

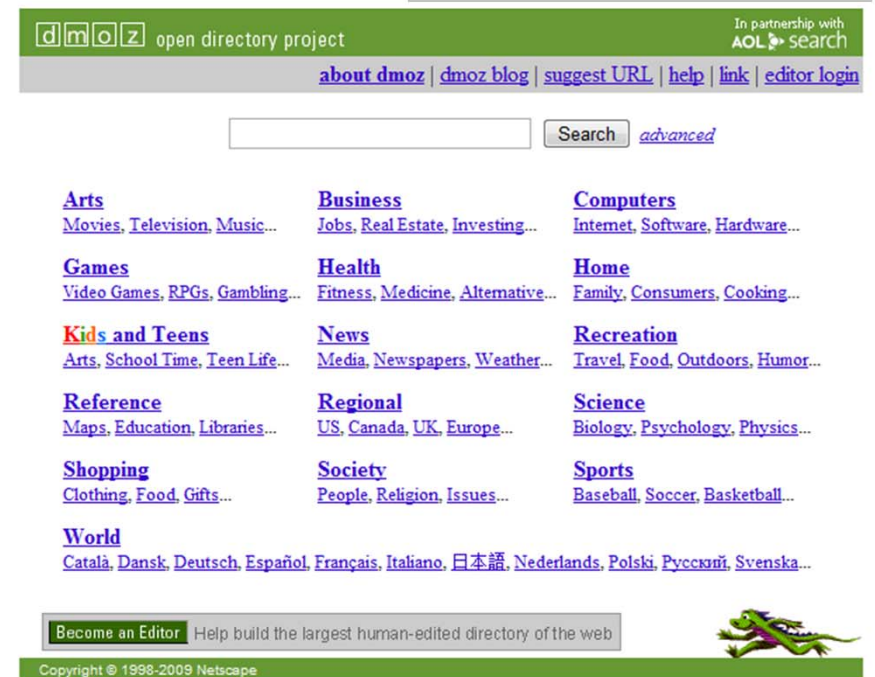
Link Analysis: PageRank and HITS

CS224W: Social and Information Network Analysis
Jure Leskovec, Stanford University
<http://cs224w.stanford.edu>



How to Organize the Web?

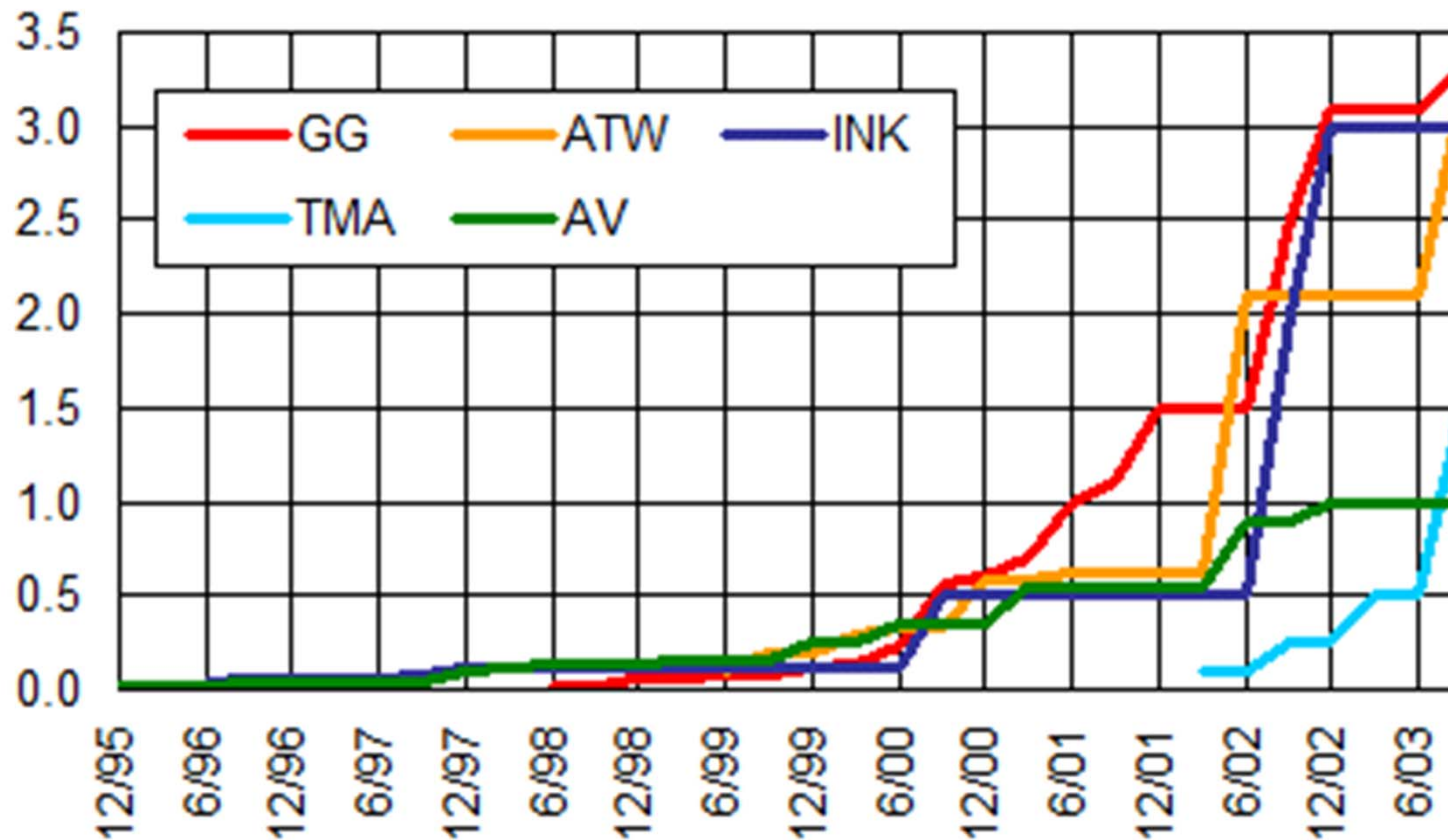
- How to organize/navigate it?
- First try: Human curated Web directories
 - Yahoo,
 - DMOZ,
 - LookSmart



Information Retrieval

- **SEARCH!**
- Find relevant docs in a small and trusted set:
 - Newspaper articles
 - Patents, etc.
- **Two traditional problems:**
 - Synonymy: buy – purchase, sick – ill
 - Polysemi: jaguar

