

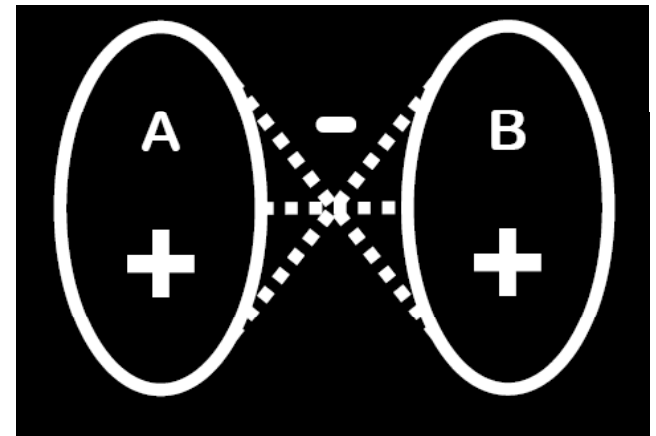
Announcements

- **Poster session:**
 - Thursday December 10 3-6 pm Gates Atrium
 - We will provide poster boards
 - 30% of project grade
- **Project writeup:**
 - Due Friday December 11
 - PDF by email to course staff list
 - Max 6 min 4 pages in ACM format
 - More info on the website
 - 70% of project grade

Coalitions in signed networks

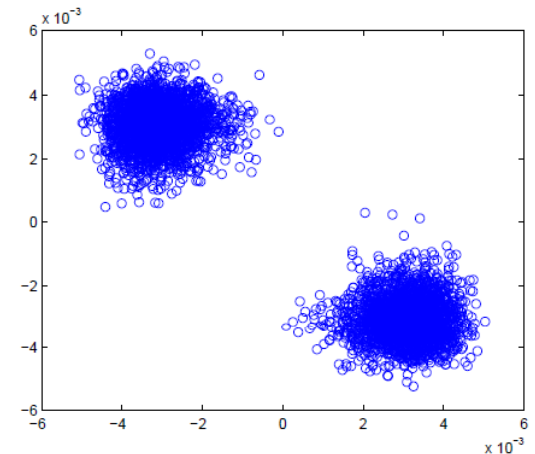
- Received 15 entries
- Top score:
 - RPL: 351,944
 - GNP: 1,150,563 (5 got the OPT)
- Top 5:

Name	Score
Shayan_Oveis_Gharan	1,502,507
Farnaz_Ronaghi_Khameneh	-2
Ying_Wang	-11
Abhijeet_Mohapatra	-92
Nipun_Dave	-162



Ying Wang's approach

- Idea: combine
 - min-cut on positive edges
 - 2nd smallest eigenvector x of Laplacian
 - max-cut on negative edges
 - Largest eigenvector y of normalized Laplacian
- So for each node 2 scores (positions):
 - Min-cut score, Max-cut score
- Now simply partition the nodes
 - GNP (6 edges from best solution): 1,150,557
 - RPL: 342,021 (and after local updates 351,939)



Link Analysis for Web Search: Hubs and Authorities & PageRank

CS 322: (Social and Information) Network Analysis
Jure Leskovec
Stanford University



Web

- Many many documents
- How to organize/navigate it?

- First try:

Web directories

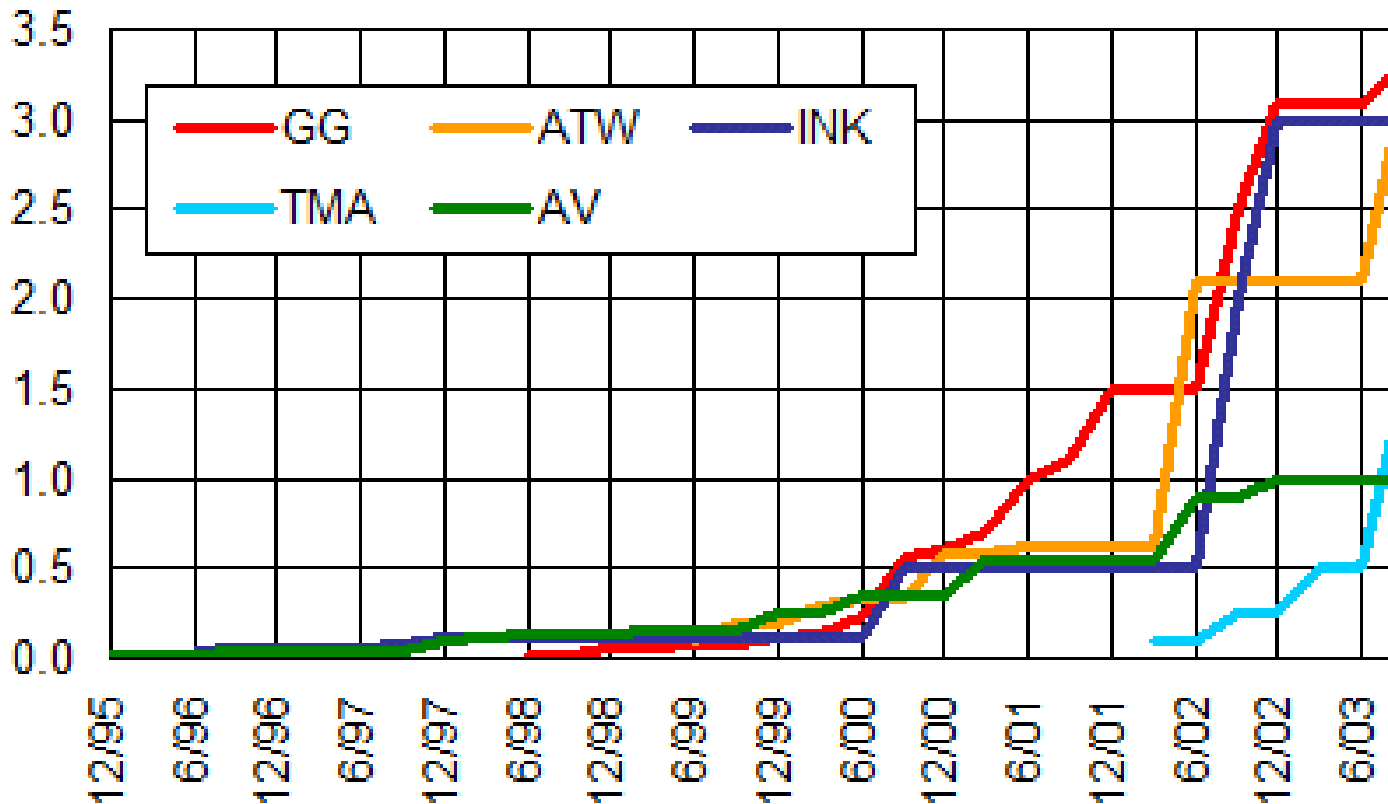
- Yahoo,
- DMOZ,
- LookSmart

The image shows two screenshots of web directories. The top screenshot is the Yahoo! homepage, featuring the 'YAHOO!' logo, a search bar, and a grid of category links such as Arts, Business and Economy, Computers and Internet, Education, Entertainment, Government, Health, News, Recreation, Reference, Regional, Science, Social Science, and Society and Culture. The bottom screenshot is the DMOZ open directory project website, which has a green header with the 'dmoz' logo and the text 'open directory project'. It features a search bar with a 'Search' button and a link to 'advanced' search. Below the search bar is a grid of category links including Arts, Business, Computers, Games, Health, Home, Kids and Teens, News, Recreation, Reference, Regional, Science, Shopping, Society, Sports, and World. At the bottom of the DMOZ page, there is a 'Become an Editor' button, a link to 'Help build the largest human-edited directory of the web', and a small green dragon logo.

Information Retrieval

- Started in 1960s
- Find relevant items in a repository of often small and trusted set:
 - Newspaper articles
 - Patents, etc.
- Two traditional problems:
 - Synonymy: buy and purchase, sick and ill
 - Polysemi: Jaguar
- Second try: **Search**

