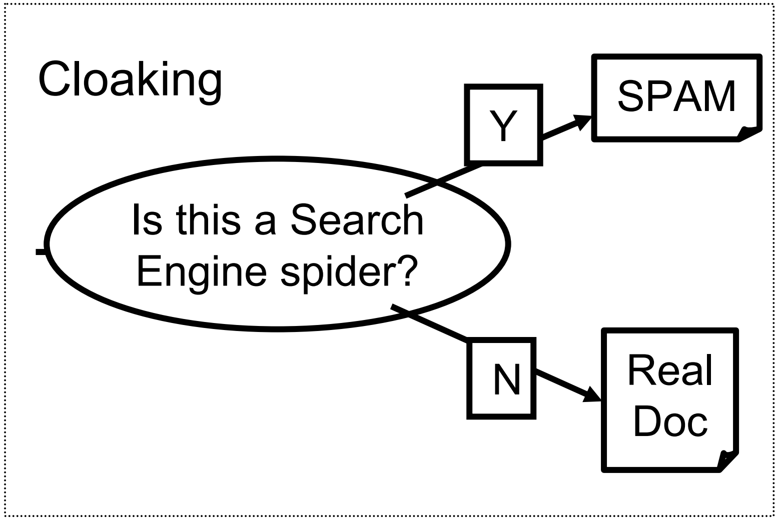
Adversarial IR (Spam)

- Motives
  - Commercial, political, religious, lobbies
  - Promotion funded by advertising budget
- Operators
  - Contractors (Search Engine Optimizers) for lobbies, companies
  - Web masters
  - Hosting services
- Forum
  - Web master world (www.webmasterworld.com)
    - Search engine specific tricks
    - Discussions about academic papers ☺

A few spam technologies

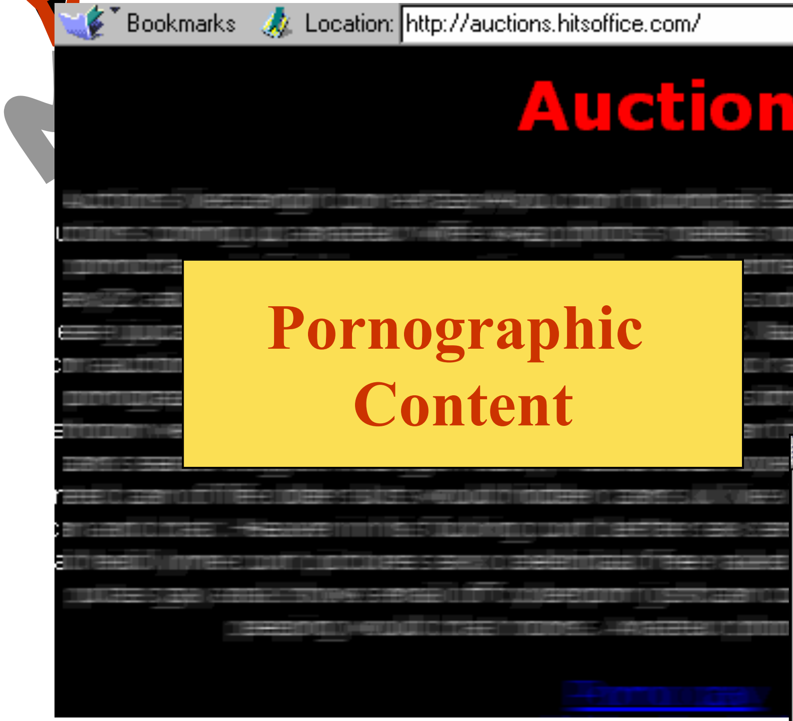
- **Cloaking**
  - Serve fake content to search engine robot
  - *DNS cloaking:* Switch IP address. Impersonate
- **Doorway pages**
  - Pages optimized for a single keyword that re-direct to the real target page
- **Keyword Spam**
  - Misleading meta-keywords, excessive repetition of a term, fake "anchor text"
  - Hidden text with colors, CSS tricks, etc.
- **Link spamming**
  - Mutual admiration societies, hidden links, awards
  - *Domain flooding:* numerous domains that point or re-direct to a target page
- **Robots**
  - Fake click stream
  - Fake query stream
  - Millions of submissions via Add-Url



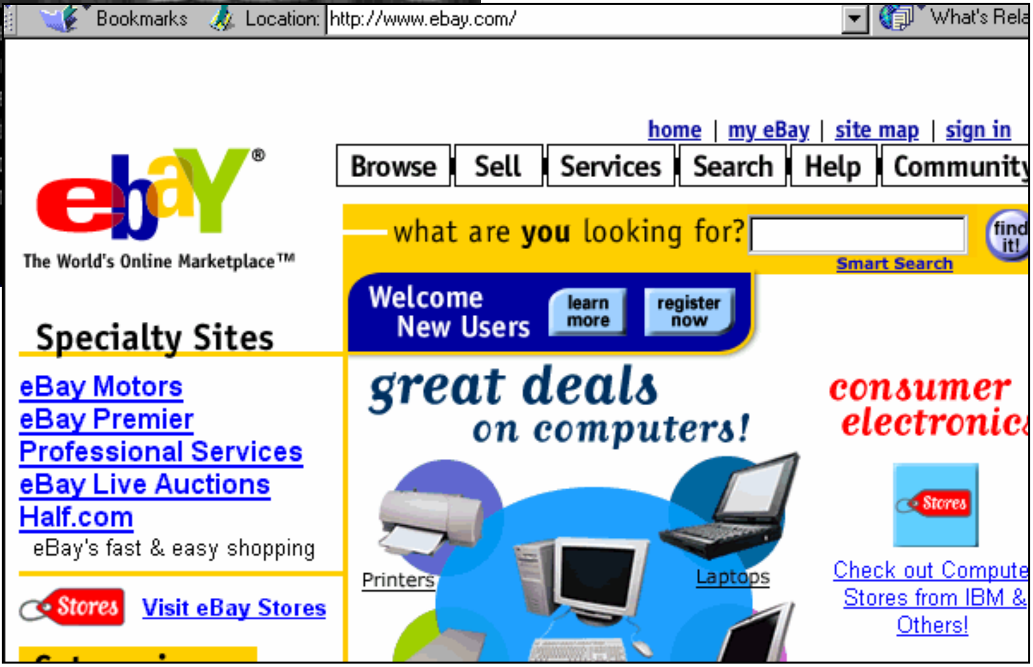
Meta-Keywords =
"… London hotels, hotel, holiday inn, hilton, discount, booking, reservation, sex, mp3, britney spears, viagra, …"

Can you trust words on the page?

auctions.hitsoffice.com/



www.ebay.com/

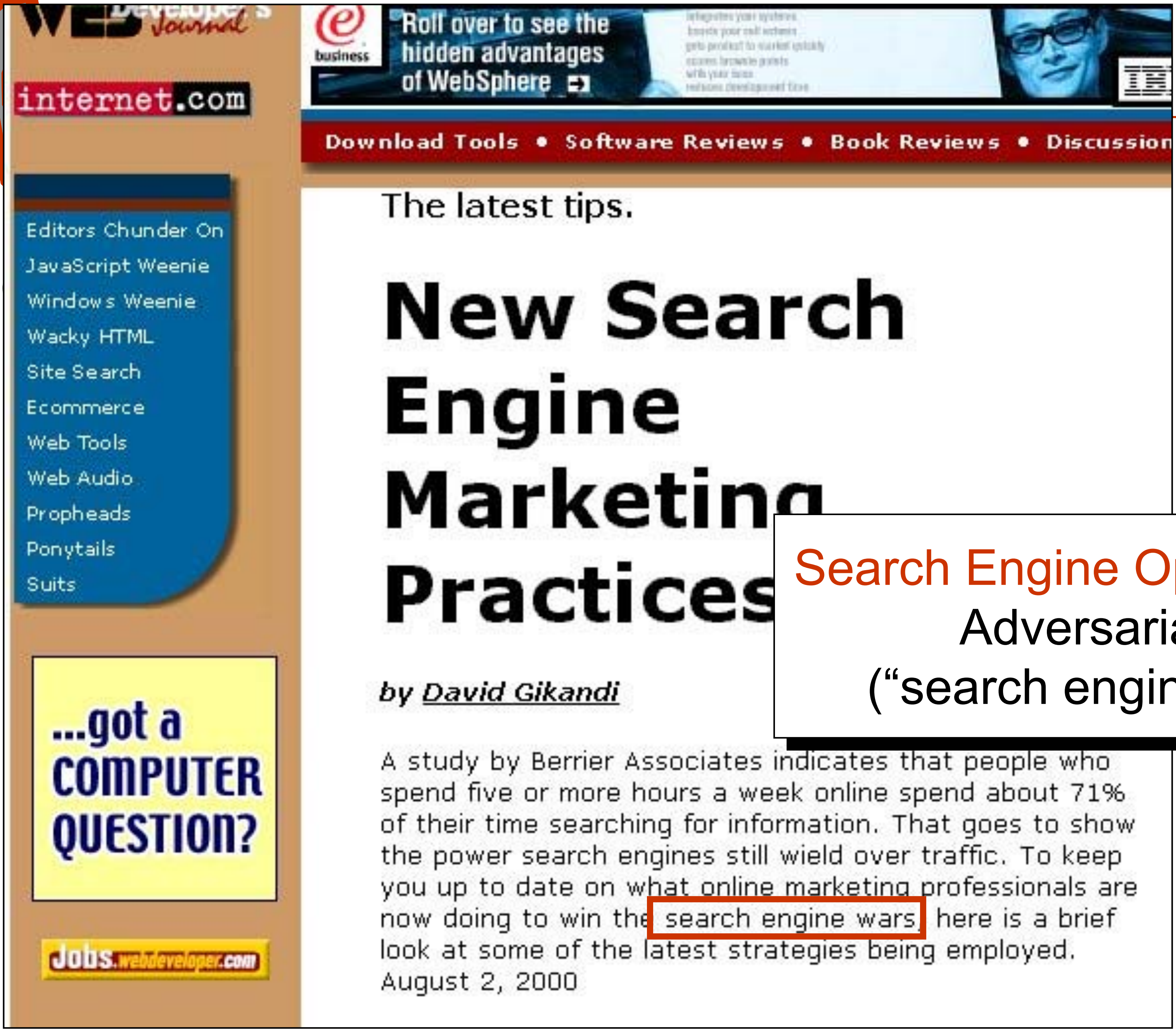


Examples from July 2002

Search Engine Optimization I

Adversarial IR

(“search engine wars”)

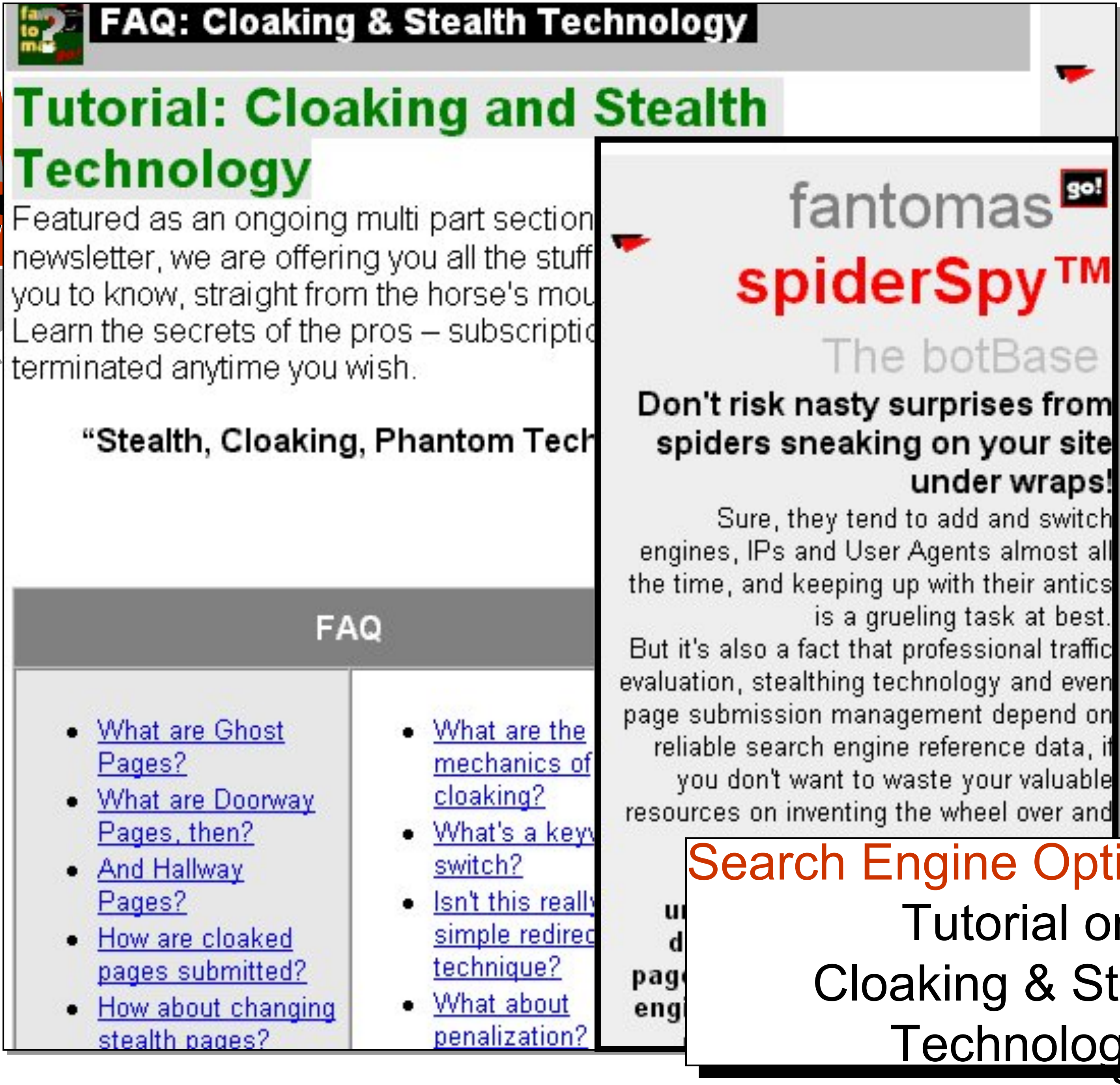


The latest tips.

New Search Engine Marketing Practices

by David Gikandi

A study by Berrier Associates indicates that people who spend five or more hours a week online spend about 71% of their time searching for information. That goes to show the power search engines still wield over traffic. To keep you up to date on what online marketing professionals are now doing to win the search engine wars here is a brief look at some of the latest strategies being employed.
August 2, 2000



FAQ: Cloaking & Stealth Technology

Tutorial: Cloaking and Stealth Technology

Featured as an ongoing multi part section
newsletter, we are offering you all the stuff
you to know, straight from the horse's mou
Learn the secrets of the pros – subscriptic
terminated anytime you wish.

"Stealth, Cloaking, Phantom Tech

FAQ

- What are Ghost Pages?
- What are Doorway Pages, then?
- And Hallway Pages?
- How are cloaked pages submitted?
- How about changing stealth pages?

- What are the mechanics of cloaking?
- What's a keyw switch?
- Isn't this really simple redirec technique?
- What about penalization?

fantomas go!

spiderSpy™

The botBase

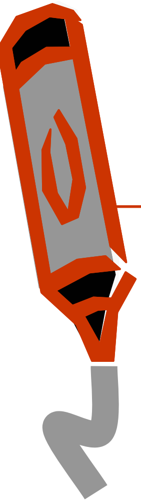
Don't risk nasty surprises from spiders sneaking on your site under wraps!

