Link Analysis: PageRank and HITS

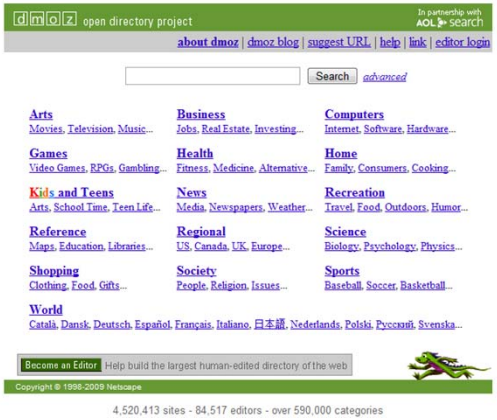
CS224W: Social and Information Network Analysis

Jure Leskovec, Stanford University

http://cs224w.stanford.edu

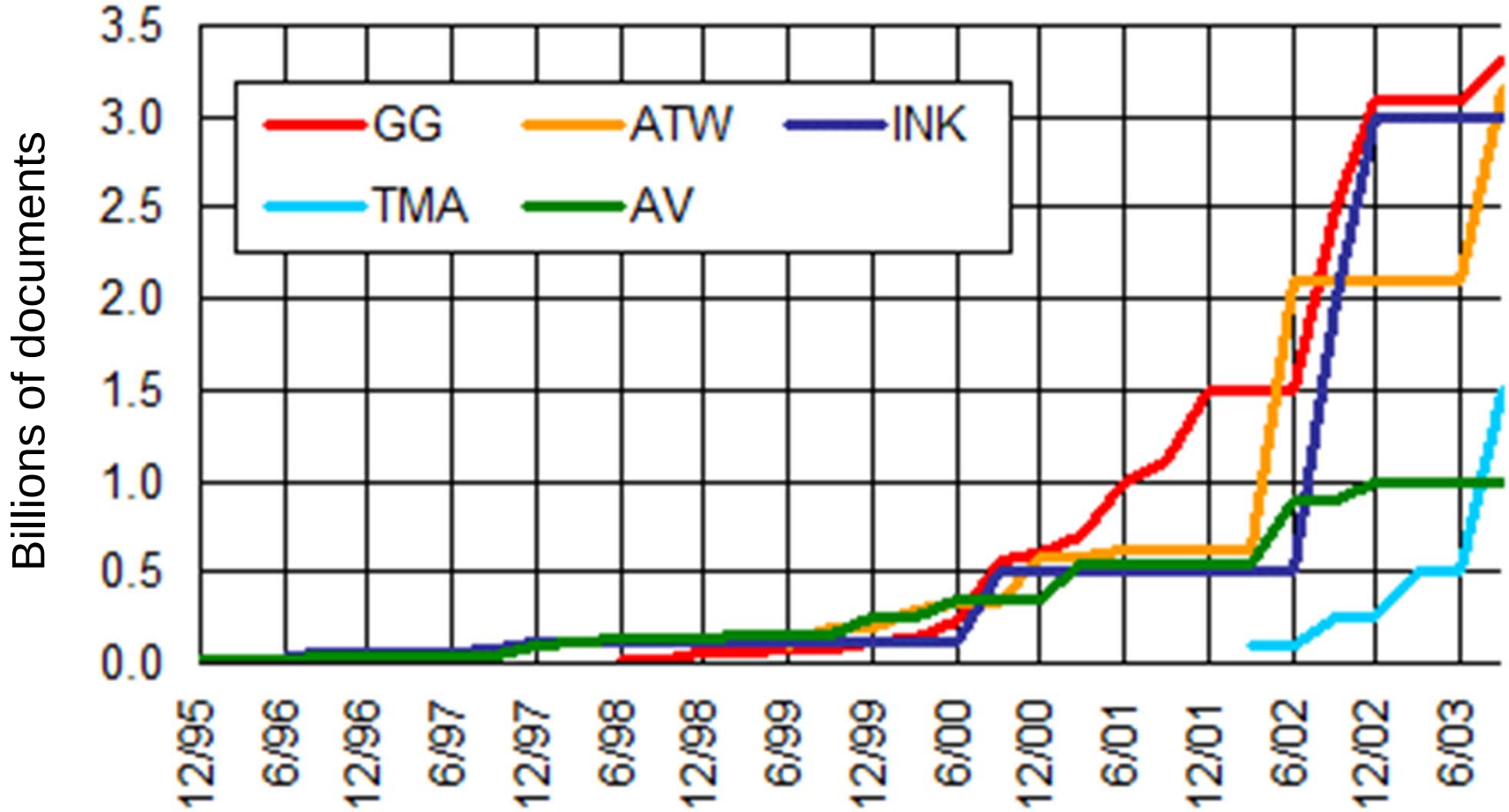
How to Organize the Web?

- **How to organize the Web?**
- **First try:** Human curated **Web directories**
  - Yahoo, DMOZ, LookSmart
- **Second try: Web Search**
  - **Information Retrieval** investigates: Find relevant docs in a small and trusted set
    - Newspaper articles, Patents, etc.
  - **But:** Web is **huge**, full of untrusted documents, random things, web spam, etc.



The Index Size Wars

Does more documents mean better results?



Search engine index size over time

Web Search: 2 Challenges

2 challenges of web search:

- **(1) Web contains many sources of information Who to "trust"?**
  - **Trick:** Trustworthy pages may point to each other!
- **(2) What is the "best" answer to query "newspaper"?**
  - No single right answer
  - **Trick:** Pages that actually know about newspapers might all be pointing to many newspapers

Ranking Nodes on the Graph

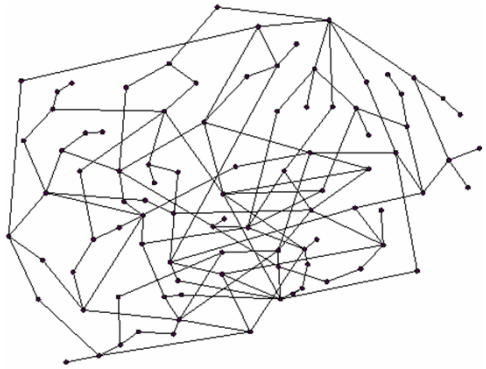
- **All web pages are not equally "important"**

  www.joe-schmoe.com vs. www.stanford.edu

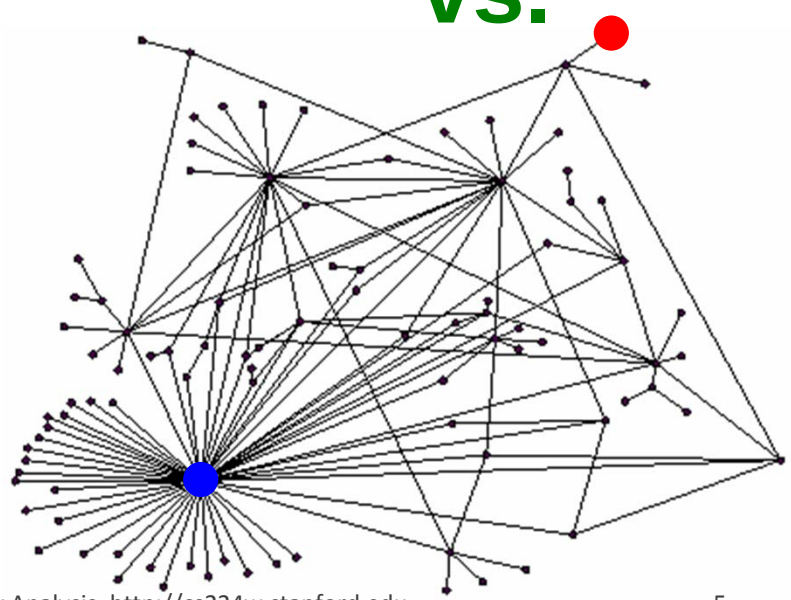
- **We already know:**

  There is large diversity in the web-graph node connectivity.

  Let's rank the pages by the link structure!



vs.



