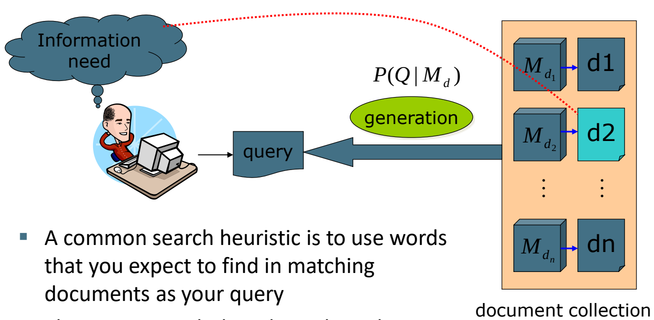
Information Retrieval & Social Web

CS 525/DS 595
Worcester Polytechnic Institute
Department of Computer Science
Instructor: Prof. Kyumin Lee

Previous Class…

- Statistical Language Models

IR based on Language Model (LM)



- A common search heuristic is to use words that you expect to find in matching documents as your query
- The LM approach directly exploits that idea!

Basic mixture model summary

- General formulation of the LM for IR

$$P(q|d) \propto \prod_{t \in q} ((1-\lambda) P(t|M_c) + \lambda P(t|M_d))$$

general language model

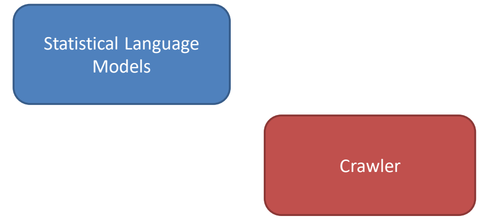
individual-document model

- The user has a document in mind, and generates the query from this document.
- The equation represents the probability that the document that the user had in mind was in fact this one.

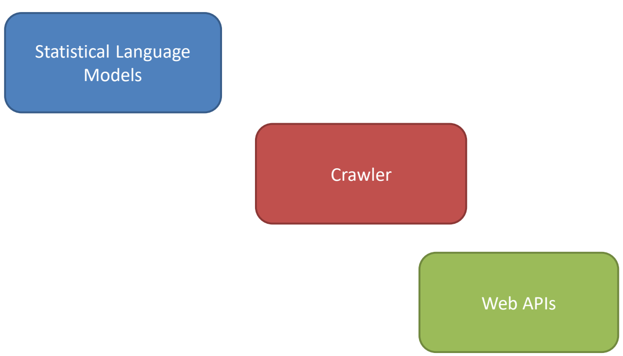
Example

- Document collection (2 documents)
  - d_1: Xerox reports a profit but revenue is down
  - d_2: Lucent narrows quarter loss but revenue decreases further
- Model: MLE unigram from documents; $\lambda = ½$
- Query: *revenue down*
  - $P(Q|d_1) = [(1/8 + 2/16)/2] \times [(1/8 + 1/16)/2]$
    $= 1/8 \times 3/32 = 3/256$
  - $P(Q|d_2) = [(1/8 + 2/16)/2] \times [(0 + 1/16)/2]$
    $= 1/8 \times 1/32 = 1/256$
- Ranking: $d_1 > d_2$

Previous Class…



Previous Class…



Available Web APIs

- Twitter: https://dev.twitter.com/
- Flickr: http://www.flickr.com/services/api/
- Google Maps: https://developers.google.com/maps/
- Facebook: http://developers.facebook.com/
- Foursquare: https://developer.foursquare.com/
- Yahoo Boss API: http://developer.yahoo.com/search/boss/
- Wikipedia API: http://www.mediawiki.org/wiki/API:Main_page
- Youtube API: http://code.google.com/apis/youtube/overview.html
- Openstreetmap API: http://wiki.openstreetmap.org/wiki/API
- Halo API: https://developer.haloapi.com/

- List of APIs: https://www.reddit.com/r/webdev/comments/3wrswc/what_are_some_fun_apis_to_play_with/

Static quality scores

- We want top-ranking documents to be both *relevant* and *authoritative*
- *Relevance* is being modeled by cosine scores
- *Authority* is typically a query-independent property of a document
- Examples of authority signals
  - Wikipedia among websites
  - Articles in certain newspapers
  - A paper with many citations
  - Many bitly's or diggs
  - (Pagerank)