The Index Size Wars

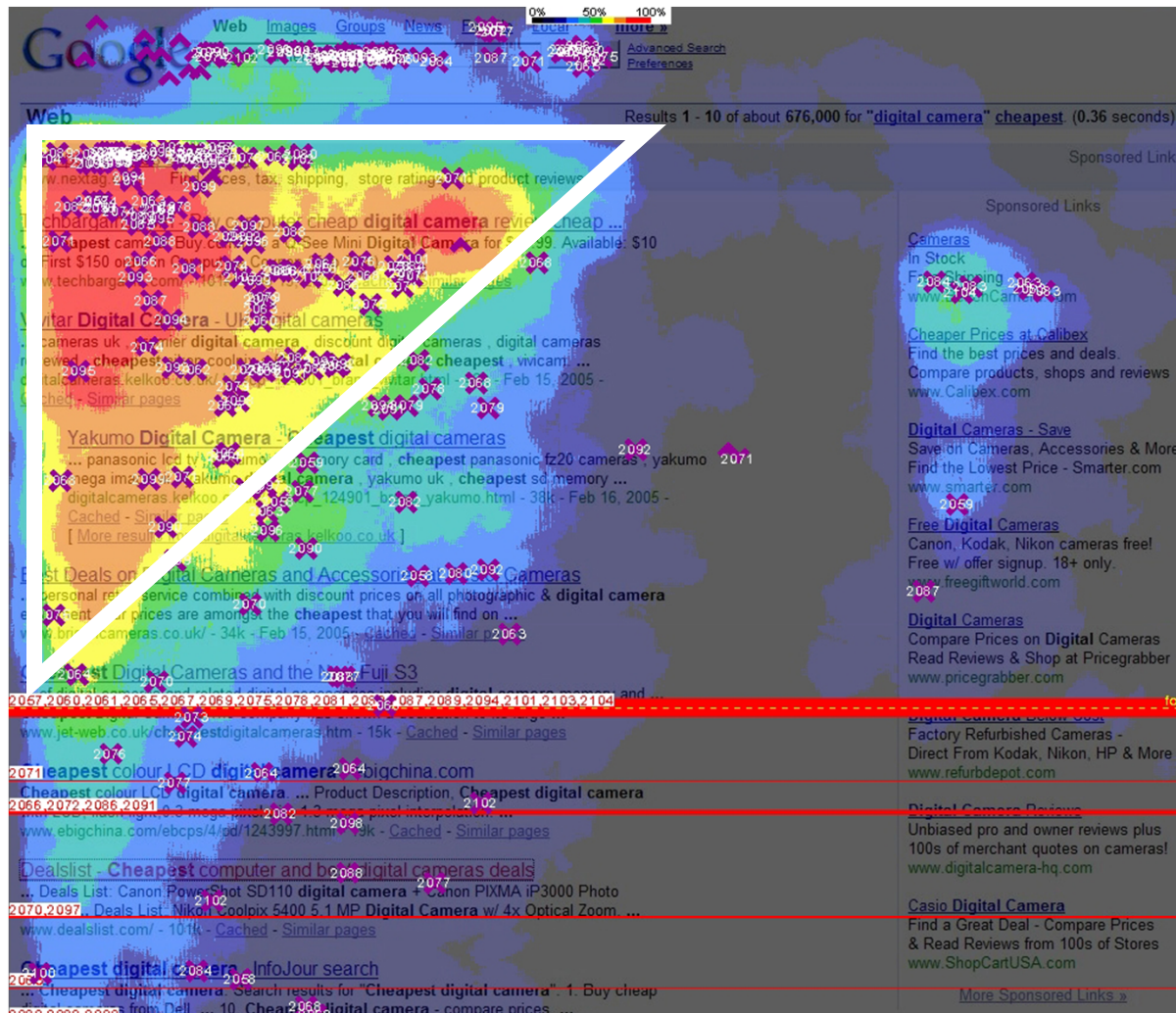


Does more documents mean better results?

Web Search vs. Inf. Retrieval

- What is “best” answer to query “Stanford”?
 - Anchor Text: I go to [Stanford](#) where I study
- What about query “newspaper”?
 - No single right answer
- **Scarcity** (IR) vs. **abundance** (Web) of information
 - Web: Many sources of information. Who to “trust”?
- **Trick**:
 - Pages that actually know about newspapers might all be pointing to many newspapers
- **Ranking!**

Ranking: Where do people look at?



the “golden triangle”

Ranking Nodes on the Graph

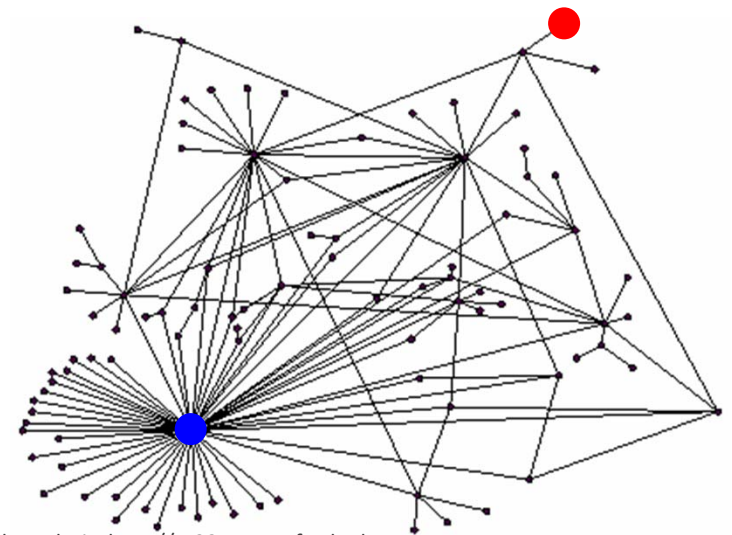
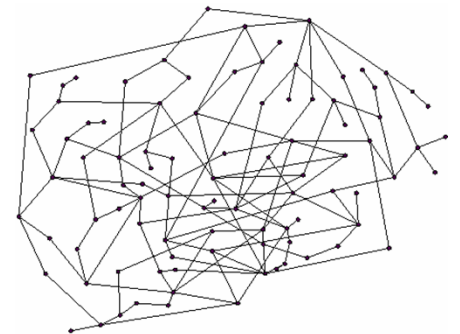
- Web pages are not equally “important”
 - www.joe-schmoe.com vs. www.stanford.edu

- **We already know:**

Since there is large diversity in the connectivity of the

webgraph we can

**rank the pages by
the link structure**



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

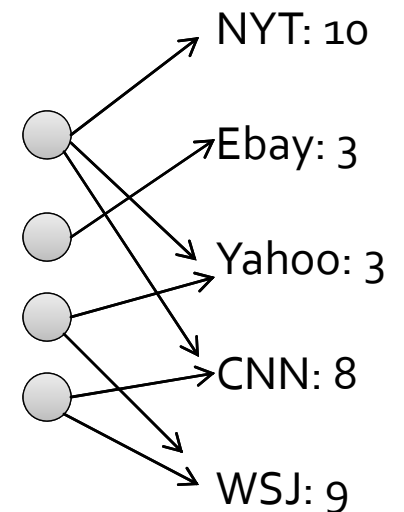
- **Degree centrality** = degree of u
- **Betweenness centrality** = #shortest paths passing through u
- **Closeness centrality** = avg. length of shortest paths from u to all other nodes
- **Eigenvector centrality** = like PageRank

Link Analysis

- **Goal** (back to the newspaper example):
 - Don't just find newspapers. Find "experts" – people who 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

