# Chapter 5: Link Analysis for Authority Scoring

**5.1 PageRank (S. Brin and L. Page 1997/1998)**

**5.2 HITS (J. Kleinberg 1997/1999)**

**5.3 Comparison and Extensions**

**5.4 Topic-specific and Personalized PageRank**

**5.5 Efficiency Issues**

**5.6 Online Page Importance**

**5.7 Spam-Resilient Authority Scoring**

# Improving Precision by Authority Scores

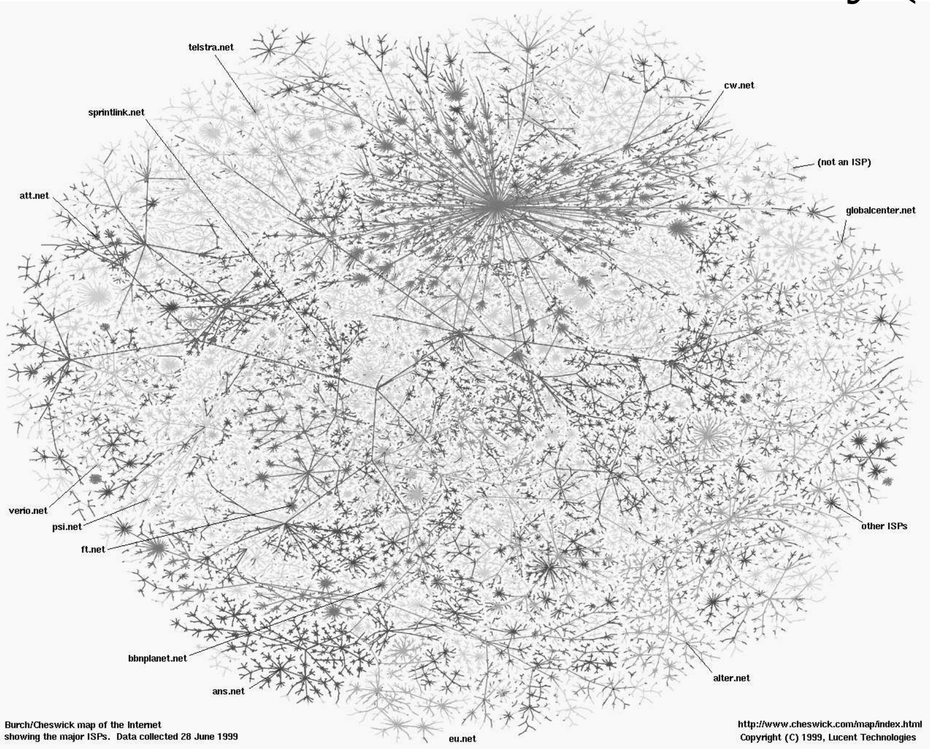
<u>Goal:</u>
Higher ranking of URLs with high authority regarding
volume, significance, freshness, authenticity of information content
$\rightarrow$ improve precision of search results

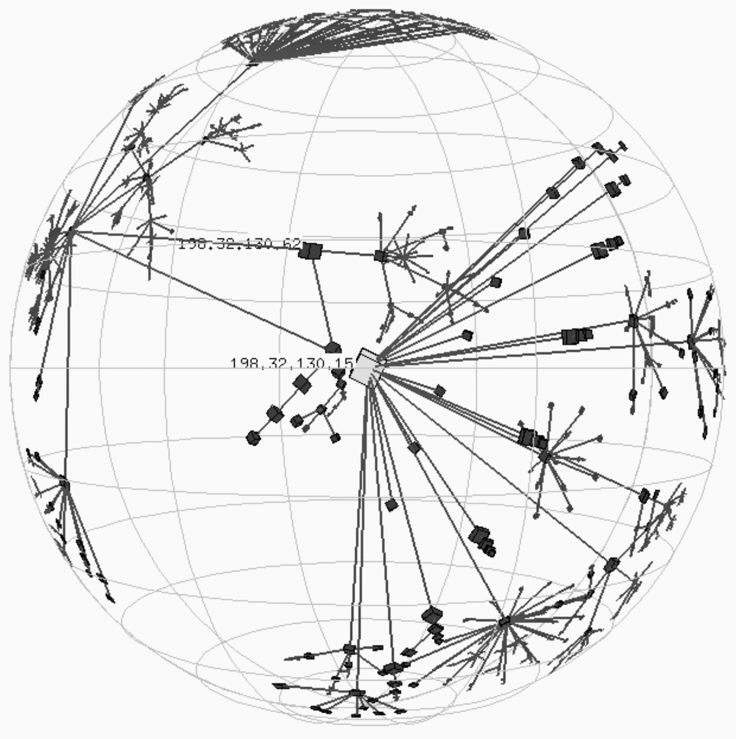
<u>Approaches (all interpreting the Web as a directed graph G):</u>
- citation or impact rank (q) ~ indegree (q)
- PageRank (by Lawrence Page)
- HITS algorithm (by Jon Kleinberg)

<u>Combining relevance and authority ranking:</u>
- by weighted sum with appropriate coefficients (Google)
- by initial relevance ranking and iterative
  improvement via authority ranking (HITS)

# Web Structure: Small Diameter

Small World Phenomenon (Milgram 1967)
Studies on Internet Connectivity (1999)



