Institut für Informationssysteme
Technische Universität Braunschweig

Information Retrieval and Web Search Engines

Lecture 12: Link Analysis

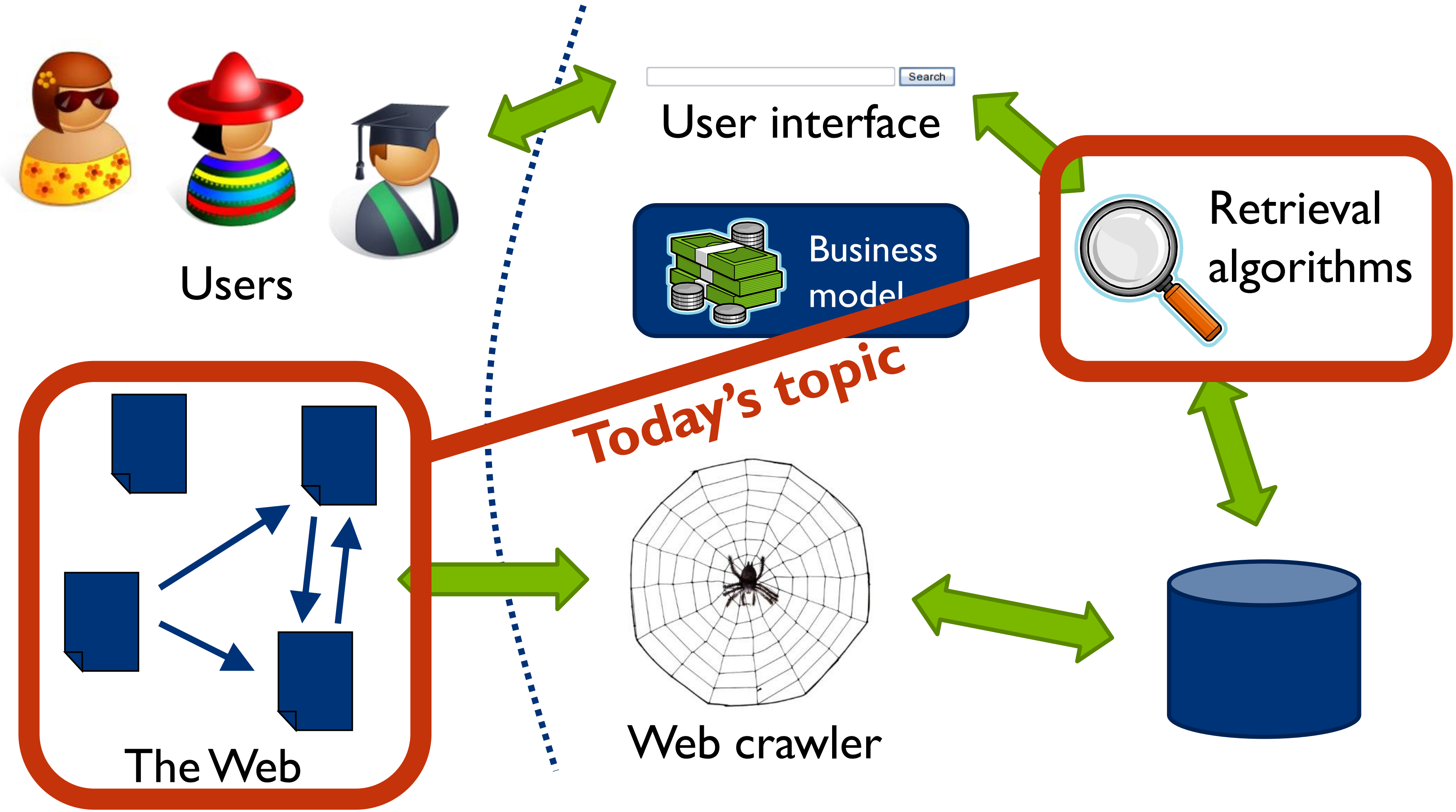
Wolf-Tilo Balke

Muhammad Usman

Institut für Informationssysteme
Technische Universität Braunschweig

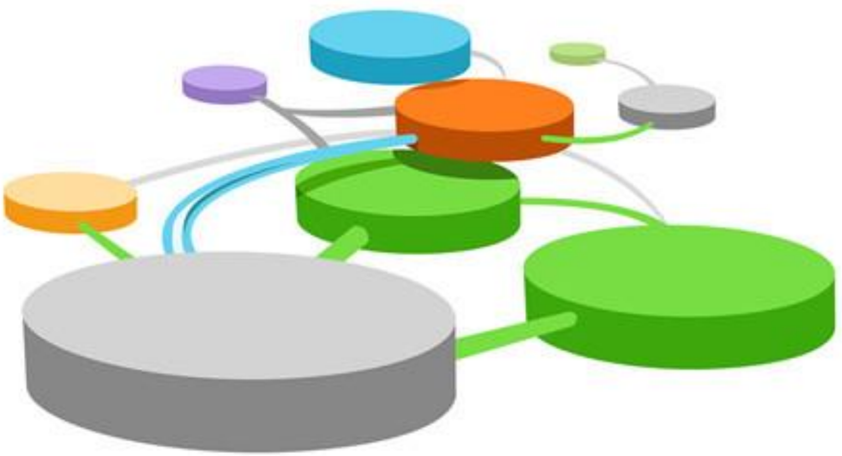
An Overview of Web Retrieval

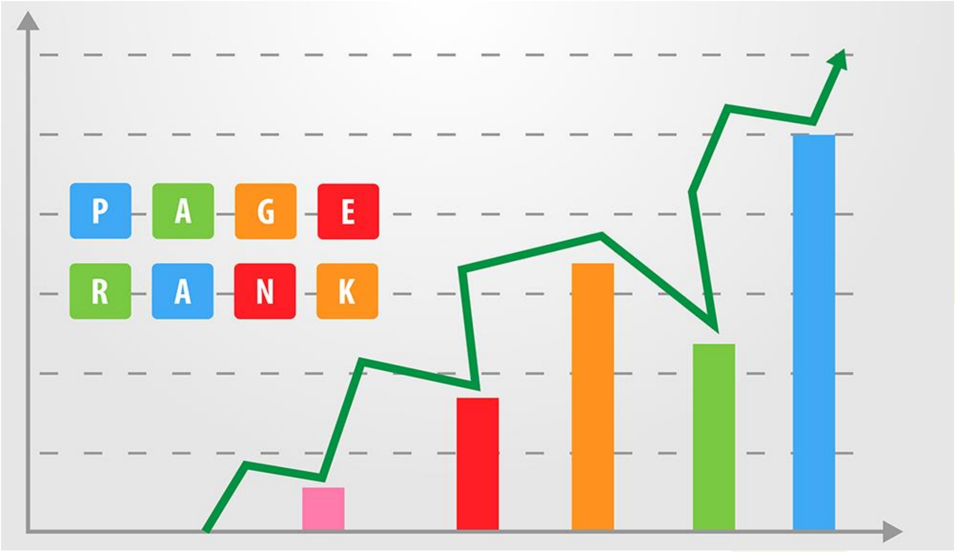
A typical Web search engine:



Link Analysis

1. **Link Structures**
2. PageRank
3. HITS





Social Networks

Networks of **social interactions** are formed…

– Between academics by **co-authoring**

Optimal Preference Elicitation for
Skyline Queries over Categorical Domains

Jongwuk Lee[1], Gae-won You[1], Seung-won Hwang[1],
Joachim Selke[2], and Wolf-Tilo Balke[2]