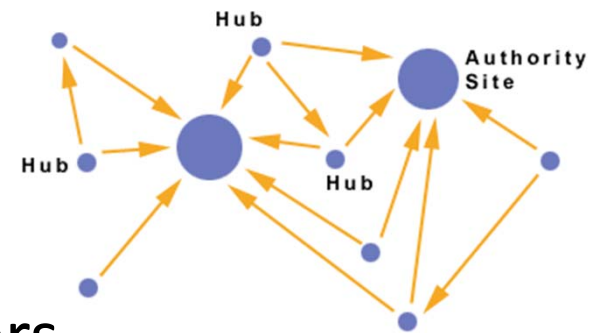


Hubs and Authorities

Interesting pages fall into two classes:

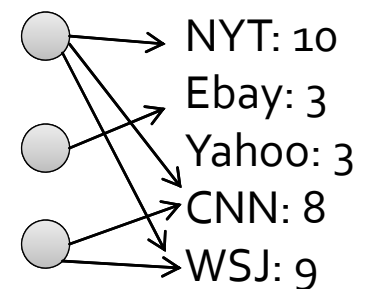
1. **Authorities** are pages containing useful information

- Newspaper home pages
- Course home pages
- Home pages of auto manufacturers

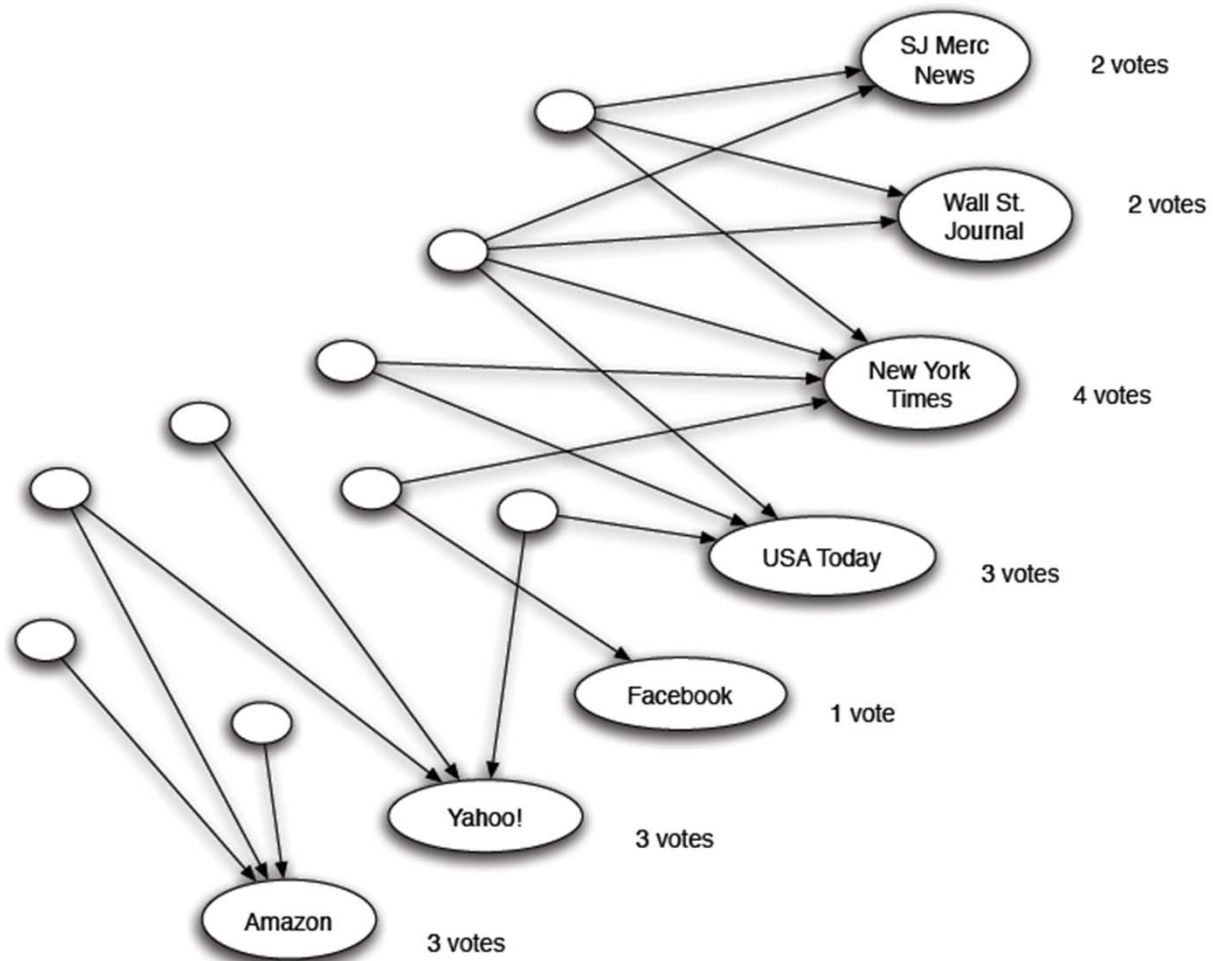


2. **Hubs** are pages that link to authorities

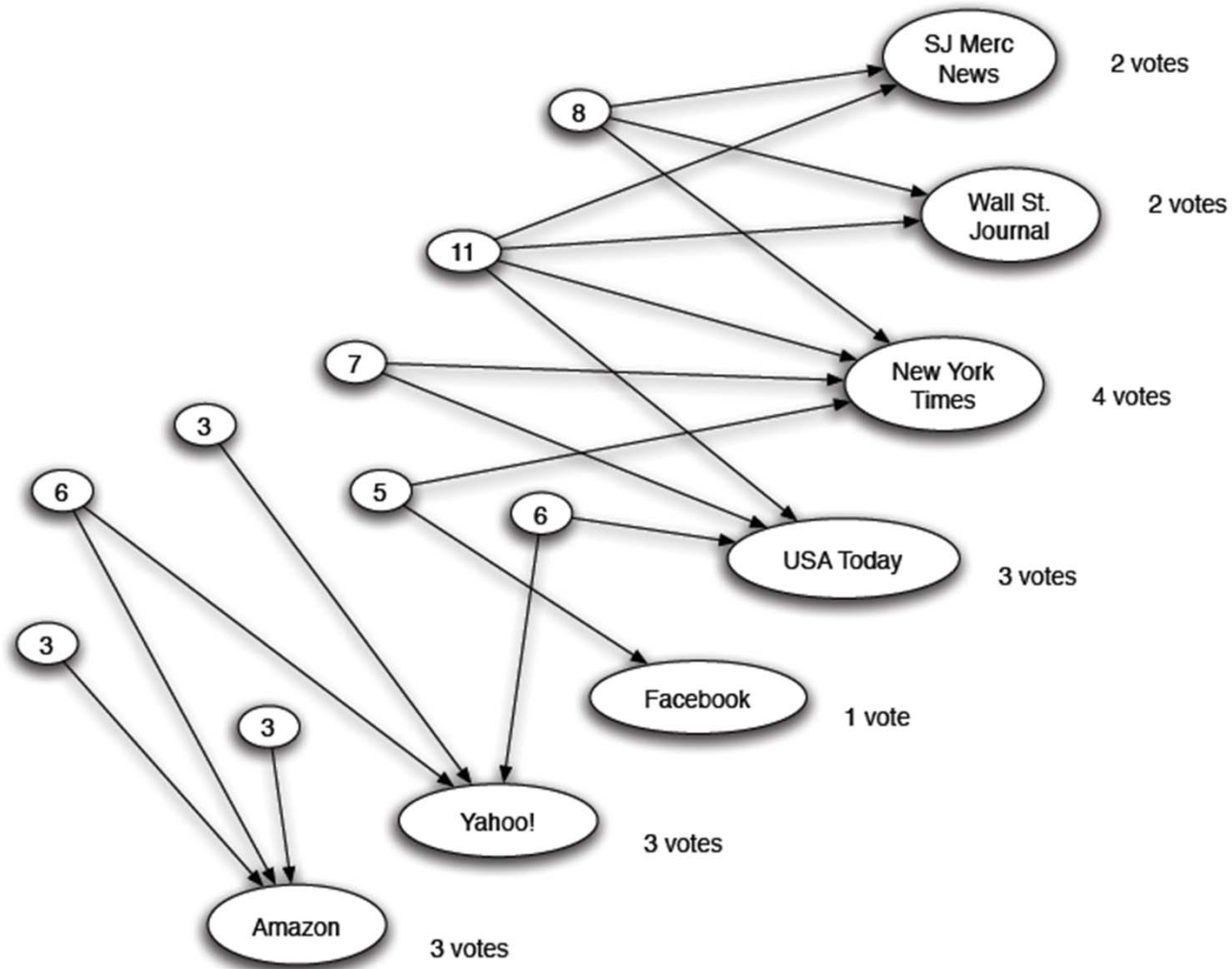
- List of newspapers
- Course bulletin
- List of US auto manufacturers