Quantitative

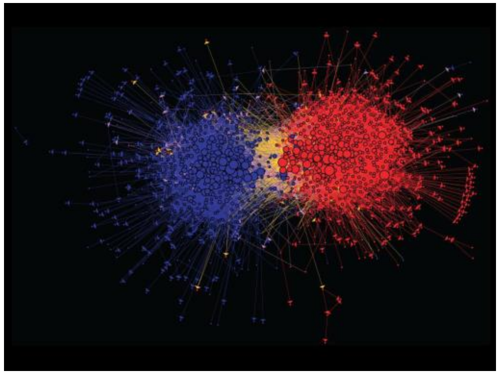
Today: Link Analysis

- Anchor text
- PageRank

Graph Data: Social Networks

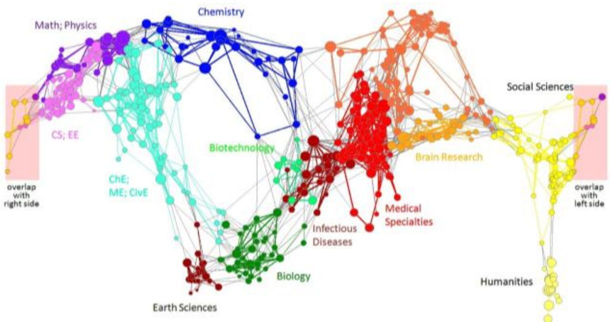
Facebook social graph
4-degrees of separation [Backstrom-Boldi-Rosa-Ugander-Vigna, 2011]

Graph Data: Media Networks



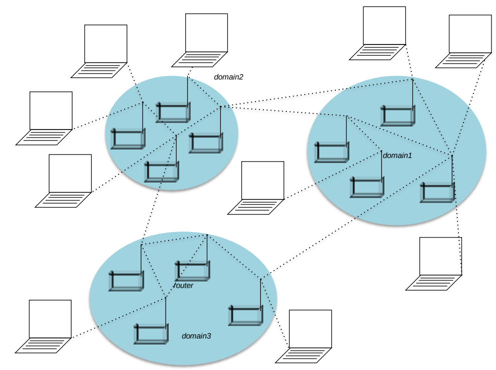
Connections between political blogs
Polarization of the network [Adamic-Glance, 2005]

Graph Data: Information Nets



Citation networks and Maps of science
[Börner et al., 2012]

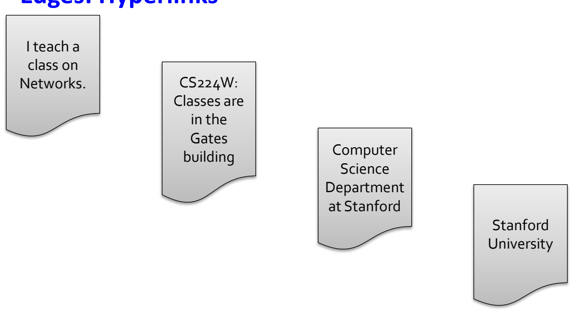
Graph Data: Communication Nets



Internet

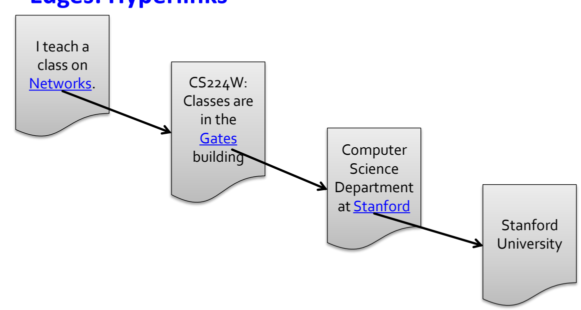
Web as a Graph

- **Web as a directed graph:**
  - **Nodes: Webpages**
  - **Edges: Hyperlinks**

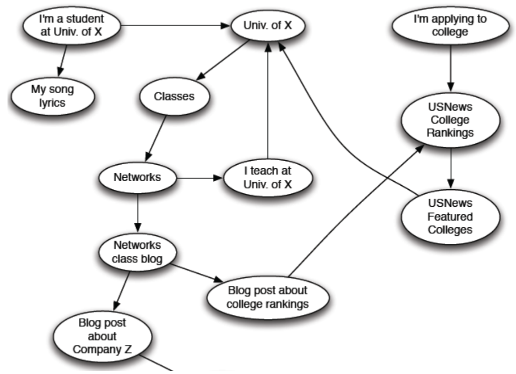


Web as a Graph

- **Web as a directed graph:**
  - **Nodes: Webpages**
  - **Edges: Hyperlinks**



Web as a Directed Graph

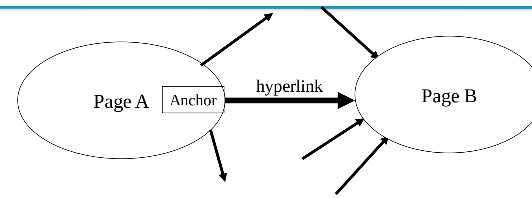


Broad Question

- **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.