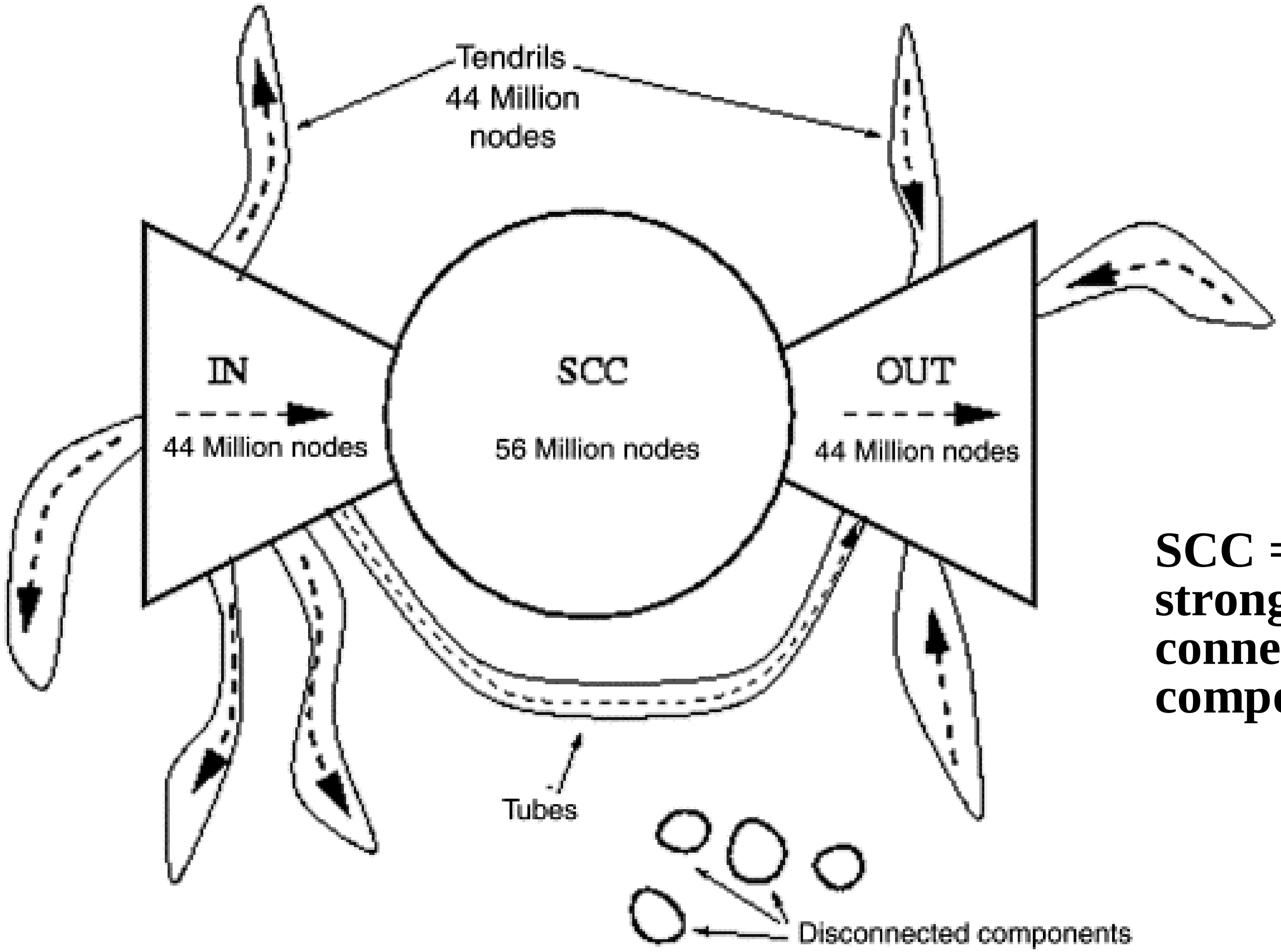

Source: Bill Cheswick and Hal Burch,
http://research.lumeta.com/ches/map/index.html

Source: KC Claffy,
http://www.caida.org/outreach/papers/1999/Nae/Nae.html

**suggested small world phenomenon: low-diameter graph**
**( diameter = max {shortest path (x,y) | nodes x and y} )**

# Web Structure: Connected Components

Study of Web Graph (Broder et al. 2000)