– Between movie personnel by **directing and acting**

Social Networks

– Between **musicians, soccer stars, friends, and relatives**

– Between **countries** via trading relations

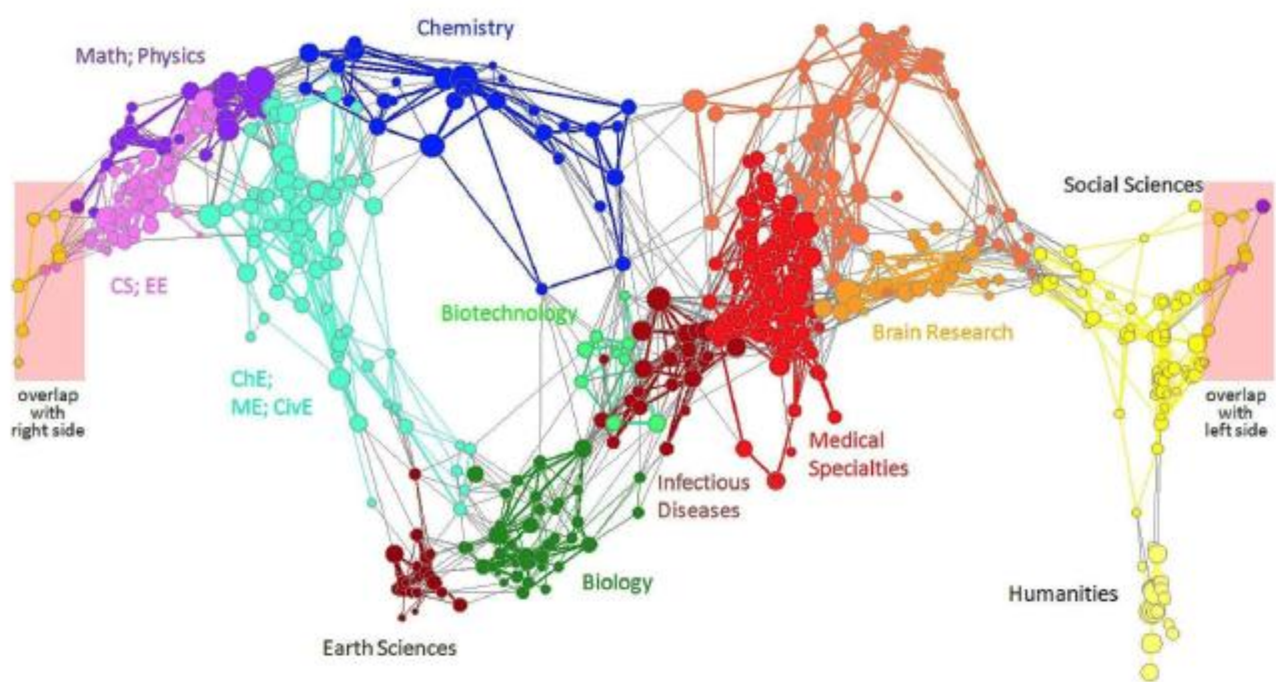
Social Networks

– Between people making **phone calls**

– Between people transmitting **infections**

Social Networks

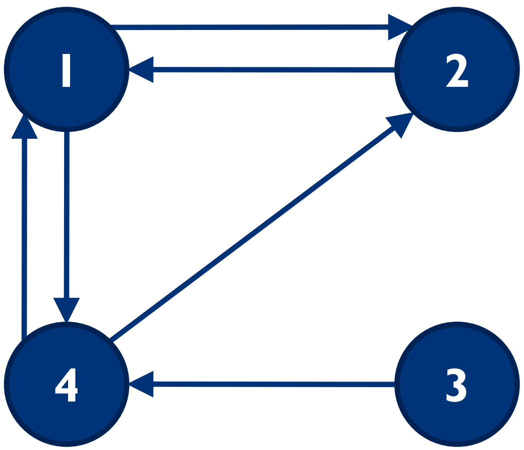
– Between scientific papers through **citations**



– **And, of course, between Web pages through links…**

Models of Social Networks

- It has been quite common for decades to model social networks using **directed graphs:**