Anchor Text

The Web as a Directed Graph

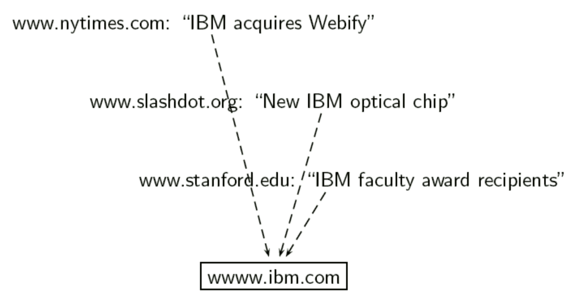


- **Assumption 1:** a hyperlink is a quality signal
  - A hyperlink between pages denotes author perceived relevance
- **Assumption 2:** The anchor text describes the target page
  - we use anchor text somewhat loosely here: the text surrounding the hyperlink. Example: "You can find cheap cars here"

[document text only] vs. [document text + anchor text]

- Searching on [document text + anchor text] is often more effective than searching on [document text only].
- Example: Query ***IBM***
  - Matches IBM's copyright page
  - Matches many spam pages
  - Matches IBM wikipedia article
  - May not match IBM home page! (if IBM home page is mostly graphical)
- Searching on anchor text is better for the query IBM.
- **Represent each page by all the anchor text pointing to it.**
- In this representation, the page with the most occurrences of IBM is www.ibm.com.

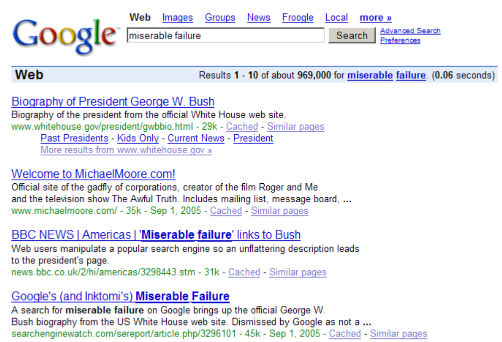
Anchor text containing ***IBM*** pointing to www.ibm.com



Indexing anchor text

- Thus: anchor text is often a better description of a page's content than the page itself
- Anchor text can be weighted more highly than document text (based on Assumptions 1 & 2)
- Indexing anchor text can have unexpected side effects - Google bombs.
- A Google bomb is a search with "bad" results due to maliciously manipulated anchor text
- Google introduced a new weighting function in January 2007 that fixed many Google bombs

Google bomb example



Low graphics | Accessibility help

BBC NEWS

Watch One-Minute World News

News services
Your news when you want it

News Front Page
Africa
Americas
Asia-Pacific
Europe
Middle East
South Asia
UK
Business
Health
Science & Environment
Technology
Entertainment
Also in the news

Video and Audio

Programmes
Have Your Say
In Pictures
Country Profiles
Special Reports

RELATED BBC SITES
SPORT
WEATHER
ON THIS DAY

Last Updated: Sunday, 7 December, 2003, 15:04 GMT

E-mail this to a friend | Printable version

'Miserable failure' links to Bush

George W Bush has been Google bombed.

Web users entering the words "miserable failure" into the popular search engine are directed to the biography of the president on the White House website.

Bush has been the target of similar pranks before

The trick is possible because Google searches more than just the contents of web pages - it also counts how often a site is linked to, and with what words.

Thus, members of an online community can affect the results of Google searches - called "Google bombing" - by linking their sites to a chosen one.

Weblogger Adam Mathes is credited with inventing the practice in 2001, when he used it to link the phrase "talentless hack" to a friend's website.

The search engine can be manipulated by a fairly small group of users, one report suggested.

Newsday newspaper says as few as 32 web pages with the words "miserable failure" link to the Bush biography.

"If you are George Bush and typed the country's name in the address bar, make sure that it is spelled correctly (IRAQ)"
Prank website

The Bush administration has

SEE ALSO:
WMD spoof is internet hit
04 Jul 03 | West Midlands
Google hit by link bombers
13 Mar 02 | Science/Nature

RELATED INTERNET LINKS:
White House
Google bombing
The BBC is not responsible for the content of external internet sites

TOP AMERICAS STORIES
US lifts lid on WikiLeaks probe
Iran scientist heads home
Argentina legalises gay marriage
| News feeds

Web Search: Pre-History

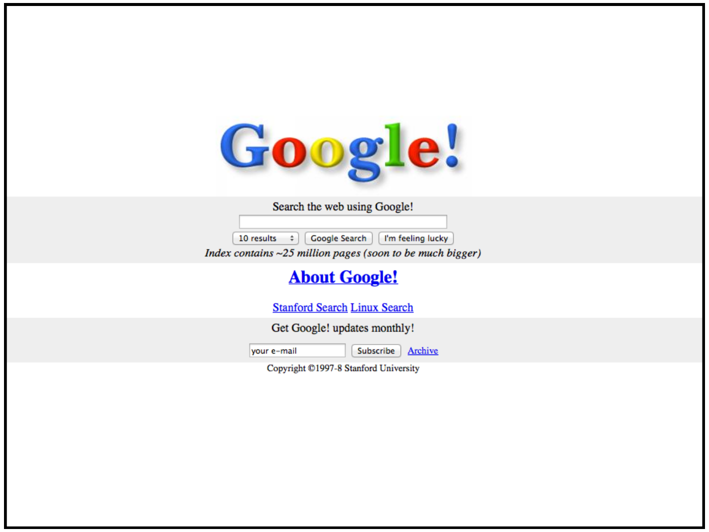
Brief (non-technical) history of Web Search