The index size war

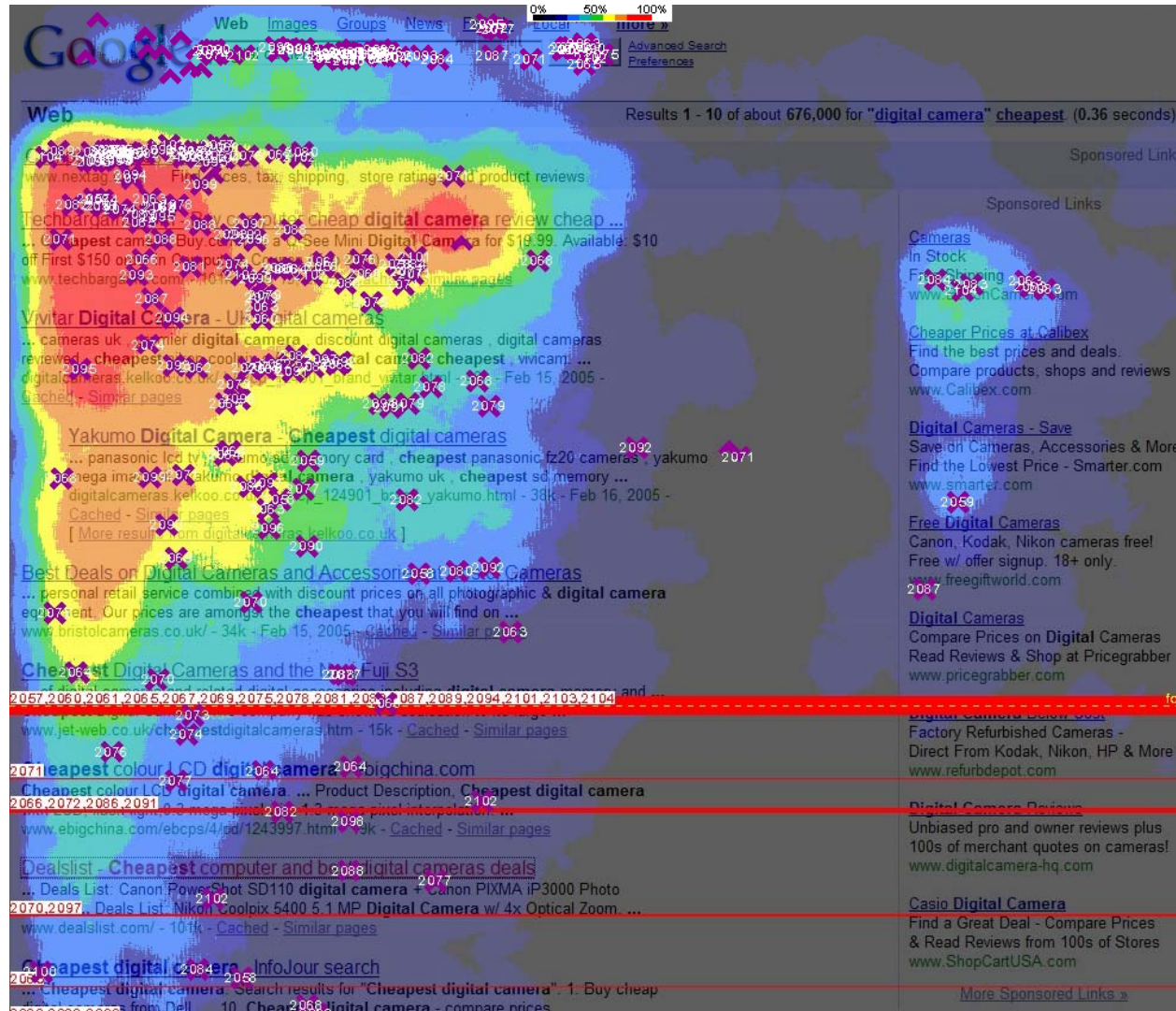


Does bigger index mean better results?

Web Search vs. IR

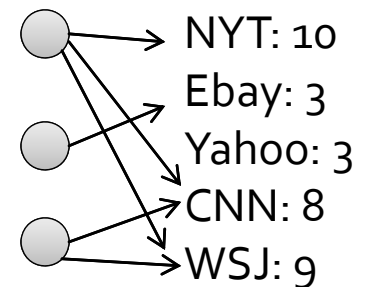
- What is “best” answer to query “Stanford”?
 - Anchor Text: I go to [Stanford](#) where I study
- What about query “newspaper”?
 - Not a single right answer
- Scarcity (IR) vs. abundance (Web)
 - Many sources of info: who to “trust”
- Trick:
 - pages that actually know about newspapers might all be pointing to many newspapers
- **Ranking!**

Where people look?