Link Analysis Algorithms

- **We will cover the following Link Analysis approaches** to computing importances of nodes in a graph:
  - Hubs and Authorities (HITS)
  - Page Rank
  - Topic-Specific (Personalized) Page Rank

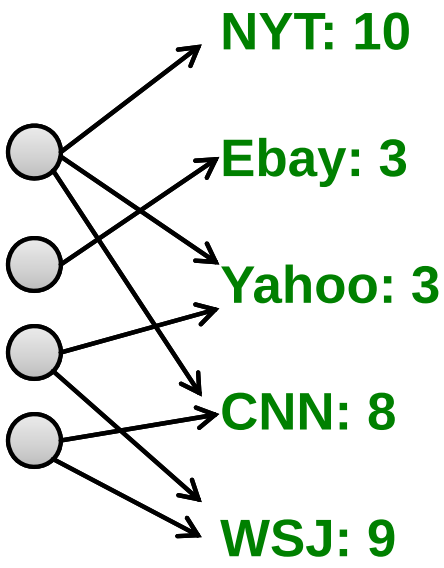
Sidenote: Various notions of **node centrality: Node u**

- **Degree dentrality** = degree of u
- **Betweenness centrality** = #shortest paths passing through u
- **Closeness centrality** = avg. length of shortest paths from u to all other nodes of the network
- **Eigenvector centrality** = like PageRank

Link Analysis

- **Goal** (back to the newspaper example):
  - Don't just find newspapers. Find "experts" – pages that link in a coordinated way to good newspapers
- **Idea: Links as votes**
  - **Page is more important if it has more links**
    - In-coming links? Out-going links?
- **Hubs and Authorities**

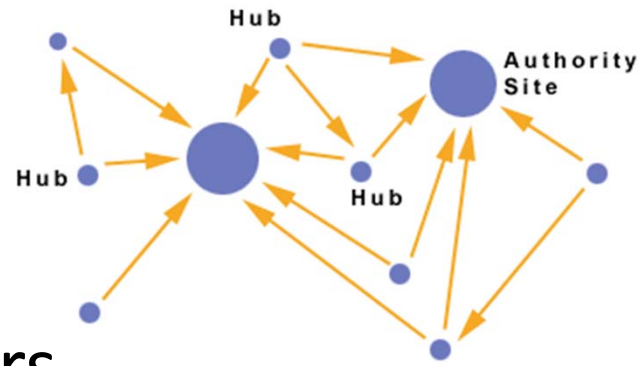
  Each page has 2 scores:
  - **Quality as an expert (hub):**
    - Total sum of votes of pages pointed to
  - **Quality as an content (authority):**
    - Total sum of votes of experts
  - Principle of repeated improvement



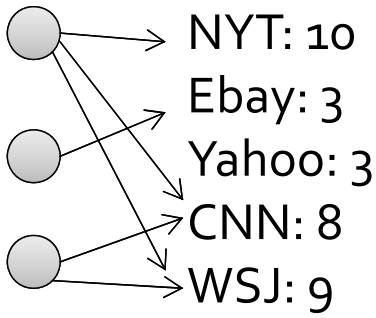
Hubs and Authorities

Interesting pages fall into two classes:

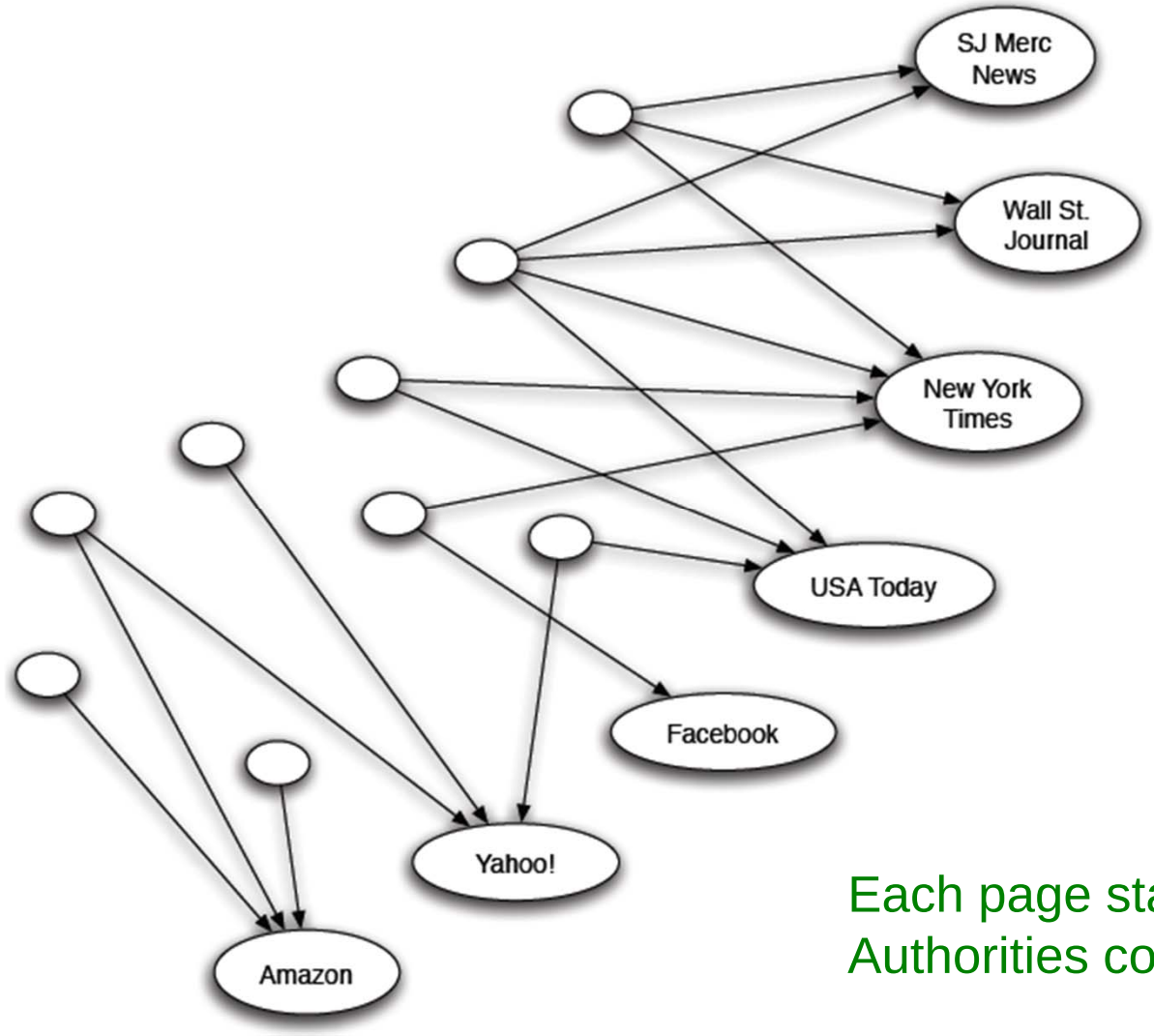
1. **Authorities** are pages containing useful information
   - Newspaper home pages
   - Course home pages
   - Home pages of auto manufacturers



2. **Hubs** are pages that link to authorities
   - List of newspapers
   - Course bulletin
   - List of US auto manufacturers



