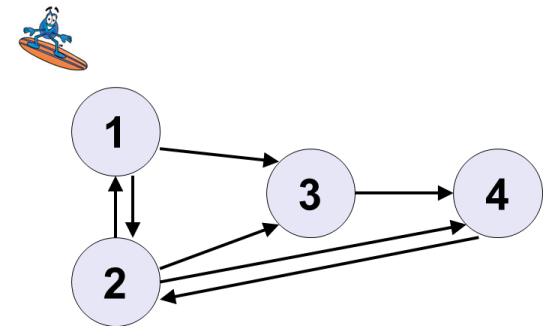
Initial probability distribution

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

$$\mathbf{q}^{(0)} = (1 \quad 0 \quad 0 \quad 0)$$

Transition probability matrix

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



Markov Chain Transitions

The way we **compute** transitions:

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

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

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

Matrix notation, the way we **talk** about it:

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

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

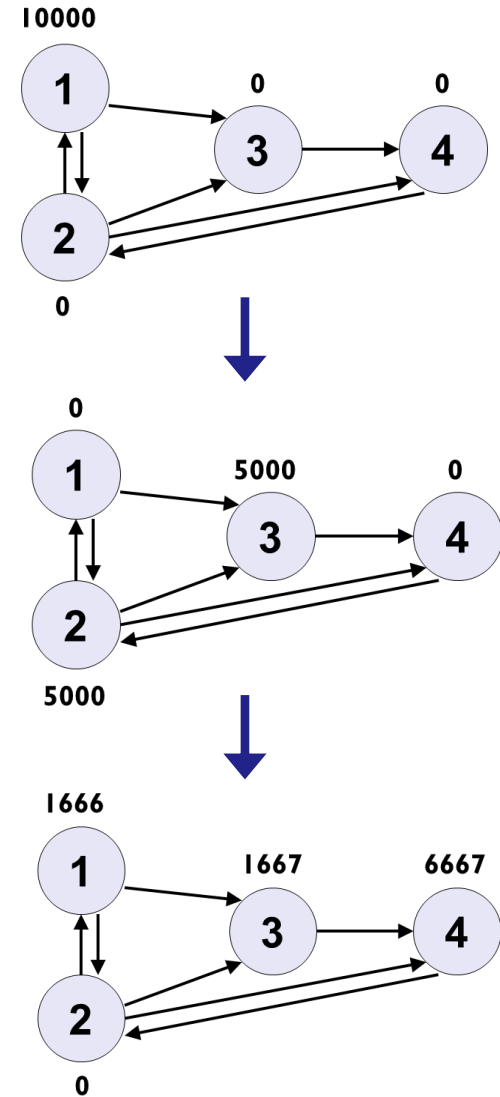
Example (cont.)

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

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

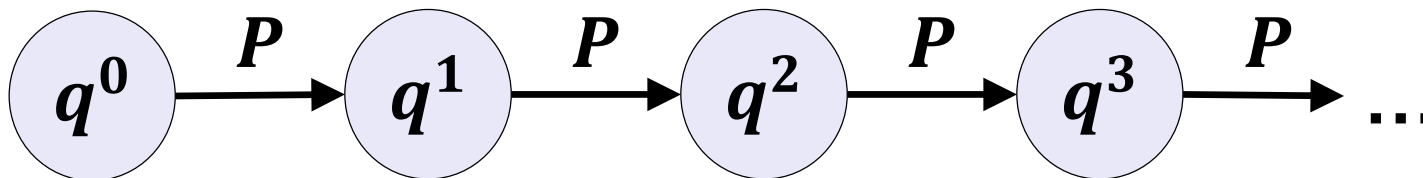
$$q^{(2)} = q^{(1)} P$$

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



...and for the Bayesian network folks!

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



PageRank – Stationary Distribution

q is stationary, if
 $q = qP$

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

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

“Eigen-Stuff”

q with $q = qP$ is also called an **Eigen-vector** of P with **Eigen-value** 1.

In fact, q is the **principal** Eigen-vector, because P is a transition probability matrix.

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

- SVD
- Power iteration – PageRank algorithm

Stationarity conditions

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

A Markov chain is ergodic if it is

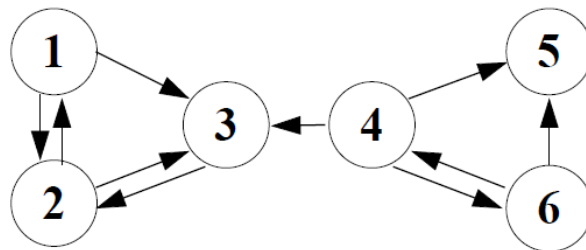
- *Irreducible*: “every node is reachable from every other node”
- *Aperiodic*: “nodes are not partitioned into sets such that all transitions occur cyclically from one set to another”

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 $(1-\alpha)$ (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