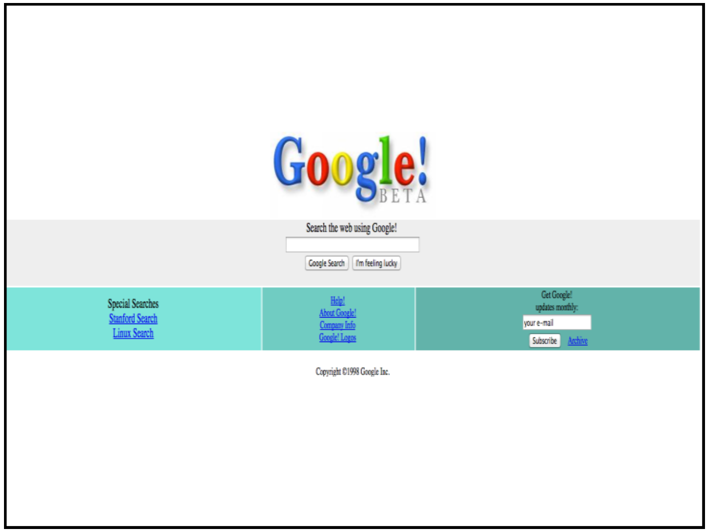
- Early keyword-based engines ca. 1995-1997
  - Altavista, Excite, Infoseek, Inktomi, Lycos,
- Paid placement ranking: Goto.com (morphed into Overture.com → Yahoo!)
  - Your search ranking depended on how much you paid
  - Auction for keywords: ***casino*** was expensive!

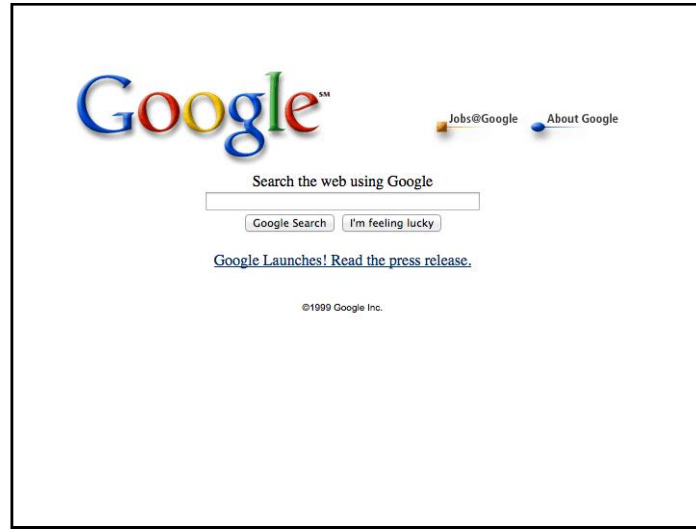

Search for information about:
in the World Wide Web
Search Now
Infoseek Guide is best viewed with:
Want personalized news? Get Personal now!
Basic Search Tips:
Click in the box above and type a few words that describe what you want to find. For example, typing growing orchids indoors will find sites about caring for orchids.
If you are looking for a person or place, type the name, starting with capital letters. For example, typing Florence Italy will find sites about this famous city.
These detailed search tips describe how to use the features of Infoseek Guide to find what you are looking for.
For the broadest results, you can search the entire World Wide Web.
To restrict your search to hand-picked and categorized sites, choose Infoseek Select Sites.
Or just search for a category within Infoseek Select by choosing Categories of Sites.
To search through Internet discussion forums (similar to bulletin boards), choose Usenet Newsgroups.
To search for someone's e-mail address, choose E-mail Addresses.
To search through news stories within the past month, choose Reuters News.
To search through answers to Frequently Asked Questions, choose Web FAQs.
Explore these popular Infoseek Select topics:
Arts & Entertainment
Business & Finance
Computers & Internet
Education
Government & Politics
Health & Medicine
Living
News
Reference
Science & Technology
Sports
Travel
Try Infoseek Personal, your personalized news service
Customize | Add Site | Help | Feedback