Sure, they tend to add and switch engines, IPs and User Agents almost all the time, and keeping up with their antics is a grueling task at best.
But it's also a fact that professional traffic evaluation, stealthing technology and even page submission management depend on reliable search engine reference data, if you don't want to waste your valuable resources on inventing the wheel over and

u
d
pag
engi

Search Engine Optimization II

Tutorial on
Cloaking & Stealth
Technology



The war against spam

- Quality signals - Prefer authoritative pages based on:
  - Votes from authors (linkage signals)
  - Votes from users (usage signals)
- Policing of URL submissions
  - Anti robot test
- Limits on meta-keywords
- Robust link analysis
  - Ignore statistically implausible linkage (or text)
  - Use link analysis to detect spammers (guilt by association)
- Spam recognition by machine learning
  - Training set based on known spam
- Family friendly filters
  - Linguistic analysis, general classification techniques, etc.
  - For images: flesh tone detectors, source text analysis, etc.
- Editorial intervention
  - Blacklists
  - Top queries audited
  - Complaints addressed

www IR

- world wide web
- google, page rank
- markov chains
- HITS link analysis
- behavior-based web search
- crawling, indexing the web
- duplicates, mirrors and spam
- www infrastructure
- www size
- cache, hardware, systems

duplicates/near duplicates detection

- *Duplication*: Exact match with fingerprints
- *Near-Duplication*: Approximate match
    - Overview
        - Compute syntactic similarity with an edit-distance measure
        - Use similarity threshold to detect near-duplicates
            - E.g., Similarity > 80% => Documents are "near duplicates"
            - Not transitive though sometimes used transitively

near similarity

- Features:
  - Segments of a document (natural or artificial breakpoints) [Brin95]
  - *Shingles* (Word N-Grams) [Brin95, Brod98]

    "a rose is a rose is a rose" =>

    a_rose_is_a

    rose_is_a_rose

    is_a_rose_is

- Similarity Measure
  - TFIDF [Shiv95]
  - Set intersection [Brod98]

    (Specifically, Size_of_Intersection / Size_of_Union)

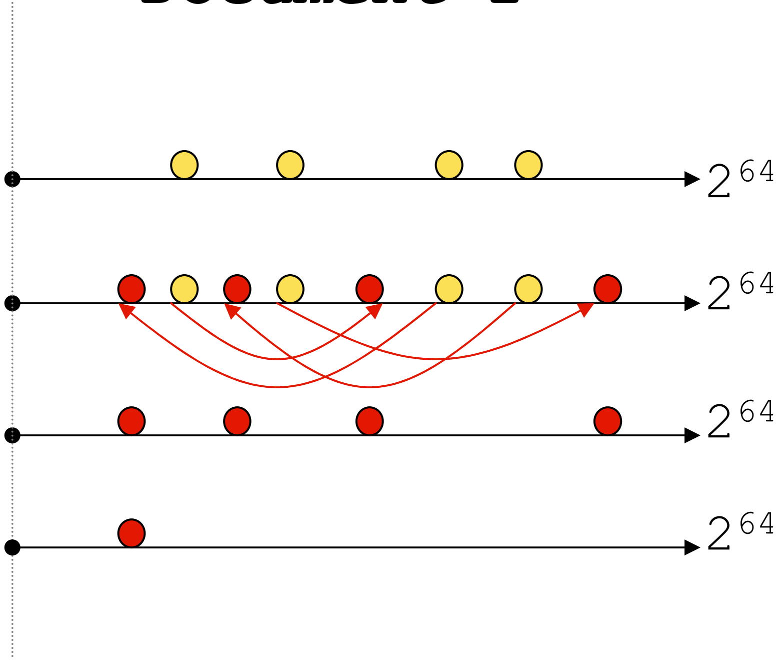
Shingles + Set Intersection

- Computing exact set intersection of shingles between all pairs of documents is expensive and infeasible
  - Approximate using a cleverly chosen subset of shingles from each (a sketch)

- Estimate size_of_intersection / size_of_union based on a short sketch ([Brod97, Brod98])
  - Create a "sketch vector" (e.g., of size 200) for each document
  - Documents which share more than t (say 80%) corresponding vector elements are similar
  - For doc D, sketch[i] is computed as follows:
    - Let f map all shingles in the universe to $0..2^m$ (e.g., f = fingerprinting)
    - Let π_i be a specific random permutation on $0..2^m$
    - Pick sketch[i] := MIN π_i (f(s)) over all shingles s in D

Computing Sketch[i] for doc1

Document 1

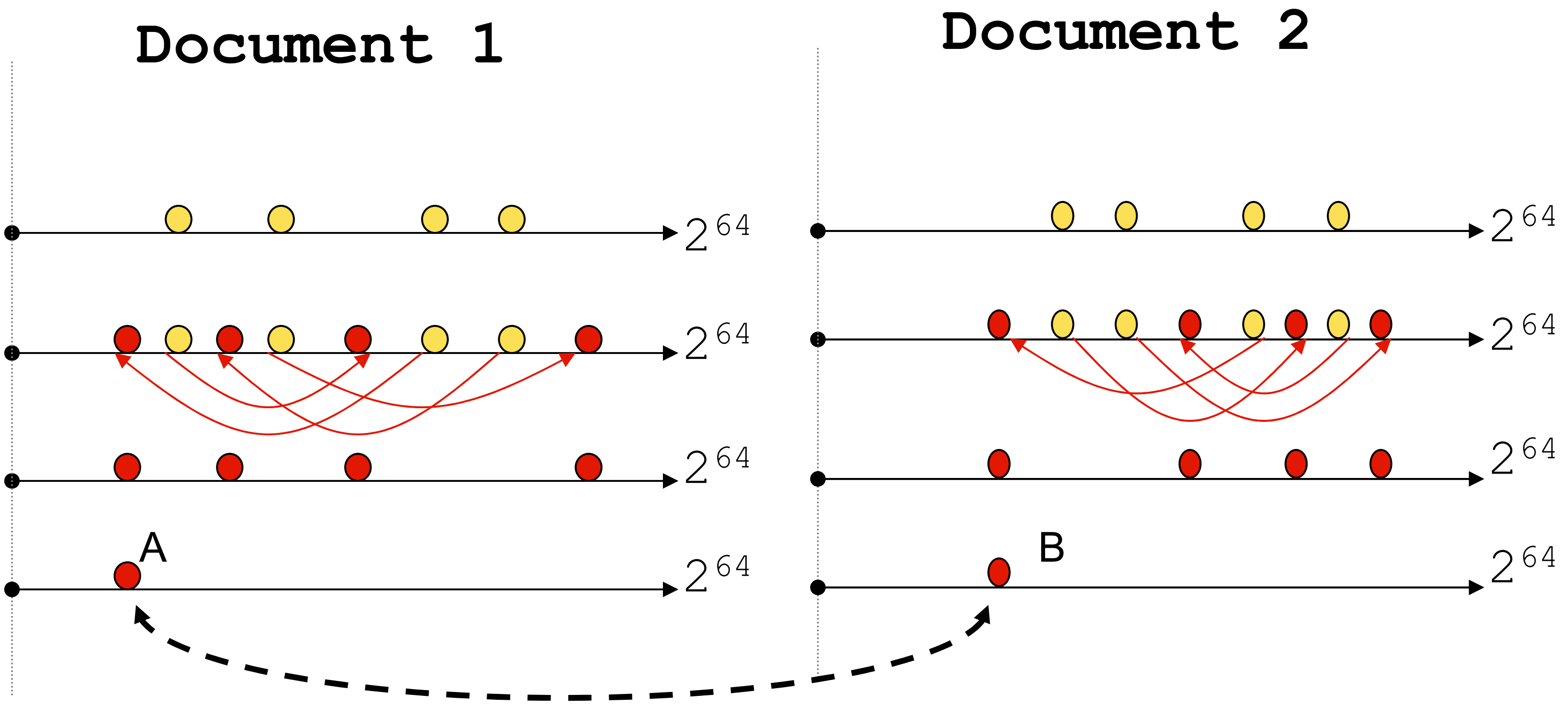


Start with 64 bit shingles

Permute on the number line with π_i

Pick the min value

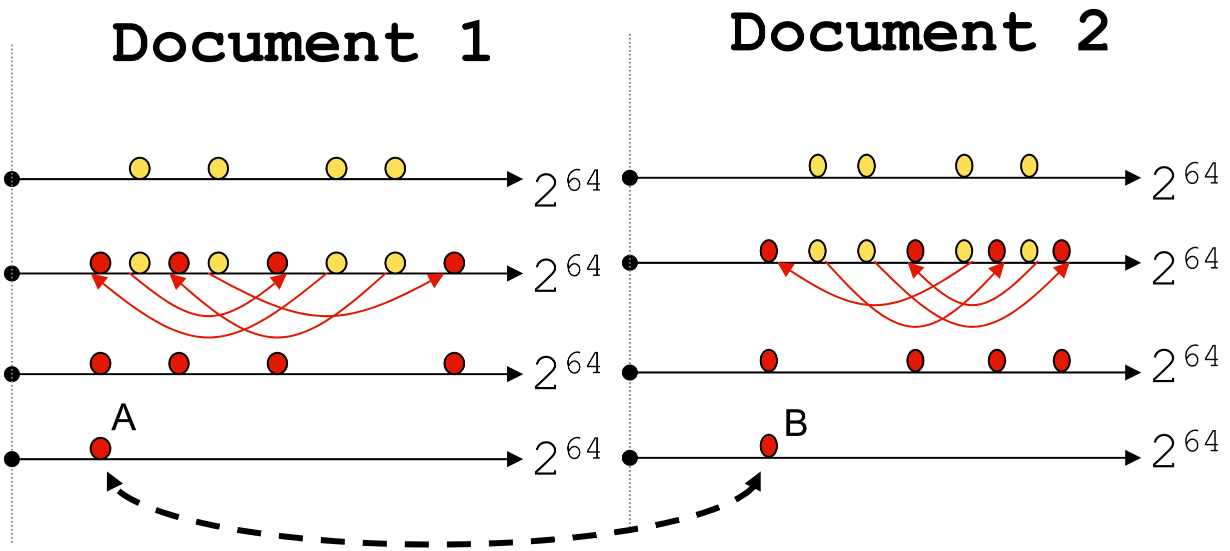
Sketch comparison



Are these equal?

Test for 200 random permutations: $\pi_1, \pi_2, \ldots \pi_{200}$

Sketch comparison



A = B iff the shingle with the MIN value in the union of Doc1 and Doc2 is common to both (I.e., lies in the intersection)

This happens with probability:

`Size_of_intersection / Size_of_union`

mirrors

- Mirroring is systematic replication of web pages across hosts.
  - Single largest cause of duplication on the web

- **Host1**/α and **Host2**/β are mirrors iff

  For all (or most) paths p such that when

  http://**Host1**/ α / p exists

  http://**Host2**/ β / p exists as well

  with identical (or near identical) content, and vice versa.

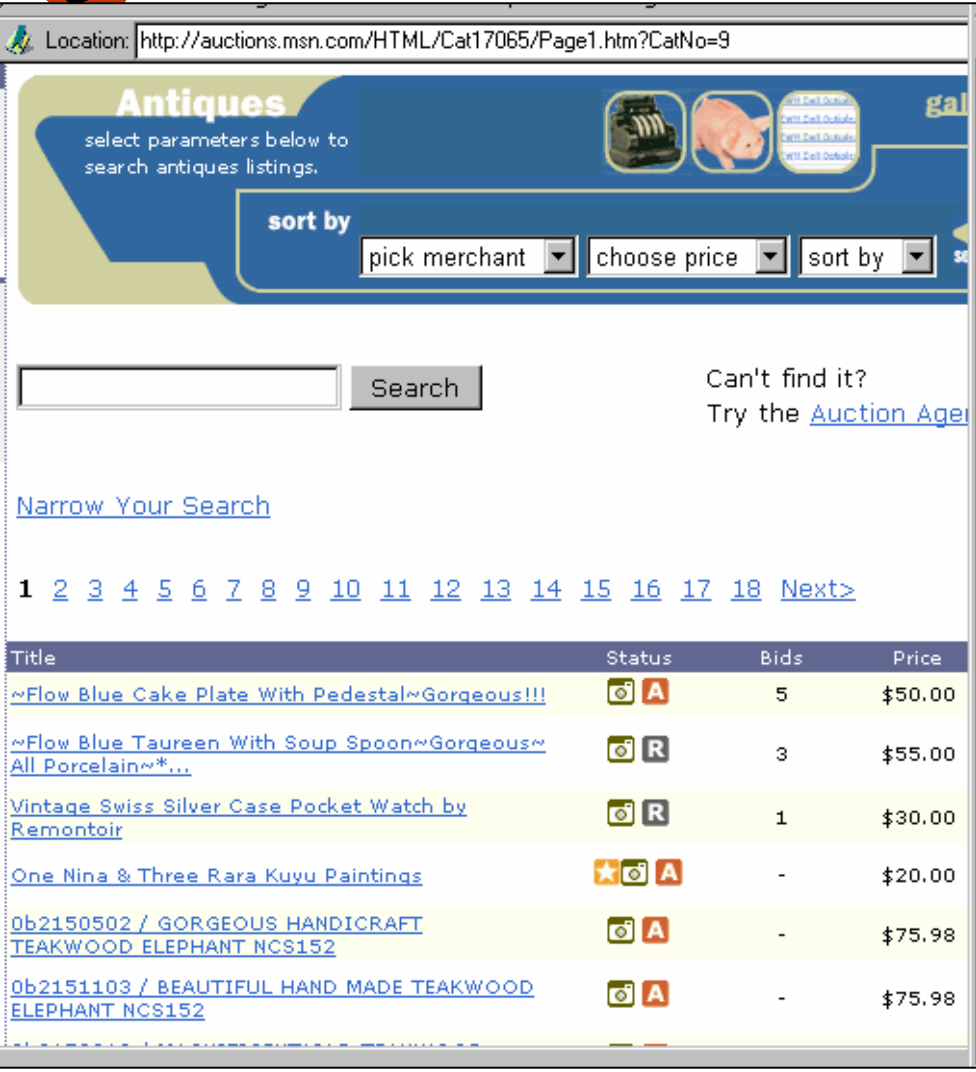
mirror detection

- http://**www.elsevier.com**/ and http://**www.elsevier.nl**/
- Structural Classification of Proteins
  - http://**scop.mrc-lmb.cam.ac.uk**/scop
  - http://**scop.berkeley.edu**/
  - http://**scop.wehi.edu.au/**scop
  - http://**pdb.weizmann.ac.il**/scop
  - http://**scop.protres.ru**/

mirrors: repackaged

Auctions.msn.com

Auctions.lycos.com



http://auctions.msn.com/HTML/Cat17065/Page1.htm?CatNo=9

Antiques

select parameters below to search antiques listings.