Directed graph

A	1	2	3	4
1	0	1	0	1
2	1	0	0	0
3	0	0	0	1
4	1	1	0	0

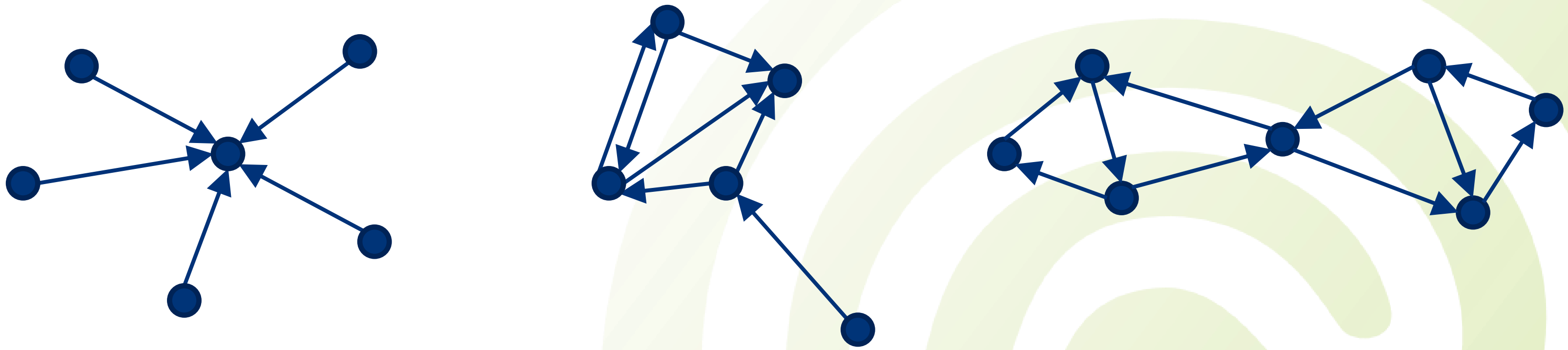
Adjacency matrix

$A_{i,j} = 1$
if and only if
node i links to node j

Models of Social Networks

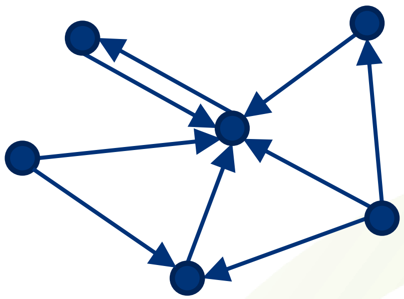
Classical research questions:

- Which authors have a high **prestige** (or status)?
- Which countries are **well-connected,** which are **isolated?**
- Which people **connect** different communities?



The Recursive Nature of Prestige

- Using the graph model, it has been clear that **in-degree is a good first-order indicator of prestige**



- In 1949, the sociologist John R. Seeley realized the **recursive nature of prestige** in a social network
  - A person's status is a function of the status of those who choose him
  - And their status is a function of those who choose them
  - And so *ad infinitum*…

A Model of Prestige

- Seeley **modeled prestige** as follows:
  - Every node u has a notion of **prestige $p(u)$** associated with it, which is simply a **positive real number**
  - **Recursive constraint:**
    The prestige of each node u should be proportional to the total sum of prestige of all nodes that link to u, i.e.

$$p(u) = \alpha \cdot \sum_{v \longrightarrow u} p(v)$$

  - Over all nodes, we represent the prestige score as a real column vector p having exactly one entry for each node
  - **Equivalent fixpoint condition:**

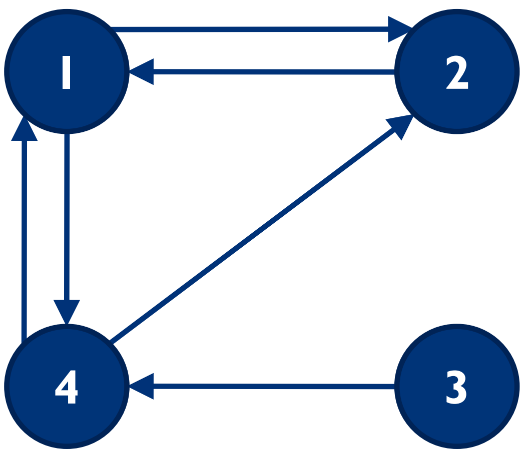
$$p = \alpha \cdot A^{\mathrm{T}} \cdot p$$

    - **Task:** Find numbers p and α such that the condition holds
    - This approach fits well to ideas from linear algebra (later)

A Model of Prestige

$$p(u) = \alpha \cdot \sum_{v \to u} p(v) \qquad p = \alpha \cdot A^T \cdot p$$

Example:



A	1	2	3	4
1	0	1	0	1
2	1	0	0	0
3	0	0	0	1
4	1	1	0	0

A^T	1	2	3	4
1	0	1	0	1
2	1	0	0	1
3	0	0	0	0
4	1	0	1	0

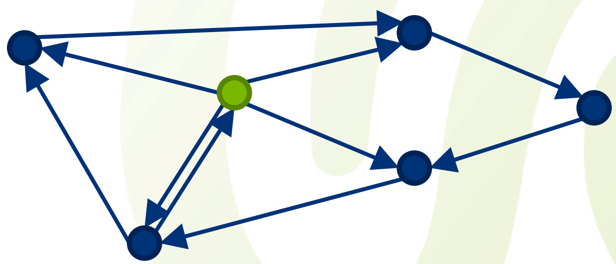
Solution:

$p = (0.65, 0.65, 0, 0.4)$

$\alpha = 0.62$

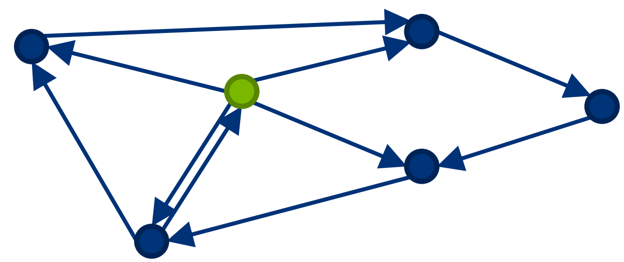
Centrality

- Another interesting notion is **centrality**
- **Definitions:**
  - The **distance** $\boldsymbol{d(u, v)}$ between two nodes u and v in a directed graph is the **smallest number of links** via which one can go **from** $\boldsymbol{u}$ **to** $\boldsymbol{v}$
  - The **radius** of a node u is $\boldsymbol{r(u) = \max_v d(u, v)}$, i.e., the distance to u's **most distant node**
  - The **center** of the graph is $\arg\min_u r(u)$, i.e., the node that has the **smallest radius**



Centrality

- The scientific **citation graph:**
  - Link a **paper** *u* to a paper *v*, i.e. set $u \rightarrow v$, if ***u* is cited by *v***
  - Papers having a **small radius** are likely to be very **influential**

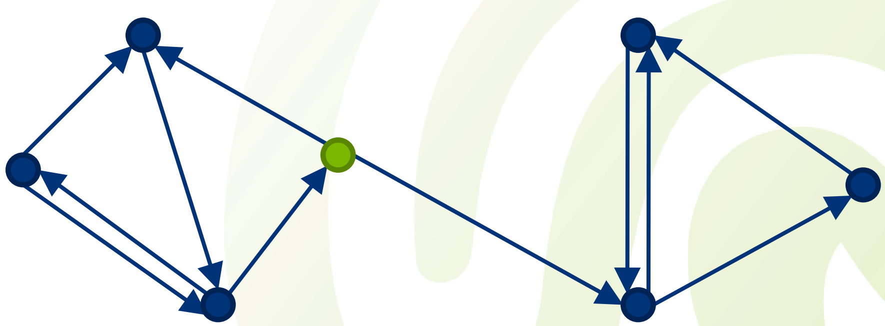


- The scientific **collaboration graph:**
  - Link two **authors** *u* and *v*, i.e. set $u \leftrightarrow v$, if they **co-authored** a paper
  - The **Erdős number** of an author *u* is his/her **distance to** the famous mathematician **Paul Erdős**

Centrality

There are many other notions of centrality, e.g., **cuts:**

- A **cut** is a (usually small) number of **edges** that, when removed, **disconnect a given pair of vertices**
- One may look for a small set of **vertices** that, when removed, will **decompose** the graph into two or more connected components
- This is useful for the study of **epidemics, espionage,** or suspected **terrorist communication** on telephone networks