Click here to try Microsoft Money 97 FREE
It's amazing where Go Get It will get you.
Find:
Go Get It
Enhance your search.
INVEST $100K
New Search • TopNews • Sites by Subject • Top 5% Sites • City Guide • Pictures & Sounds
PeopleFind • Point Review • Road Maps • Software • About Lycos • Club Lycos • Help
Add Your Site to Lycos
Copyright © 1996 Lycos™, Inc. All Rights Reserved.
Lycos is a trademark of Carnegie Mellon University.
Questions & Comments


excite
search reviews city.net live! tours
people finder maps yellow pages news
"Turbo Search!"
Download Excite Direct
Excite Search: twice the power of the competition.
What:
Where: World Wide Web
[Help]
[Advanced Search]
INTEGRATED BROWSING, EMAIL, NEWSGROUPS AND PAGE CREATION.
Take an ExciteSeeing Tour
Excite on TV
Excite Reviews: site reviews by the web's best editorial team.
Arts
Business
Computing
Education
Entertainment
Health
Hobbies
Life & Style
Money
News & Reference
Personal Pages
Politics & Law
Regional
Science
Shopping
Sports
Excite City.Net
Plan your weekend, your travels.
Find-A-Destination
Take me there!
Maps
Top Cities
Concierge
ExciteSeeing Tours
Choose from hundreds.
X-Files: The truth is out there!
Dr. Ruth's guide to safer sex
Windows 95 shareware and freeware
Celebrating Thanksgiving
Investing in high-tech stocks
New to the Net?
Make your website searchable, FREE!
Excite Live!
Your news, your way.
Latest news
Sports scores
Local weather
Movie reviews
Stock quotes
TV listings
Horoscopes
Site reviews
Excite Reference
Just the facts, ma'am.
Yellow Pages
People Finder
Email Lookup
Maps
Shareware
Dictionary

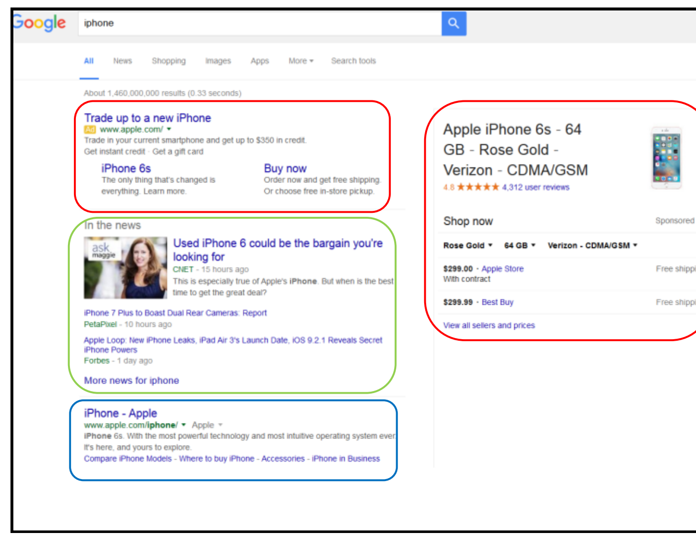

Google!
Search the web using Google!
10 results
Google Search
I'm feeling lucky
Index contains ~25 million pages (soon to be much bigger)
About Google!
Stanford Search Linux Search
Get Google! updates monthly!
your e-mail
Subscribe
Archive
Copyright ©1997-8 Stanford University