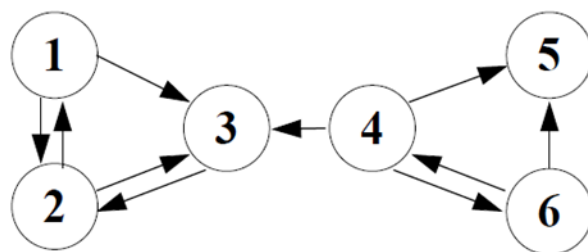
Result of teleporting:

- 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 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 $\mathbf{P}_{PageRank}$

- We use Power Iterations to approximate the PageRank by iterating

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

until $\mathbf{q}^{(t)}$ 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:

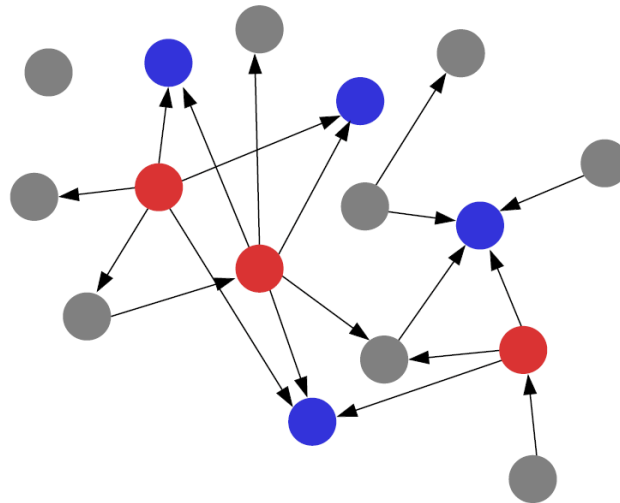
Amy N. Langville and Carl D. Meyer: Deeper Inside PageRank. Internet Mathematics Vol. 1, No. 3: 335-380.

<http://www.internetmathematicsjournal.com/article/1388-deeper-inside-pagerank>

HITS

(Hyperlink-Induced Topic Search)

Authorities and Hubs



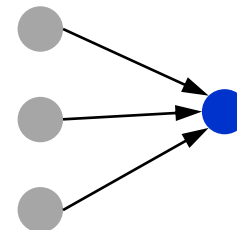
Authorities: Web pages linked to by many other pages.
Example: important company homepages

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

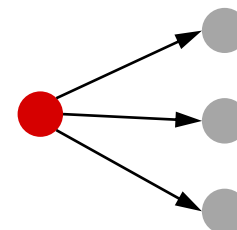
HITS – The Goal

Given a query, find

- Good sources of content
(authorities)



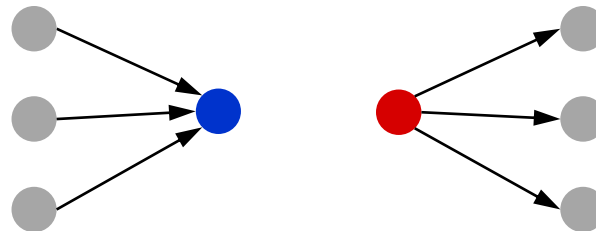
- Good sources of links
(hubs)



Intuition

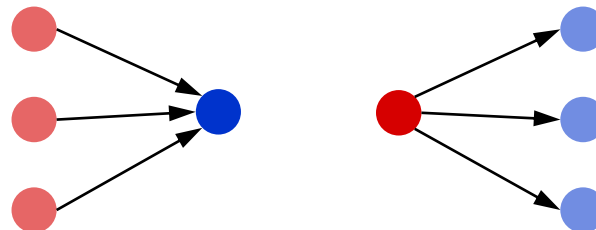
Authority comes from in-edges.

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



Better Authority comes from in-edges from **good hubs**.

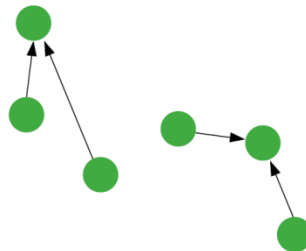
Being a **better hub** comes from out-edges to good authorities.



HITS Algorithm

Step 1 (of 3)

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

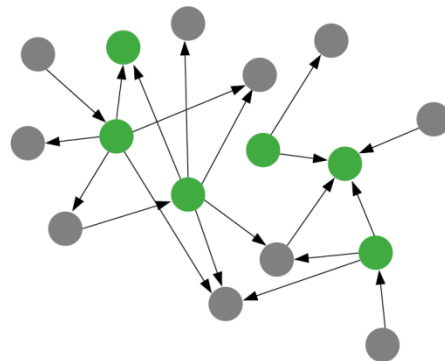


Result: the **root** set

HITS Algorithm (cont.)