sort by

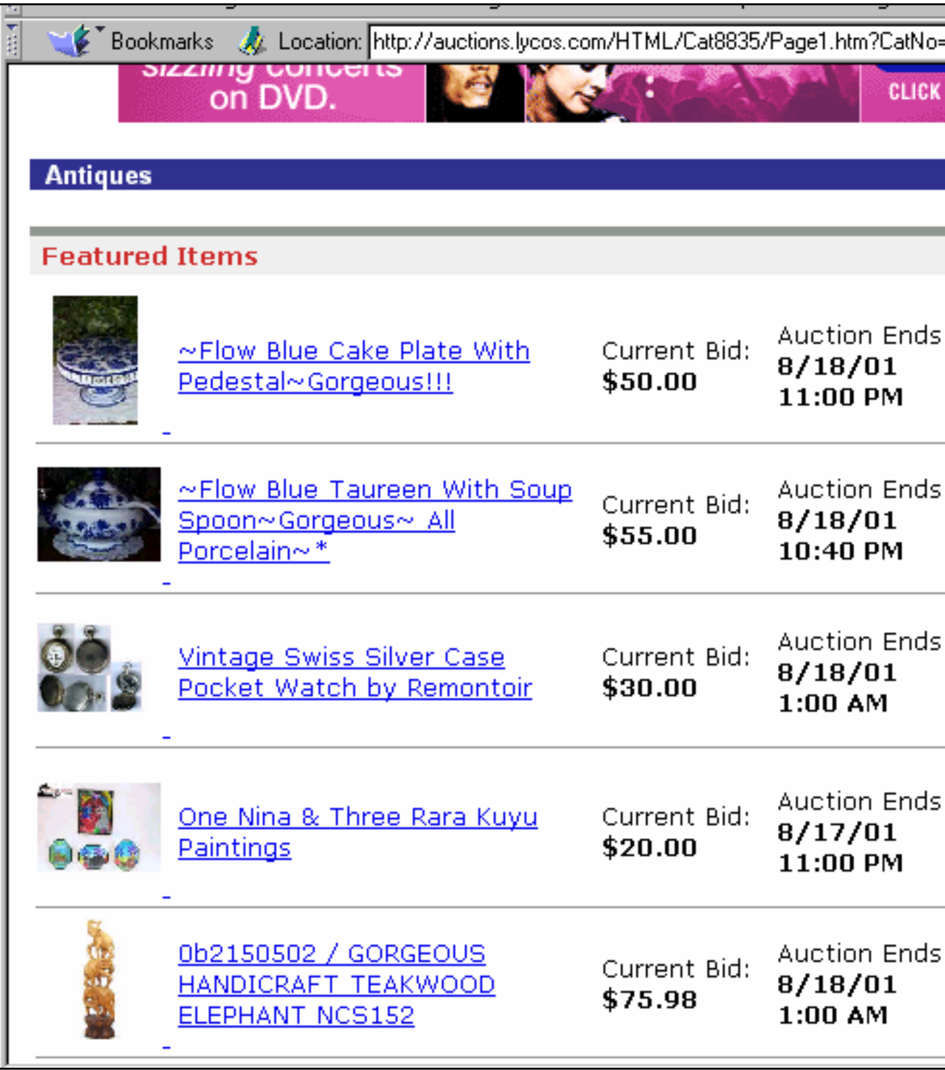
pick merchant ▾ | choose price ▾ | sort by ▾

Can't find it?
Try the Auction Age

Narrow Your Search

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 Next>

Title	Status	Bids	Price
~Flow Blue Cake Plate With Pedestal~Gorgeous!!!	📷 A	5	$50.00
~Flow Blue Taureen With Soup Spoon~Gorgeous~ All Porcelain~*...	📷 R	3	$55.00
Vintage Swiss Silver Case Pocket Watch by Remontoir	📷 R	1	$30.00
One Nina & Three Rara Kuyu Paintings	★ 📷 A	-	$20.00
0b2150502 / GORGEOUS HANDICRAFT TEAKWOOD ELEPHANT NCS152	📷 A	-	$75.98
0b2151103 / BEAUTIFUL HAND MADE TEAKWOOD ELEPHANT NCS152	📷 A	-	$75.98

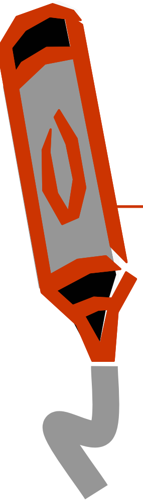


http://auctions.lycos.com/HTML/Cat8835/Page1.htm?CatNo=

Antiques

Featured Items

~Flow Blue Cake Plate With Pedestal~Gorgeous!!!	Current Bid: **$50.00**	Auction Ends **8/18/01** **11:00 PM**
~Flow Blue Taureen With Soup Spoon~Gorgeous~ All Porcelain~*	Current Bid: **$55.00**	Auction Ends **8/18/01** **10:40 PM**
Vintage Swiss Silver Case Pocket Watch by Remontoir	Current Bid: **$30.00**	Auction Ends **8/18/01** **1:00 AM**
One Nina & Three Rara Kuyu Paintings	Current Bid: **$20.00**	Auction Ends **8/17/01** **11:00 PM**
0b2150502 / GORGEOUS HANDICRAFT TEAKWOOD ELEPHANT NCS152	Current Bid: **$75.98**	Auction Ends **8/18/01** **1:00 AM**

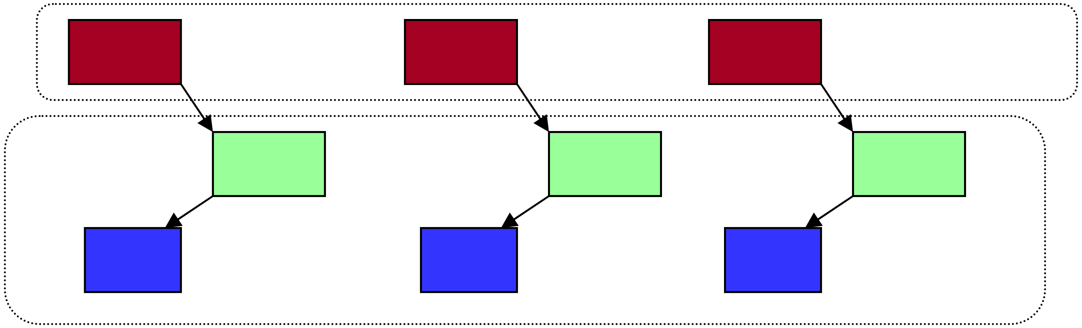


mirrors

- Why detect mirrors?
  - Smart crawling
    - Fetch from the fastest or freshest server
    - Avoid duplication
  - Better connectivity analysis
    - Combine inlinks
    - Avoid double counting outlinks
  - Redundancy in result listings
    - “If that fails you can try: <mirror>/samepath”
  - Proxy caching

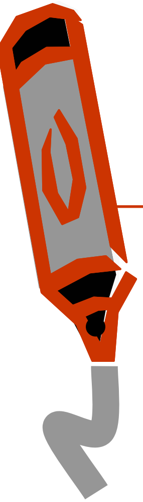
bottom up mirror detection

- Maintain clusters of subgraphs
- Initialize clusters of trivial subgraphs
  - Group near-duplicate single documents into a cluster
- Subsequent passes
  - Merge clusters of the same cardinality and corresponding linkage



  - Avoid decreasing cluster cardinality
- To detect mirrors we need:
  - Adequate path overlap
  - Contents of corresponding pages within a small time range

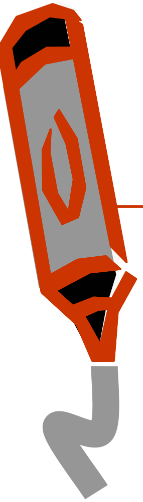
top down mirror detection



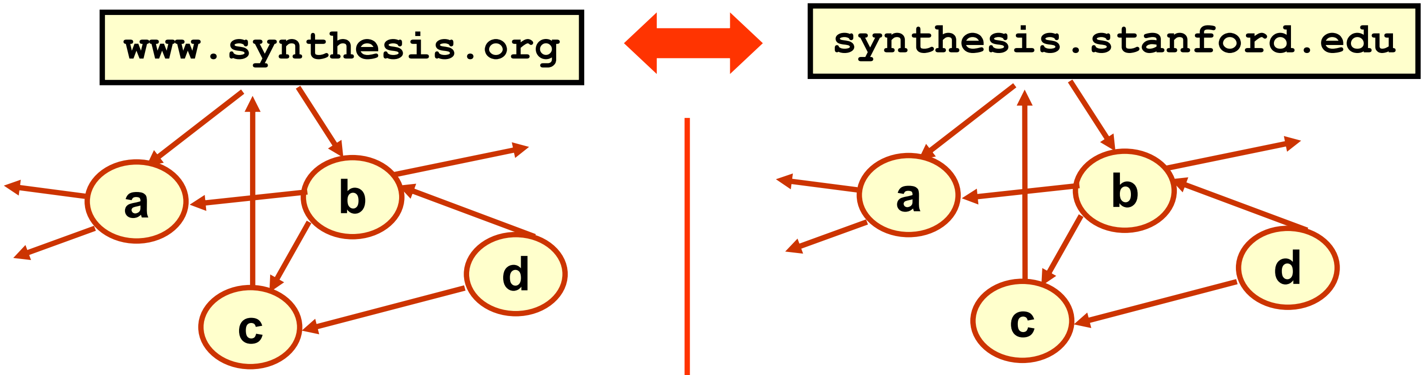
- E.g.,

  www.synthesis.org/Docs/ProjAbs/synsys/synalysis.html
  synthesis.stanford.edu/Docs/ProjAbs/synsys/quant-dev-new-teach.html

- What features could indicate mirroring?
  - Hostname similarity:
    - word unigrams and bigrams: { www, www.synthesis, synthesis, …}
  - Directory similarity:
    - Positional path bigrams { 0:Docs/ProjAbs, 1:ProjAbs/synsys, … }
  - IP address similarity:
    - 3 or 4 octet overlap
    - Many hosts sharing an IP address => virtual hosting by an ISP
  - Host outlink overlap
  - Path overlap
    - Potentially, path + sketch overlap



mirror detection by urls



www.synthesis.org

synthesis.stanford.edu

www.synthesis.org/Docs/ProjAbs/synsys/synalysis.html
www.synthesis.org/Docs/ProjAbs/synsys/visual-semi-quan
www.synthesis.org/Docs/annual.report96.final.html
www.synthesis.org/Docs/cicee-berlin-paper.html
www.synthesis.org/Docs/myr5
www.synthesis.org/Docs/myr5/cicee/bridge-gap.html
www.synthesis.org/Docs/myr5/cs/cs-meta.html
www.synthesis.org/Docs/myr5/mech/mech-intro-mechatro
www.synthesis.org/Docs/myr5/mech/mech-take-home.htm
www.synthesis.org/Docs/myr5/synsys/experiential-learning
www.synthesis.org/Docs/myr5/synsys/mm-mech-dissec.ht
www.synthesis.org/Docs/yr5ar
www.synthesis.org/Docs/yr5ar/assess
www.synthesis.org/Docs/yr5ar/cicee
www.synthesis.org/Docs/yr5ar/cicee/bridge-gap.html
www.synthesis.org/Docs/yr5ar/cicee/comp-integ-analysis.h

synthesis.stanford.edu/Docs/ProjAbs/deliv/high-tech-…
synthesis.stanford.edu/Docs/ProjAbs/mech/mech-enhanced…
synthesis.stanford.edu/Docs/ProjAbs/mech/mech-intro-…
synthesis.stanford.edu/Docs/ProjAbs/mech/mech-mm-case-…
synthesis.stanford.edu/Docs/ProjAbs/synsys/quant-dev-new-…
synthesis.stanford.edu/Docs/annual.report96.final.html
synthesis.stanford.edu/Docs/annual.report96.final_fn.html
synthesis.stanford.edu/Docs/myr5/assessment
synthesis.stanford.edu/Docs/myr5/assessment/assessment-…
synthesis.stanford.edu/Docs/myr5/assessment/mm-forum-kiosk-…
synthesis.stanford.edu/Docs/myr5/assessment/neato-ucb.html
synthesis.stanford.edu/Docs/myr5/assessment/not-available.html
synthesis.stanford.edu/Docs/myr5/cicee
synthesis.stanford.edu/Docs/myr5/cicee/bridge-gap.html
synthesis.stanford.edu/Docs/myr5/cicee/cicee-main.html
synthesis.stanford.edu/Docs/myr5/cicee/comp-integ-analysis.html

3 announcements:

- Thanks for filling out the HW1 poll
- HW2 is due today 5pm (scans must be readable)
- HW3 will be posted today

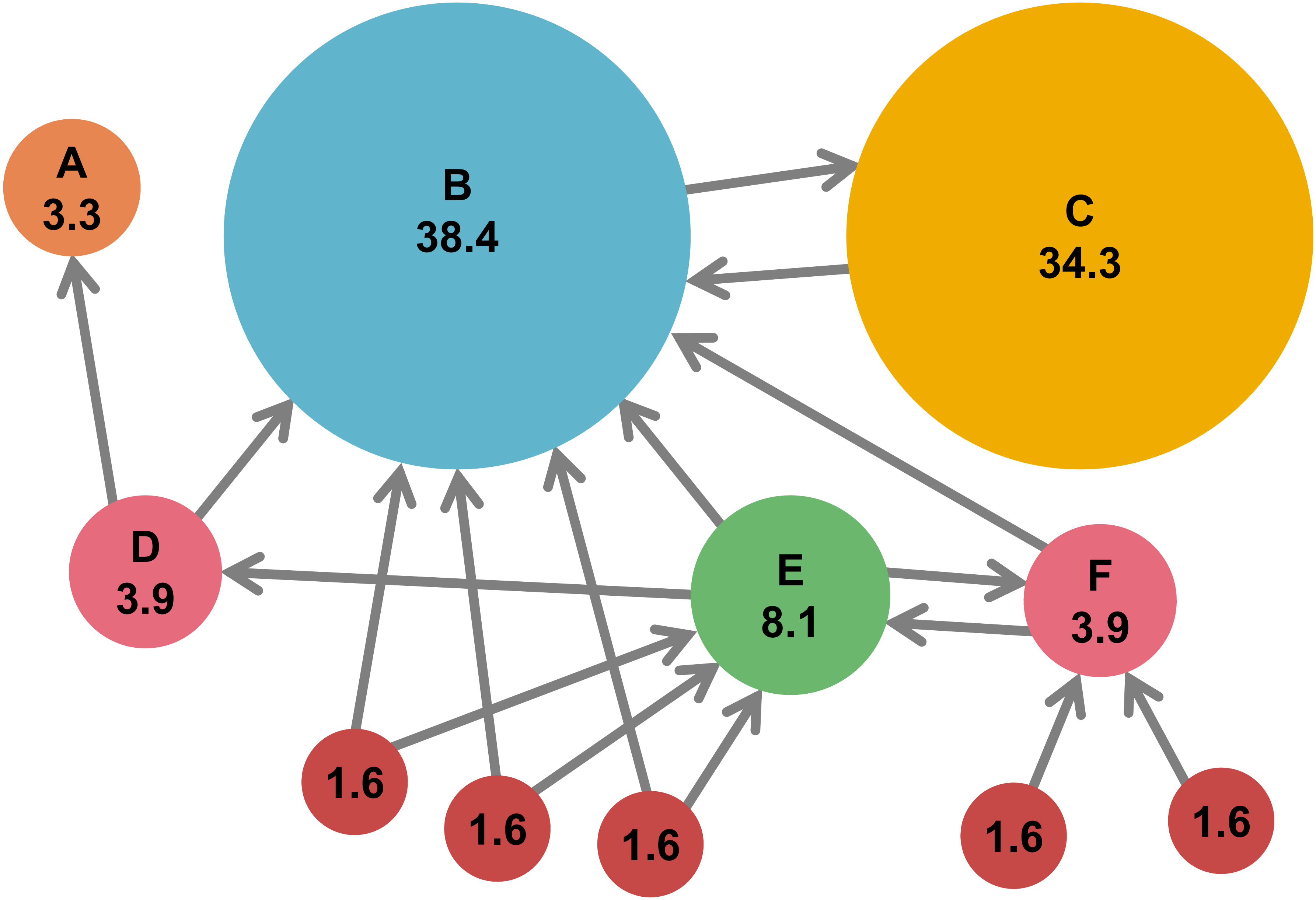
Link Analysis: TrustRank and WebSpam

CS246: Mining Massive Datasets
Jure Leskovec, Stanford University
http://cs246.stanford.edu

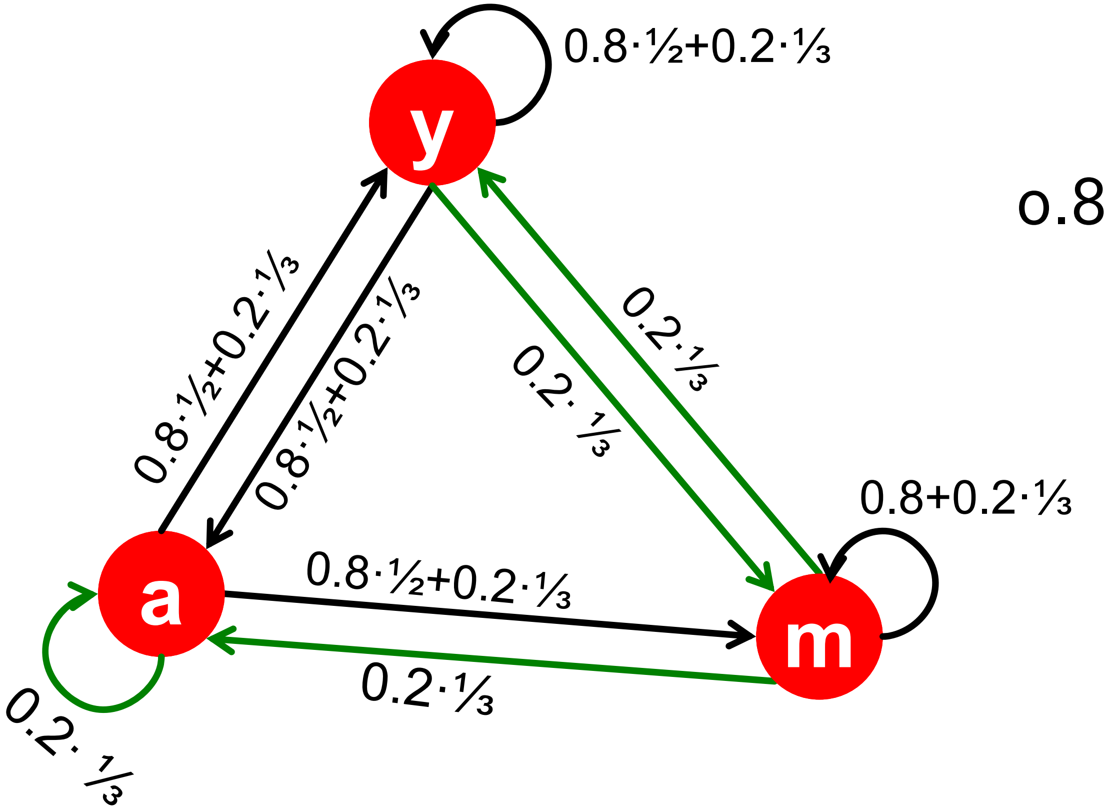
New Topic: Graph Data!