Co-Citation

- Another important measure is **co-citation**
  - If document *u* cites documents *v* and *w*, then *v* and *w* are said to be co-cited by *u*
- If documents *v* and *w* are **co-cited** by many documents, then *v* and *w* are somehow **related** to each other
- In terms of the adjacency matrix *A*:
  - Link a **document** *u* to a paper *v*, i.e. set $u \rightarrow v$, if *u* cites *v*
  - The number of documents co-citing *v* and *w* is the entry corresponding to *v* and *w* in the matrix $A^T A$:

$$A^T A[v, w] = \sum_u A^T[v, u] A[u, w]$$

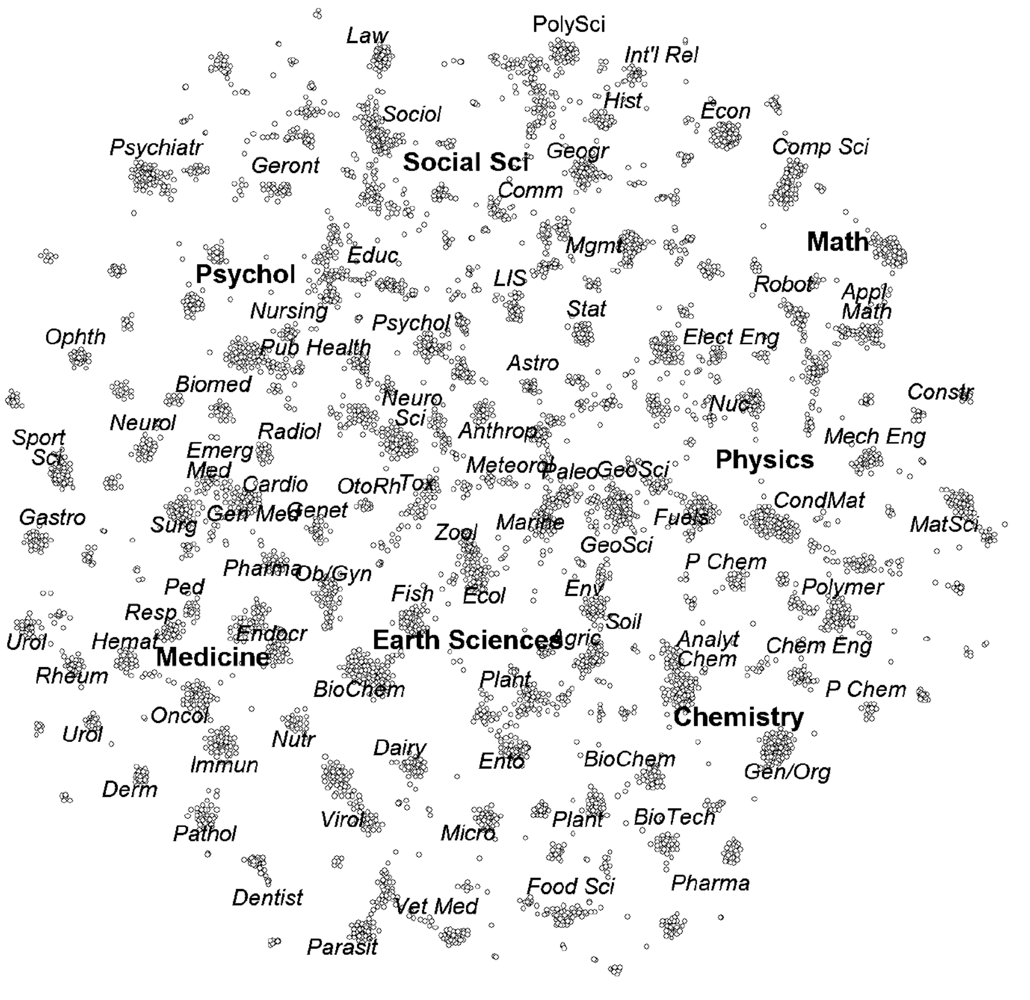
$$= \sum_u A[u, v] A[u, w] = \left| \{ u \mid u \rightarrow v \text{ and } u \rightarrow w \} \right|$$

Co-Citation

- The entry in the $A^T A$ matrix corresponding to [*v*, *w*] is the **co-citation index** of *v* and *w* and a **measure of relatedness** between *v* and *w*
- One may use this pairwise relatedness measure in a clustering algorithm, such as **multidimensional scaling**
- MDS is similar to the **singular value decomposition**
- It uses a similarity matrix to **embed** the documents into a **low-dimensional Euclidean space** (e.g. a plane)
- **Visualizing clusters** based on co-citation reveals important **social structures** between and within link communities