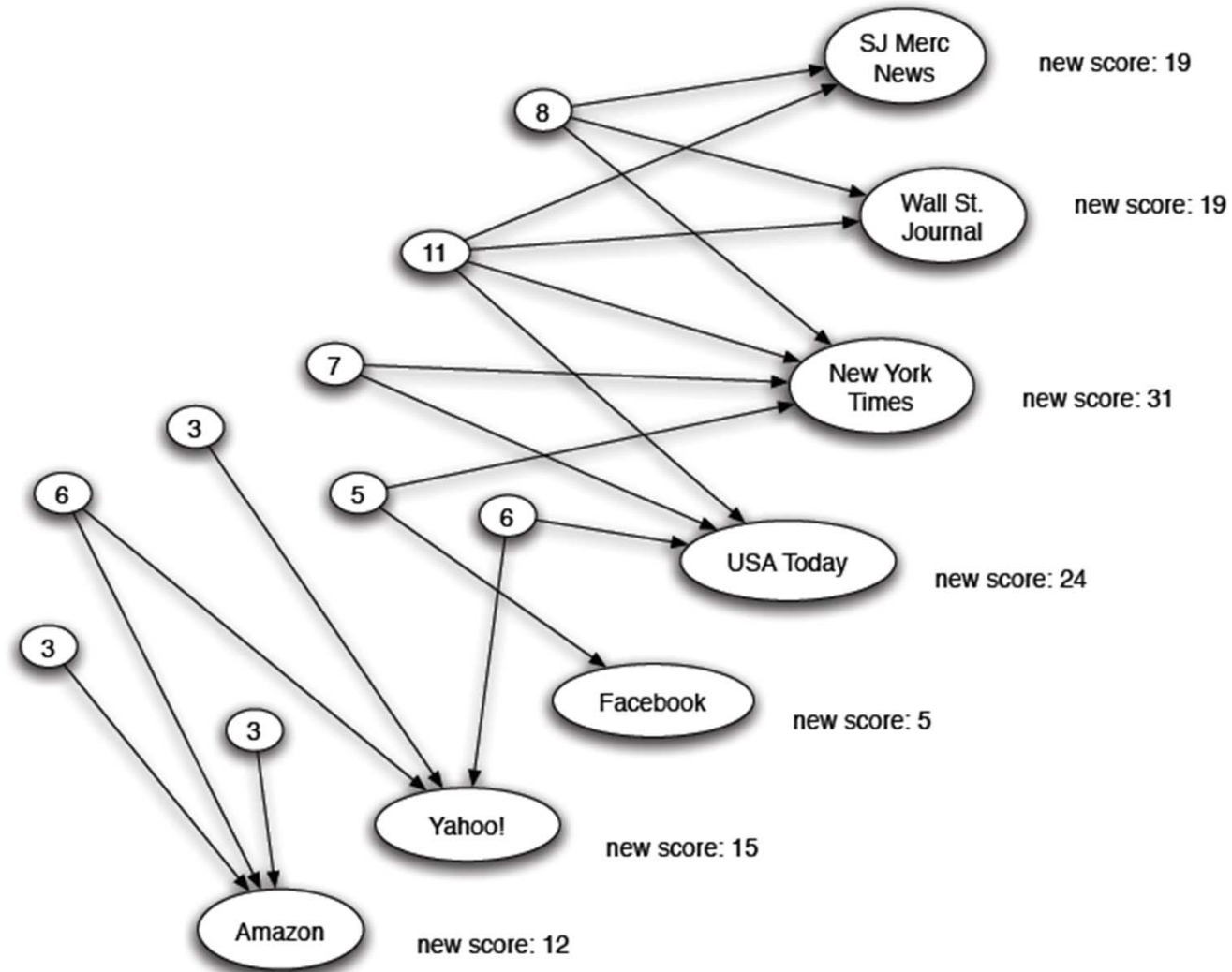
Counting in-links: Authority



Expert quality: Hub



Reweighting



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

Hubs and Authorities

- Each page i has 2 scores:

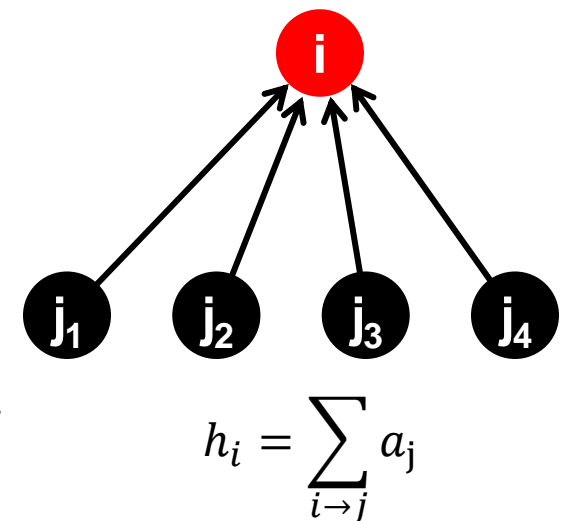
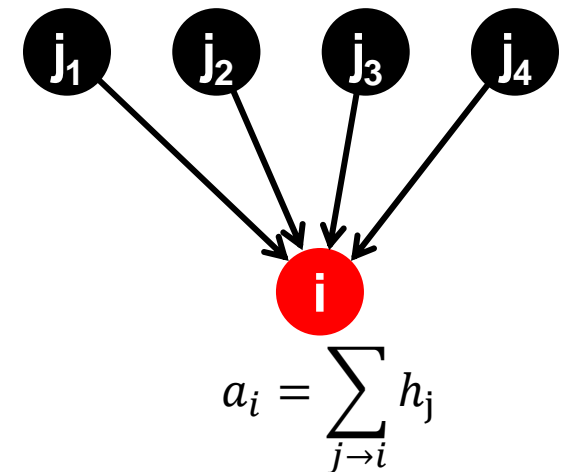
- Authority score: a_i
- Hub score: h_i