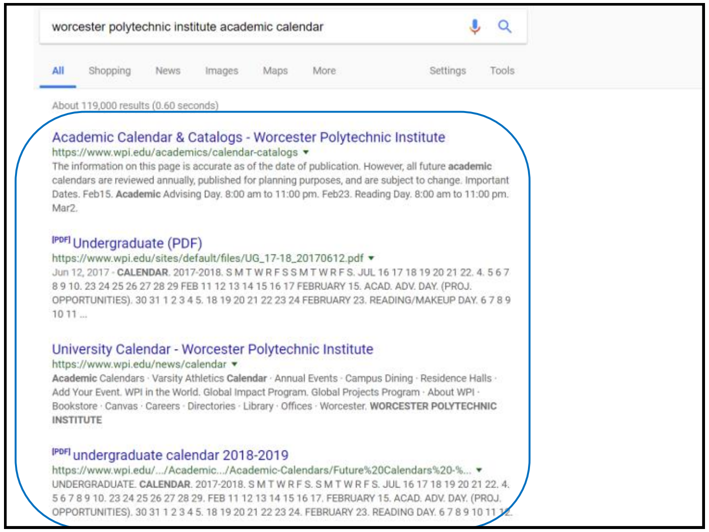




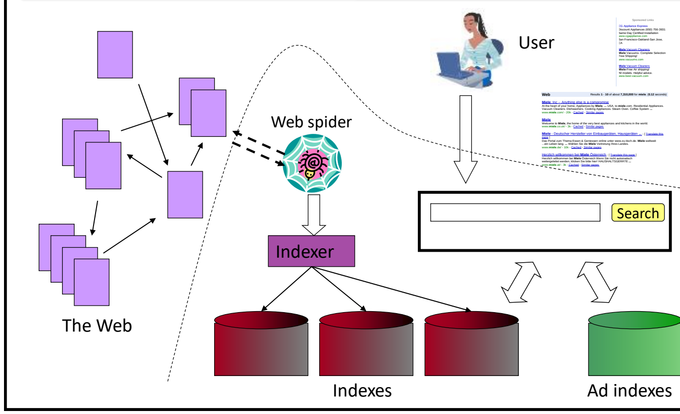
Brief (non-technical) history

- 1998+: Link-based ranking pioneered by Google
  - Blew away all early engines
  - Great user experience in search of a business model
  - Meanwhile Goto/Overture's annual revenues were nearing $1 billion
- Result: Google added paid-placement "ads" to the side, independent of search results
  - Yahoo follows suit, acquiring Overture (for paid placement) and Inktomi (for search)
- 2005+: Google gains search share, dominating in Europe and very strong in North America
  - 2009: Yahoo! and Microsoft propose combined paid search offering





Web search basics



PageRank

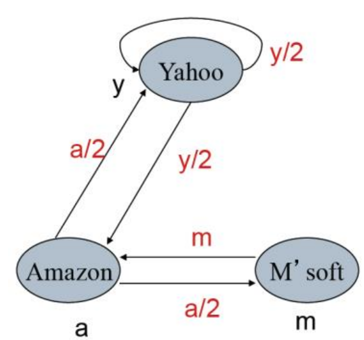
Link-based ranking

- Query processing with link-based ranking:
  - First retrieve all pages meeting the query (say **venture capital**)
  - Order these by their link popularity (= citation frequency, first generation)
  - . . . or by Pagerank (second generation)

- Simple link popularity (= number of inlinks of a page) is easy to spam.
- Why?

PageRank: Recursive formulation

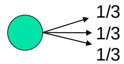
- Each link's vote is proportional to the **importance of its source page**
- If page P with importance x has n outlines, each link gets x/n votes
- Page P's own importance is the sum of the vote on its inlinks



$$y = y/2 + a/2$$
$$a = y/2 + m$$
$$m = a/2$$

PageRank basics

- Imagine a web surfer doing a random walk on the web
  – Start at a random page



  – At each step, go out of the current page along one of the links on that page, equiprobably
- "In the steady state" each page has a long-term visit rate - use this as the page's score.
- **PageRank = steady state probability**

  = long-term visit rate

Markov chains

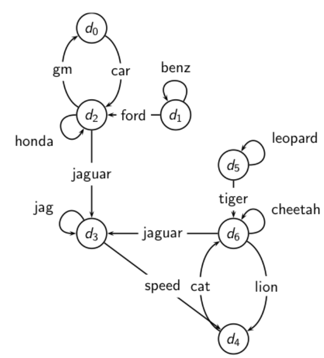
- A Markov chain consists of n states, plus an n×n <u>transition probability matrix</u> **P**.
- **state = page**
- At each step, we are on exactly one of the states.
- For $1 \leq i, j \leq n$, the matrix entry P_{ij} tells us the probability of j being the next state (page), given we are currently on page (state) i.

$$d_i \xrightarrow{P_{ij}} d_j$$

Markov chains

- Clearly, for all i, $\sum_{j=1}^{N} P_{ij} = 1$
- Markov chains are abstractions of random walks.

Example web graph



And the corresponding link matrix

	d_0	d_1	d_2	d_3	d_4	d_5	d_6
d_0	0	0	1	0	0	0	0
d_1	0	1	1	0	0	0	0
d_2	1	0	1	1	0	0	0
d_3	0	0	0	1	1	0	0
d_4	0	0	0	0	0	0	1
d_5	0	0	0	0	0	1	1
d_6	0	0	0	1	1	0	1

Transition probability matrix P

	d_0	d_1	d_2	d_3	d_4	d_5	d_6
d_0	0	0	1	0	0	0	0
d_1	0	1	1	0	0	0	0
d_2	1	0	1	1	0	0	0
d_3	0	0	0	1	1	0	0
d_4	0	0	0	0	0	0	1
d_5	0	0	0	0	0	1	1
d_6	0	0	0	1	1	0	1

Transition probability matrix

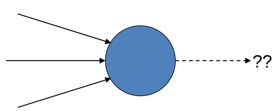
	d_0	d_1	d_2	d_3	d_4	d_5	d_6
d_0	0.00	0.00	1.00	0.00	0.00	0.00	0.00
d_1	0.00	0.50	0.50	0.00	0.00	0.00	0.00
d_2	0.33	0.00	0.33	0.33	0.00	0.00	0.00
d_3	0.00	0.00	0.00	0.50	0.50	0.00	0.00
d_4	0.00	0.00	0.00	0.00	0.00	0.00	1.00
d_5	0.00	0.00	0.00	0.00	0.00	0.50	0.50
d_6	0.00	0.00	0.00	0.33	0.33	0.00	0.33

Long-term visit rate

- Recall: PageRank = long-term visit rate
- Long-term visit rate of page ***d*** is the probability that a web surfer is at page ***d*** at a given point in time.
- Next: what properties must hold of the web graph for the long-term visit rate to be well defined?