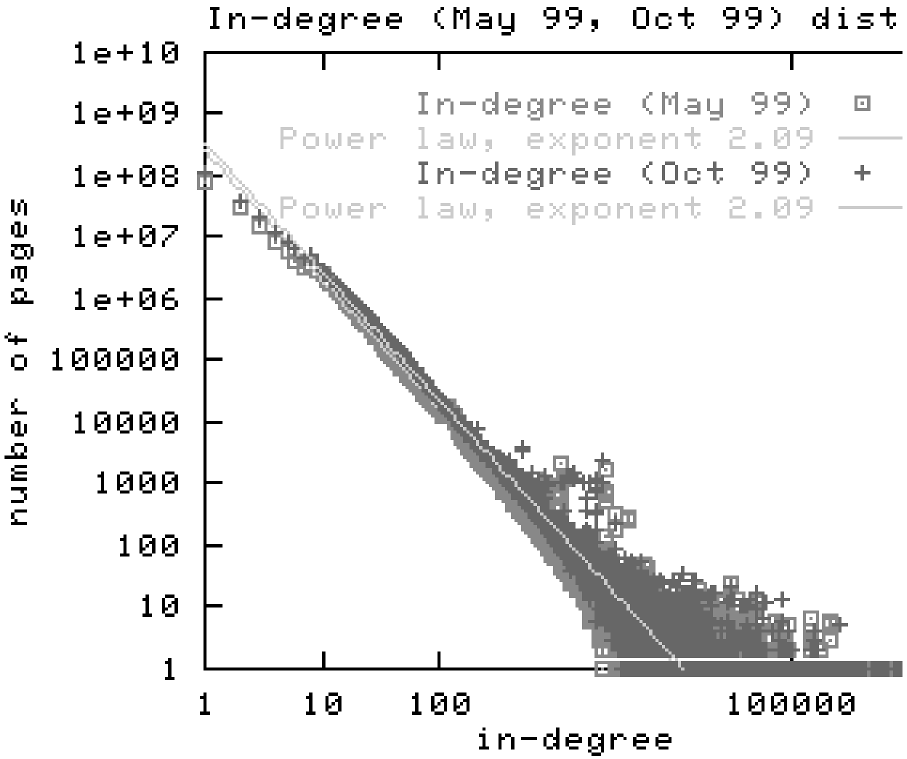

**SCC = strongly connected component**

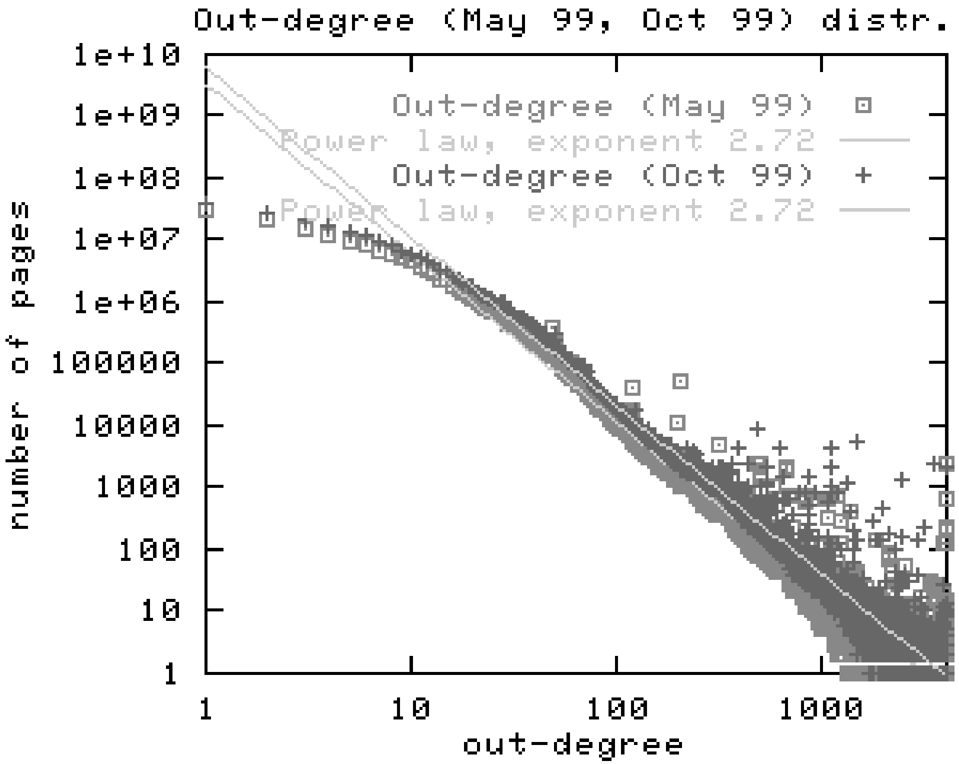
Source: A.Z. Broder et al., WWW 2000

- strongly connected core tends to have small diameter

# Web Structure: Power-Law Degrees

Study of Web Graph (Broder et al. 2000)




- power-law distributed degrees: P[degree=k] ~ $(1/k)^{\alpha}$ with $\alpha \approx 2.1$ for indegrees and $\alpha \approx 2.7$ for outdegrees