High dim. data	Graph data	Infinite data	Machine learning	Apps
Locality sensitive hashing	**PageRank,** SimRank	Filtering data streams	SVM	Recommender systems
Clustering	Community Detection	Web advertising	Decision Trees	Association Rules
Dimensionality reduction	Spam Detection	Queries on streams	Perceptron, kNN	Duplicate document detection

Example: PageRank Scores



Random Teleports ($\beta = 0.8$)



$$\begin{array}{c} \mathbf{M} \\ 0.8 \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 0 \\ 0 & 1/2 & 1 \end{bmatrix} \end{array} + 0.2 \begin{array}{c} [1/N]_{NxN} \\ \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \end{bmatrix} \end{array}$$

$$\begin{array}{c|ccc|} y & 7/15 & 7/15 & 1/15 \\ a & 7/15 & 1/15 & 1/15 \\ m & 1/15 & 7/15 & 13/15 \end{array}$$

A

$$\begin{array}{ccccccccc} y & & 1/3 & 0.33 & 0.24 & 0.26 & & 7/33 \\ a & = & 1/3 & 0.20 & 0.20 & 0.18 & \ldots & 5/33 \\ m & & 1/3 & 0.46 & 0.52 & 0.56 & & 21/33 \end{array}$$

r = A r

Equivalently: $r = \beta\, M \cdot r + \left[\frac{1-\beta}{N}\right]_N$

PageRank: The Complete Algorithm

- **Input: Graph G and parameter β**
  - Directed graph G with **spider traps** and **dead ends**
  - Parameter β
- **Output: PageRank vector r**
  - **Set:** $r_j^{(0)} = \frac{1}{N}, \quad t = 1$
  - **do:**
    - $\forall j$: $\boldsymbol{r'}_j^{(t)} = \sum_{i \to j} \boldsymbol{\beta} \frac{\boldsymbol{r}_i^{(t-1)}}{\boldsymbol{d}_i}$

      $\boldsymbol{r'}_j^{(t)} = \mathbf{0}$ if in-degree of $\boldsymbol{j}$ is **0**
    - **Now re-insert the leaked PageRank:**

      $\forall \boldsymbol{j}$: $\boldsymbol{r}_j^{(t)} = \boldsymbol{r'}_j^{(t)} + \frac{\mathbf{1-S}}{\boldsymbol{N}}$
    - $\boldsymbol{t} = \boldsymbol{t} + \mathbf{1}$ **where:** $S = \sum_j r'^{(t)}_j$
  - **while** $\sum_j \left| r_j^{(t)} - r_j^{(t-1)} \right| > \varepsilon$

If the graph has no dead-ends then the amount of leaked PageRank is **1-β**. But since we have dead-ends the amount of leaked PageRank may be larger. We have to explicitly account for it by computing **S**.

Some Problems with PageRank

- **Measures generic popularity of a page**
  - Will ignore/miss topic-specific authorities
  - **Solution:** Topic-Specific PageRank (**next**)
- **Uses a single measure of importance**
  - Other models of importance
  - **Solution:** Hubs-and-Authorities
- **Susceptible to Link spam**
  - Artificial link topographies created in order to boost page rank
  - **Solution:** TrustRank

Topic-Specific PageRank

Topic-Specific PageRank

- **Instead of generic popularity, can we measure popularity within a topic?**
- **Goal:** Evaluate Web pages not just according to their popularity, but by how close they are to a particular topic, e.g. “sports” or “history”
- **Allows search queries to be answered based on interests of the user**
  - **Example:** Query “Trojan” wants different pages depending on whether you are interested in sports, history, or computer security

Topic-Specific PageRank

- Random walker has a small probability of teleporting at any step
- **Teleport can go to:**
  - **Standard PageRank: Any page with equal probability**
    - To avoid dead-end and spider-trap problems
  - **Topic Specific PageRank: A topic-specific set of "relevant" pages (teleport set)**
- **Idea: Bias the random walk**
  - When walker teleports, he pick a page from a set $\boldsymbol{S}$
  - $\boldsymbol{S}$ contains only pages that are relevant to the topic
    - E.g., Open Directory (DMOZ) pages for a given topic/query
  - For each teleport set $\boldsymbol{S}$, we get a different vector $\boldsymbol{r_S}$

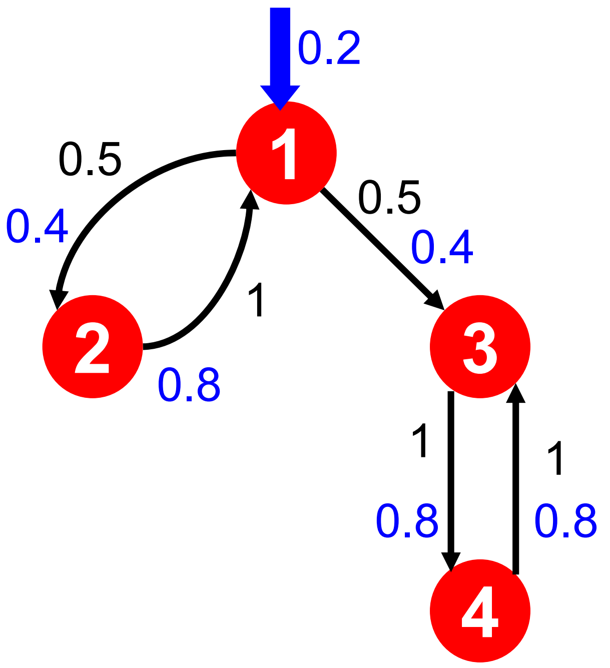
Matrix Formulation

- **To make this work all we need is to update the teleportation part of the PageRank formulation:**

$$A_{ij} = \begin{cases} \beta\, M_{ij} + (1-\beta)/|S| & \textbf{if } i \in S \\ \beta\, M_{ij} + 0 & \textbf{otherwise} \end{cases}$$

  - ***A*** is a stochastic matrix!
- We weighted all pages in the teleport set ***S*** equally
  - **Could also assign different weights to pages!**
- **Compute as for regular PageRank:**
  - Multiply by ***M***, then add a vector
  - Maintains sparseness

Example: Topic-Specific PageRank



Suppose $S = \{1\}$, $\beta = 0.8$

Node	Iteration				
	0	**1**	**2**	**…**	**stable**
1	0.25	0.4	0.28		0.294
2	0.25	0.1	0.16		0.118
3	0.25	0.3	0.32		0.327
4	0.25	0.2	0.24		0.261

S={1}, β=0.9:
r=[0.17, 0.07, 0.40, 0.36]
S={1} , β=0.8:
r=[0.29, 0.11, 0.32, 0.26]
S={1}, β=0.7:
r=[0.39, 0.14, 0.27, 0.19]

S={1,2,3,4}, β=0.8:
r=[0.13, 0.10, 0.39, 0.36]
S={1,2,3} , β=0.8:
r=[0.17, 0.13, 0.38, 0.30]
S={1,2} , β=0.8:
r=[0.26, 0.20, 0.29, 0.23]
S={1} , β=0.8:
r=[0.29, 0.11, 0.32, 0.26]

Discovering the Topic Vector S

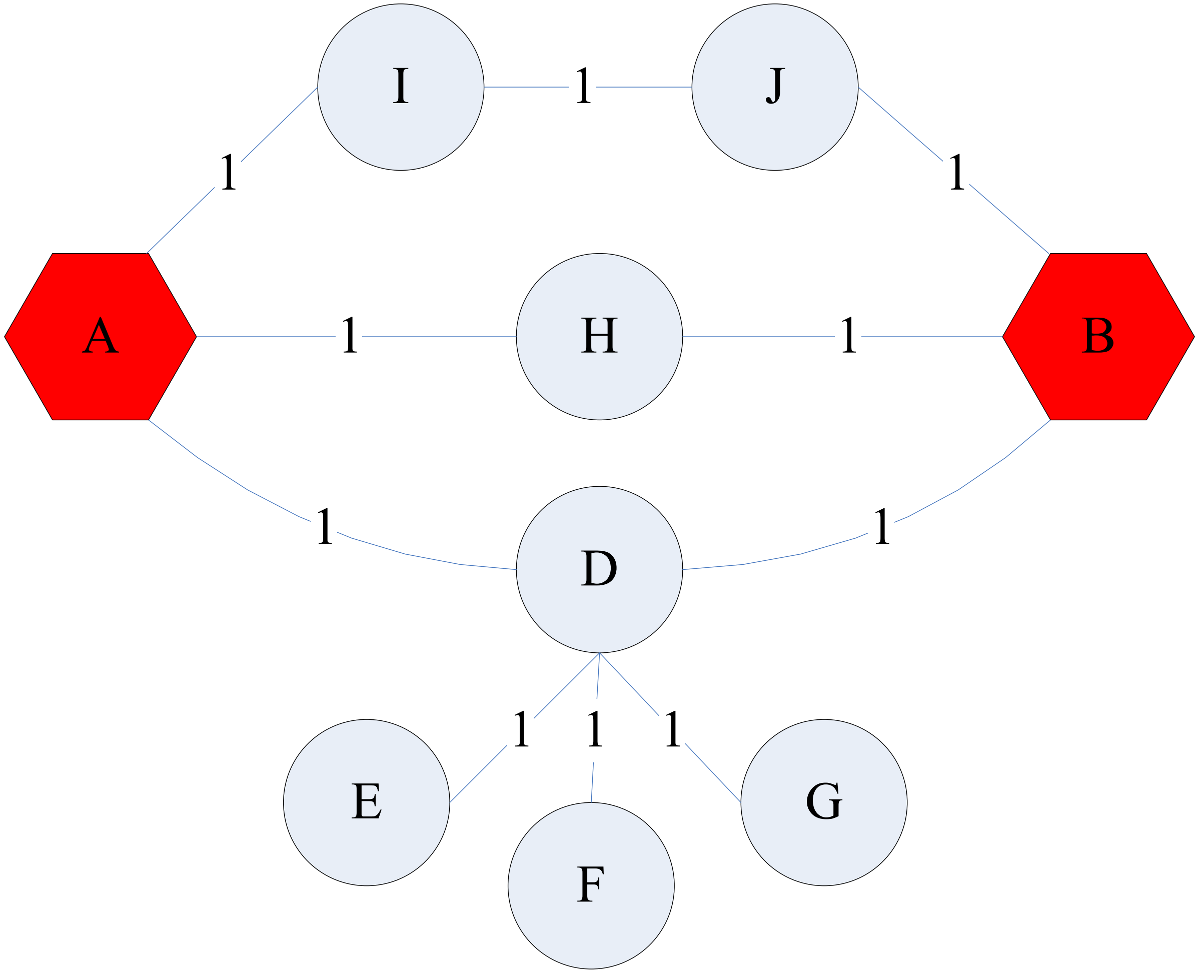
- **Create different PageRanks for different topics**
  - The 16 DMOZ top-level categories:
    - arts, business, sports,…
- **Which topic ranking to use?**
  - User can pick from a menu
  - Classify query into a topic
  - Can use the **context** of the query
    - E.g., query is launched from a web page talking about a known topic
    - History of queries e.g., "basketball" followed by "Jordan"
  - User context, e.g., user's bookmarks, …

Application to Measuring Proximity in Graphs

Random Walk with Restarts: set *S* is a single node

[Tong-Faloutsos, ‘06]

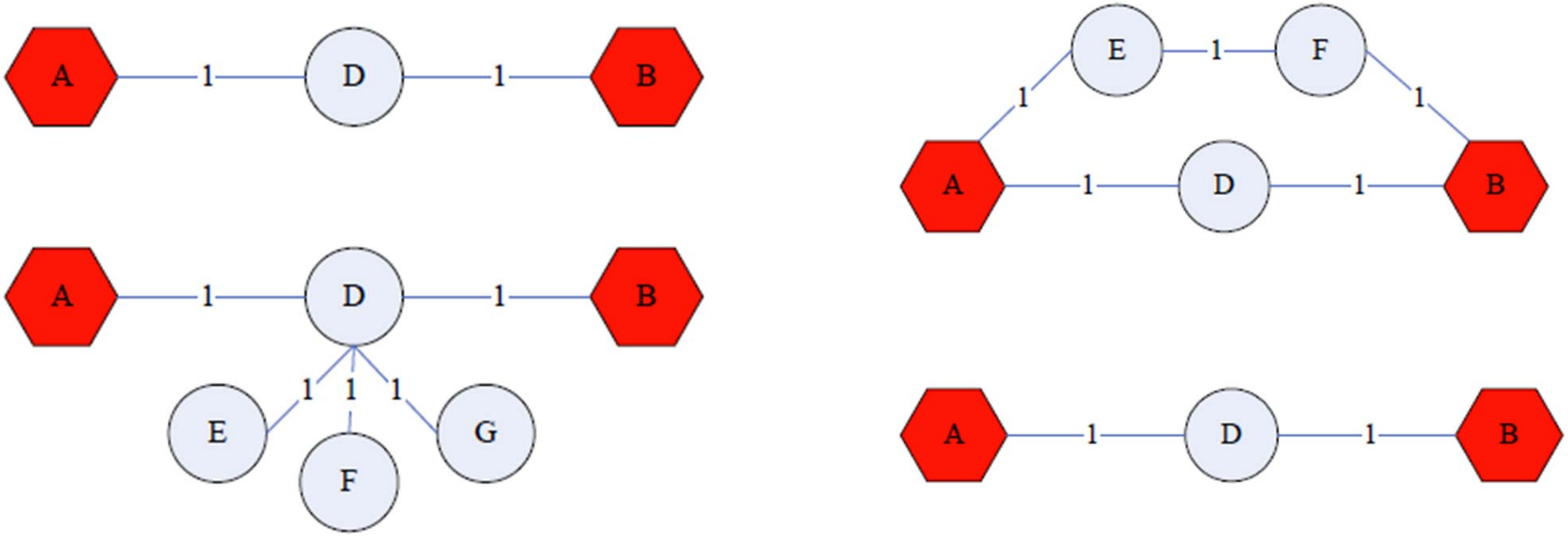
Proximity on Graphs



a.k.a.: Relevance, Closeness, ‘Similarity’…

Good proximity measure?