Not quite enough

- The web is full of dead-ends.
  - Random walk can get stuck in dead-ends.
  - Makes no sense to talk about long-term visit rates.



Teleporting

- At a dead end, jump to a random web page.
- At any non-dead end, with probability 10%, jump to a random web page.
  - With remaining probability (90%), go out on a random link.
  - 10% - a parameter.

Teleporting Matrix

- Recall: At a dead end, jump to a random web page

	d_0	d_1	d_2	d_3	d_4	d_5	d_6
d_0	1/7	1/7	1/7	1/7	1/7	1/7	1/7
d_1	1/7	1/7	1/7	1/7	1/7	1/7	1/7
d_2	1/7	1/7	1/7	1/7	1/7	1/7	1/7
d_3	1/7	1/7	1/7	1/7	1/7	1/7	1/7
d_4	1/7	1/7	1/7	1/7	1/7	1/7	1/7
d_5	1/7	1/7	1/7	1/7	1/7	1/7	1/7
d_6	1/7	1/7	1/7	1/7	1/7	1/7	1/7

Result of teleporting

- With teleporting, we cannot get stuck in a dead end
- There is a long-term rate at which any page is visited (not obvious, will show this).
- How do we compute this visit rate?

Formalization of "visit": Probability vectors

- A probability (row) vector $\mathbf{x} = (x_1, \ldots x_n)$ tells us where the walk is at any point.
- E.g., $(\underset{1}{0}00\ldots\underset{i}{1}\ldots 00\underset{n}{0})$ means we're in state *i*.
- More generally, the vector $\mathbf{x} = (x_1, \ldots x_n)$ means the walk is in state *i* with probability x_i.

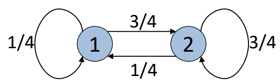
$$\sum_{i=1}^{n} x_i = 1.$$

Change in probability vector

- If the probability vector is $\mathbf{x} = (x_1, \ldots x_n)$ at this step, what is it at the next step?
- Recall that row *i* of the transition prob. Matrix **P** tells us where we go next from state *i*.
- So from **x**, our next state is distributed as **xP**.

Steady state example

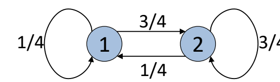
- The steady state looks like a vector of probabilities **a** = *(a_1, … a_n):*
- a_i is the probability that we are in state *i*.



What is the steady state in this example?

Steady state example

- The steady state looks like a vector of probabilities **a** = *(a_1, … a_n):*
- a_i is the probability that we are in state *i*.



For this example, $a_1=1/4$ and $a_2=3/4$.

How to compute the steady-state?

- Recall, regardless of where we start, we eventually reach the steady state **a**.
- Start with any distribution (say **x**=(*10…0*)).
- After one step, we're at **xP**;
- after two steps at $\mathbf{xP}^2$, then $\mathbf{xP}^3$ and so on.
- "Eventually" means for "large" *k*, $\mathbf{xP}^k = \mathbf{a}$.
- Algorithm: multiply **x** by increasing powers of **P** until the product looks stable.
- This is called the power method

Power method: example

Two-node example: $\vec{x} = (0.5, 0.5)$, $P = \begin{pmatrix} 0.25 & 0.75 \\ 0.25 & 0.75 \end{pmatrix}$

$\vec{x}P = (0.25, 0.75) = \vec{x}_2$

$\vec{x}_2 P = (0.25, 0.75)$

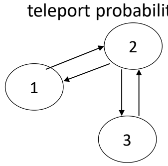
Convergence in one iteration!

Exercise on PageRank

Transition probability matrix of a surfer's walk with teleportation:

P = (1- α) * transition matrix + α * teleporting matrix

- Consider a Web graph with three nodes 1, 2, and 3. The links are as follows: 1->2, 3->2, 2->1, 2->3. Write down the transition probability matrices P and pagerank scores for the surfer's walk with teleporting, with the value of teleport probability α=0.5.