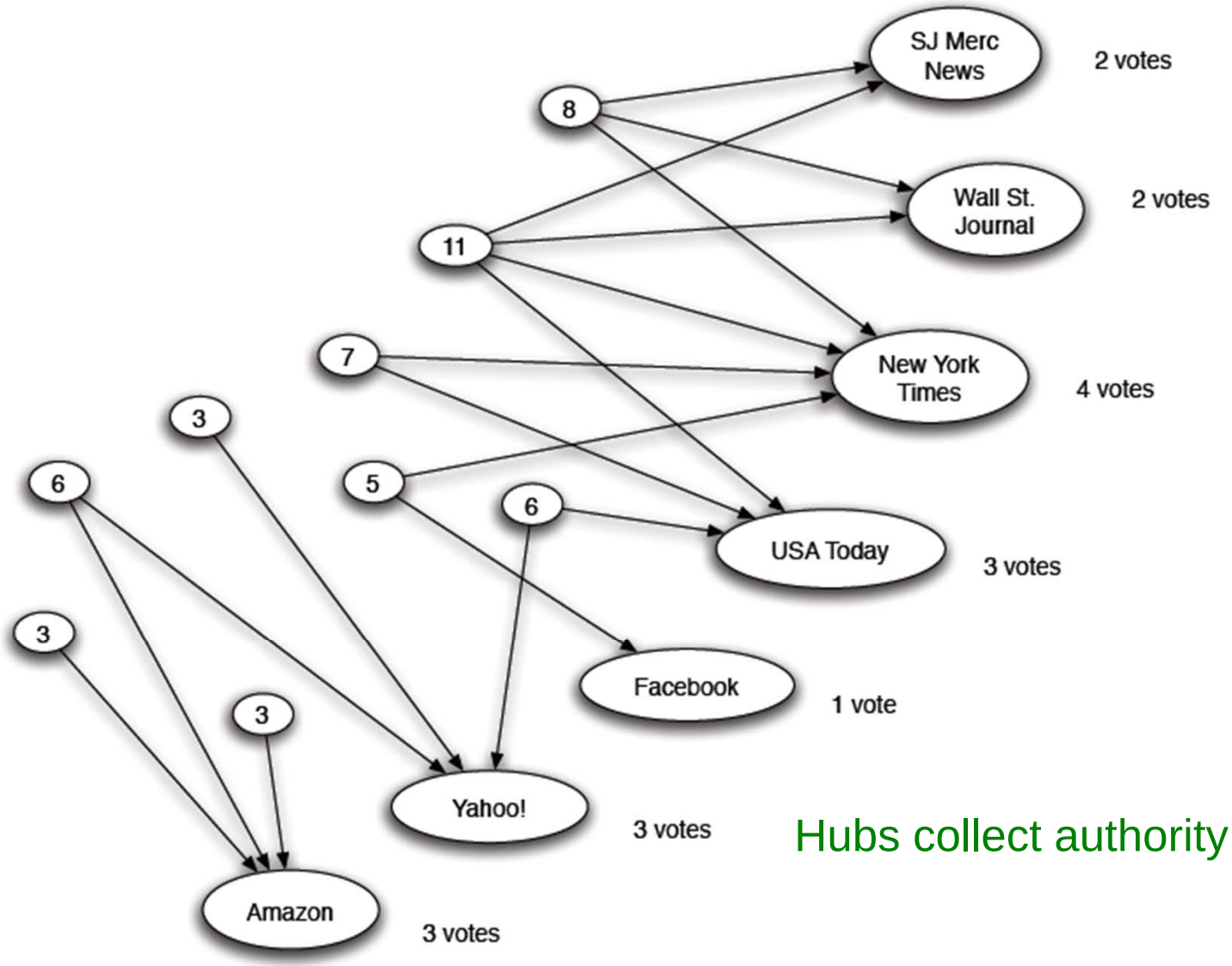
Counting in-links: Authority



Each page starts with hub score 1
Authorities collect their votes

(Note this is idealized example. In reality graph is not bipartite and each page has both the hub and authority score)

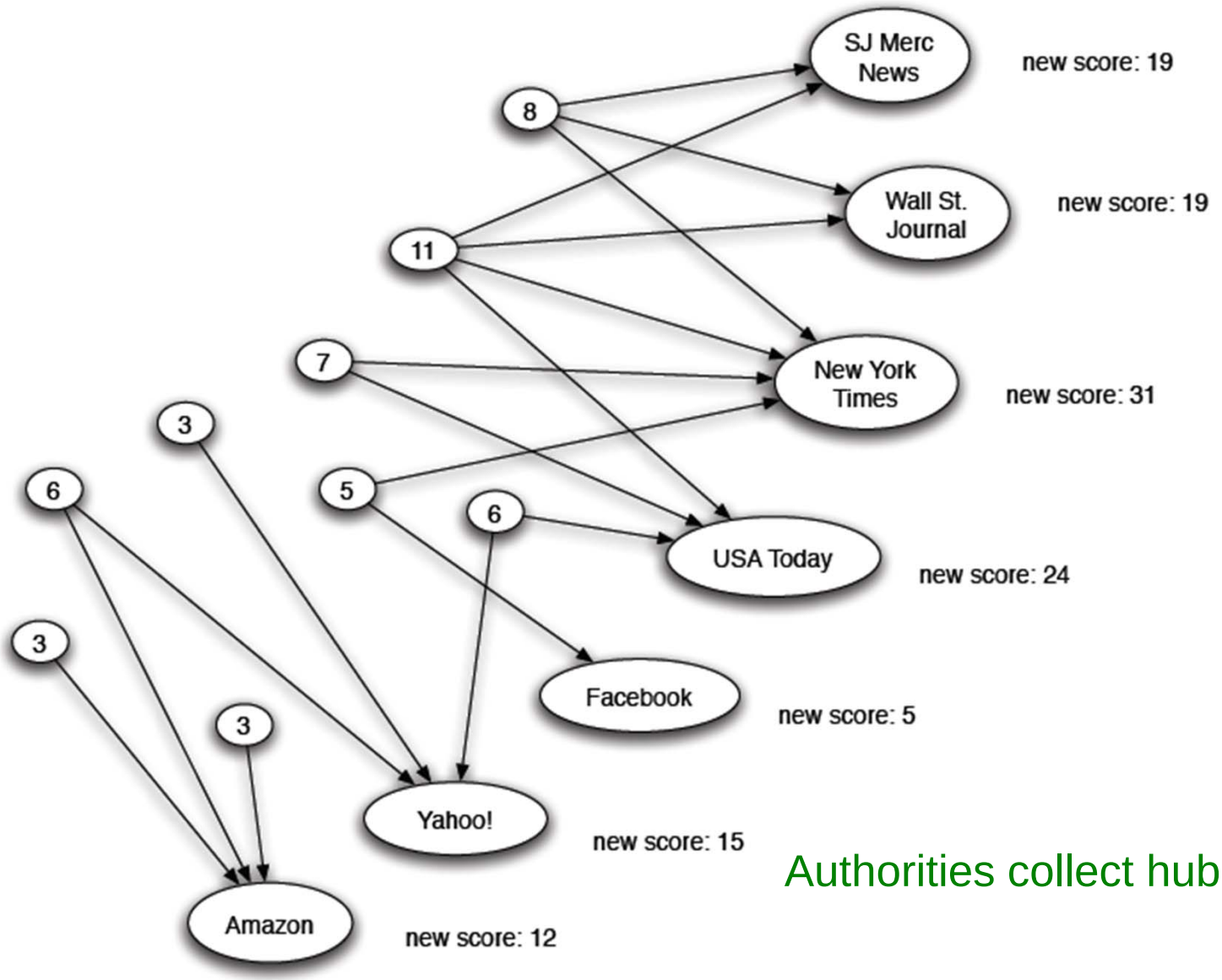
Expert Quality: Hub



Hubs collect authority scores

(Note this is idealized example. In reality graph is not bipartite and each page has both the hub and authority score)

Reweighting



Authorities collect hub scores

(Note this is idealized example. In reality graph is not bipartite and each page has both the hub and authority score)

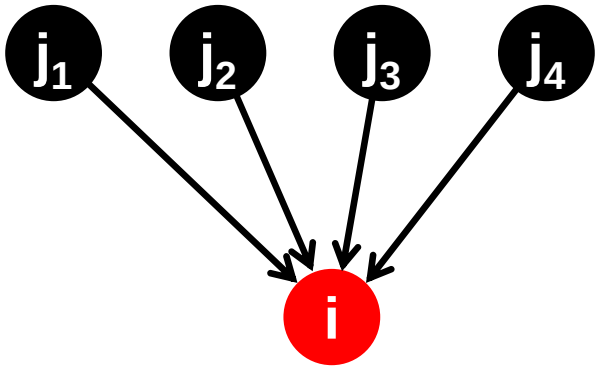
Mutually Recursive Definition

- **A good hub links to many good authorities**
- **A good authority is linked from many good hubs**
- **Model using two scores for each node:**
  - **Hub** score and **Authority** score
  - Represented as vectors h and a