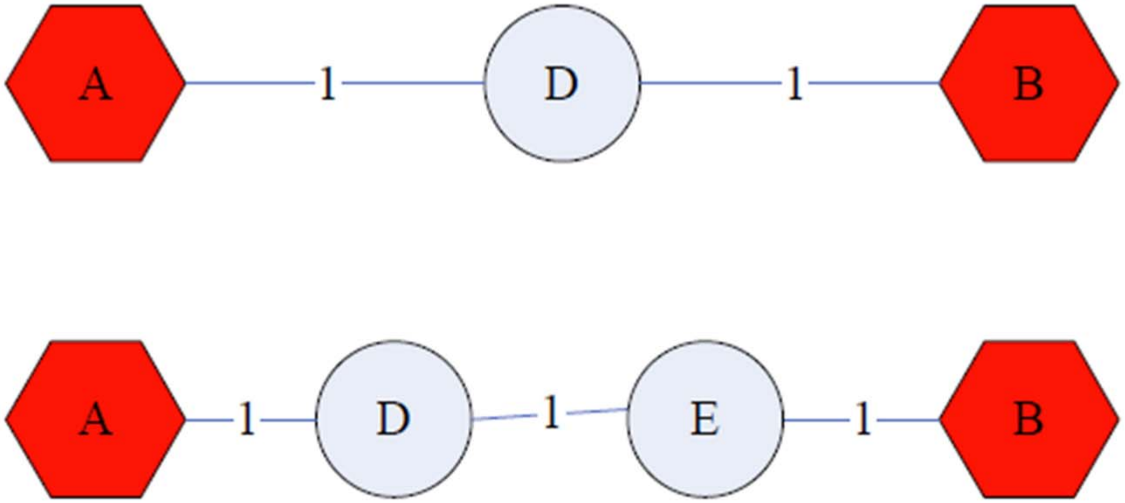
- **Shortest path is not good:**



- **No effect of degree-1 nodes (E, F, G)!**
- Multi-faceted relationships

Good proximity measure?

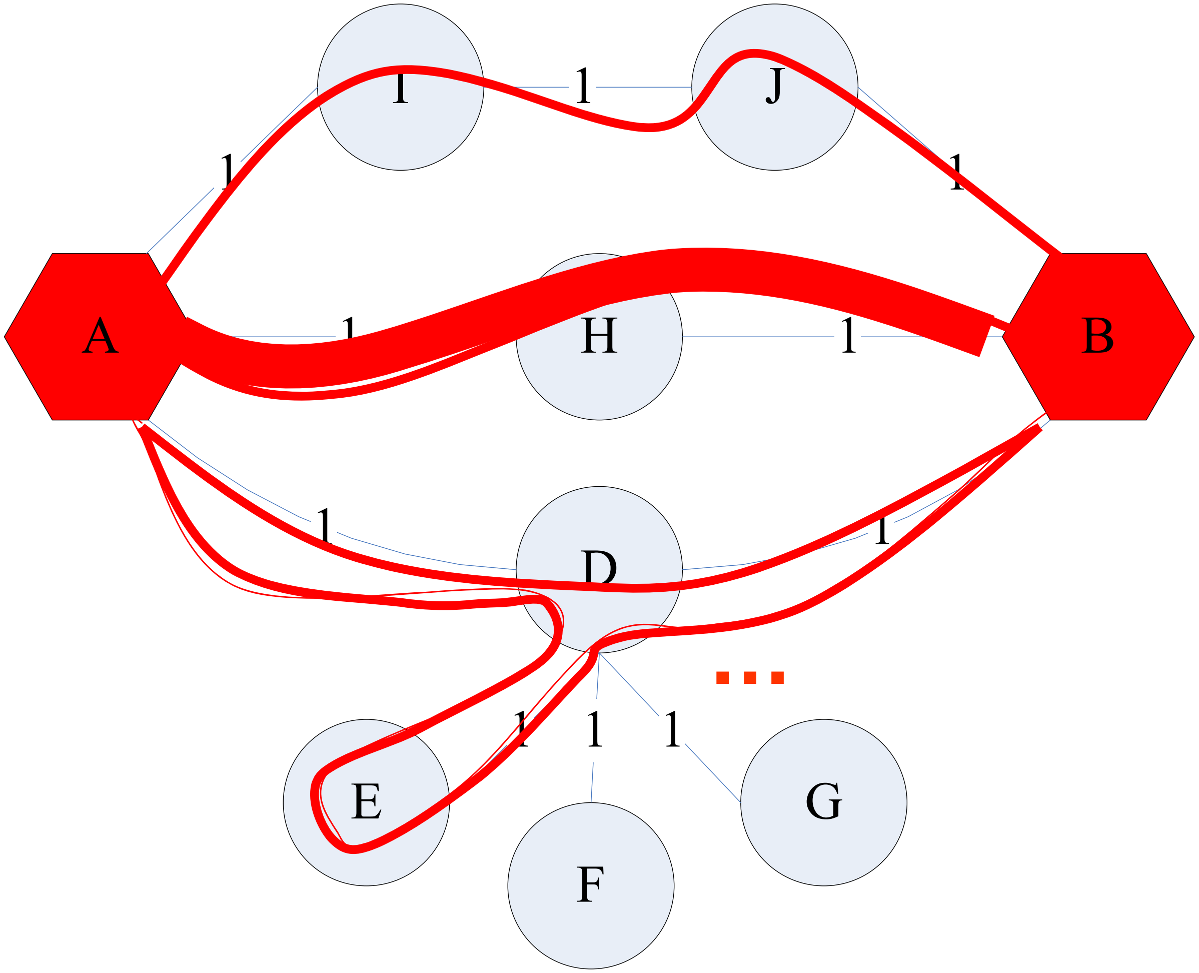
- Network flow is not good:



- Does not punish long paths

[Tong-Faloutsos, '06]

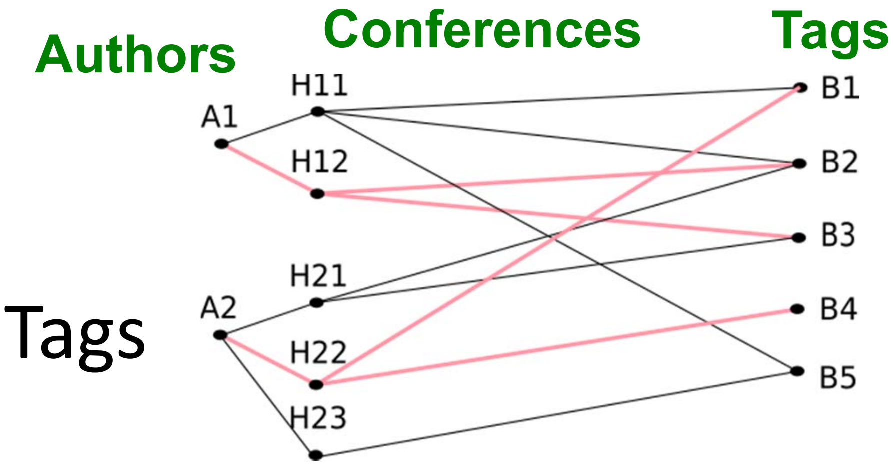
What is good notion of proximity?



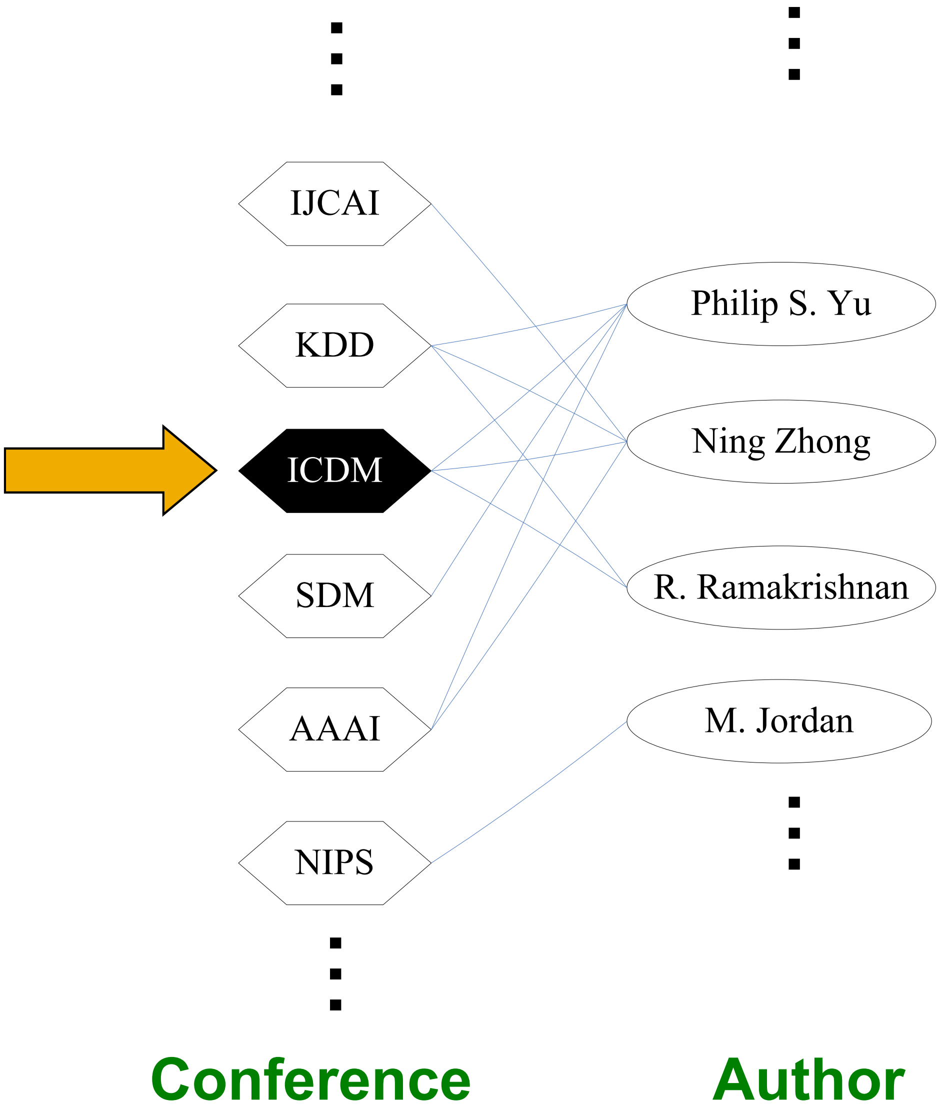
- Multiple connections
- Quality of connection
    - Direct & Indirect connections
    - Length, Degree, Weight…

SimRank: Idea

- **SimRank:** Random walks from a **fixed node** on ***k***-partite graphs
- **Setting:** ***k***-partite graph with ***k*** types of nodes
  - E.g.: Authors, Conferences, Tags
- **Topic Specific PageRank** from node ***u***: **teleport set *S* = {*u*}**
- Resulting scores measure similarity/proximity to node ***u***
- **Problem:**
  - Must be done once for each node ***u***
  - Suitable for sub-Web-scale applications



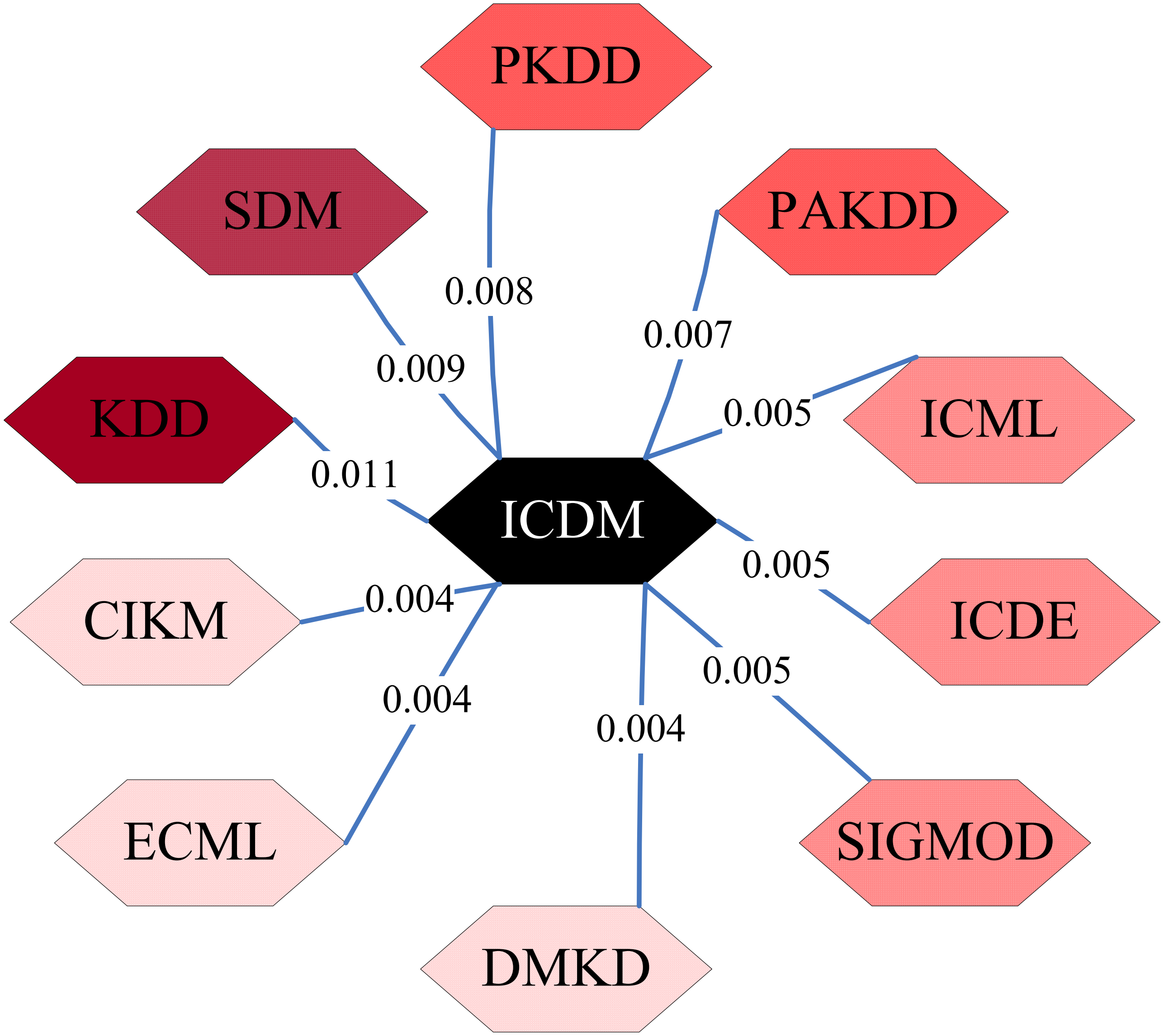
SimRank: Example



Q: What is most related conference to **ICDM**?

A: Topic-Specific PageRank with teleport set S={ICDM}

SimRank: Example



PageRank: Summary

- **“Normal” PageRank:**
  - Teleports uniformly at random to any node
  - All nodes have the same probability of surfer landing there: **S =** [0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1]
- **Topic-Specific PageRank** also known as **Personalized PageRank:**
  - Teleports to a topic specific set of pages
  - Nodes can have different probabilities of surfer landing there: **S =** [0.1, 0, 0, 0.2, 0, 0, 0.5, 0, 0, 0.2]
- **Random Walk with Restarts:**
  - Topic-Specific PageRank where teleport is always to the same node. S=[0, 0, 0, 0, **1**, 0, 0, 0, 0, 0, 0]

TrustRank:
Combating the Web Spam

What is Web Spam?