HITS algorithm:

- Initialize: $a_j = 1, h_i = 1$

- Then keep iterating:

- $\forall i$: Authority: $a_i = \sum_{j \rightarrow i} h_j$
- $\forall i$: Hub: $h_i = \sum_{i \rightarrow j} a_j$
- $\forall i$: normalize: $\sum_j a_j = 1, \sum_j h_j = 1$



Hubs and Authorities

- **HITS converges to a single stable point**
- Slightly change the notation:
 - Vector $a = (a_1, \dots, a_n)$, $h = (h_1, \dots, h_n)$
 - Adjacency matrix ($n \times n$): $M_{ij} = 1$ if $i \rightarrow j$
- **Then:**

$$h_i = \sum_{i \rightarrow j} a_j \Leftrightarrow h_i = \sum_j M_{ij} a_j$$

- So: $h = Ma$
- And likewise: $a = M^T h$

Hubs and Authorities


- HITS algorithm in new notation:

- Set: $a = h = 1^n$

- Repeat:

- $h = Ma, a = M^T h$

- Normalize

- Then: $a = M^T (Ma)$


- Thus, in 2k steps:

$$a = (M^T M)^k a$$

$$h = (M M^T)^k h$$

a is being updated (in 2 steps):

$$M^T (M a) = (M^T M) a$$

h is updated (in 2 steps):

$$M (M^T h) = (M M^T) h$$

Repeated matrix powering

Eigenvalues & Eigenvectors

- Definition:

- Let $Ax = \lambda x$ for some scalar λ , vector x , matrix A
- Then x is an eigenvector, and λ is its eigenvalue

- **Fact:**

- If A is symmetric ($A_{ij} = A_{ji}$)
(in our case $M^T M$ and $M M^T$ are symmetric)
- Then A has n orthogonal unit eigenvectors $w_1 \dots w_n$ that form a basis (coordinate system) with eigenvalues $\lambda_1 \dots \lambda_n$ ($|\lambda_i| \geq |\lambda_{i+1}|$)

How to Think About $A \cdot x$?

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

$$x = \sum_i \alpha_i w_i$$

- x has coordinates $(\alpha_1, \dots, \alpha_n)$

- **Suppose:** $\lambda_1 \dots \lambda_n$ ($|\lambda_1| \geq \dots \geq |\lambda_n|$)

- $A^k x = \lambda^k x = \sum_i \lambda_i^k \alpha_i w_i$

$$Ax = \lambda x$$

- As $k \rightarrow \infty$, if we normalize

$$A^k x \rightarrow \lambda_1 \alpha_1 w_1$$

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

$$\lim_{k \rightarrow \infty} \frac{\lambda_1^k}{\lambda_2^k} = \left(\frac{\lambda_1}{\lambda_2}\right)^k \rightarrow \infty$$

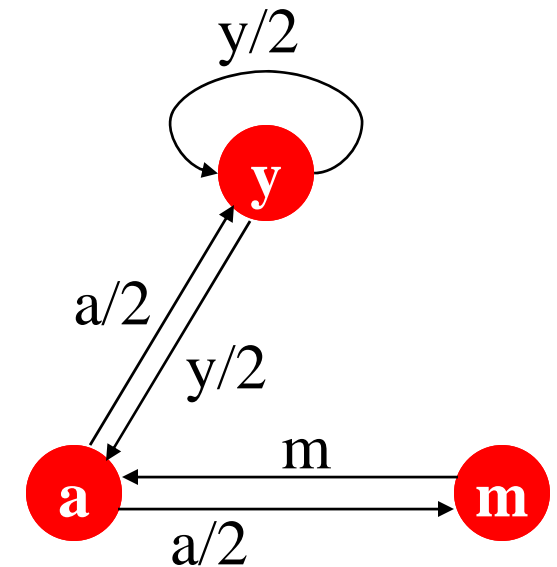
- So **authority** a is eigenvector of $M^T M$ associated with largest eigenvalue λ_1
- Similarly: **hub** h is eigenvector of $M M^T$

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

$$r_j = \sum_{i \rightarrow j} \frac{r_i}{d_{\text{out}}(i)}$$

The web in 1839



Flow equations:

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

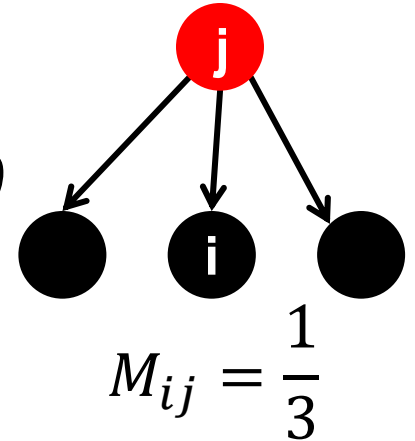
$$r_a = r_y/2 + r_m$$

$$r_m = r_a/2$$

PageRank: Matrix Formulation

- **Stochastic adjacency matrix M**

- Let page j has d_j out-links
- If $j \rightarrow i$, then $M_{ij} = 1/d_j$ else $M_{ij} = 0$
 - M is a **column stochastic matrix**
 - Columns sum to 1



- **Rank vector 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**

$$r = M r$$

Random Walk Interpretation

- Imagine a **random web surfer**:
 - At any time t , surfer is on some page u
 - At time $t+1$, the surfer follows an out-link from u uniformly at random
 - Ends up on some page v linked from u
 - Process repeats indefinitely
- **Let:**
 - $\mathbf{p}(t)$... vector whose i^{th} coordinate is the prob. that the surfer is at page i at time t
 - $\mathbf{p}(t)$ is a probability distribution over pages

The Stationary Distribution

- Where is the surfer at time $t+1$?

- Follows a link uniformly at random

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

- Suppose the random walk reaches a state

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

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

- Our rank vector \mathbf{r} satisfies $\mathbf{r} = \mathbf{M}\mathbf{r}$

- So, it 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
 - calculate the page rank of each node

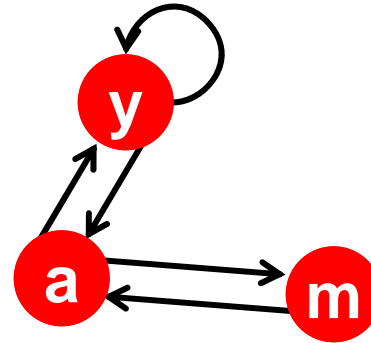
$$r_j^{(t+1)} = \sum_{i \rightarrow j} \frac{r_i^{(t)}}{d_i}$$

d_i out-degree of node i

PageRank: How to solve?