# Power-Law Distributions

**Zipf distribution**
**for $0 \le k \le n$ :**

$$f(k) \sim \frac{1}{k}$$

***frequently observed for ranks in socio-economic systems***

**discrete**
**Pareto distribution**
**for $0 \le k$:**

$$f(k) \sim \frac{1}{k^{\alpha}}$$

***frequently observed for absolute values in socio-economic systems***

**continuous**
**Pareto distribution**
**for $x_0 \le x$:**

$$f(k) = \frac{\alpha - 1}{x_0} \left( \frac{x_0}{x} \right)^{\alpha}$$

**Pareto distribution is heavy-tailed**
**($E[X^k]$ defined if and only if $\alpha > k+1$)**

# Example Zipf Distribution

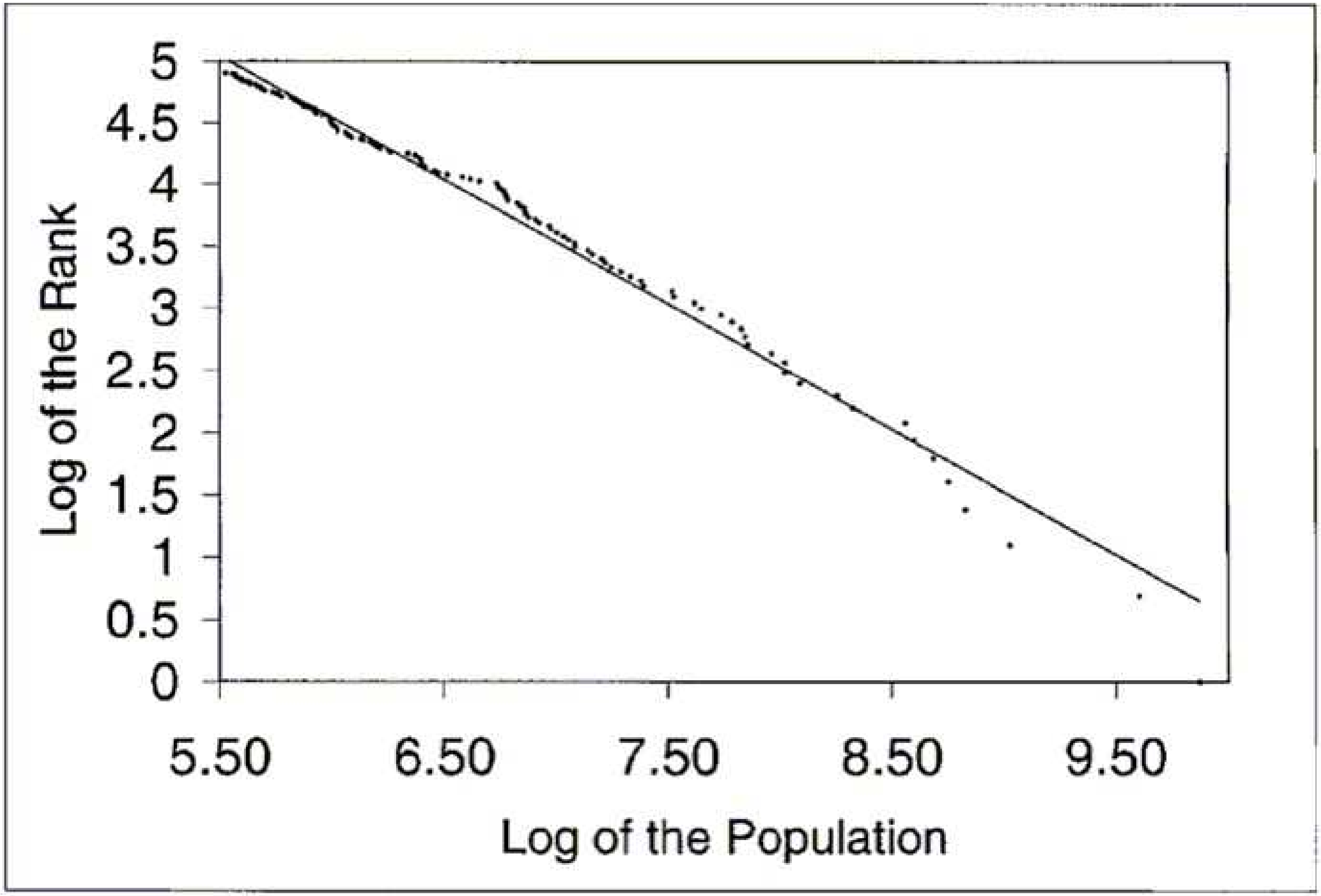
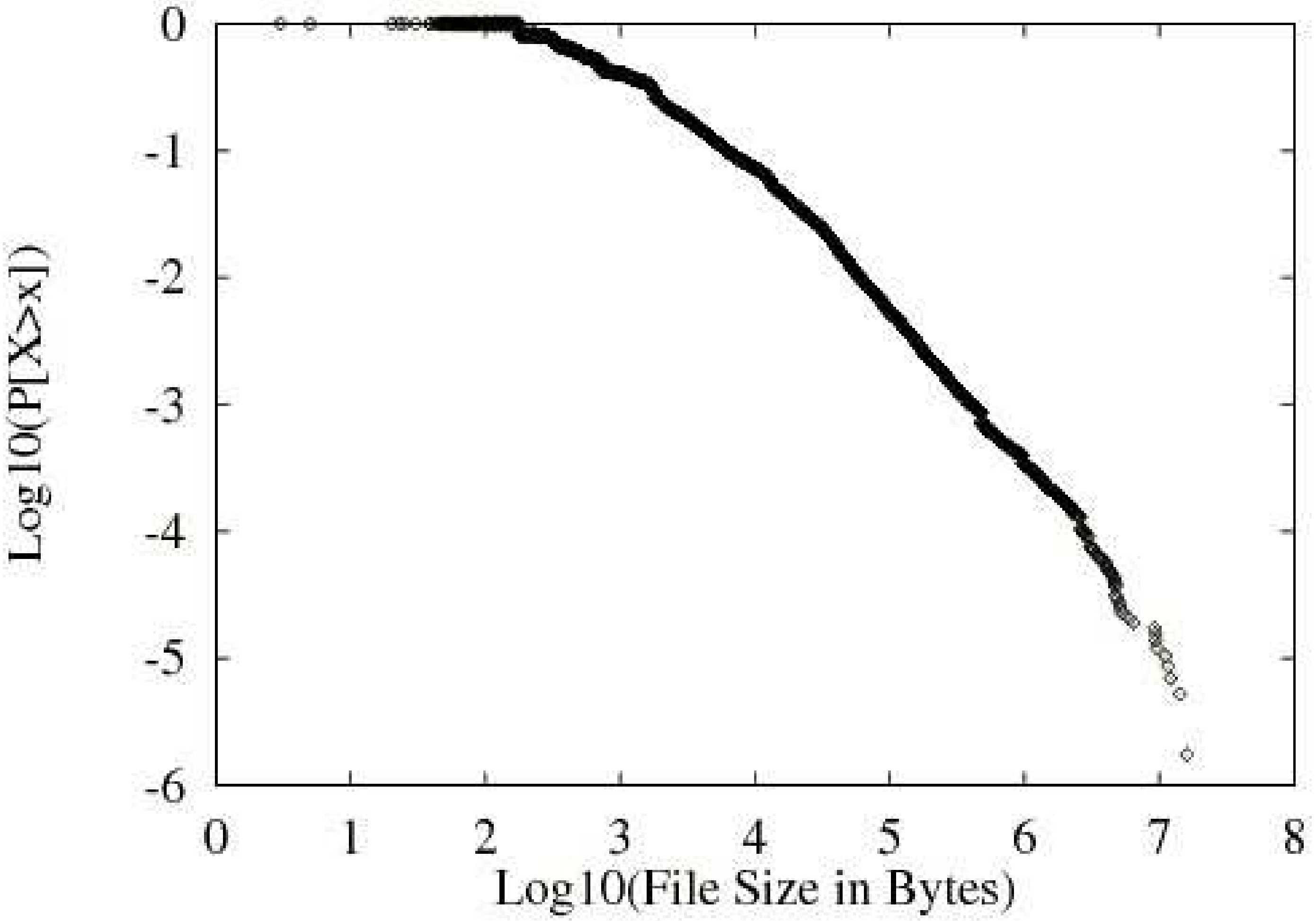



FIGURE I
Log Size versus Log Rank of the 135 largest U. S. Metropolitan Areas in 1991
Source: Statistical Abstract of the United States [1993].

**size of cities**

**Source: Denise Pumain, Scaling Laws and Urban Distributions, 2003**

# Example Pareto Distribution




**size of file requests**

**Source: Mark Crovella et al., Heavy-tailed Probability Distributions in the World Wide Web, 1998**

# Page Rank r(q)

<u>given:</u> directed Web graph G=(V,E) with |V|=n and adjacency matrix A: $A_{ij} = 1$ if $(i,j) \in E$, 0 otherwise

**<u>Idea:</u>** $$r(q) \approx k \sum_{(p,q) \in G} r(p) / outdegree(p)$$

**<u>Def.:</u>** $$r(q) = \varepsilon / n + (1-\varepsilon) \sum_{(p,q) \in G} r(p) / outdegree(p)$$

with $0 < \varepsilon \leq 0.2$

**<u>Theorem:</u>** With $A'_{ij} = 1/outdegree(j)$ if $(j,i) \in E$, 0 otherwise:

$$\vec{r} = \frac{\vec{\varepsilon}}{n} + (1-\varepsilon) A' \vec{r} \quad \Leftrightarrow \quad \vec{r} = \left( \frac{\vec{\varepsilon}}{n} \vec{1}^T + (1-\varepsilon) A' \right) \vec{r}$$