- **Spamming:**
  - Any deliberate action to boost a web page’s position in search engine results, incommensurate with page’s real value
- **Spam:**
  - Web pages that are the result of spamming
- This is a very broad definition
  - **SEO** industry might disagree!
  - SEO = search engine optimization
- Approximately **10-15%** of web pages are spam

Web Search

- **Early search engines:**
  - Crawl the Web
  - Index pages by the words they contained
  - Respond to search queries (lists of words) with the pages containing those words
- **Early page ranking:**
  - Attempt to order pages matching a search query by "importance"
  - **First search engines considered:**
    - **(1)** Number of times query words appeared
    - **(2)** Prominence of word position, e.g. title, header

First Spammers

- As people began to use search engines to find things on the Web, those with commercial interests tried to **exploit search engines** to bring people to their own site – whether they wanted to be there or not
- **Example:**
  - Shirt-seller might pretend to be about “movies”
- **Techniques for achieving high relevance/importance for a web page**

First Spammers: Term Spam

- **How do you make your page appear to be about movies?**
  - **(1)** Add the word movie 1,000 times to your page
  - Set text color to the background color, so only search engines would see it
  - **(2)** Or, run the query “movie” on your target search engine
  - See what page came first in the listings
  - Copy it into your page, make it “invisible”
- **These and similar techniques are term spam**

Google's Solution to Term Spam

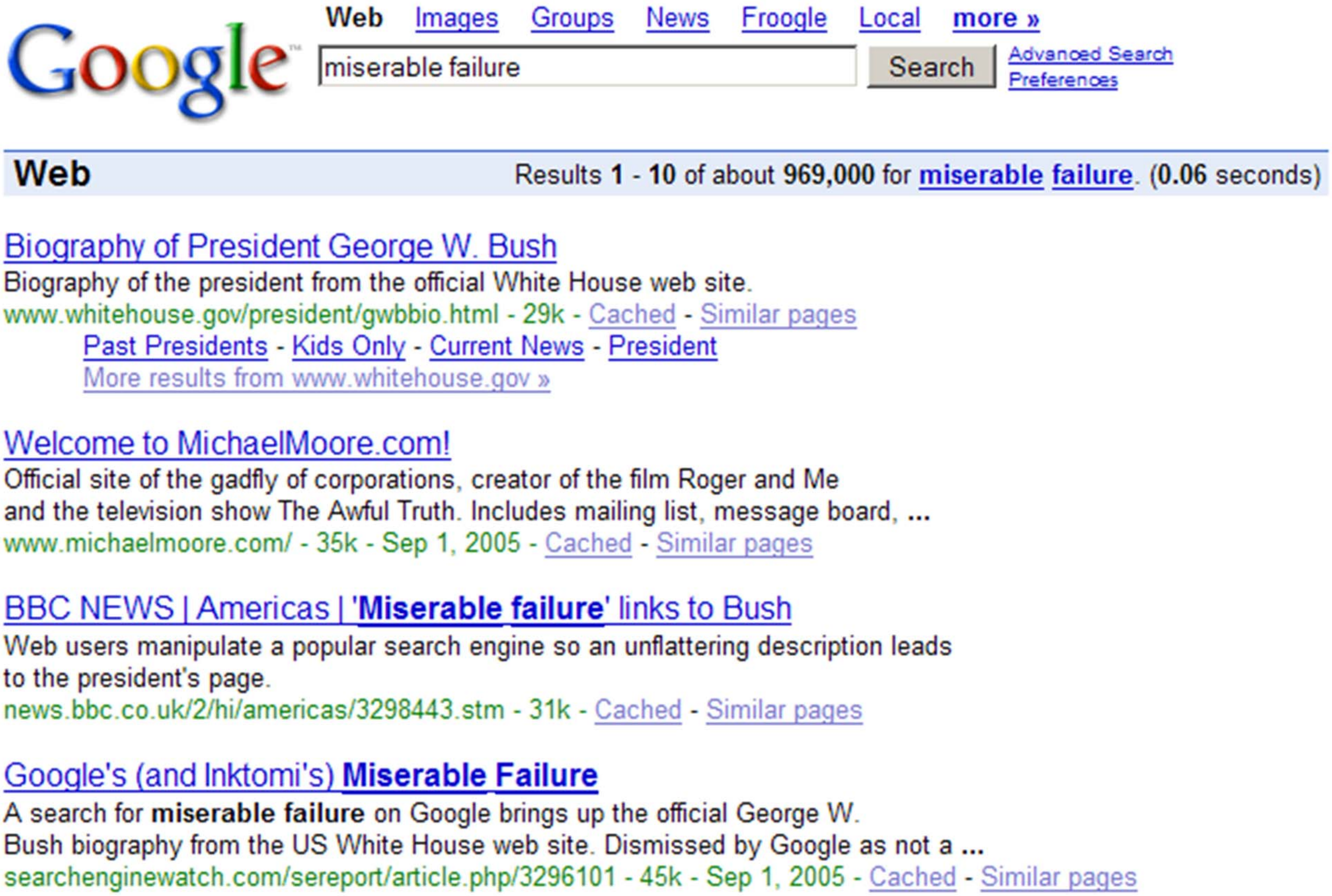
- **Believe what people say about you, rather than what you say about yourself**
  - Use words in the anchor text (words that appear underlined to represent the link) and its surrounding text

- **PageRank as a tool to measure the "importance" of Web pages**

Why It Works?

- **Our hypothetical shirt-seller looses**
  - Saying he is about movies doesn’t help, because others don’t say he is about movies
  - His page isn’t very important, so it won’t be ranked high for shirts or movies
- **Example:**
  - Shirt-seller creates 1,000 pages, each links to his with “movie” in the anchor text
  - These pages have no links in, so they get little PageRank
  - So the shirt-seller can’t beat truly important movie pages, like IMDB

Why it does not work?



miserable failure

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 from the US White House web site. Dismissed by Google as not a ...
searchenginewatch.com/sereport/article.php/3296101 - 45k - Sep 1, 2005 - Cached - Similar pages

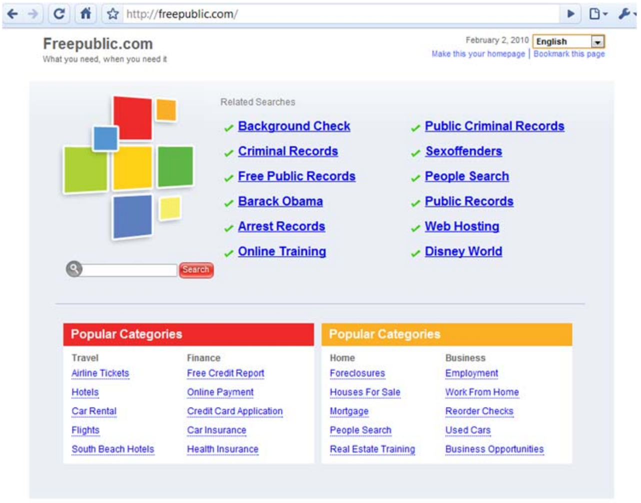
SPAM FARMING

Google vs. Spammers: Round 2!

- Once Google became the dominant search engine, spammers began to work out ways to fool Google

- **Spam farms** were developed to concentrate PageRank on a single page

- **Link spam:**
  - Creating link structures that boost PageRank of a particular page



http://freepublic.com/

Freepublic.com

What you need, when you need it

February 2, 2010 English

Make this your homepage | Bookmark this page

Related Searches

- ✓ Background Check
- ✓ Criminal Records
- ✓ Free Public Records
- ✓ Barack Obama
- ✓ Arrest Records
- ✓ Online Training
- ✓ Public Criminal Records
- ✓ Sexoffenders
- ✓ People Search
- ✓ Public Records
- ✓ Web Hosting
- ✓ Disney World

Popular Categories

Travel	Finance
Airline Tickets	Free Credit Report
Hotels	Online Payment
Car Rental	Credit Card Application
Flights	Car Insurance
South Beach Hotels	Health Insurance

Popular Categories

Home	Business
Foreclosures	Employment
Houses For Sale	Work From Home
Mortgage	Reorder Checks
People Search	Used Cars
Real Estate Training	Business Opportunities

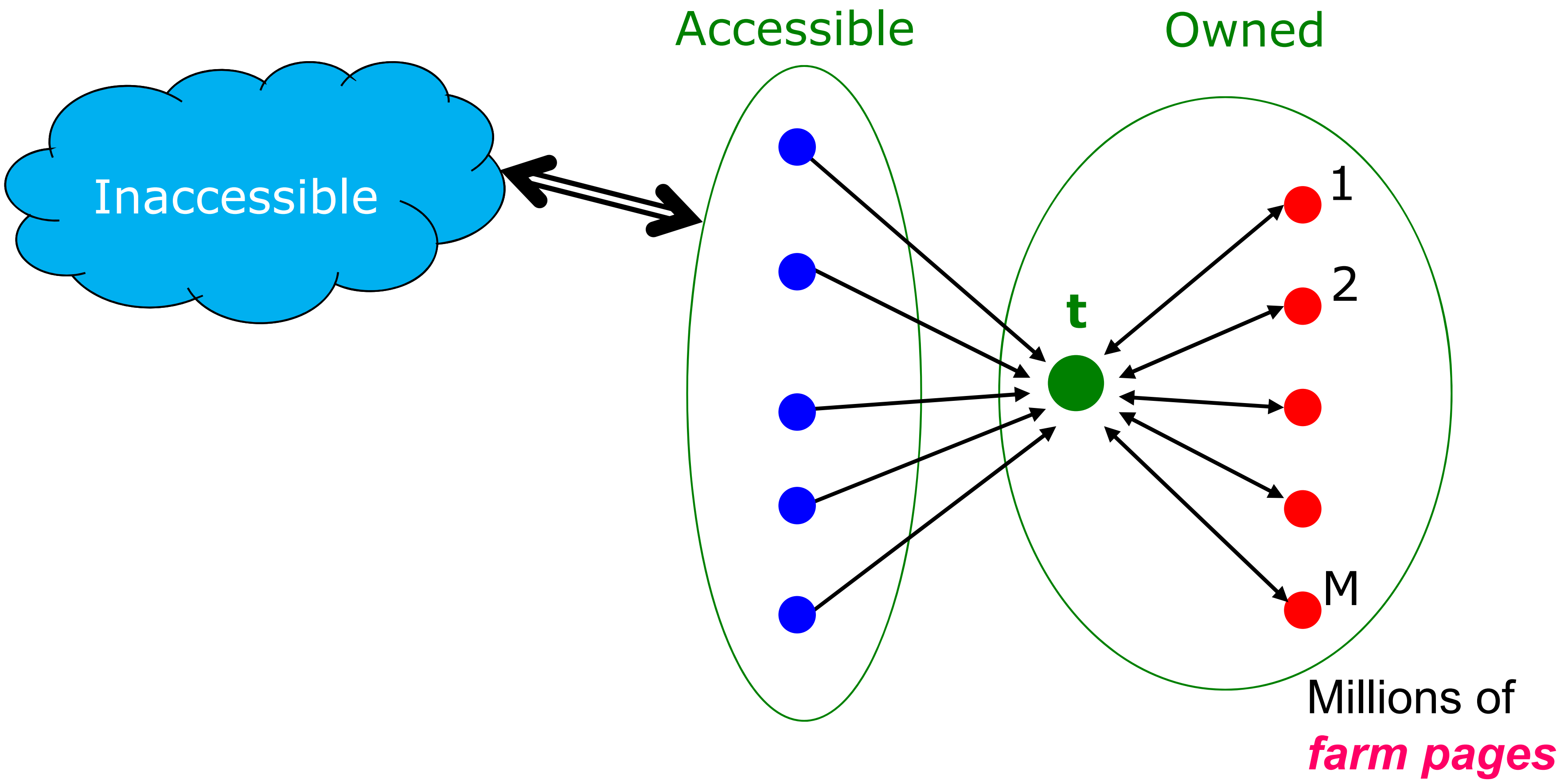
Link Spamming

- **Three kinds of web pages from a spammer's point of view**
  - **Inaccessible pages**
  - **Accessible pages**
    - e.g., blog comments pages
    - spammer can post links to his pages
  - **Owned pages**
    - Completely controlled by spammer
    - May span multiple domain names

Link Farms

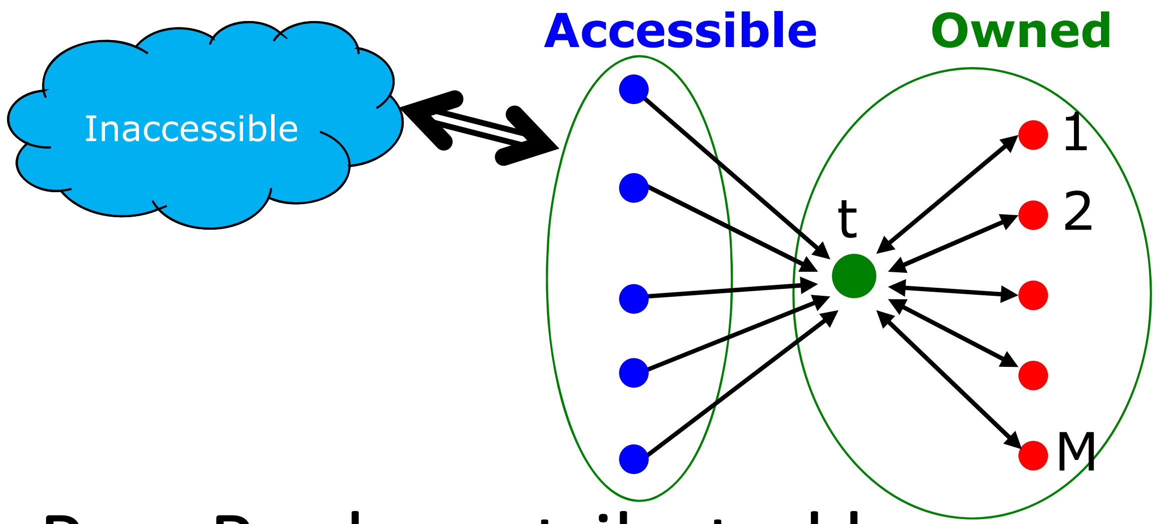
- **Spammer's goal:**
  - Maximize the PageRank of target page ***t***
- **Technique:**
  - Get as many links from accessible pages as possible to target page ***t***
  - Construct "link farm" to get PageRank multiplier effect

Link Farms



One of the most common and effective organizations for a link farm

Analysis



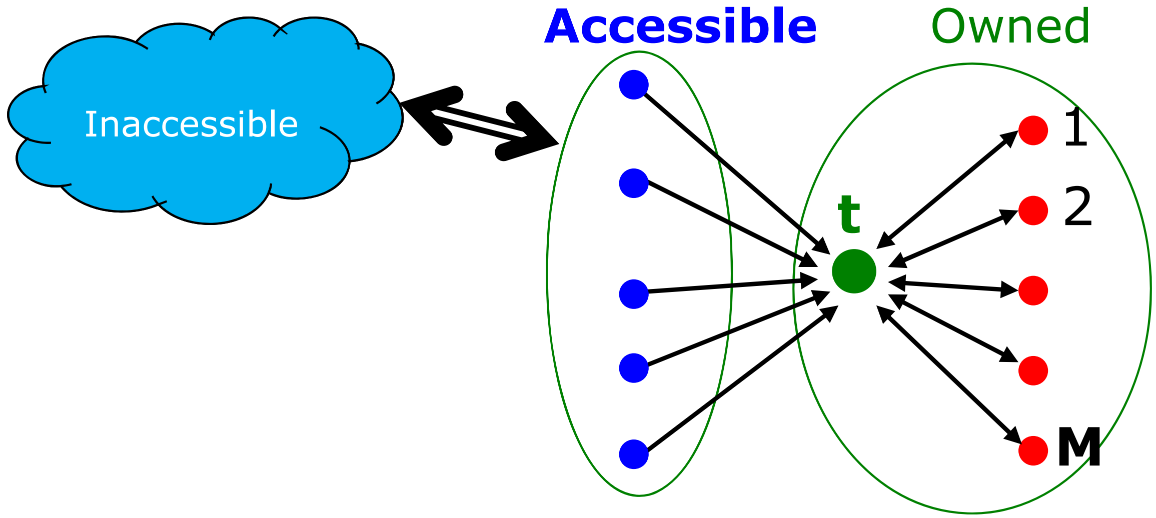
N…# pages on the web
M…# of pages spammer owns