[Kleinberg '98]

Hubs and Authorities

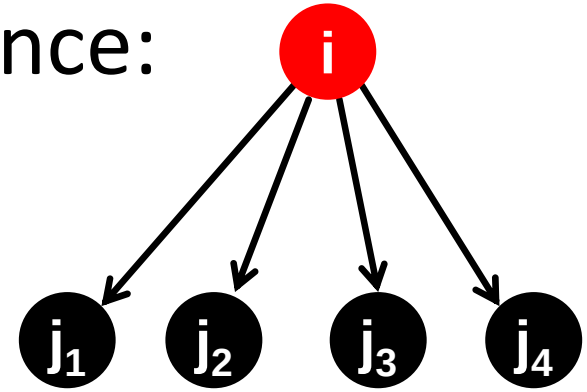
- **Each page i has 2 scores:**
  - Authority score: a_i
  - Hub score: h_i



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

HITS algorithm:

- Initialize: $a_j = 1/\sqrt{n},\ h_i = 1/\sqrt{n}$
- Then keep iterating until convergence:
  - $\forall i$: Authority: $a_i = \sum_{j \to i} h_j$
  - $\forall i$: Hub: $h_i = \sum_{i \to j} a_j$
  - $\forall i$: Normalize:
    $\sum_i a_i^2 = 1, \sum_j h_j^2 = 1$



$$h_i = \sum_{i \to j} a_j$$

[Kleinberg '98]

Hubs and Authorities

- **HITS converges to a single stable point**
- **Notation:**
  - Vector $a = (a_1 \dots, a_n)$, $h = (h_1 \dots, h_n)$
  - Adjacency matrix $\boldsymbol{A}$ (n x n): $A_{ij} = 1$ if $i \rightarrow j$
- **Then** $h_i = \sum_{i \rightarrow j} a_j$
  can be rewriten as $h_i = \sum_j A_{ij} \cdot a_j$

- **So:** $\boldsymbol{h} = \boldsymbol{A} \cdot \boldsymbol{a}$
- **And likewise:** $\boldsymbol{a} = \boldsymbol{A}^{\boldsymbol{T}} \cdot \boldsymbol{h}$

Hubs and Authorities

- **HITS algorithm in vector notation:**
  - Set: $a_i = h_i = \frac{1}{\sqrt{n}}$

  Repeat until convergence:
  - $h = A \cdot a$
  - $a = A^T \cdot h$
  - Normalize a and h
- **Then:** $\boldsymbol{a} = \underbrace{\boldsymbol{A}^{\boldsymbol{T}} \cdot (\underbrace{\boldsymbol{A} \cdot \boldsymbol{a}}_{\text{new } \boldsymbol{h}})}_{\text{new } \boldsymbol{a}}$
- **Thus, in $2k$ steps:**

$$a = (A^T \cdot A)^k \cdot a$$
$$h = (A \cdot A^T)^k \cdot h$$

Convergence criterion:

$$\sum_i \left(h_i^{(t)} - h_i^{(t-1)}\right)^2 < \varepsilon$$
$$\sum_i \left(a_i^{(t)} - a_i^{(t-1)}\right)^2 < \varepsilon$$

a is updated (in 2 steps):

$$a = A^T (A\, a) = (A^T A)\, a$$

h is updated (in 2 steps):

$$h = A\,(A^T h) = (A\, A^T)\, h$$

Repeated matrix powering

Eigenvalues & Eigenvectors

- **Definition:**
  - Let $R \cdot x = \lambda \cdot x$
    for some scalar λ, vector x, matrix R
  - Then $\boldsymbol{x}$ **is an eigenvector**, and λ **is its eigenvalue**
- **Fact:**
  - If R is symmetric ($R_{ij} = R_{ji}$)
    (in our case $R = A^T \cdot A$ and $\mathrm{R} = A \cdot A^T$ are symmetric)
  - Then R has n orthogonal unit eigenvectors $w_1 \ldots w_n$ that form a basis (coordinate system) with eigenvalues $\lambda_1 \ldots \lambda_n$ $\quad (|\lambda_i| \geq |\lambda_{i+1}|)$

How to Think About A·x?

$Rx = \lambda x$

- Let's write x in coordinate system $w_1 \dots w_n$

  $x = \lambda_i \, \alpha_i \, w_i$

  - x has coordinates $(\alpha_1, \dots, \alpha_n)$
- **Suppose:** $\lambda_1 \dots \lambda_n \quad (|\lambda_1| \geq \dots \geq |\lambda_n|)$
- **Then:** $R^k x = \lambda^k x = \sum_i \lambda_i^k \, \alpha_i \, w_i$
- **As** $k \to \infty$, if we normalize

  $R^k x \to \lambda_1 \, \alpha_1 \, w_1$

  (contribution of all other coordinates $\to 0$)

  $\lim_{k\to\infty} \frac{{\lambda_1}^k}{{\lambda_2}^k} = \left(\frac{\lambda_1}{\lambda_2}\right)^k \to \infty$
- So, **authority** a is eigenvector of $\mathrm{R} = A^T A$ associated with largest eigenvalue λ_1
- Similarly: **hub** h is eigenvector of $\mathrm{R} = A \, A^T$

PageRank

Links as Votes

- **Still the same idea: Links as votes**
  - **Page is more important if it has more links**
    - In-coming links? Out-going links?
- **Think of in-links as votes:**
  - www.stanford.edu has 23,400 in-links
  - www.joe-schmoe.com has 1 in-link
- **Are all in-links are equal?**
  - Links from important pages count more
  - Recursive question!

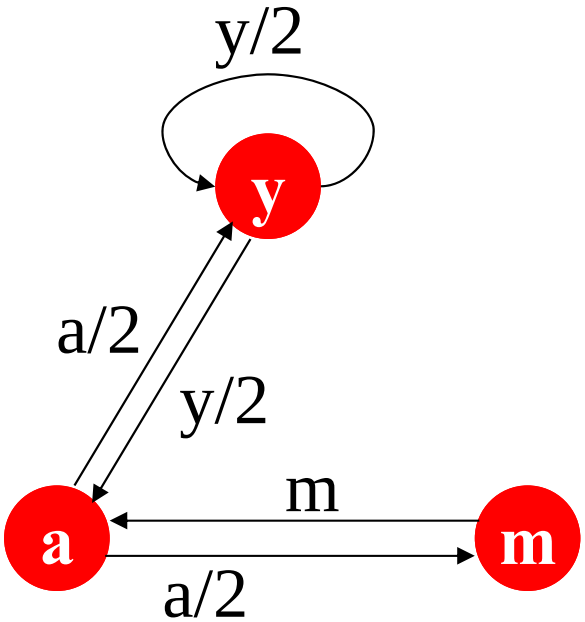
PageRank: The "Flow" Model

- A "vote" from an important page is worth more
- A page is important if it is pointed to by other important pages
- Define a "rank" r_j for node j

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

d_i … out-degree of node i

The web in 1839



"Flow" equations:

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

PageRank: Matrix Formulation

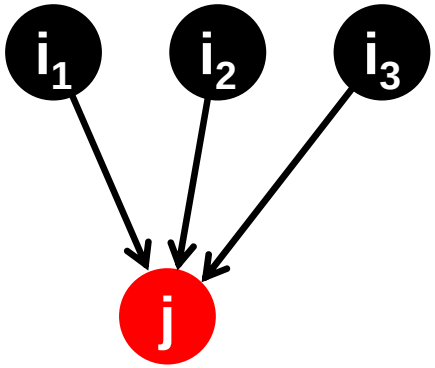
- **Stochastic adjacency matrix $\boldsymbol{M}$**
  - Let page j has d_j out-links
  - If $j \rightarrow i$, then $M_{ij} = \frac{1}{d_j}$ else $M_{ij} = 0$
    - $\boldsymbol{M}$ is a **column stochastic matrix**
      - Columns sum to 1
- **Rank vector $\boldsymbol{r}$:** vector with an entry per page
  - r_i is the importance score of page i
  - $\sum_i r_i = 1$
- **The flow equations can be written**