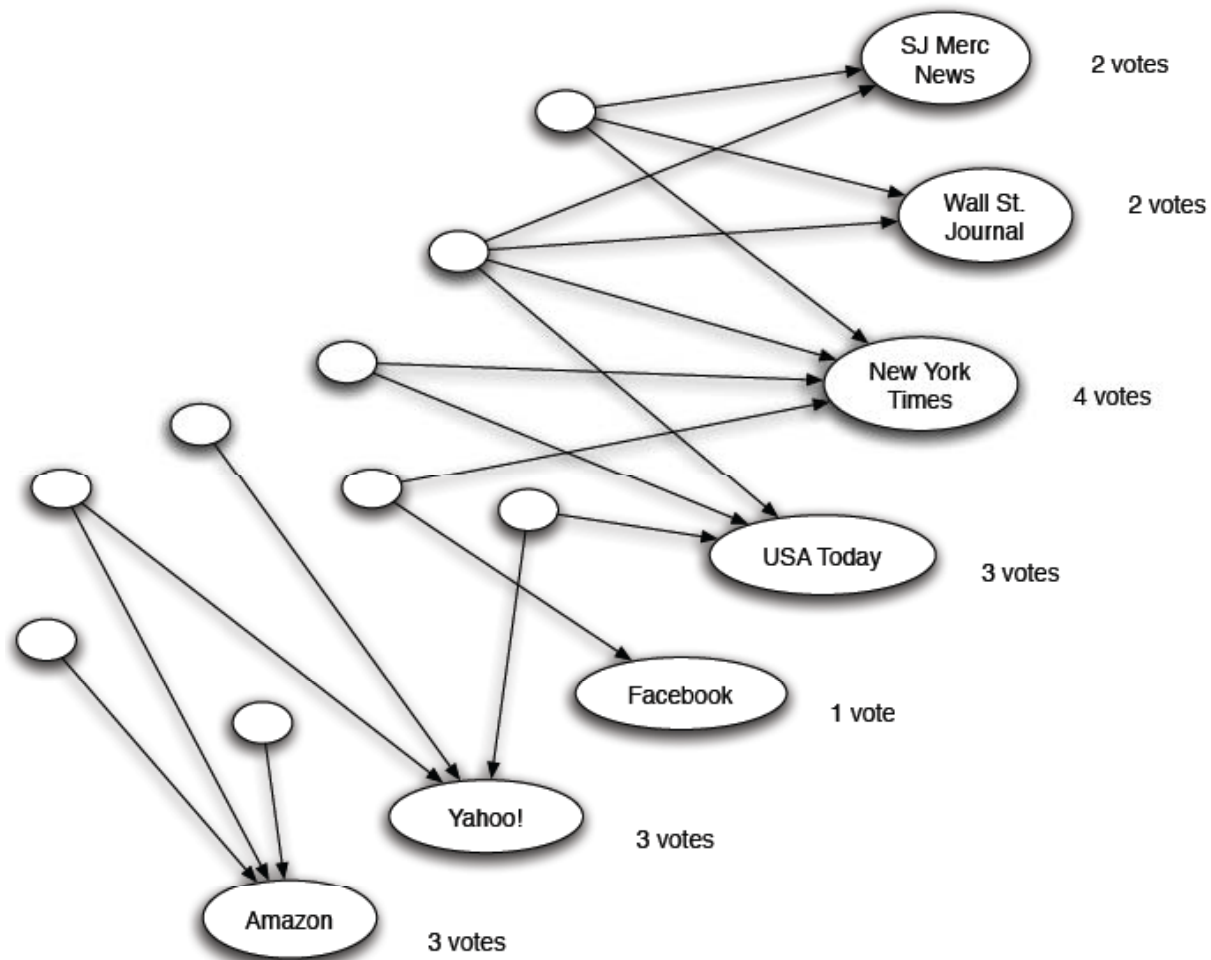


Link analysis

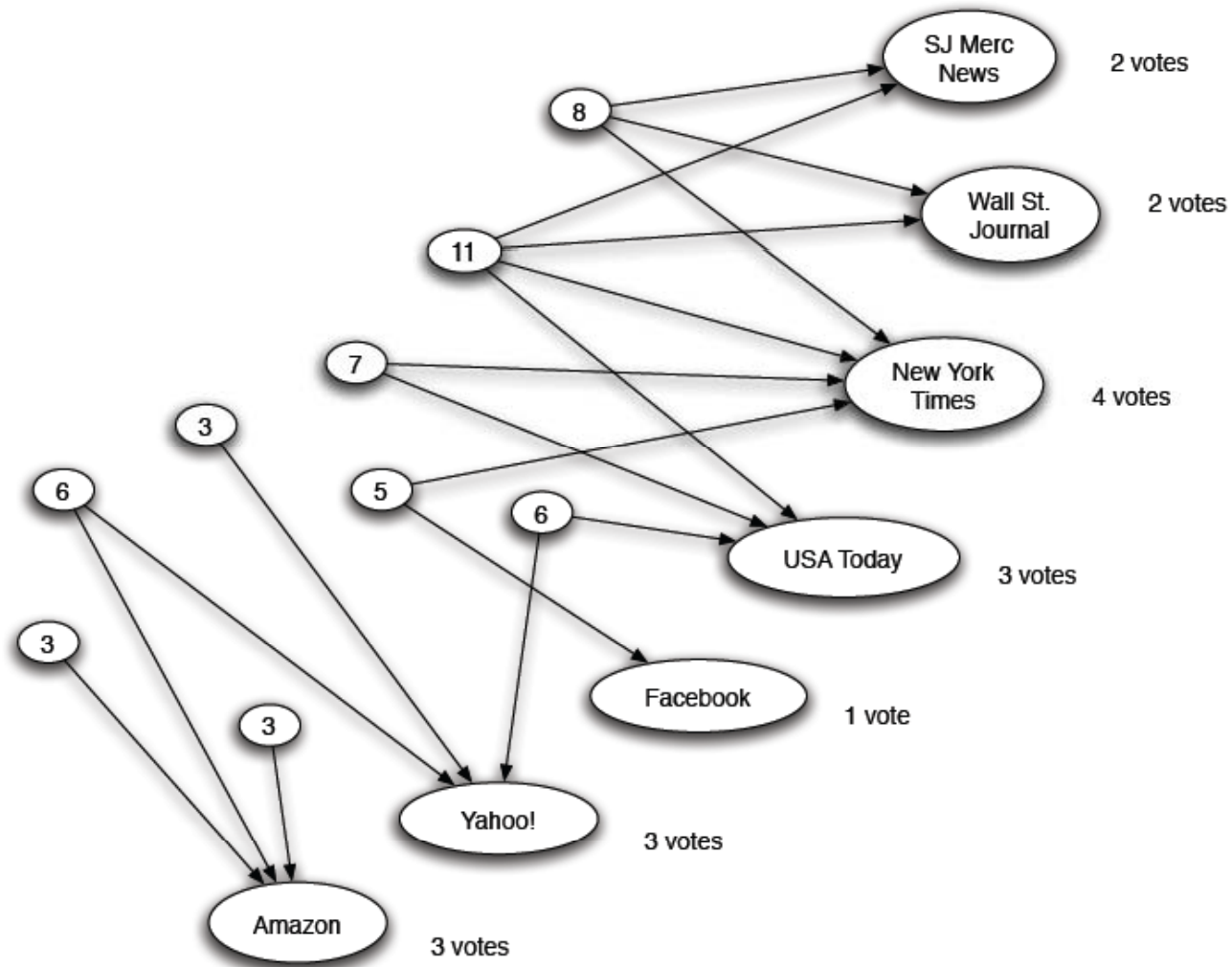
- **Goal** (back to newspaper example):
 - Don't just find newspapers but also find “experts”
 - people who link in a coordinated way to many good newspapers
- **Idea: link voting**
 - Quality as an expert (**hub**):
 - Total sum of votes of pages pointed to
 - Quality as an content (**authority**):
 - Total sum of votes of experts
 - Principle of repeated improvement