Co-Citation

(Boyack *et al.*, 2005) visualized similarity data based on co-citations created from over **1 million journal articles** published in 2000:



Each point represents a journal

Back to the Web

- **Classical IR:**
  - The **worth of a document** with regard to a query is **intrinsic** to the document
  - **Documents are self-contained units,** and are generally descriptive and **truthful** about their contents

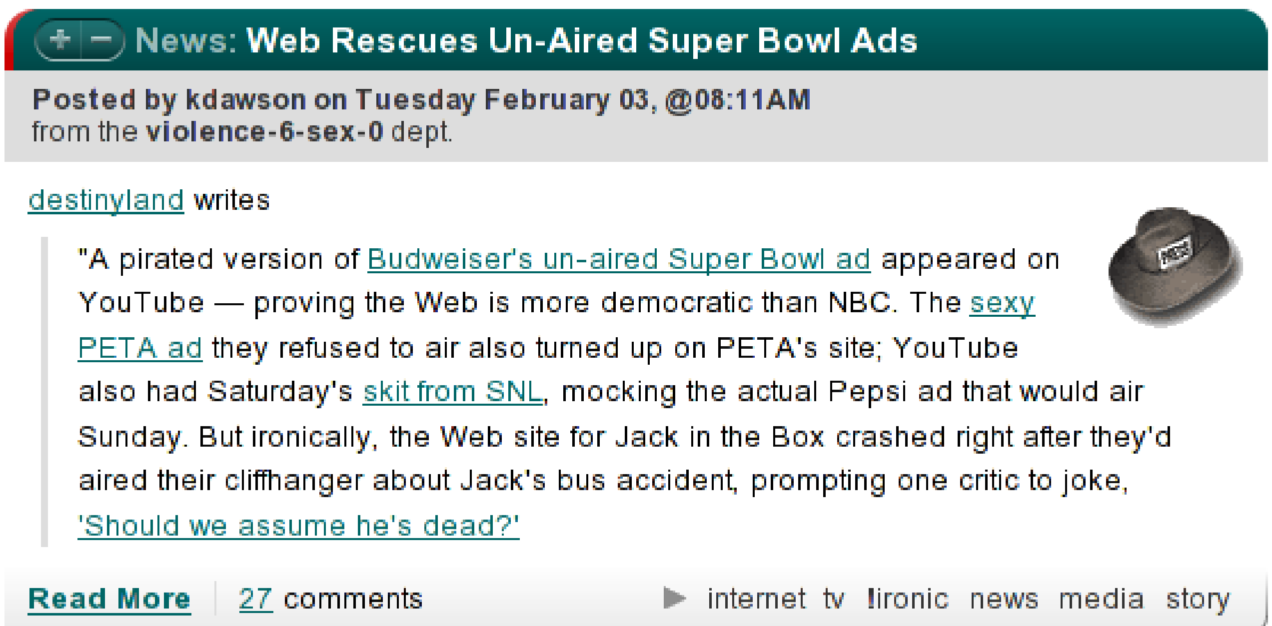
- **Modern Web search:**
  - Apply ideas from network analysis to the **Web graph…**
  - **Links are recommendations**
  - **Anchor texts** can be used as document descriptions

Back to the Web

Assumption 1:

A hyperlink is signal of quality or popular interest

– In some sense, a link is a democratic vote


News: Web Rescues Un-Aired Super Bowl Ads

Posted by kdawson on Tuesday February 03, @08:11AM
from the **violence-6-sex-0** dept.

destinyland writes

"A pirated version of Budweiser's un-aired Super Bowl ad appeared on YouTube — proving the Web is more democratic than NBC. The sexy PETA ad they refused to air also turned up on PETA's site; YouTube also had Saturday's skit from SNL, mocking the actual Pepsi ad that would air Sunday. But ironically, the Web site for Jack in the Box crashed right after they'd aired their cliffhanger about Jack's bus accident, prompting one critic to joke, 'Should we assume he's dead?'