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

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

Random Walk Interpretation

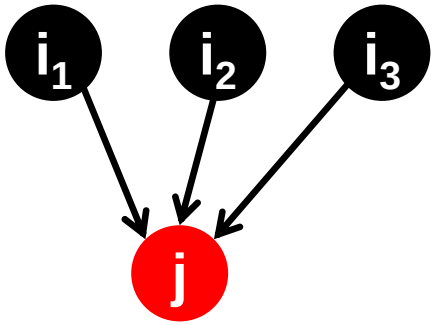
- **Imagine a random web surfer:**
  - At any time t, surfer is on some page i
  - At time $t+1$, the surfer follows an out-link from i uniformly at random
  - Ends up on some page j linked from i
  - Process repeats indefinitely



$$r_j = \sum_{i \to j} \frac{r_i}{\mathrm{d}_{\mathrm{out}}(\mathrm{i})}$$

- **Let:**
  - $\boldsymbol{p}(\boldsymbol{t})$ … vector whose i^{th} coordinate is the prob. that the surfer is at page i at time t
  - So, $\boldsymbol{p}(\boldsymbol{t})$ is a probability distribution over pages

The Stationary Distribution



$$p(t+1) = \mathrm{M} \cdot p(t)$$

- **Where is the surfer at time *t+1*?**
  - Follows a link uniformly at random

$$p(t+1) = M \cdot p(t)$$

- Suppose the random walk reaches a state

$$p(t+1) = M \cdot p(t) = p(t)$$

then $\boldsymbol{p}(t)$ is stationary distribution of a random walk

- **Our original rank vector** $\boldsymbol{r}$ satisfies $\boldsymbol{r} = \boldsymbol{M} \cdot \boldsymbol{r}$
  - **So, r is a stationary distribution for the random walk**

PageRank: How to solve?

Given a web graph with n nodes, where the nodes are pages and edges are hyperlinks

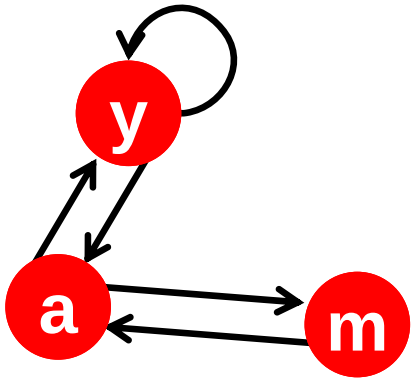
- Assign each node an initial page rank
- Repeat until convergence ($\Sigma_i |\mathbf{r}_i^{(t+1)} - \mathbf{r}_i^{(t)}| < \varepsilon$)
  - Calculate the page rank of each node

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

d_i …. out-degree of node i

PageRank: How to solve?

- **Power Iteration:**
  - Set $r_j = 1$
  - $r_j = \sum_{i \to j} \frac{r_i}{d_i}$
    - And iterate



	y	a	m
y	½	½	0
a	½	0	1
m	0	½	0

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

- **Example:**

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

Iteration 0, 1, 2, …

PageRank: Three Questions

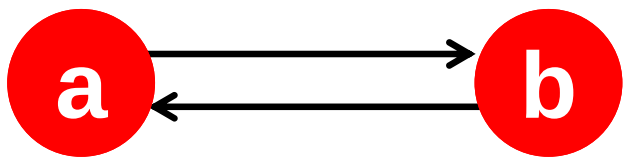
$$r_j^{(t+1)} = \sum_{i \to j} \frac{r_i^{(t)}}{\mathrm{d_i}}$$

or equivalently

$$r = Mr$$

- **Does this converge?**
- **Does it converge to what we want?**
- **Are results reasonable?**

Does this converge?



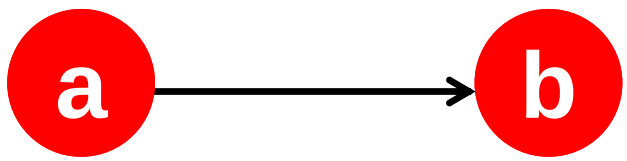
$$r_j^{(t+1)} = \sum_{i \to j} \frac{r_i^{(t)}}{\mathrm{d_i}}$$

- **Example:**

r_a	=	1	0	1	0
r_b		0	1	0	1

Iteration 0, 1, 2, …

Does it converge to what we want?



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

- **Example:**

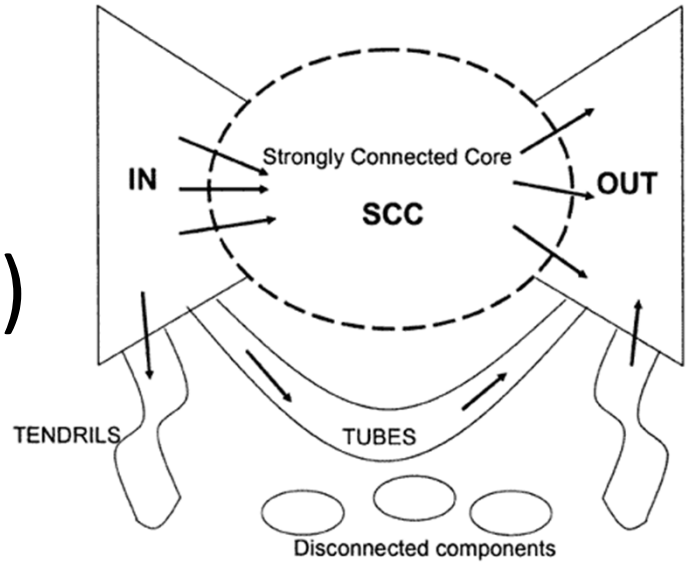
$$\begin{matrix} r_a \\ r_b \end{matrix} = \begin{matrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{matrix}$$

Iteration 0, 1, 2, …

RageRank: Problems

2 problems:

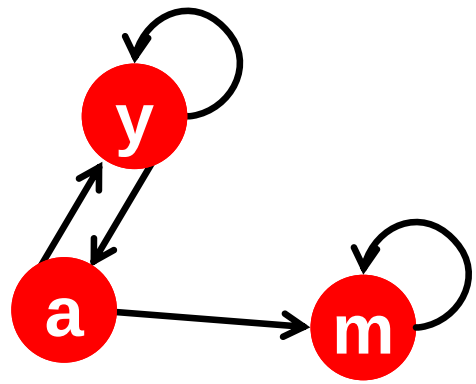
- **(1)** Some pages are **dead ends** (have no out-links)
  - Such pages cause importance to "leak out"



- **(2) Spider traps** (all out-links are within the group)
  - Eventually spider traps absorb all importance

Problem: Spider Traps

- **Power Iteration:**
  - Set $r_j = 1$
  - $r_j = \sum_{i \to j} \frac{r_i}{d_i}$
    - And iterate



	y	a	m
y	½	½	0
a	½	0	0
m	0	½	1

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

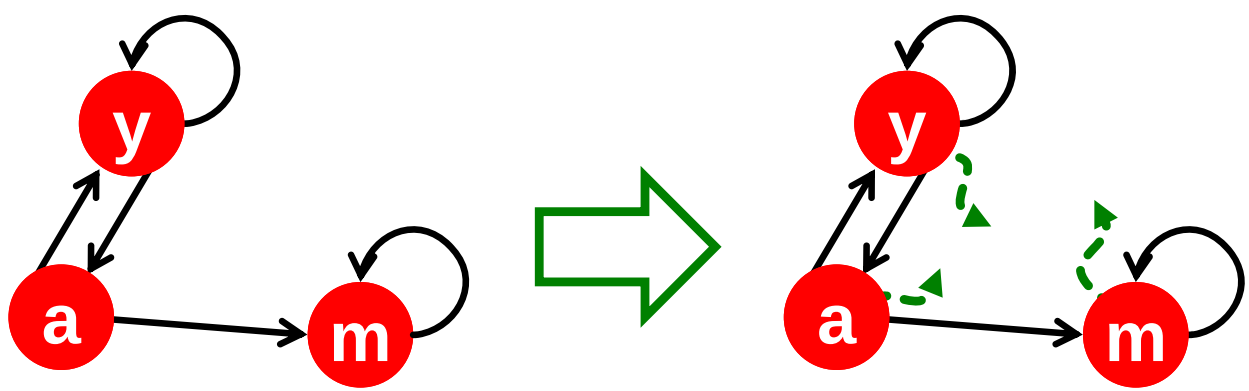
- **Example:**

$$\begin{pmatrix} r_y \\ r_a \\ r_m \end{pmatrix} = \begin{matrix} 1/3 & 2/6 & 3/12 & 5/24 & & 0 \\ 1/3 & 1/6 & 2/12 & 3/24 & \dots & 0 \\ 1/3 & 3/6 & 7/12 & 16/24 & & 1 \end{matrix}$$