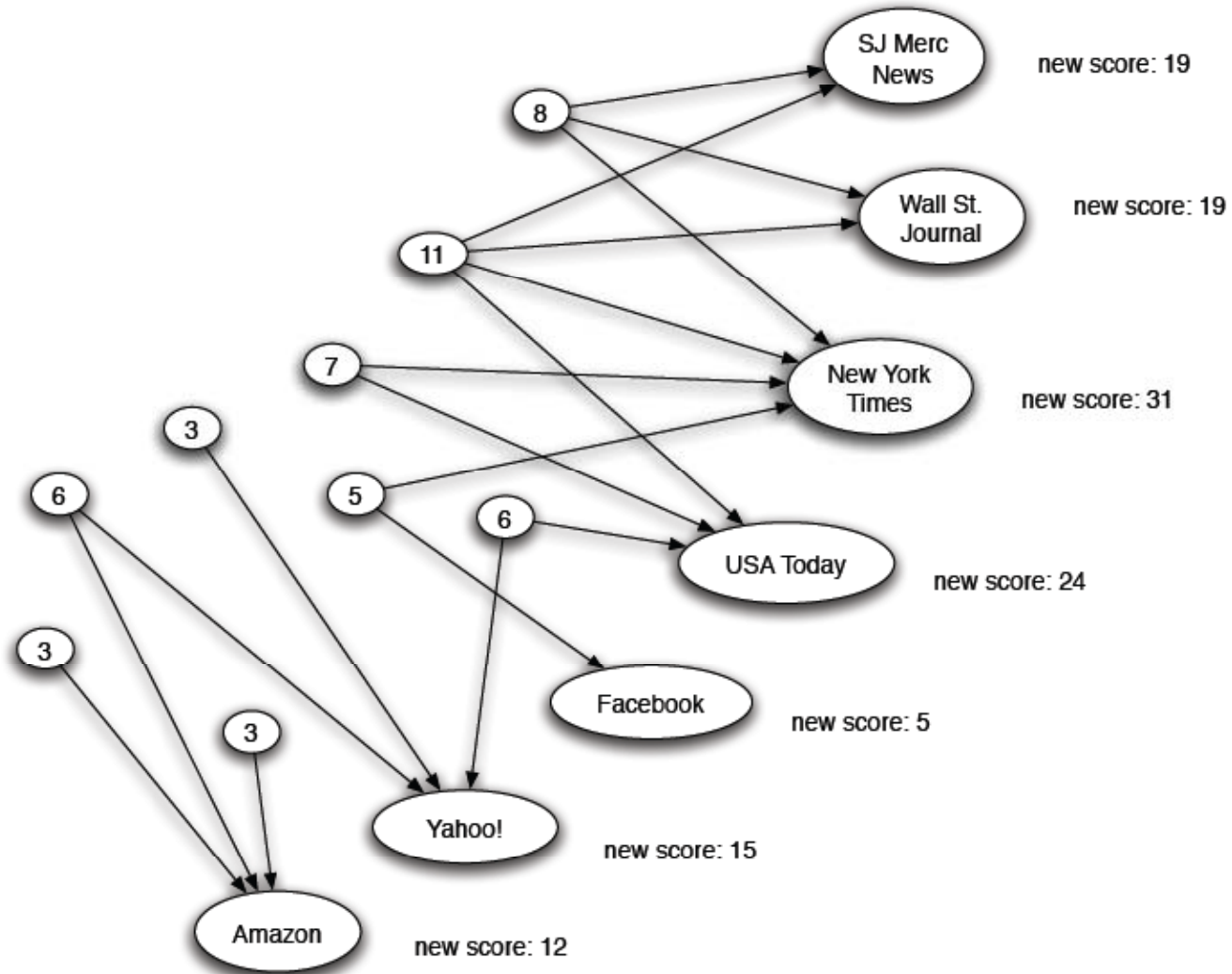
Counting in-links: Authority



Expert quality: Hub

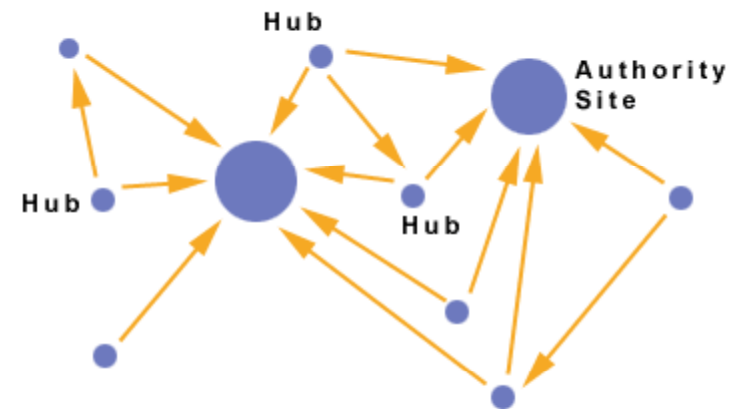


Reweighting



Hubs and Authorities

- Each page i has 2 kinds of scores:
 - Hub score: h_i
 - Authority score: a_i
- Algorithm:
 - Initialize: $a_i=h_i=1$
 - Then keep iterating:
 - Authority: $a_j = \sum_{i \rightarrow j} h_i$
 - Hub: $h_i = \sum_{i \rightarrow j} a_j$
 - Normalize:
 $\sum a_i=1, \sum h_i=1$



Hubs and Authorities

- This will converge 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

- Algorithm in new notation:

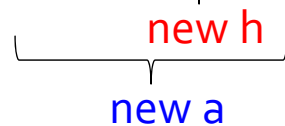
- Set: $a = h = \mathbf{1}^n$

- 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):

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

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 and matrix A
- then x is an eigenvector, and λ is its eigenvalue

- **Fact:**

- If A is symmetric ($A_{ij} = A_{ji}$)
(note 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 Ax ?