■ Power Iteration:

- Set $r_j = 1$
- $r_j = \sum_{i \rightarrow j} \frac{r_i}{d_i}$
 - And iterate



	y	a	m
y	1/2	1/2	0
a	1/2	0	1
m	0	1/2	0

$$\begin{aligned} \mathbf{r}_y &= \mathbf{r}_y/2 + \mathbf{r}_a/2 \\ \mathbf{r}_a &= \mathbf{r}_y/2 + \mathbf{r}_m \\ \mathbf{r}_m &= \mathbf{r}_a/2 \end{aligned}$$

■ Example:

$$\begin{pmatrix} \mathbf{r}_y \\ \mathbf{r}_a \\ \mathbf{r}_m \end{pmatrix} = \begin{matrix} 1/3 & 1/3 & 5/12 & 9/24 & & 6/15 \\ 1/3 & 3/6 & 1/3 & 11/24 & \dots & 6/15 \\ 1/3 & 1/6 & 3/12 & 1/6 & & 3/15 \end{matrix}$$

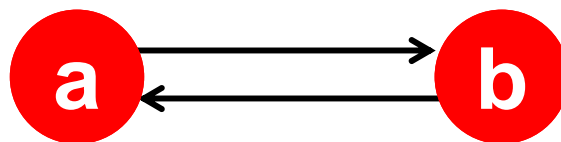
Iteration 0, 1, 2, ...

PageRank: Three Questions

$$r_j^{(t+1)} = \sum_{i \rightarrow j} \frac{r_i^{(t)}}{d_i} \quad \text{or equivalently} \quad \mathbf{r} = \mathbf{M}\mathbf{r}$$

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

Does This Converge?

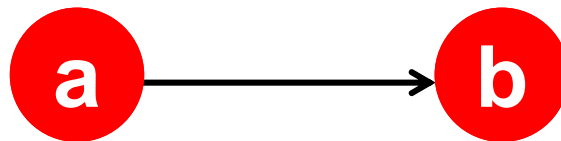


■ Example:

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

Iteration 0, 1, 2, ...

Does it Converge to What We Want?



■ Example:

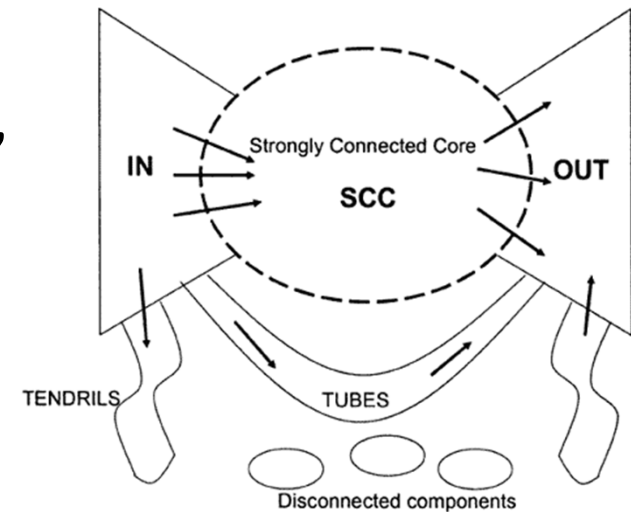
$$\begin{matrix} r_a \\ r_b \end{matrix} = \begin{matrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{matrix}$$

Iteration 0, 1, 2, ...

RageRank: Problems

2 problems:

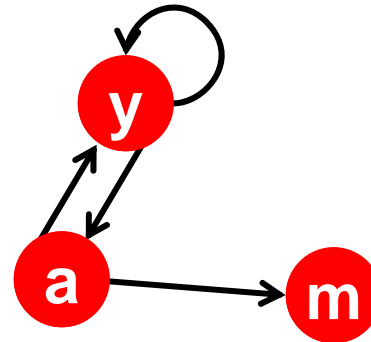
- Some pages are “**dead ends**” (have no out-links)
 - Such pages cause importance to “leak out”
- **Spider traps** (all out links are within the group)
 - Eventually spider traps absorb all importance



Problems: Dead Ends

Power Iteration:

- Set $r_j = 1$
- $r_j = \sum_{i \rightarrow j} \frac{r_i}{d_i}$
 - And iterate



	y	a	m
y	1/2	1/2	0
a	1/2	0	0
m	0	1/2	0

$$\mathbf{r}_y = \mathbf{r}_y/2 + \mathbf{r}_a/2$$

$$\mathbf{r}_a = \mathbf{r}_y/2$$

$$\mathbf{r}_m = \mathbf{r}_a/2$$

Example:

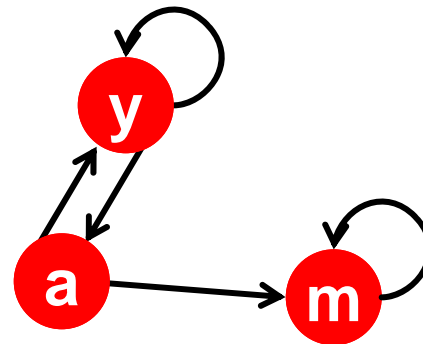
$$\begin{pmatrix} \mathbf{r}_y \\ \mathbf{r}_a \\ \mathbf{r}_m \end{pmatrix} = \begin{matrix} 1/3 & 2/6 & 3/12 & 5/24 & & 0 \\ 1/3 & 1/6 & 2/12 & 3/24 & \dots & 0 \\ 1/3 & 1/6 & 1/12 & 2/24 & & 0 \end{matrix}$$

Iteration 0, 1, 2, ...

Problems: Spider Traps

■ Power Iteration:

- Set $r_j = 1$
- $r_j = \sum_{i \rightarrow j} \frac{r_i}{d_i}$
 - And iterate



	y	a	m
y	1/2	1/2	0
a	1/2	0	0
m	0	1/2	1

$$\mathbf{r}_y = \mathbf{r}_y/2 + \mathbf{r}_a/2$$

$$\mathbf{r}_a = \mathbf{r}_y/2$$

$$\mathbf{r}_m = \mathbf{r}_a/2 + \mathbf{r}_m$$

■ Example:

$$\begin{pmatrix} \mathbf{r}_y \\ \mathbf{r}_a \\ \mathbf{r}_m \end{pmatrix} = \begin{matrix} 1/3 & 2/6 & 3/12 & 5/24 & & 0 \\ 1/3 & 1/6 & 2/12 & 3/24 & \dots & 0 \\ 1/3 & 3/6 & 7/12 & 16/24 & & 1 \end{matrix}$$

Iteration 0, 1, 2, ...

Looks a Lot Like...

$$r^{(t+1)} = Mr^{(t)}$$

Markov Chains

- Set of states X
- Transition matrix P where $P_{ij} = P(X_t=i \mid X_{t-1}=j)$
- π specifying the probability of being at each state $x \in X$
- Goal is to find π such that $\pi = \pi P$

Why is This Analogy Useful?