Iteration 0, 1, 2, …

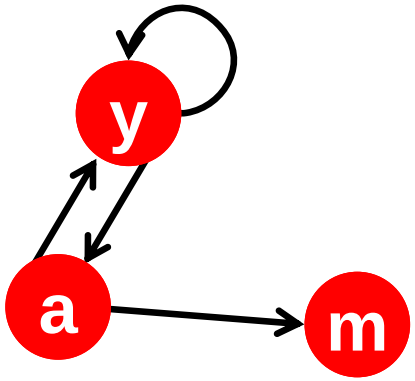
Solution: Random Teleports

- **The Google solution for spider traps: At each time step, the random surfer has two options**
  - With prob. β, follow a link at random
  - With prob. 1-β, jump to some page uniformly at random
  - Common values for β are in the range 0.8 to 0.9
- **Surfer will teleport out of spider trap within a few time steps**



Problem: Dead Ends

- **Power Iteration:**
  - Set $r_j = 1$
  - $r_j = \sum_{i \to j} \frac{r_i}{d_i}$
    - And iterate



	y	a	m
y	½	½	0
a	½	0	0
m	0	½	0

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

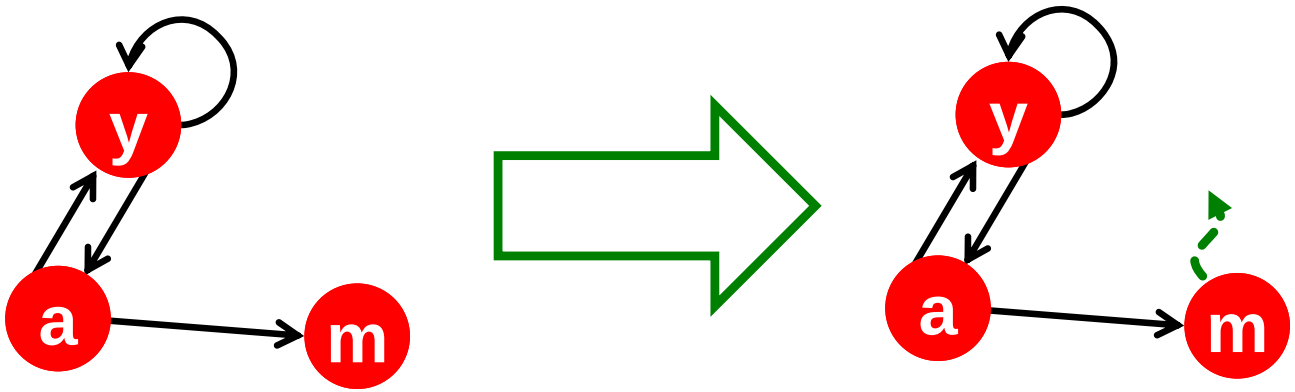
- **Example:**

$$\begin{pmatrix} r_y \\ r_a \\ r_m \end{pmatrix} = \begin{matrix} 1/3 & 2/6 & 3/12 & 5/24 & & 0 \\ 1/3 & 1/6 & 2/12 & 3/24 & \dots & 0 \\ 1/3 & 1/6 & 1/12 & 2/24 & & 0 \end{matrix}$$

Iteration 0, 1, 2, …

Solution: Always Teleport

- **Teleports:** Follow random teleport links with probability 1.0 from dead-ends
  - Adjust matrix accordingly



	y	a	m
y	½	½	0
a	½	0	0
m	0	½	0

	y	a	m
y	½	½	⅓
a	½	0	⅓
m	0	½	⅓

Solution: Random Jumps

- **Google's solution: At each step, random surfer has two options:**
  - With probability $1-\beta$, follow a link at random
  - With probability β, jump to some random page
- **PageRank equation** [Brin-Page, 98]

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

d_i … out-degree of node i

The above formulation assumes that M has no dead ends. We can either preprocess matrix M **(bad!)** or explicitly follow random teleport links with probability 1.0 from dead-ends. See P. Berkhin, *A Survey on PageRank Computing*, Internet Mathematics, 2005.

PageRank & Eigenvectors

- **PageRank as a principal eigenvector**

  $r = M \cdot r$ or equivalently $r_j = \sum_i \frac{r_i}{d_i}$

- **But we really want:**

  $$r_j = \beta \sum_i \frac{r_i}{d_i} + (1-\beta)\frac{1}{n}$$

- **Let's define:**

  $$M'_{ij} = \beta M_{ij} + (1-\beta)\frac{1}{n}$$

- **Now we get what we want:**

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

- **What is $1-\beta$?**
  - In practice 0.15 (5 links and jump)

d_i ... out-degree of node i

Note: M is a sparse matrix but M' is dense (all entries $\neq 0$). In practice we never "materialize" M but rather we use the "sum" formulation

PageRank: The Complete Algorithm

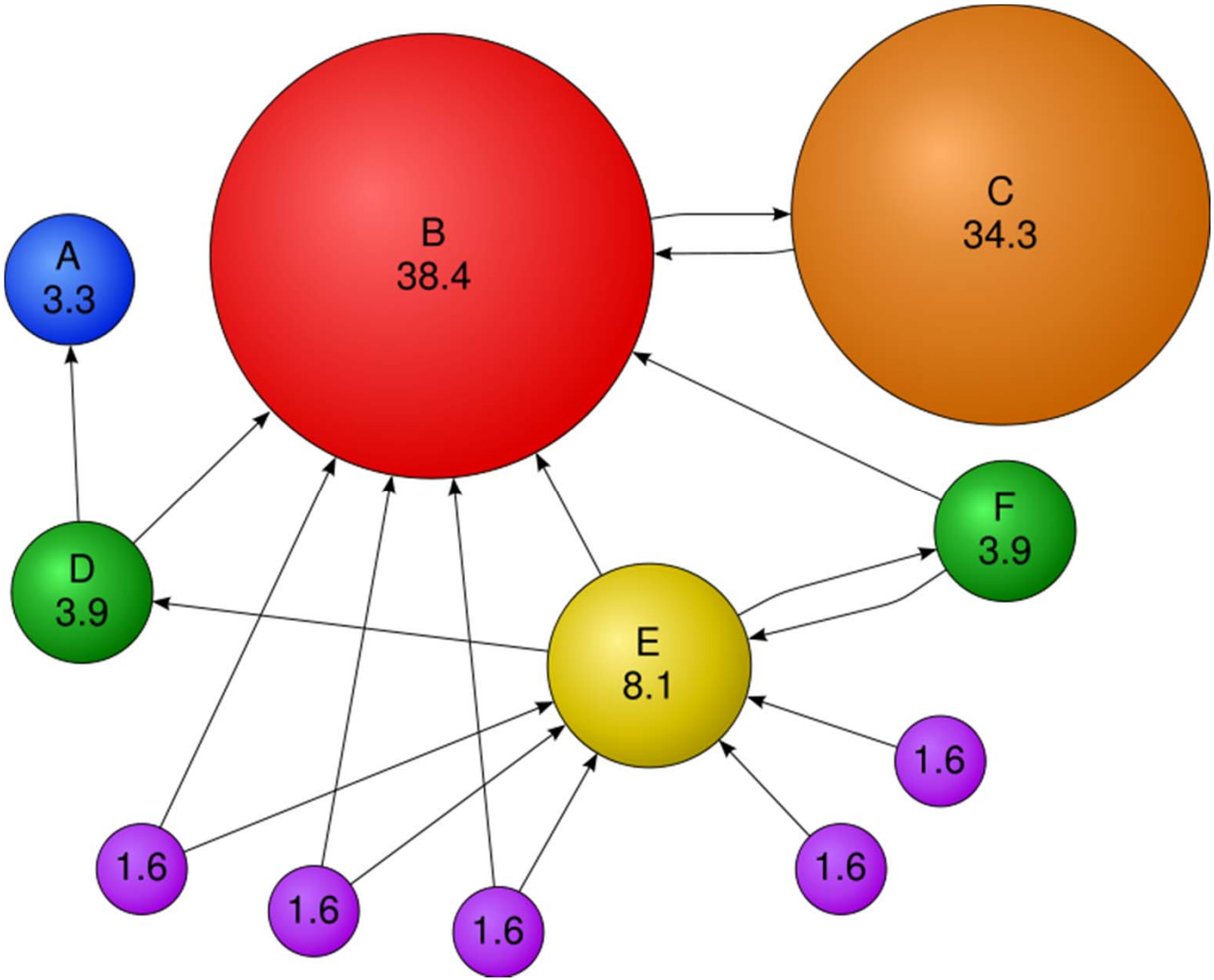
See P. Berkhin, *A Survey on PageRank Computing*, Internet Mathematics, 2005.