- 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:** $\lambda_1 \dots \lambda_n$ ($|\lambda_1| \geq |\lambda_2| \geq \dots \geq |\lambda_n|$)

- $A^k x = (\lambda_1^k \alpha_1, \lambda_2^k \alpha_2, \dots, \lambda_n^k \alpha_n) = \sum \lambda_i^k \alpha_i w_i$

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

$$A^k x \rightarrow \lambda_1 \alpha_1 w_1 \quad (\text{all other coordinates} \rightarrow 0)$$

- So **authority** a is eigenvector of $M^T M$ associated with largest eigenvalue λ_1 (need $|\lambda_1| > |\lambda_2|$)

PageRank

- 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 should be proportional to:

$$\sum_{i \rightarrow j} \frac{r_i}{\text{outdegree of } i}$$

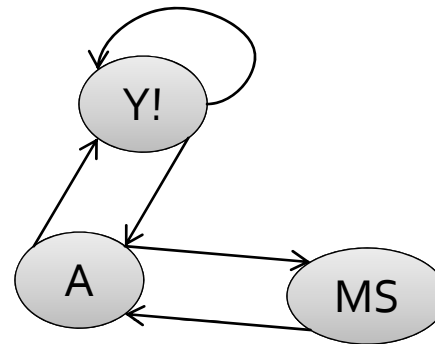
PageRank: alternate interpretation

- r_j ... probability I'm currently at j in a random walk
$$r_j = \sum \Pr[\text{at } i] \Pr[i \rightarrow j]$$
- **But** $r_j = \sum r_i / (\text{out-degree of } i)$
prob. of being at j after one step of a random walk
- **Define:**
 - $N_{ij} = M_{ij} / d_i = 1 / d_i$
 - $M_{ij} = 1$ if node i links to j
 - out-degree of i is d_i
 - N_{ij} is prob. we will be at j if we are currently at i
- **Then in the limit:** $r = Nr$
 - i.e., r is principal eigenvector of N

PageRank: Example

- Power iteration:

- Set $r_i = 1$
- $r_j = \sum_i r_i / d_i$
- And iterate



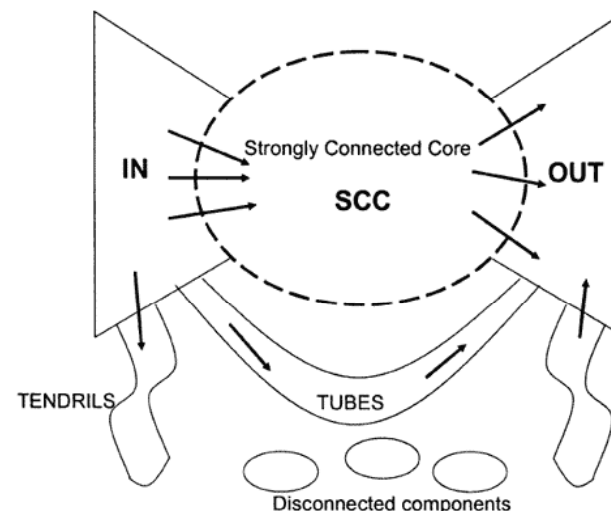
	Y!	A	MS
Y!	1/2	1/2	0
A	1/2	0	1
MS	0	1/2	0

- Example:

y	=	1	1	5/4	9/8	6/5
a		1	3/2	1	11/8	6/5
m		1	1/2	3/4	1/2	3/5

Problems

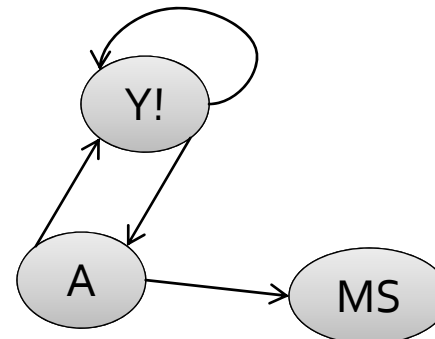
- Some pages are “**dead ends**” (have no out-links)
 - Such pages cause importance to leak out
- **Spider traps** (all out links are within the group)
 - Eventually spider traps absorb all importance



Dead ends

- Power iteration:

- Set $r_i = 1$
- $r_j = \sum_i r_i / d_i$
- And iterate



	Y!	A	MS
Y!	1/2	1/2	0
A	1/2	0	0
MS	0	1/2	0

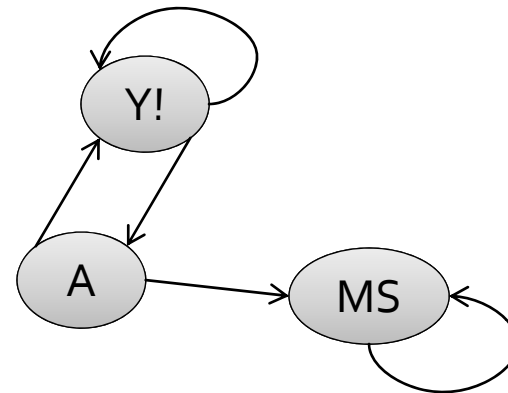
- Example:

y	=	1	1	3/4	5/8	0
a		1	1/2	1/2	3/8	...
m		1	1/2	1/4	1/4	0

Spider traps

- Power iteration:

- Set $r_i=1$
- $r_j = \sum_i r_i / d_j$
- And iterate



	Y!	A	MS
Y!	1/2	1/2	0
A	1/2	0	0
MS	0	1/2	1

- Example:

y	=	1	1	3/4	5/8	...	0
a		1	1/2	1/2	3/8	...	0
m		1	3/2	7/4	2	...	3

Solution: the real PageRank

- “Tax” each page by at each iteration
- Add a fixed constant to all pages
- Models a random walk with a fixed probability of jumping to a random page

- We really want:

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

- Random walk that follows a link with prob. $1 - \varepsilon$ and randomly jumps with prob. ε

d_i ... outdegree
of node i

PageRank & eigenvectors

- PageRank as a principal eigenvector

$$r = N^T r \Leftrightarrow r_j = \sum_i r_i / d_i$$

- But we really want:

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

- Define:

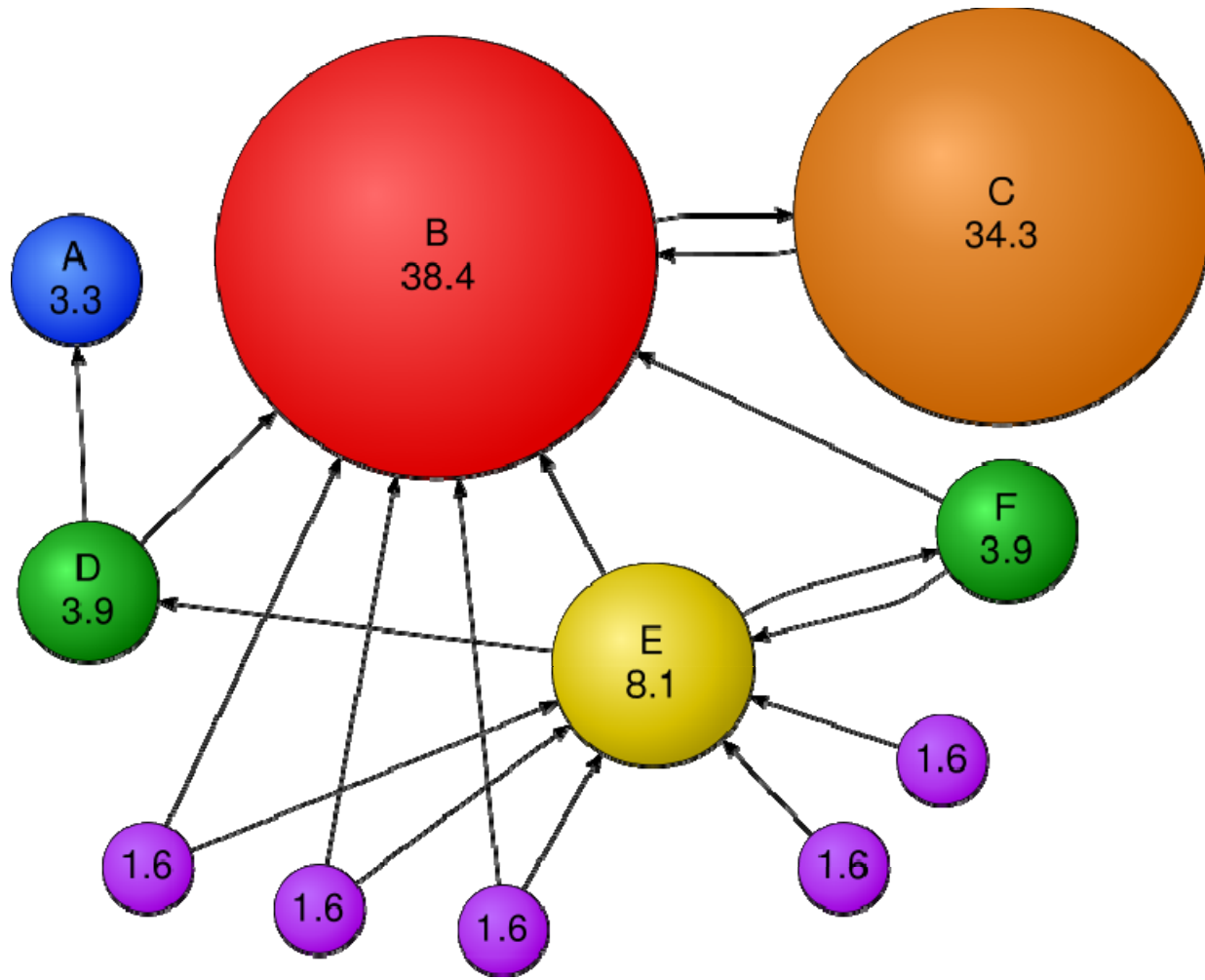
$$N'_{ij} = (1-\varepsilon) N_{ij} + \varepsilon 1/n$$

- Then: $r = N'^T r$

- What is ε ?

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

d_i ... outdegree
of node i

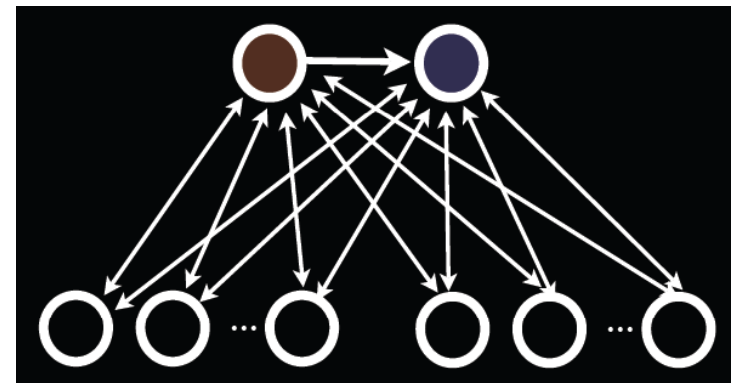


Personalized PageRank

- Topic-specific PageRank
- Goal: evaluate pages not just by popularity but by how close they are to the topic
- Walker has a small teleporting probability
- 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
 - $N'_{ij} = (1-\varepsilon)N_{ij} + \varepsilon c$ (where c is a vector)