- **x**: PageRank contributed by accessible pages
- **y**: PageRank of target page ***t***
- Rank of each "farm" page $= \frac{\beta y}{M} + \frac{1-\beta}{N}$
- $y = x + \beta M\left[\frac{\beta y}{M} + \frac{1-\beta}{N}\right] + \frac{1-\beta}{N}$

  $= x + \beta^2 y + \frac{\beta(1-\beta)M}{N} + \boxed{\frac{1-\beta}{N}}$

  Very small; ignore
  Now we solve for ***y***
- $y = \frac{x}{1-\beta^2} + c\frac{M}{N}$ where $c = \frac{\beta}{1+\beta}$

Analysis



N…# pages on the web
M…# of pages spammer owns

- $y = \frac{x}{1-\beta^2} + c\frac{M}{N}$ where $c = \frac{\beta}{1+\beta}$
- For $\beta = 0.85$, $1/(1-\beta^2) = 3.6$

- Multiplier effect for acquired PageRank
- By making ***M*** large, we can make ***y*** as **large as we want**

TrustRank: Combating the Web Spam

Combating Spam

- **Combating term spam**
  - Analyze text using statistical methods
  - Similar to email spam filtering
  - Also useful: Detecting approximate duplicate pages
- **Combating link spam**
  - **Detection and blacklisting of structures that look like spam farms**
    - Leads to another war – hiding and detecting spam farms
  - **TrustRank** = topic-specific PageRank with a teleport set of **trusted pages**
    - Example: .edu domains, similar domains for non-US schools

TrustRank: Idea

- **Basic principle: Approximate isolation**
  - It is rare for a “good” page to point to a “bad” (spam) page
- Sample a set of **seed pages** from the web
- Have an **oracle** (**human**) to identify the good pages and the spam pages in the seed set
  - **Expensive task,** so we must make seed set as small as possible

Trust Propagation

- Call the subset of seed pages that are identified as **good** the **trusted pages**

- Perform a topic-sensitive PageRank with **teleport set = trusted pages**
  - **Propagate trust through links:**
    - Each page gets a trust value between **0** and **1**

- **Solution 1:** **Use a threshold value and mark all pages below the trust threshold as spam**

Simple Model: Trust Propagation

- **Set trust of each trusted page to 1**
- Suppose trust of page ***p*** is $\boldsymbol{t_p}$
  - Page ***p*** has a set of out-links $\boldsymbol{o_p}$
- For each $\boldsymbol{q \in o_p}$, ***p*** **confers the trust** to ***q***
  - $\beta\, \boldsymbol{t_p / |o_p|}$ for $0 < \beta < 1$
- **Trust is additive**
  - Trust of ***p*** is the sum of the trust conferred on ***p*** by all its in-linked pages
- **Note similarity to Topic-Specific PageRank**
  - Within a scaling factor, **TrustRank = PageRank** with trusted pages as teleport set

Why is it a good idea?

- **Trust attenuation:**
  - The degree of trust conferred by a trusted page decreases with the distance in the graph
- **Trust splitting:**
  - The larger the number of out-links from a page, the less scrutiny the page author gives each out-link
  - Trust is **split** across out-links

Picking the Seed Set

- **Two conflicting considerations:**
  - Human has to inspect each seed page, so seed set must be as small as possible
  - Must ensure every **good page** gets adequate trust rank, so need make all good pages reachable from seed set by short paths

Approaches to Picking Seed Set

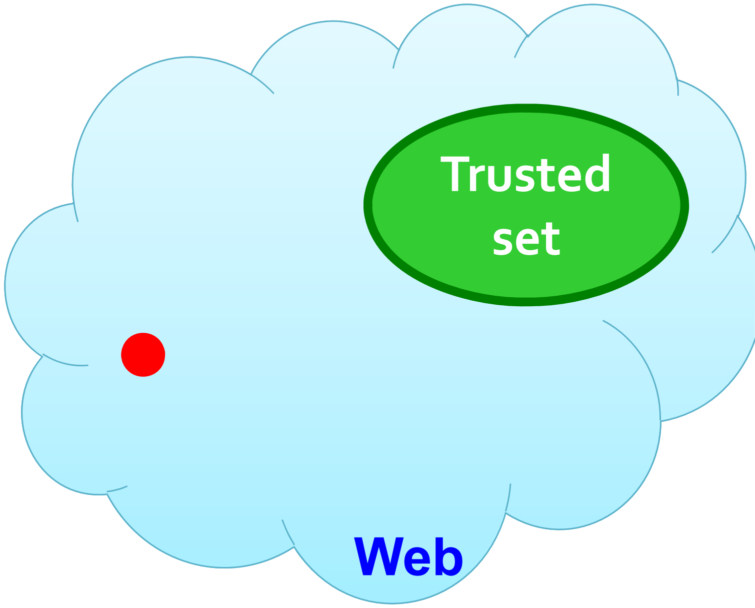
- Suppose we want to pick a seed set of ***k*** pages
- **How to do that?**
- **(1) PageRank:**
  - Pick the top ***k*** pages by PageRank
  - Theory is that you can't get a bad page's rank really high
- **(2) Use trusted domains** whose membership is controlled, like .edu, .mil, .gov

Spam Mass

- In the **TrustRank** model, we start with good pages and propagate trust

- **Complementary view:**
  What fraction of a page's PageRank comes from **spam** pages?

- In practice, we don't know all the spam pages, so we need to estimate



Spam Mass Estimation

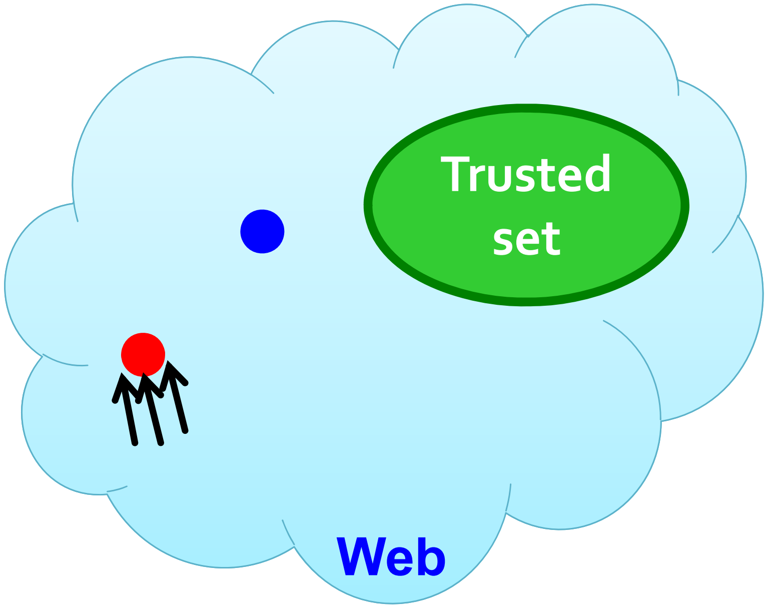
Solution 2:

- r_p = PageRank of page ***p***
- r_p^+ = PageRank of ***p*** with teleport into **trusted** pages only

- **Then:** What fraction of a page's PageRank comes from **spam** pages?

$$r_p^- = r_p - r_p^+$$

- **Spam mass of *p* =** $\frac{r_p^-}{r_p}$
  - Pages with high spam mass are spam.



HITS: Hubs and Authorities

Hubs and Authorities

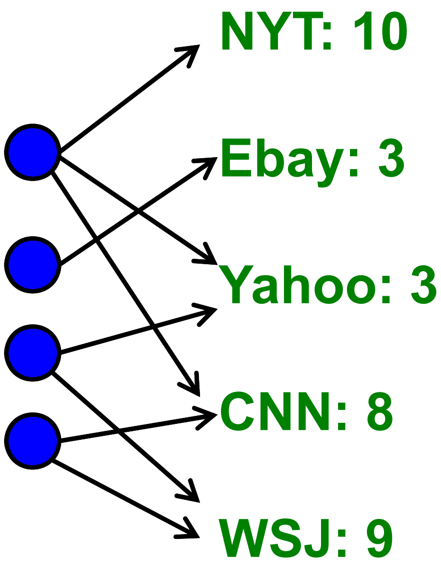
- HITS (Hypertext-Induced Topic Selection)
  - **Is a measure of importance of pages or documents, similar to PageRank**
  - Proposed at around same time as PageRank (‘98)
- **Goal**: Say we want to find good newspapers
  - Don’t just find newspapers. Find “experts” – people who link in a coordinated way to good newspapers
- **Idea: Links as votes**
  - **Page is more important if it has more links**
    - In-coming links? Out-going links?

Finding newspapers

- **Hubs and Authorities**

  Each page has 2 scores:

  - **Quality as an expert (hub):**
    - Total sum of votes of authorities pointed to
  - **Quality as a content (authority):**
    - Total sum of votes coming from experts



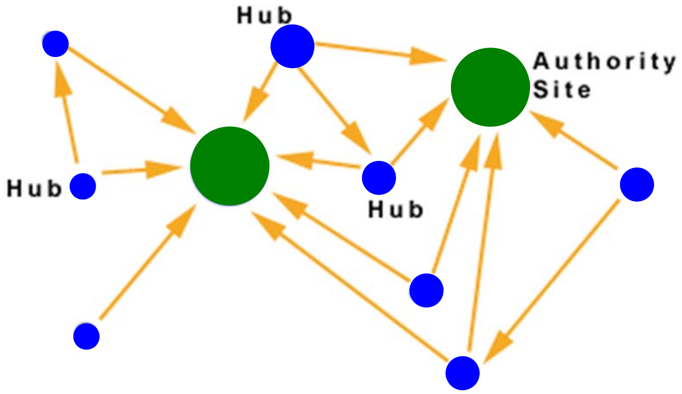
- **Principle of repeated improvement**

Hubs and Authorities

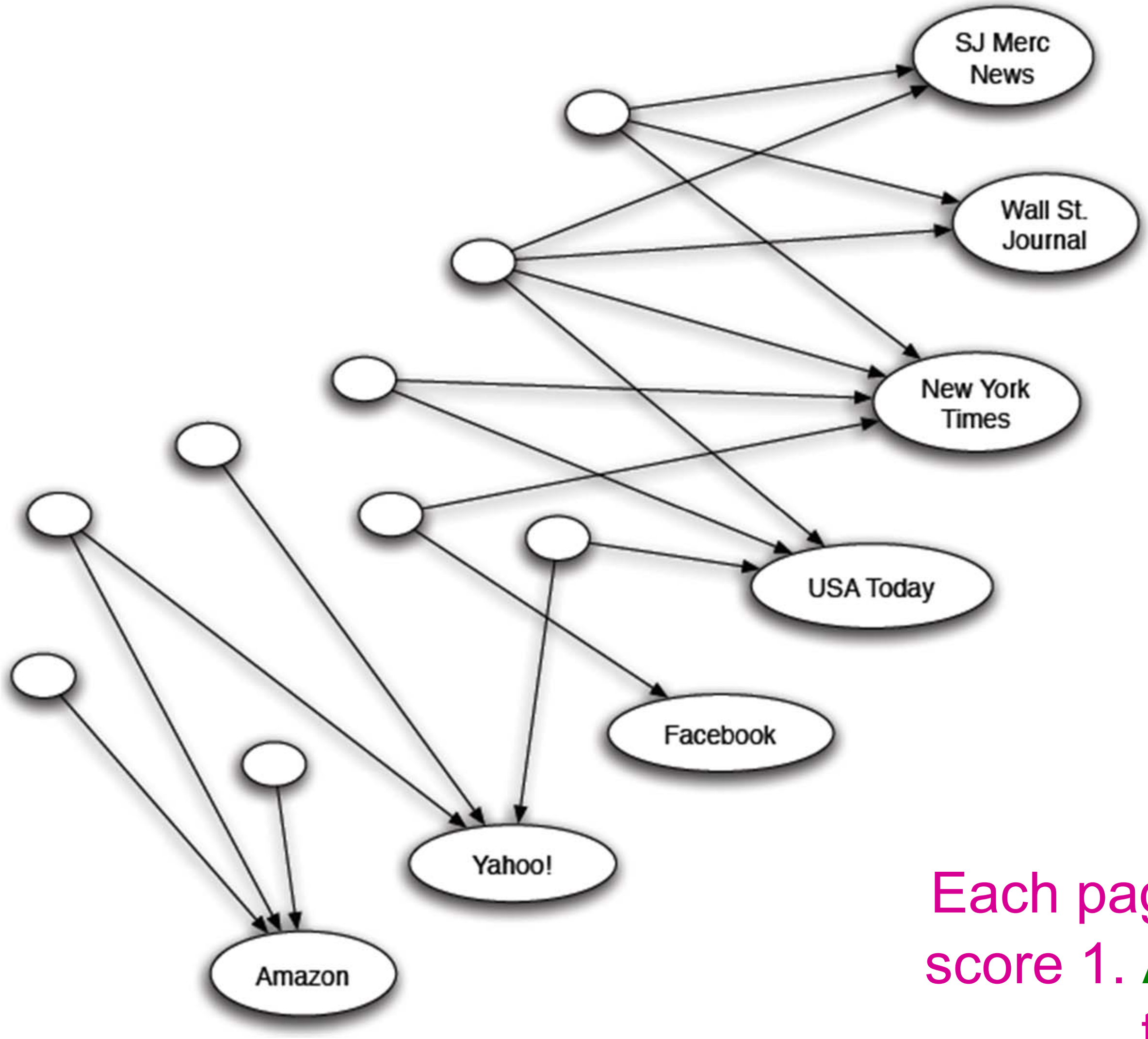
Interesting pages fall into two classes:

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



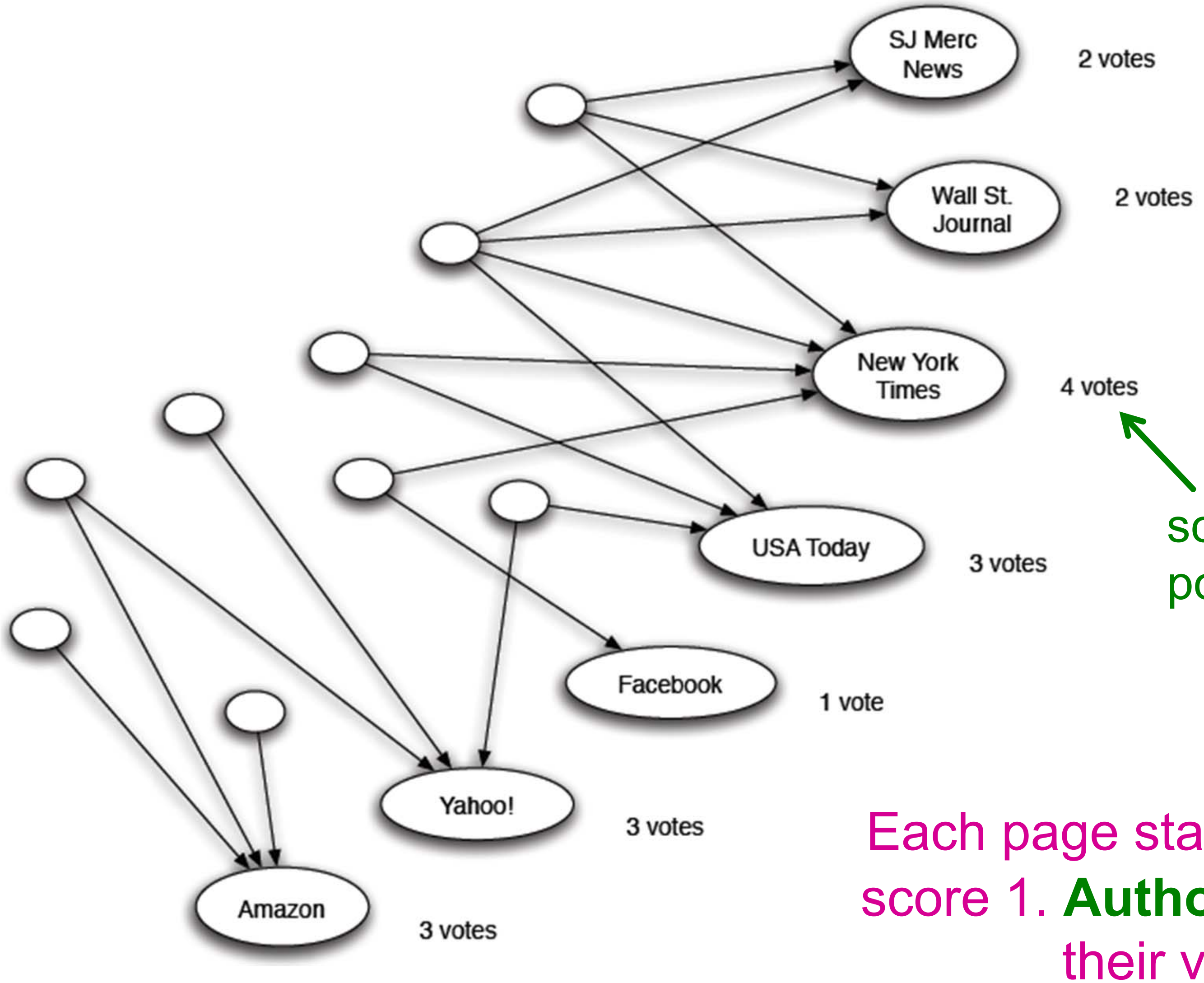
Counting in-links: Authority



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)

Counting in-links: Authority



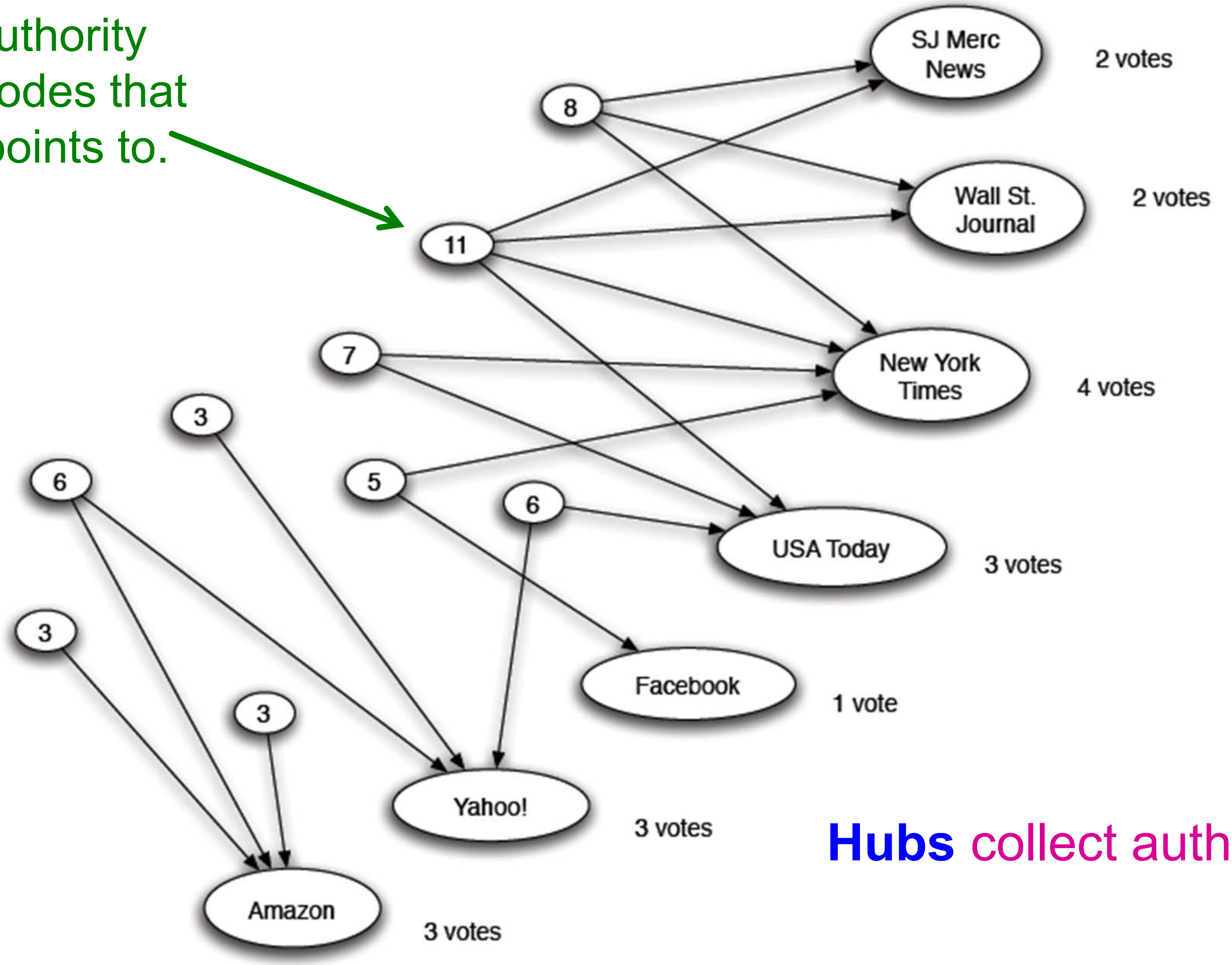
Sum of **hub** scores of nodes pointing to NYT.

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

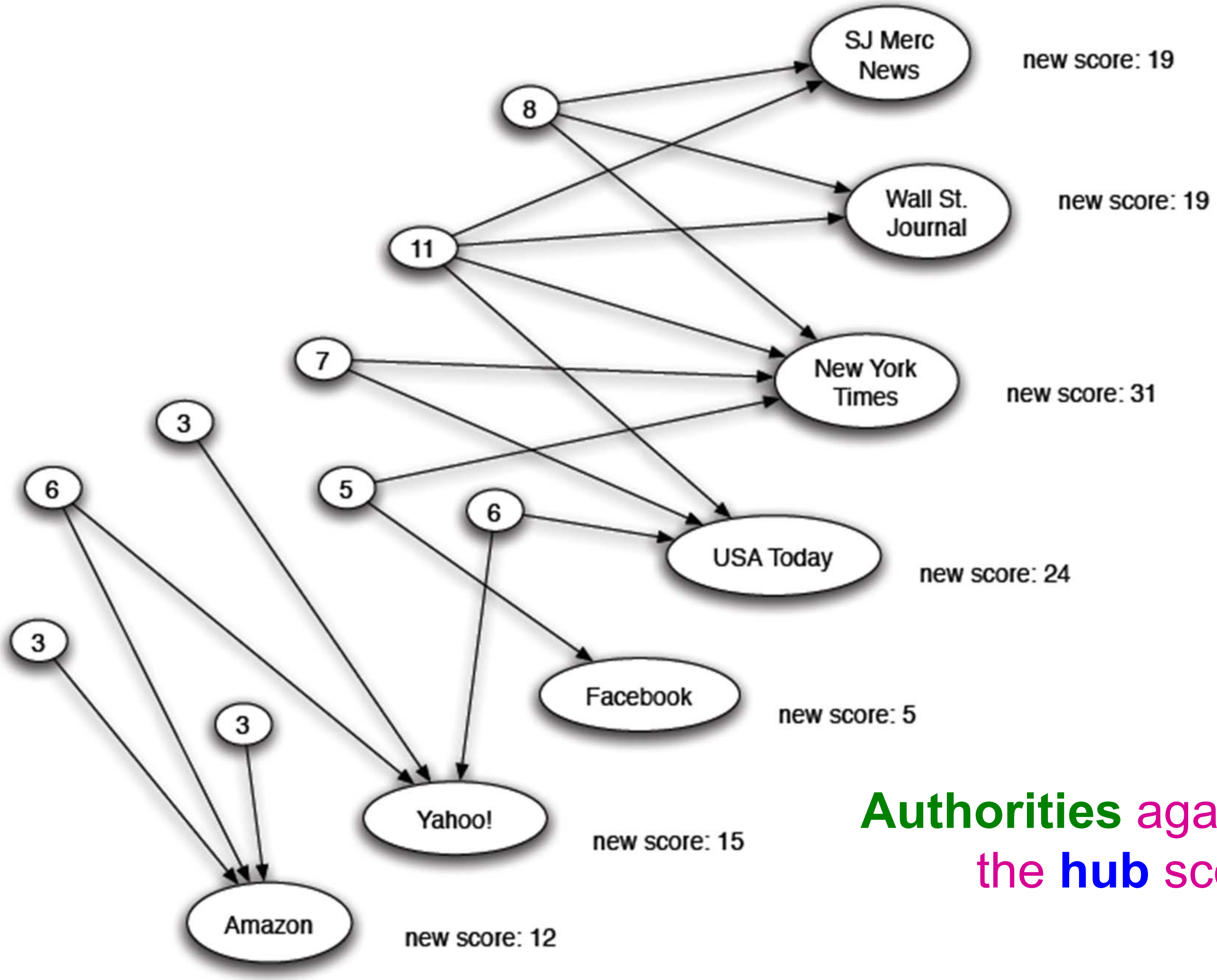
Sum of authority scores of nodes that the node points to.



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 again collect the **hub** scores

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

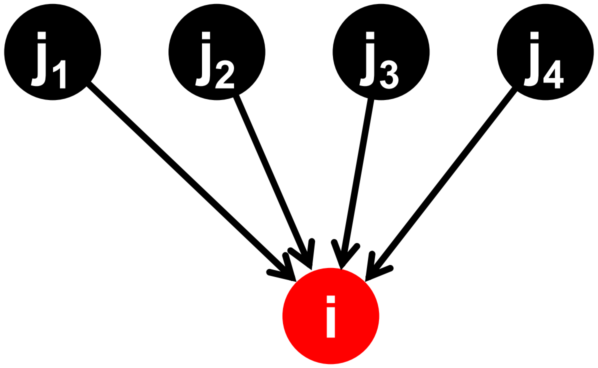
Mutually Recursive Definition

- **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 $\boldsymbol{h}$ and $\boldsymbol{a}$

[Kleinberg '98]

Hubs and Authorities

- **Each page i has 2 scores:**
  - Authority score: $\boldsymbol{a_i}$
  - Hub score: $\boldsymbol{h_i}$

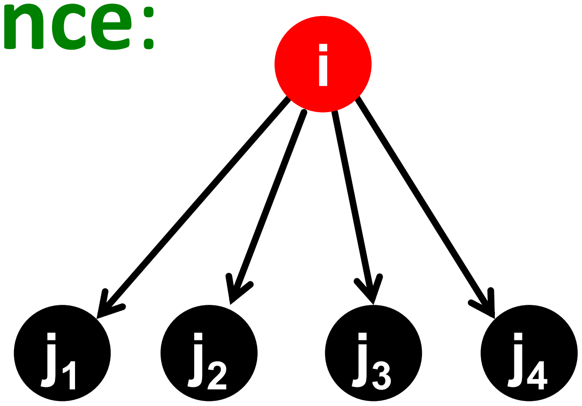


$$\boldsymbol{a_i = \sum_{j \to i} h_j}$$

HITS algorithm:

- Initialize: $a_j^{(0)} = 1/\sqrt{\mathrm{N}}, \ \ \mathrm{h}_\mathrm{j}^{(0)} = 1/\sqrt{\mathrm{N}}$
- Then keep iterating until **convergence**:
  - $\forall \boldsymbol{i}$: Authority: $a_i^{(t+1)} = \sum_{\boldsymbol{j \to i}} h_j^{(t)}$
  - $\forall \boldsymbol{i}$: Hub: $h_i^{(t+1)} = \sum_{\boldsymbol{i \to j}} a_j^{(t)}$
  - $\forall \boldsymbol{i}$: Normalize:

$$\sum_i \left(a_i^{(t+1)}\right)^2 = 1, \sum_j \left(h_j^{(t+1)}\right)^2 = 1$$



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

[Kleinberg '98]

Hubs and Authorities

- HITS converges to a single stable point
- **Notation:**
  - Vector $\boldsymbol{a} = (a_1 \dots, a_n)$, $\boldsymbol{h} = (h_1 \dots, h_n)$
  - Adjacency matrix $\boldsymbol{A}$ (NxN): $\boldsymbol{A_{ij}} = 1$ if $i \to j$, 0 otherwise
- **Then $\boldsymbol{h_i} = \sum_{i \to j} \boldsymbol{a_j}$**

  can be rewritten as $\boldsymbol{h_i} = \sum_j \boldsymbol{A_{ij}} \cdot \boldsymbol{a_j}$

  So: $\boldsymbol{h} = \boldsymbol{A} \cdot \boldsymbol{a}$
- **Similarly, $\boldsymbol{a_i} = \sum_{j \to i} \boldsymbol{h_j}$**

  can be rewritten as $\boldsymbol{a_i} = \sum_j \boldsymbol{A_{ji}} \cdot \boldsymbol{h_j} = \boldsymbol{A^T} \cdot \boldsymbol{h}$

Hubs and Authorities

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

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

Convergence criterion:

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

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

$\boldsymbol{a}$ is updated (in 2 steps):

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

$\boldsymbol{h}$ is updated (in 2 steps):

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

Repeated matrix powering

Existence and Uniqueness

- $h = \lambda\, A\, a$
- $a = \mu\, A^T\, h$
- $h = \lambda\, \mu\, A\, A^T\, h$
- $a = \lambda\, \mu\, A^T A\, a$

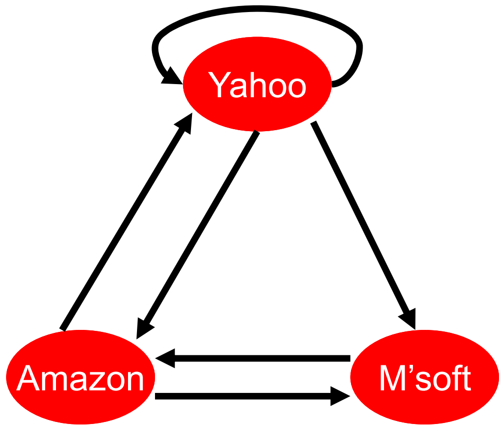
$\lambda = 1 / \sum h_i$

$\mu = 1 / \sum a_i$

- Under reasonable assumptions about **A**, HITS **converges to vectors *h^** and *a^****:
  - ***h^**** is the **principal eigenvector** of matrix ***$A\, A^T$***
  - ***a^**** is the **principal eigenvector** of matrix ***$A^T A$***

Example of HITS

$$A = \boxed{\begin{matrix} 1 & 1 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{matrix}} \qquad A^T = \boxed{\begin{matrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{matrix}}$$



$$\begin{array}{lcllllcl}
h(\text{yahoo}) & = & .58 & .80 & .80 & .79 & \cdots & .788 \\
h(\text{amazon}) & = & .58 & .53 & .53 & .57 & \cdots & .577 \\
h(\text{m'soft}) & = & .58 & .27 & .27 & .23 & \cdots & .211
\end{array}$$

$$\begin{array}{lcllllcl}
a(\text{yahoo}) & = & .58 & .58 & .62 & .62 & \cdots & .628 \\
a(\text{amazon}) & = & .58 & .58 & .49 & .49 & \cdots & .459 \\
a(\text{m'soft}) & = & .58 & .58 & .62 & .62 & \cdots & .628
\end{array}$$

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**