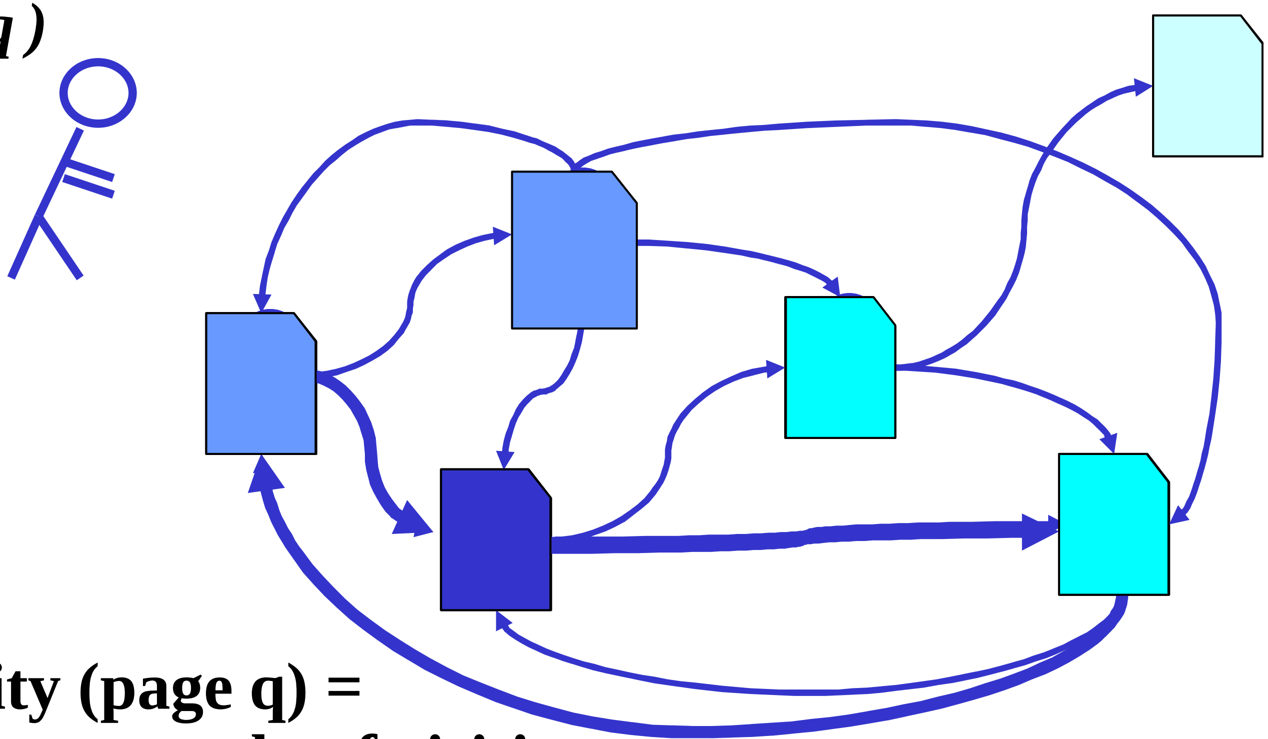
i.e. **r is Eigenvector** of a modified adjacency matrix

**Iterative computation** of r(q) (after large Web crawl):
- Initialization: r(q) := 1/n
- Improvement by evaluating recursive equation of definition; typically converges after about 100 iterations

# Google‘s PageRank

<u>Idea:</u> incoming links are endorsements & increase page authority, authority is higher if links come from high-authority pages

$$PR(q) = \varepsilon \cdot j(q) + (1-\varepsilon) \cdot \sum_{p \in IN(q)} PR(p) \cdot t(p,q)$$



**Authority (page q) = stationary prob. of visiting q**

**random walk: uniformly random choice of links + random jumps**

# PageRank as Eigenvector of Stochastic Matrix

**A stochastic matrix is an $n \times n$ matrix M with row sum $\Sigma_{j=1..n} M_{ij} = 1$ for each row i**

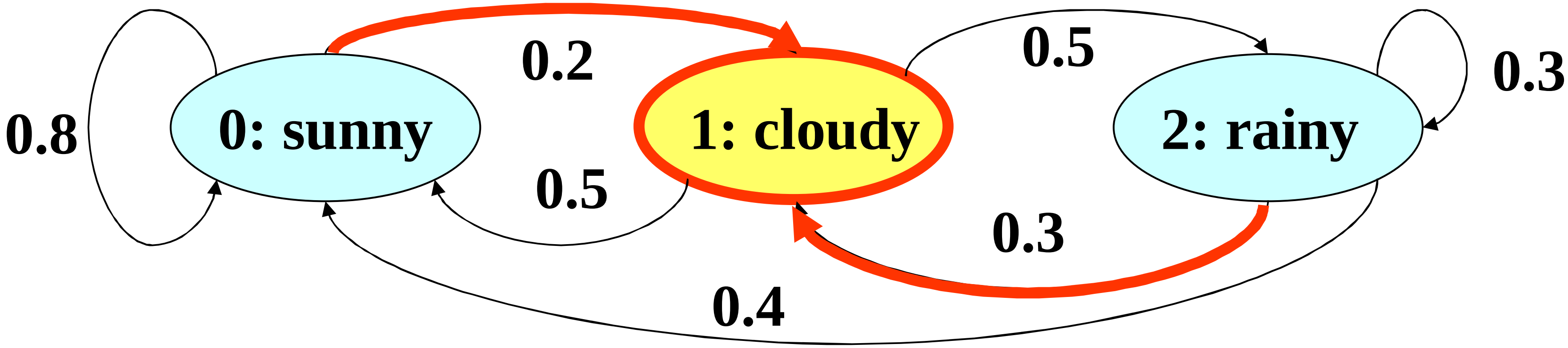
**Random surfer follows a stochastic matrix**

**<u>Theorem:</u>**
**For every stochastic matrix M all Eigenvalues $\lambda$ have the property $|\lambda| \le 1$ and there is an Eigenvector x with Eigenvalue 1 s.t. $x \ge 0$ and $||x||_1 = 1$**

**Suggests power iteration $x^{(i+1)} = M^T x^{(i)}$**

**But: real Web graph has sinks, may be periodic, is not strongly connected**

# Markov Chains in a Nutshell




p0 = 0.8 p0 + 0.5 p1 + 0.4 p2
p1 = 0.2 p0 + 0.3 p2
p2 = 0.5 p1 + 0.3 p2
p0 + p1 + p2 = 1

$\Rightarrow$ p0 ≈ 0.657, p1 = 0.2, p2 ≈ 0.143

**state set: finite or infinite** **time: discrete or continuous**

**state transition prob‘s:** $p_{ij}$ **state prob‘s in step t:** $p_i^{(t)} = P[S(t)=i]$

**Markov property:** $P[S(t)=i \mid S(0), \ldots, S(t-1)] = P[S(t)=i \mid S(t-1)]$

**interested in stationary state probabilities:**

$$p_j := \lim_{t\to\infty} p_j^{(t)} = \lim_{t\to\infty} \sum_k p_k^{(t-1)} p_{kj} \qquad p_j = \sum_k p_k \, p_{kj} \qquad \sum_j p_j = 1$$

**guaranteed to exist for irreducible, aperiodic, finite Markov chains**

# Digression: Markov Chains

A **stochastic process** is a family of random variables $\{X(t) \mid t \in T\}$.
T is called parameter space, and the domain M of X(t) is called state space. T and M can be discrete or continuous.