Step 2 (of 3)

Add all neighbors of the pages in the root set:



Result: the **base** set

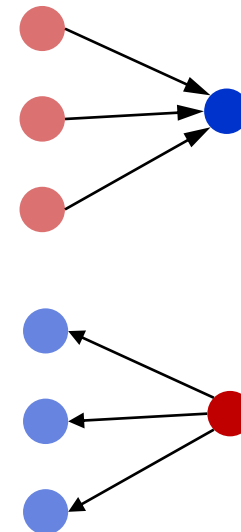
HITS Algorithm (cont.)

Step 3 (of 3)

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

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

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



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

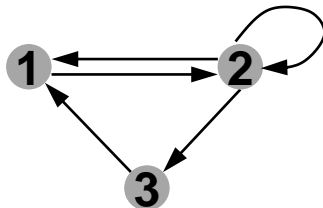
Math...

$n \times n$ link-matrix 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 $= 0$.

$n \times n$ transposed link-matrix L^T :

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



		j		
	L	1	2	3
i	1	0	1	0
	2	1	1	1
	3	1	0	0

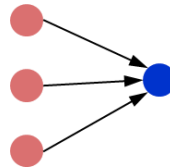
		j		
	L^T	1	2	3
i	1	0	1	1
	2	1	1	0
	3	0	1	0

HITS Algorithm (cont.)

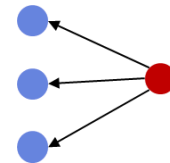
Step 3 (of 3) – the way we compute it

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

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

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

Eigen-stuff

HITS - Step 3 (of 3)

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

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

Substituting \Rightarrow

$$\mathbf{a} = \mathbf{L}^T \mathbf{L} \mathbf{a}$$

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

What does
this look like?

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

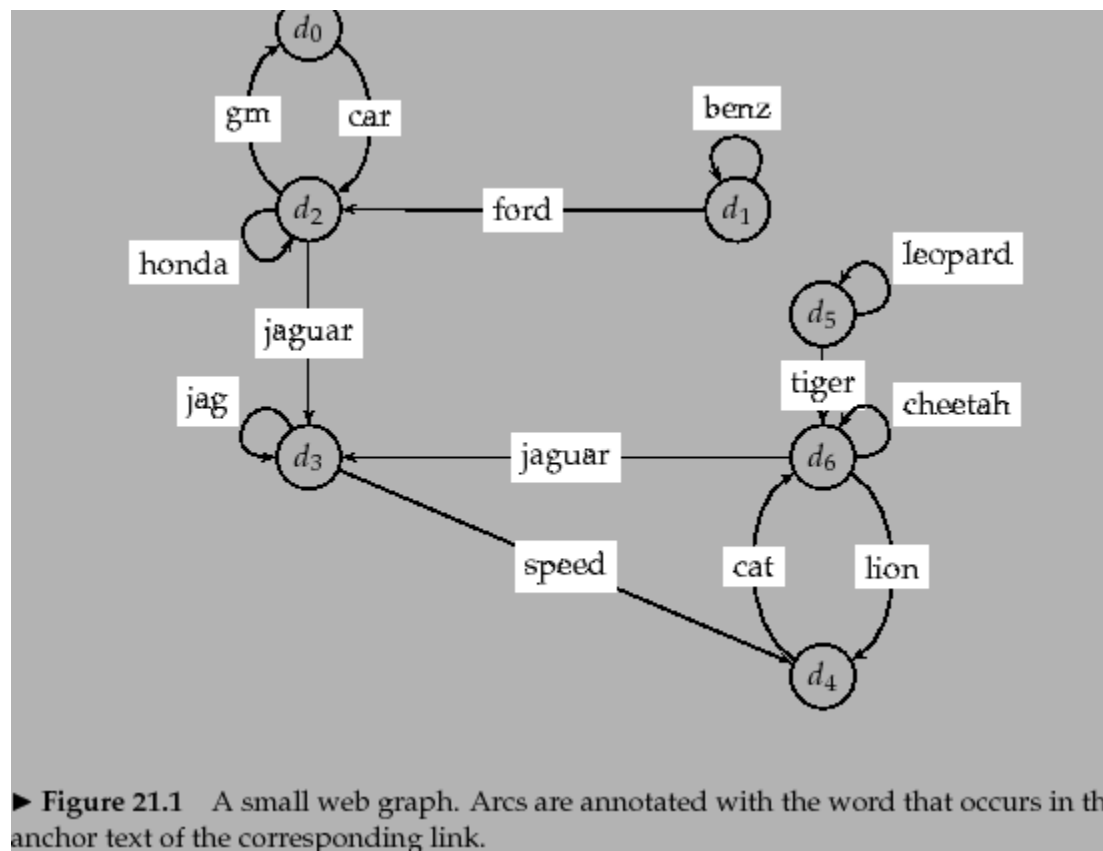
The HITS algorithm is a particular, known algorithm for computing eigenvectors: the *power iteration* method.

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



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

PageRank vs. HITS – Example (cont.)

$P_{PageRank}$ 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

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

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)