- **Input: A and β**
  - Adjacency matrix A of a directed graph with spider traps and dead ends
  - Parameter β
- **Output: PageRank vector r**
  - Set: $r_j^{(0)} = 1/n$
  - **Repeat until:** $\sum_j \left| r_j^{(t)} - r_j^{(t-1)} \right| < \varepsilon$
    - $\forall j$: $r'^{(t)}_j = \sum_{i \to j} \beta \frac{r_i^{(t-1)}}{d_i}$, if in-deg. of j is 0 then $r'^{(t)}_j = 0$
    - Now re-insert the leaked PageRank:

$$\forall j: r_j^{(t)} = r'^{(t)}_j + (1 - S)/n$$

Where: $S = \sum_j r'^{(t)}_j$

Example



PageRank and HITS

- **PageRank and HITS are two solutions to the same problem:**
  - **What is the value of an in-link from *u* to *v*?**
  - In the PageRank model, the value of the link depends on the links **into** *u*
  - In the HITS model, it depends on the value of the other links **out of** *u*
- **The destinies of PageRank and HITS post-1998 were very different**

Personalized PageRank and Applications

Personalized PageRank

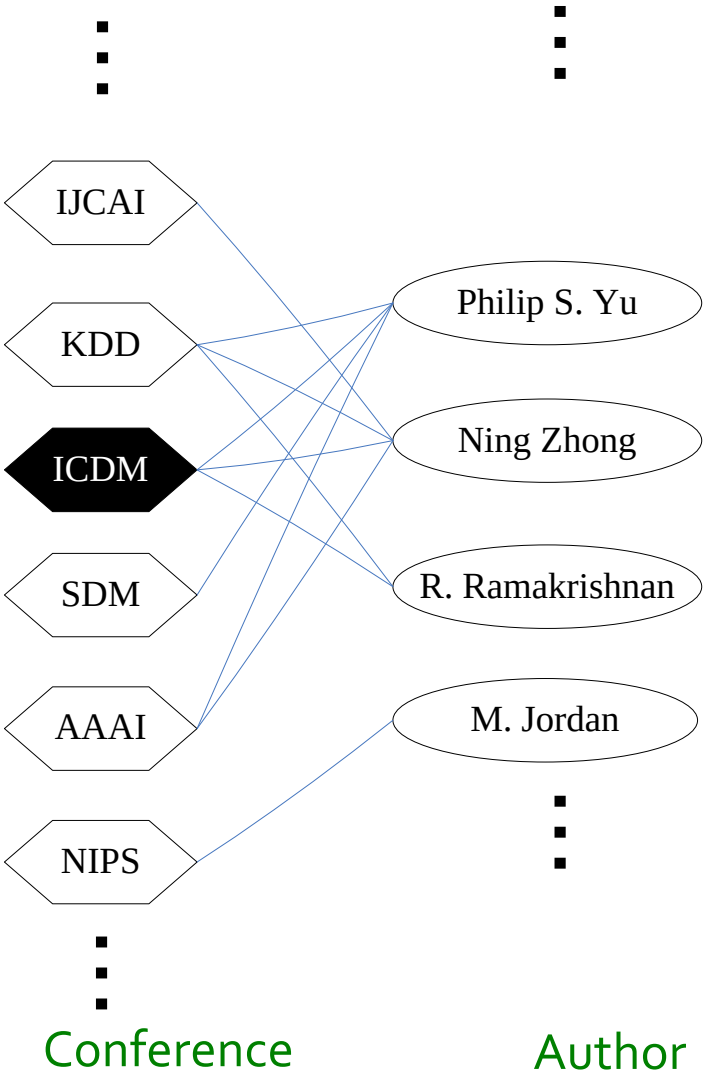
- **Goal:** Evaluate pages not just by popularity but by how close they are to the topic
- **Teleporting can go to:**
  - **Any page with equal probability**
    - (we used this so far)
  - **A topic-specific set of "relevant" pages**
    - Topic-specific (personalized) PageRank (S ...teleport set)

$$M'_{ij} = (1-\beta)\, M_{ij} + \beta / |S| \quad \text{if } i \in S$$
$$= (1-\beta)\, M_{ij} \quad \text{otherwise}$$