A stochastic process is called **Markov process** if for every choice of $t_1, ..., t_{n+1}$ from the parameter space and every choice of $x_1, ..., x_{n+1}$ from the state space the following holds:

$$P[\, X(t_{n+1}) = x_{n+1} \mid X(t_1) = x_1 \wedge X(t_2) = x_2 \wedge ... \wedge X(t_n) = x_n \,]$$
$$= P[\, X(t_{n+1}) = x_{n+1} \mid X(t_n) = x_n \,]$$

A Markov process with discrete state space is called **Markov chain**.
A canonical choice of the state space are the natural numbers.
Notation for Markov chains with discrete parameter space:
$X_n$ rather than $X(t_n)$ with n = 0, 1, 2, ...

# Properties of Markov Chains with Discrete Parameter Space (1)

The Markov chain Xn with discrete parameter space is

**homogeneous** if the transition probabilities
$p_{ij} := P[X_{n+1} = j \mid X_n = i]$ are independent of n

**irreducible** if every state is reachable from every other state with positive probability:

$$\sum_{n=1}^{\infty} P[X_n = j \mid X_0 = i] > 0 \quad \text{for all i, j}$$

**aperiodic** if every state i has period 1, where the period of i is the gcd of all (recurrence) values n for which

$$P[X_n = i \wedge X_k \neq i \ for\ k = 1,\ldots,n-1 \mid X_0 = i] > 0$$

# Properties of Markov Chains with Discrete Parameter Space (2)

The Markov chain Xn with discrete parameter space is

**positive recurrent** if for every state i the recurrence probability is 1 and the mean recurrence time is finite:

$$\sum_{n=1}^{\infty} P[X_n = i \wedge X_k \neq i \ for\ k = 1,\ldots,n-1 \mid X_0 = i] = 1$$

$$\sum_{n=1}^{\infty} n\, P[X_n = i \wedge X_k \neq i \ for\ k = 1,\ldots,n-1 \mid X_0 = i] < \infty$$

**ergodic** if it is homogeneous, irreducible, aperiodic, and positive recurrent.

# Results on Markov Chains with Discrete Parameter Space (1)

For the **n-step transition probabilities**

$p_{ij}^{(n)} := P[\, X_n = j \,|\, X_0 = i \,]$ the following holds:

$$p_{ij}^{(n)} = \sum_k p_{ik}^{(n-1)} \, p_{kj} \quad \text{with} \quad p_{ij}^{(1)} := p_{ik}$$

$$= \sum_k p_{ik}^{(n-l)} \, p_{kj}^{(l)} \quad \mathit{for}\; 1 \le l \le n-1$$

in matrix notation: $P^{(n)} = P^n$

For the **state probabilities after n steps**

$\pi_j^{(n)} := P[\, X_n = j \,]$ the following holds:

$$\pi_j^{(n)} = \sum_i \pi_i^{(0)} \, p_{ij}^{(n)}$$ with initial state probabilities $\pi_i^{(0)}$

in matrix notation: $\Pi^{(n)} = \Pi^{(0)} P^{(n)}$ *(Chapman-Kolmogorov equation)*

# Results on Markov Chains with Discrete Parameter Space (2)

Every homogeneous, irreducible, aperiodic Markov chain with a finite number of states is positive recurrent and ergodic.

For every ergodic Markov chain there exist **stationary state probabilities**

$$\pi_j := \lim_{n \to \infty} \pi_j^{(n)}$$

These are independent of $\Pi^{(0)}$
and are the solutions of the following system of linear equations:

$$\pi_j = \sum_i \pi_i \, p_{ij} \quad \text{for all } j$$

*(balance equations)*

$$\sum_j \pi_j = 1$$

in matrix notation: $\Pi = \Pi P$

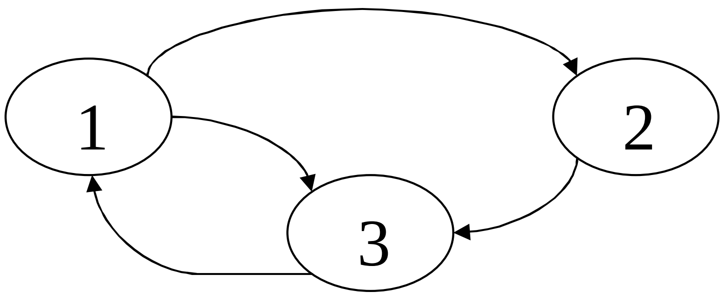
(with $1 \times n$ row vector $\Pi$) $\Pi \vec{1} = 1$

# Page Rank as a Markov Chain Model

Model a **random walk** of a Web surfer as follows:
• follow outgoing hyperlinks with uniform probabilities
• perform „random jump“ with probability ε
→ ergodic Markov chain
  The **PageRank** of a URL is the **stationary visiting probability** of URL in the above Markov chain.
Further generalizations have been studied
(e.g. random walk with back button etc.)