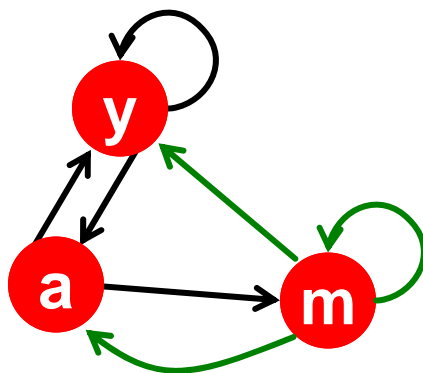
- **Markov chains theory**
- Fact: For **any start vector**, the power method applied to a Markov transition matrix P will **converge** to a **unique** positive stationary vector as long as P is **stochastic**, **irreducible** and **aperiodic**.

Make M Stochastic

- **Stochastic:** every column sums to 1

$$S = M + a\left(\frac{1}{n}e\right)$$

e...vector
of all 1



	y	a	m
y	1/2	1/2	1/3
a	1/2	0	1/3
m	0	1/2	1/3

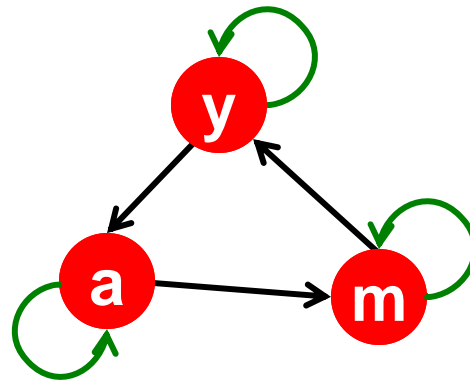
$$r_y = r_y/2 + r_a/2 + r_m/3$$

$$r_a = r_y/2 + r_m/3$$

$$r_m = r_a/2 + r_m/3$$

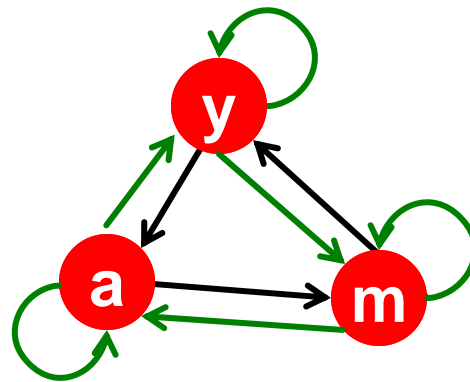
Make M Aperiodic

- A chain is periodic if there exists $k > 1$ such that the interval between two visits to some state s is always a multiple of k .



Make M Irreducible

- From any state, there is a non-zero probability of going from any one state to any another.



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 page uniformly at random

- **PageRank equation** [Brin-Page, 98]

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

Assuming we follow random teleport links
with probability 1.0 from dead-ends

d_i ... outdegree
of node i

The Google Matrix

- The Google Matrix:

$$G = (1 - \beta)S + \beta \frac{1}{n} e e^T$$

- G is stochastic, aperiodic and irreducible.

$$r^{t+1} = G r^t$$

- G is dense but computable using sparse matrix H

- $$\begin{aligned} G &= (1 - \beta)S + \beta \frac{1}{n} e e^T = \\ &= (1 - \beta)(M + \frac{1}{n} a^T e) + \beta \frac{1}{n} e e^T = \\ &= (1 - \beta)M + ((1 - \beta)a^T + \beta e^T) \frac{1}{n} e \end{aligned}$$

PageRank & Eigenvectors

- PageRank as a principal eigenvector

$$r = Mr \Leftrightarrow r_j = \sum_i r_i / d_i$$

- But we really want:

$$r_j = (1 - \beta) \sum_{i \rightarrow j} r_i / d_i + \beta$$

d_i ... out-degree
of node i

- Define:

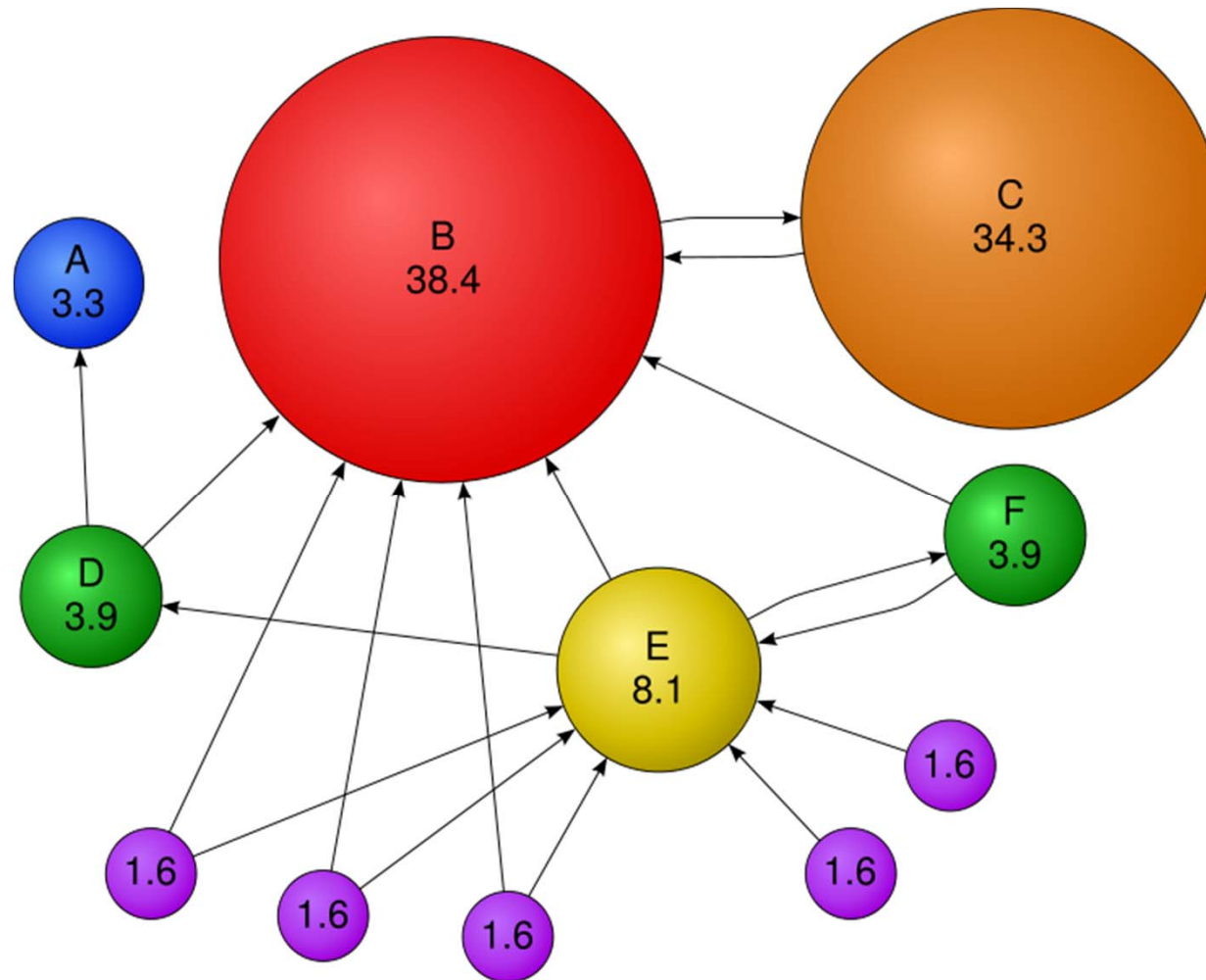
$$M'_{ij} = (1 - \beta) M_{ij} + \beta / n$$

- Then: $r = M'r$

- What is β ?