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



affordablecellphonerates.com

Search

Results for "card phone prepaid"

[Online prepaid phone card](#)
Over 95 prepaid calling cards sale Great quality instant ph
www.zscomm.com

[US 1¢/min - World 2¢/min](#)
Use from Home, Office, Hotel & Cell Pre-Paid Card. Easy Ne
PennyTalk.com

[Prepaid Phone Cards](#)
Your Guide To Consumer Electronics. Find Prepaid Phone C
GizmoCafe.com

[Prepaid Phone](#)
Compare prepaid phones and plans. Boost Mobile® Officia
boostmobile.com

[Prepaid Phone Cards](#)
Looking For Prepaid Phone Cards? See Our Prepaid Phone
kellyscornerstore.com

[Phone Card](#)
Find Providers of Prepaid Phone Cards on the Business.coi
www.business.com

[Phone card](#)
Unlimited local and long distance Int'l rates as low as 1¢ p
www.Vonage.com

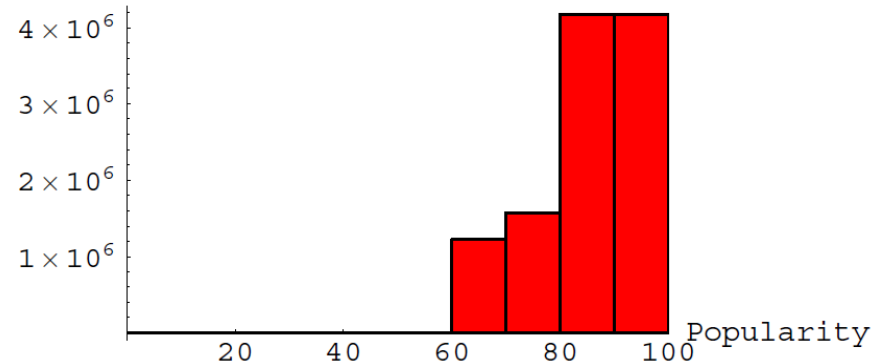
Related Searches

- [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](#)
- [Buy a Calling Card](#)

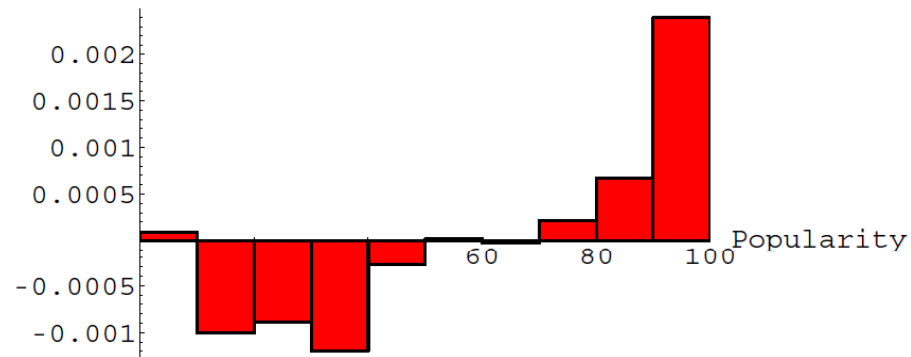
Issues with PageRank

- Rich get richer

Absolute increase in
the no. of In-Links



Absolute increase in
the PageRank values



Google bombs (1)



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](#)

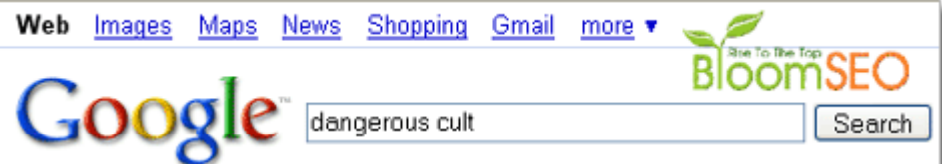
Web users manipulate a popular search engine so an unflattering description leads to the president's page.

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. A search for [miserable failure](#) on Google brings up the official George W. Bush biography. A search for [miserable failure](#) on Google brings up the official George W. Bush biography. A search for [miserable failure](#) on Google brings up the official George W. Bush biography.



Web Results 1 - 10 of about 252,000 for [dangerous cult](#). (0.06 seconds)

[Scientology - Church of Scientology Official Site](#)

Living in a **Dangerous** Environment · Drug and Alcohol Problems · Personalities, Emotions 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 much as 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/rektop/budcomp.htm - 10k - [Cached](#) - [Similar pages](#) - [Note this](#)

Google bombs (2)

The image displays two side-by-side screenshots of a Netscape browser window, illustrating a Google bomb search for the word "Jew".

Left Window (English): The browser title is "Netscape: Google Search: Jew". The address bar shows "http://www.google.com/search?q=Jew". The search results page shows "Results 1 - 10 of about 1,810,000 for Jew [definition]. (0.08 sec)". The first result is "Offensive Search Results" from google.com/explanation. Other results include "News results for Jew", "Jew Watch", "Jew - Wikipedia", and "Judaism 101: Who Is a Jew?".

Right Window (German): The browser title is "Netscape: Google-Suche: Jew". The address bar shows "http://www.google.de/search?q=Jew". The search results page shows "Ergebnisse 1 - 10 von ungefähr 1,340,000 für Jew. (0.12 S)". The first result is "Jew - Wikipedia". Other results include "Judaism 101", "CJF Ministries - Welcome!", and "Amazon.com: Books: To Be a Jew: A Guide to Jewish Observance in ...".