Drawback of Page rank method:
Page rank is query-independent and orthogonal to relevance

# Example: Page Rank Computation




$\varepsilon = 0.2$

$$P = \begin{pmatrix} 0.0 & 0.5 & 0.5 \\ 0.1 & 0.0 & 0.9 \\ 0.9 & 0.1 & 0.0 \end{pmatrix}$$

$$\Pi^{(0)} \approx \begin{pmatrix} 0.333 \\ 0.333 \\ 0.333 \end{pmatrix}^T \Rightarrow \Pi^{(1)} \approx \begin{pmatrix} 0.333 \\ 0.200 \\ 0.466 \end{pmatrix}^T \Rightarrow \Pi^{(2)} \approx \begin{pmatrix} 0.439 \\ 0.212 \\ 0.346 \end{pmatrix}^T \Rightarrow \Pi^{(3)} \approx \begin{pmatrix} 0.332 \\ 0.253 \\ 0.401 \end{pmatrix}^T$$

$$\Rightarrow \Pi^{(4)} \approx \begin{pmatrix} 0.385 \\ 0.176 \\ 0.527 \end{pmatrix}^T \Rightarrow \Pi^{(5)} \approx \begin{pmatrix} 0.491 \\ 0.244 \\ 0.350 \end{pmatrix}^T$$

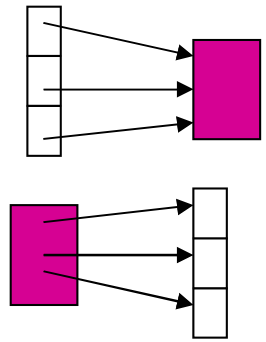
$\pi 1 = 0.1\ \pi 2 + 0.9\ \pi 3$
$\pi 2 = 0.5\ \pi 1 + 0.1\ \pi 3$
$\pi 3 = 0.5\ \pi 1 + 0.9\ \pi 2$
$\pi 1 + \pi 2 + \pi 3 = 1$

$\Rightarrow \pi 1 \approx 0.3776,\ \pi 2 \approx 0.2282,\ \pi 3 \approx 0.3942$

# 5.2 HITS Algorithm: Hyperlink-Induced Topic Search (1)

Idea:

Determine
- good content sources: **Authorities** (high indegree)
- good link sources: **Hubs** (high outdegree)



Find
- better authorities that have good hubs as predecessors
- better hubs that have good authorities as successors

For Web graph G=(V,E) define for nodes p, q $\in$ V

**authority score** $x_q = \sum_{(p,q) \in E} y_p$ and

**hub score** $y_p = \sum_{(p,q) \in E} x_q$

# HITS Algorithm (2)

Authority and hub scores in matrix notation:

$$\vec{x} = A^T \vec{y} \qquad \vec{y} = A \vec{x}$$

Iteration with adjacency matrix A:

$$\vec{x} := A^T \vec{y} := A^T A \vec{x} \qquad \vec{y} := A \vec{x} := A A^T \vec{y}$$

x and y are **Eigenvectors** of $A^T A$ and $AA^T$, resp.

<u>Intuitive interpretation:</u>

$M^{(auth)} := A^T A$ is the cocitation matrix: $M^{(auth)}_{ij}$ is the number of nodes that point to both i and j

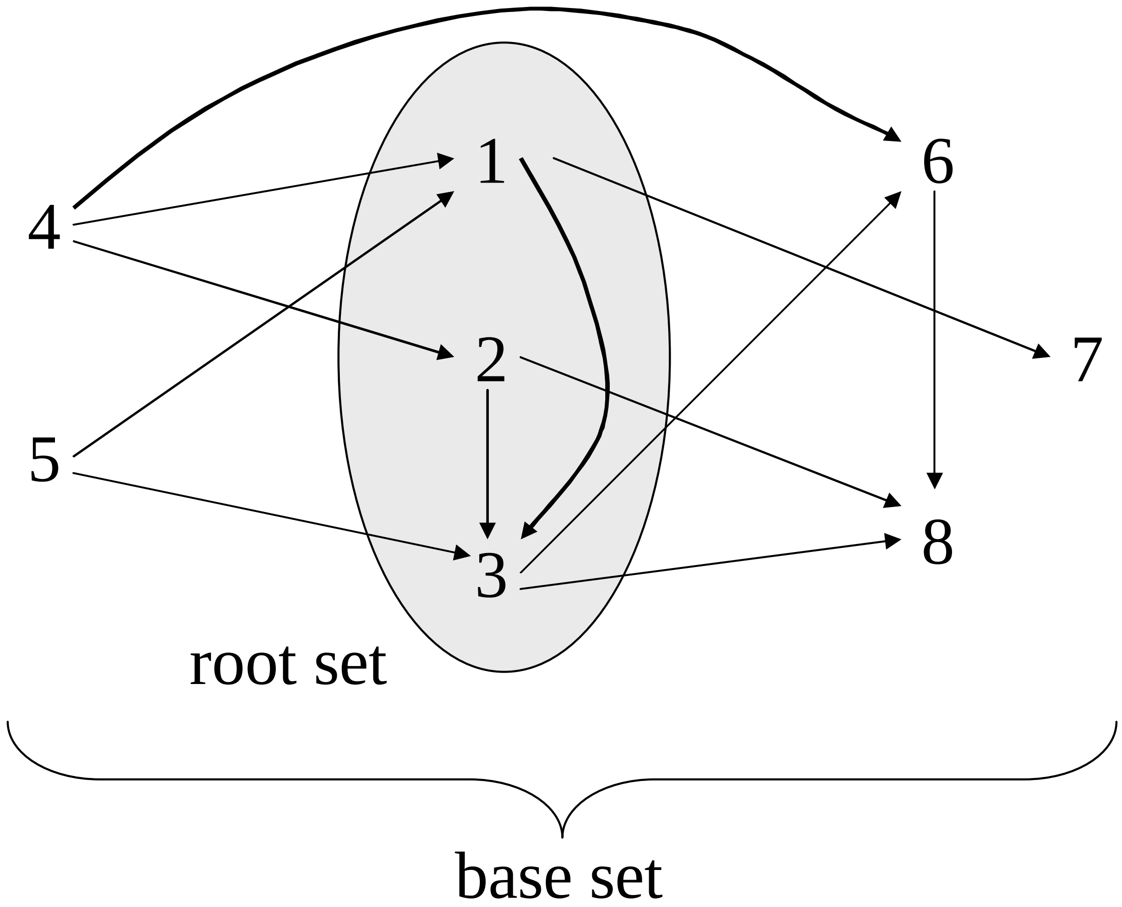
$M^{(hub)} := AA^T$ is the bibliographic-coupling matrix: $M^{(hub)}_{ij}$ is the number of nodes to which both i and j point