P =

0	1	0
1	0	1
0	1	0

Each 1 divied by the number of ones in this row

(1- α)*

0	1	0
½	0	½
0	1	0

\+

α*

1/3	1/3	1/3
1/3	1/3	1/3
1/3	1/3	1/3

=

1/6	2/3	1/6
5/12	1/6	5/12
1/6	2/3	1/6

Exercise on PageRank (Cont'd)

Remember

$\vec{x}_1 = \vec{x}_0 P$

$\vec{x}_2 = \vec{x}_1 P$

$\vec{x}_2 = \vec{x}_1 P$

…

…

…

Until converged

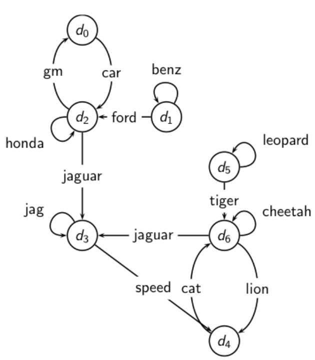
P=

1/6	2/3	1/6
5/12	1/6	5/12
1/6	2/3	1/6

$\vec{x}_0 =$	1	0	0
$\vec{x}_1 =$	1/6	2/3	1/6
$\vec{x}_2 =$	1/3	1/3	1/3
$\vec{x}_3 =$	1/4	1/2	1/4
…			
…			
$\vec{x}_k =$	5/18	4/9	5/18

converged

Example web graph



And the corresponding link matrix

	d_0	d_1	d_2	d_3	d_4	d_5	d_6
d_0	0	0	1	0	0	0	0
d_1	0	1	1	0	0	0	0
d_2	1	0	1	1	0	0	0
d_3	0	0	0	1	1	0	0
d_4	0	0	0	0	0	0	1
d_5	0	0	0	0	0	1	1
d_6	0	0	0	1	1	0	1

Transition matrix with teleporting

	d_0	d_1	d_2	d_3	d_4	d_5	d_6
d_0	0.00	0.00	1.00	0.00	0.00	0.00	0.00
d_1	0.00	0.50	0.50	0.00	0.00	0.00	0.00
d_2	0.33	0.00	0.33	0.33	0.00	0.00	0.00
d_3	0.00	0.00	0.00	0.50	0.50	0.00	0.00
d_4	0.00	0.00	0.00	0.00	0.00	0.00	1.00
d_5	0.00	0.00	0.00	0.00	0.00	0.50	0.50
d_6	0.00	0.00	0.00	0.33	0.33	0.00	0.33

α = 0.14

P =

	d_0	d_1	d_2	d_3	d_4	d_5	d_6
d_0	0.02	0.02	0.88	0.02	0.02	0.02	0.02
d_1	0.02	0.45	0.45	0.02	0.02	0.02	0.02
d_2	0.31	0.02	0.31	0.31	0.02	0.02	0.02
d_3	0.02	0.02	0.02	0.45	0.45	0.02	0.02
d_4	0.02	0.02	0.02	0.02	0.02	0.02	0.88
d_5	0.02	0.02	0.02	0.02	0.02	0.45	0.45
d_6	0.02	0.02	0.02	0.31	0.31	0.02	0.31

Power method convergence

	x	xP^1	xP^2	xP^3	xP^4	xP^5	xP^6	xP^7	xP^8	xP^9	xP^{10}	xP^{11}	xP^{12}	xP^{13}
d_0	0.14	0.06	0.09	0.07	0.07	0.06	0.06	0.06	0.06	0.05	0.05	0.05	0.05	0.05
d_1	0.14	0.08	0.06	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04
d_2	0.14	0.25	0.18	0.17	0.15	0.14	0.13	0.12	0.12	0.12	0.12	0.11	0.11	0.11
d_3	0.14	0.16	0.23	0.24	0.24	0.24	0.24	0.25	0.25	0.25	0.25	0.25	0.25	0.25
d_4	0.14	0.12	0.16	0.19	0.19	0.20	0.21	0.21	0.21	0.21	0.21	0.21	0.21	0.21
d_5	0.14	0.08	0.06	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04
d_6	0.14	0.25	0.23	0.25	0.27	0.28	0.29	0.29	0.30	0.30	0.30	0.30	0.31	0.31

Pagerank summary

- Preprocessing:
  - Given graph of links, build matrix **P**.
  - From it compute **a**.
  - The entry a_i is a number between 0 and 1: the pagerank of page *i*.
- Query processing:
  - Retrieve pages meeting query.
  - Rank them by their pagerank.
  - Order is **query-*independent***.

PageRank issues

- Real surfers are not random surfers – Markov model is not a good model of surfing.
  - Issues: back button, short vs. long paths, bookmarks, directories – and search!
- Simple PageRank ranking (as described on previous slide) produces bad results for many pages.
  - Consider the query ***video service***
  - The Yahoo home page (i) has a very high PageRank and (ii) contains both words.
  - If we rank all Boolean hits according to PageRank, then the Yahoo home page would be top-ranked.
  - Clearly not desirable
- In practice: rank according to weighted combination of many factors, including raw text match, anchor text match, PageRank and many other factors

How important is PageRank?

- Frequent claim: PageRank is the most important component of web ranking.
- The reality:
  - There are several components that are at least as important: e.g., anchor text, indexing , zone weighting, phrases ...
- Rumor has it that PageRank in his original form (as presented here) now has a negligible impact on ranking!
- However, variants of a page’s PageRank are still an essential part of ranking.
- Addressing link spam is difficult and crucial.

What is PageRank?