- In practice $\beta = 0.15$ (5 links and jump)

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

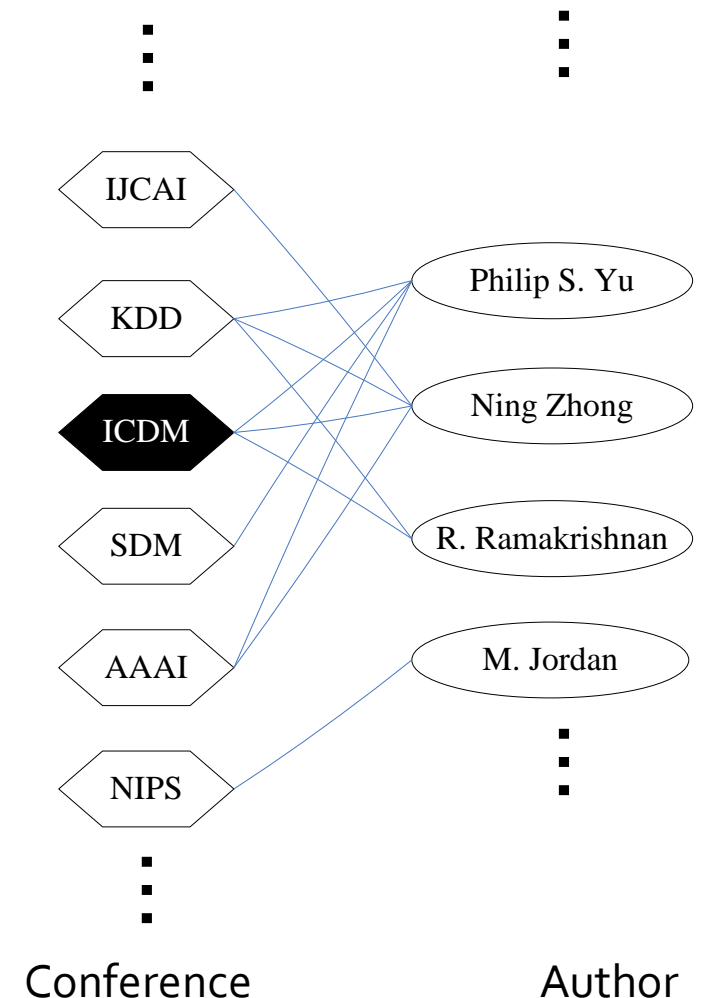
Personalized PageRank

- 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

$$\begin{aligned} M'_{ij} &= (1-\beta) M_{ij} + \beta / |S| && \text{if } i \text{ in } S \quad (S \dots \text{teleport set}) \\ &= (1-\beta) M_{ij} && \text{otherwise} \end{aligned}$$

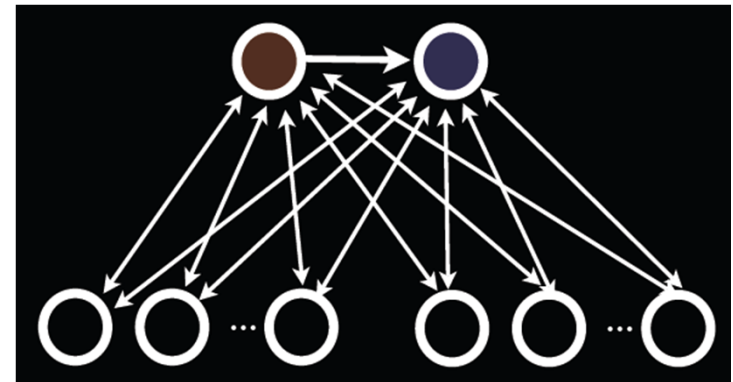
PageRank: Applications

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



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



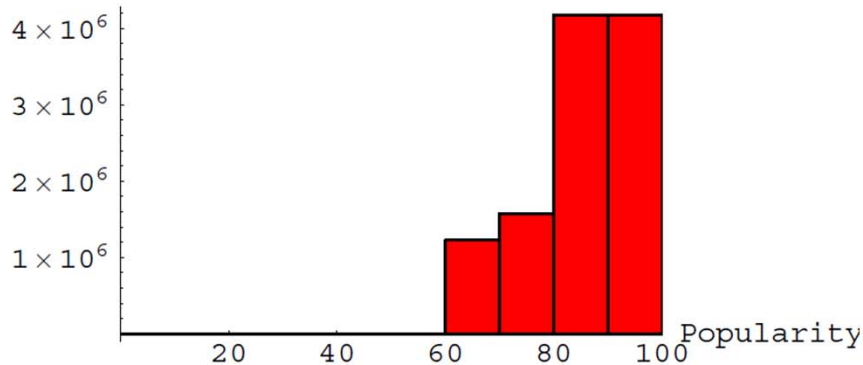
The screenshot shows the website **affordablecellphonerates.com**. It features a search bar with the text "Search" and a "Search" button. Below the search bar is a "Related Searches" section with links to "Free Prepaid Calling Card", "Refill", "International Call", "Internet Phone Card", "Calling Cards from To", "Calling Cards for India", "Cellular Phone Prepaid Phone Card", "Long Distance Card", "Cheap International Calling Cards", "Instant Calling Card Pin", "Calling Card Costa Rica", "South Africa Calling Card", and "Buy a Calling Card". The main content area displays search results for "card phone prepaid", including links to "Online prepaid phone card", "US 1¢/min - World 2¢/min", "Prepaid Phone Cards", "Prepaid Phone", "Prepaid Phone Cards", "Phone Card", and "Phone card".

PageRank: Problems

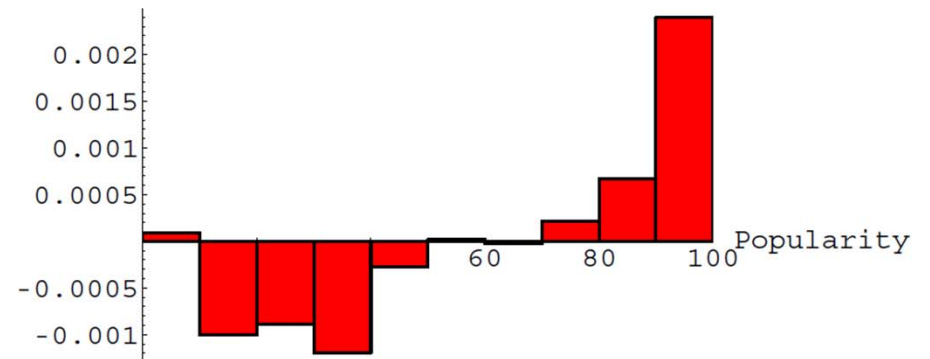
Issues with PageRank

- **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

Absolute increase in the no. of In-Links



Absolute increase in the PageRank values



Google Bombs



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 biography page. [not a ...](#)
[searchengine](#) [Cached](#) - [Simi](#)