# Implementation of the HITS Algorithm

1) Determine sufficient number (e.g. 50-200) of „root pages“ via relevance ranking (e.g. using tf*idf ranking)
2) Add all successors of root pages
3) For each root page add up to d predecessors
4) Compute iteratively the authority and hub scores of this „base set“ (of typically 1000-5000 pages) with initialization $x_q := y_p := 1 / |\text{base set}|$ and normalization after each iteration
   $\rightarrow$ converges to principal Eigenvector (Eigenvector with largest Eigenvalue (in the case of multiplicity 1)
5) Return pages in descending order of authority scores (e.g. the 10 largest elements of vector x)

Drawback of HITS algorithm:
relevance ranking within root set is not considered

# Example: HITS Algorithm



# Improved HITS Algorithm

Potential weakness of the HITS algorithm:
• irritating links (automatically generated links, spam, etc.)
• topic drift (e.g. from „Jaguar car“ to „car“ in general)

Improvement:
• Introduce **edge weights**:
  0 for links within the same host,
  1/k with k links from k URLs of the same host to 1 URL (xweight)
  1/m with m links from 1 URL to m URLs on the same host (yweight)
• Consider **relevance weights** w.r.t. query topic (e.g. tf*idf)

$\rightarrow$ Iterative computation of

authority score $$x_q = \sum_{(p,q) \in E} y_p * topic\,score(p) * xweight(p,q)$$

hub score $$y_p = \sum_{(p,q) \in E} x_q * topic\,score(q) * yweight(p,q)$$

# Finding Related URLs

**Cocitation algorithm:**

- Determine up to B predecessors of given URL u
- For each predecessor p determine up to BF successors $\neq$ u
- Determine among all siblings s of u those with the largest number of predecessors that point to both s and u (degree of cocitation)

**Companion algorithm:**

- Determine appropriate base set for URL u („vicinity“ of u)
- Apply HITS algorithm to this base set

# Companion Algorithm for Finding Related URLs

1) Determine **base set:** u plus
   - up to B predecessors of u and
     for each predecessor p up to BF successors $\neq$ u plus
   - up to F successors of u and
     for each successor c up to FB predecessors $\neq$ u

   with elimination of stop URLs (e.g. www.yahoo.com)
2) **Duplicate elimination:**
   Merge nodes both of which have more than 10 successors
   and have 95 % or more overlap among their successors
3) Compute **authority scores**
   using the improved HITS algorithm

# SimRank [Jeh/Widom 2002]

**Idea: pages x and y are similar if referenced by similar pages**

$$SR(x,y) = \frac{c}{|In(x)| \cdot |In(y)|} \sum_{p \in In(x)} \sum_{q \in In(y)} SR(p,q)$$

**with constant c < 1 and SR(x,y)=1 for x=y and 0 otherwise,**
**or SR(x,y) set to content similarity of x and y**

**solved by iteration procedure,**
**conceptually operating on $G^2$ graph of all node pairs**
**with edge $(a,b) \rightarrow (c,d)$ if G has edges $a \rightarrow c$ and $b \rightarrow d$**

**can be extended to bipartite graphs (e.g. customers and products)**
**or even more general typed graphs**

# HITS Algorithm for „Community Detection“

Root set may contain multiple topics or „communities“,
e.g. for queries „jaguar“, „Java“, or „randomized algorithm“

Approach:

- Compute k largest Eigenvalues of $A^T A$ and the corresponding Eigenvectors x
- For each of these k Eigenvectors x the largest authority scores indicate a densely connected „community“

# SALSA: Random Walk on Hubs and Authorities

View each node v of the link graph as two nodes $v_h$ and $v_a$

Construct bipartite undirected graph G'(V',E') from link graph G(V,E):

$V' = \{v_h \mid v \in V \text{ and outdegree}(v)>0\} \cup \{v_a \mid v \in V \text{ and indegree}(v)>0\}$

$E' = \{(v_h, w_a) \mid (v,w) \in E\}$

**Stochastic hub matrix H:**

$$h_{ij} = \sum_k \frac{1}{\deg ree(i_h)} \frac{1}{\deg ree(k_a)}$$

for hubs i, j and k ranging over all nodes with $(i_h, k_a), (k_a, j_h) \in E'$

**Stochastic authority matrix A:**

$$a_{ij} = \sum_k \frac{1}{\deg ree(i_a)} \frac{1}{\deg ree(k_h)}$$

for authorities i, j and k ranging over all nodes with $(i_a, k_h), (k_h, j_a) \in E'$

The corresponding Markov chains are ergodic on connected component

The stationary solutions for these Markov chains are:

$\pi[v_h]$ ~ outdegree(v) for H and $\pi[v_a]$ ~ indegree(v) for A

# Additional Literature for Chapter 5

Link Analysis Principles & Algorithms:

- Chakrabarti, Chapter 7
- S. Chakrabarti: Using Graphs in Unstructured and Semistructured Data Mining, Tutorial Slides, ADFOCS Summer School 2004
- J.M. Kleinberg: Authoritative Sources in a Hyperlinked Environment, JACM 46(5), 1999
- S Brin, L. Page: The Anatomy of a Large-Scale Hypertextual Web Search Engine, WWW 1998
- K. Bharat, M. Henzinger: Improved Algorithms for Topic Distillation in a Hyperlinked Environment, SIGIR 1998
- J. Dean, M. Henzinger: Finding Related Pages in the WorldWideWeb, WWW 1999
- R. Lempel, S. Moran: SALSA: The Stochastic Approach for Link-Structure Analysis, ACM TOIS 19(2), 2001.
- A. Borodin, G.O. Roberts, J.S. Rosenthal, P. Tsaparas: Finding Authorities and Hubs from Link Structures on the World Wide Web, WWW 2001
- G. Jeh, J. Widom: SimRank Measure of Structural-Context Similarity, KDD 2002
- C. Ding, X. He, P. Husbands, H. Zha, H. Simon: PageRank, HITS, and a Unified Framework for Link Analysis, SIAM Int. Conf. on Data Mining, 2003.
- A. Borodin, G.O. Roberts, J.S. Rosenthal, P. Tsaparas: Link analysis ranking: algorithms, theory, and experiments. ACM TOIT 5(1), 2005
- M. Bianchini, M. Gori, F. Scarselli: Inside PageRank.ACM TOIT 5(1), 2005
- A.N. Langville, C.D. Meyer: Deeper inside PageRank. Internet Math., 1(3), 2004