    - Useful for measuring "proximity" of other nodes to S

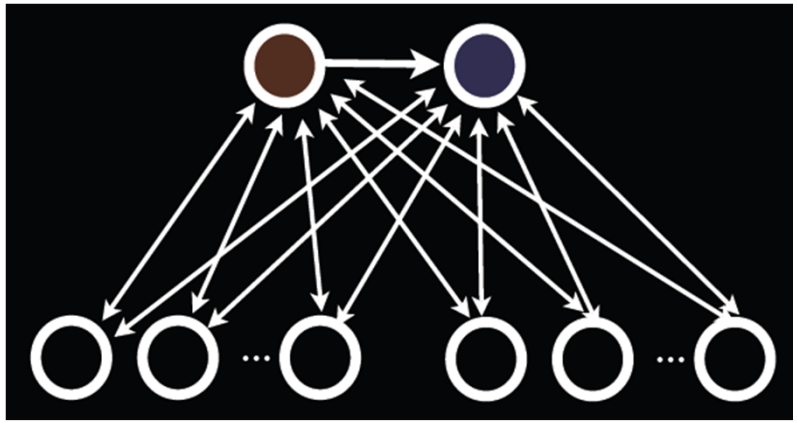
PageRank: Applications

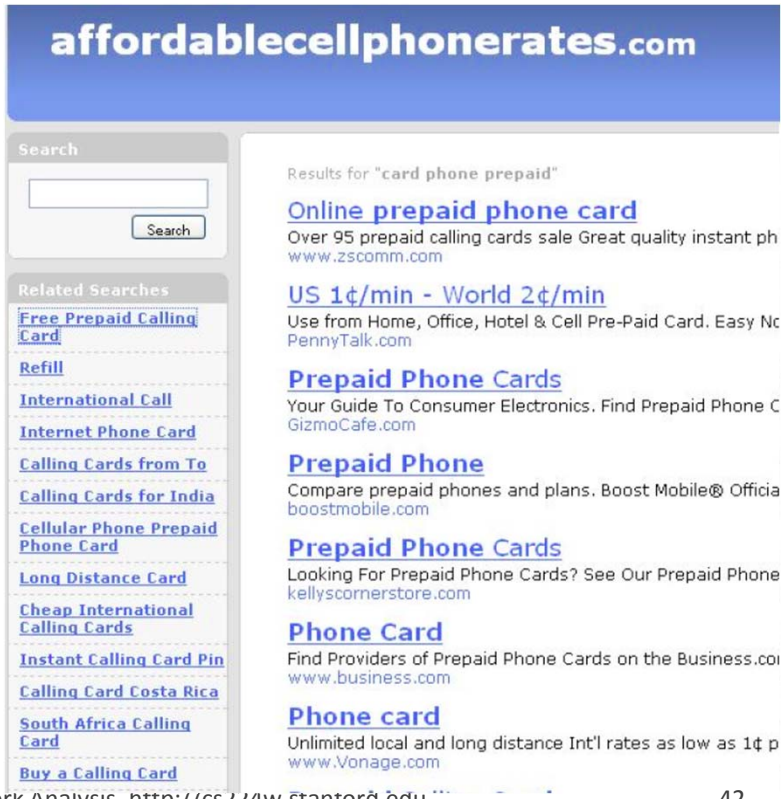
- **Graphs and web search:**
  - Ranks nodes by "importance"
- **Personalized PageRank:**
  - Ranks proximity of nodes to the teleport nodes $\boldsymbol{S}$
- **Proximity on graphs:**
  - **Q:** What is most related conference to ICDM?
  - **Random Walks with Restarts**
    - Teleport back to the starting node: $\boldsymbol{S}$ = { **single node** }



Application: TrustRank

- **Link Farms:** networks of millions of pages design to focus PageRank on a few undeserving webpages
- To minimize their influence use a teleport set of trusted webpages
  - E.g., homepages of universities





PageRank: Problems

Google Bombs

Web Images Groups News Froogle Local more »

Google

miserable failure | Search | Advanced Search | Preferences

Web Results **1** - **10** of about **969,000** for **miserable failure**. (**0.06** seconds)

Biography of President George W. Bush
Biography of the president from the official White House web site.
www.whitehouse.gov/president/gwbbio.html - 29k - Cached - Similar pages
Past Presidents - Kids Only - Current News - President
More results from www.whitehouse.gov »

Welcome to MichaelMoore.com!
Official site of the gadfly of corporations, creator of the film Roger and Me and the television show The Awful Truth. Includes mailing list, message board, ...
www.michaelmoore.com/ - 35k - Sep 1, 2005 - Cached - Similar pages

BBC NEWS | Americas | '**Miserable failure**' links to Bush
Web users manipulate a popular search engine so an unflattering description leads to the president's page.
news.bbc.co.uk/2/hi/americas/3298443.stm - 31k - Cached - Similar pages

Google's (and Inktomi's) **Miserable Failure**
A search for **miserable failure** on Google brings up the official George W. Bush biogra ... not a ...
searchengin ... ached - Simi

Google | Advanced Search | Preferences | Language Tools
french military victories | Google S
Search: the web
Web | Images | Groups | Directory | News

Did you mean: french military *defeats*

No standard web pages containing all your search terms were found.

Your search - **french military victories** - did not match any documents.

Suggestions:

- Make sure all words are spelled correctly.
- Try different keywords.
- Try more general keywords.
- Try fewer keywords.

Web Images Maps News Shopping Gmail more ▾

BloomSEO

Google

dangerous cult | Search

Web Results **1** - **10** of about **252,000** for **dangerous cult**. (**0.06** se

Scientology - Church of Scientology Official Site
Living in a **Dangerous** Environment · Drug and Alcohol Problems · Personalities, Emot and How to Deal with Others ...
www.scientology.org/ - 73k - Cached - Similar pages - Note this

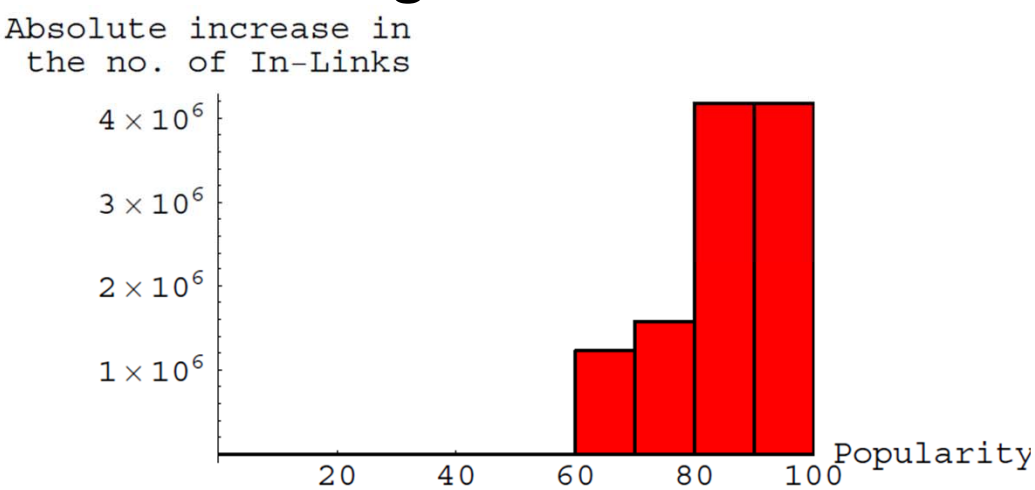
The Most **Dangerous Cult** in The World by Laura Knight-Jadczyk
There's a new religious **cult** in America. It's not composed of so-called "crazies" so mu mainstream, middle to upper-middle class Americans. ...
www.cassiopaea.org/cass/Laura-Knight-Jadczyk/fastest_growing_**cult**.htm - 144k - Cached - Similar pages - Note this

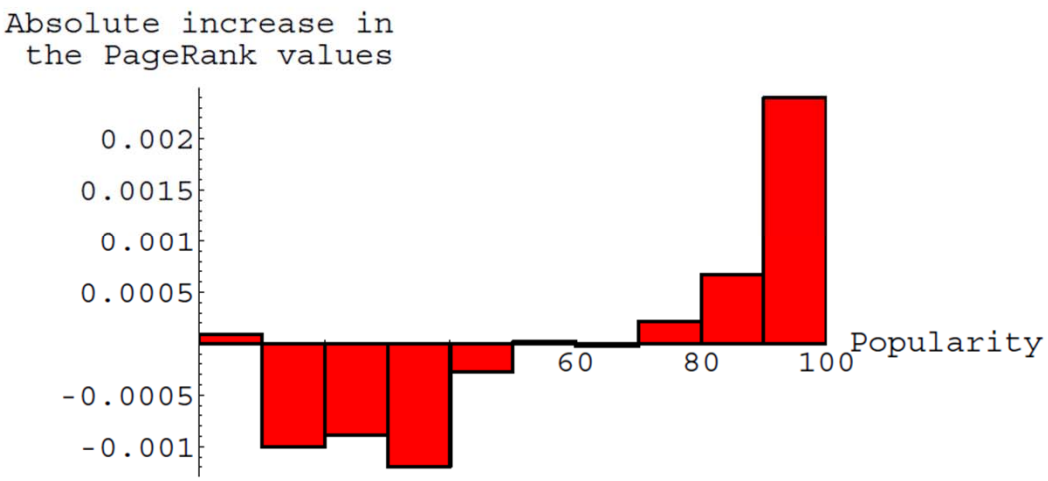
Dangerous Cult Warning Signs
If you, or a loved one, are in a **dangerous cult**, as determined by the above checklist, must do everything you possibly can to remove the potential ...
www.vistech.net/users/rsturge/**cults**.html - 4k - Cached - Similar pages - Note this

The Watchman Expositor: The Most **Dangerous Cult** in America
However, when the world's final chapter is written, which will prove to be "THE most **dangerous cult** in America?" One of the cults mentioned above? ...
www.watchman.org/reltop/budcomp.htm - 10k - Cached - Similar pages - Note this

Issues with PageRank

- **Rich get richer** [Cho et al., WWW '04]
  - Two snapshots of the web-graph at two different time points
  - Measure the change:
    - In the number of in-links
    - PageRank





http://oak.cs.ucla.edu/~cho/papers/cho-bias.pdf