Read More | 27 comments ▶ internet tv !ironic news media story

Back to the Web

Assumption 2:

The anchor text of a link (or its surrounding text) describes the target page

- [IBM](#)
  IBM manufactures and sells computer services, hardware, and software. Also provides financing services in support of its computer business.
  www.ibm.com

Excerpt from
Yahoo! Directory

IBM's home page
(does not contain the term computer!)

Link Analysis

- Both assumptions clearly do not hold always
- But exploiting them has proved to be much better than not exploiting them

- **We will concentrate on the first assumption:**
  "Links are quality signals"

- **Two highly popular algorithms:**
  - PageRank (Page *et al.*, 1998)
  - HITS (Kleinberg, 1999)

PageRank and HITS

- **PageRank**
  - Developed around the fall of 1996 at Stanford University by Larry Page and Sergey Brin, the founders of **Google**
  - **Idea:** Assign a **query-independent** measure of **prestige** to each Web resource

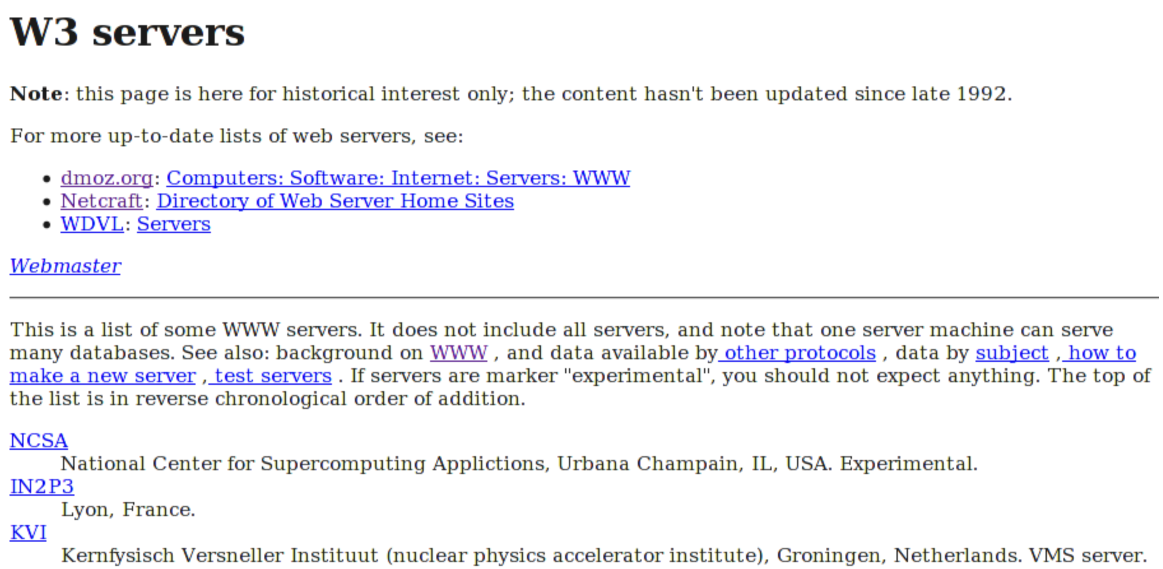
- **HITS**
  - Developed at the same time at IBM Almaden Research Center by Jon Kleinberg, a famous computer scientist
  - **Idea:** For any given query, assign **two measures** to each Web resource, a **hub score** and an **authority score**
    - **Hub:** A compilation of links to relevant Web resources
    - **Authority:** A resource that is relevant in itself

History of Web Search

Detour

Before 1993:

- There are **no search engines…**
  - Archie (ftp indexing) at McGill University
- Tim Berners-Lee maintains a **list of Web servers:**



W3 servers

Note: this page is here for historical interest only; the content hasn't been updated since late 1992.

For more up-to-date lists of web servers, see:

- dmoz.org: Computers: Software: Internet: Servers: WWW
- Netcraft: Directory of Web Server Home Sites
- WDVL: Servers

Webmaster

This is a list of some WWW servers. It does not include all servers, and note that one server machine can serve many databases. See also: background on WWW , and data available by other protocols , data by subject , how to make a new server , test servers . If servers are marker "experimental", you should not expect anything. The top of the list is in reverse chronological order of addition.

NCSA
National Center for Supercomputing Applictions, Urbana Champain, IL, USA. Experimental.
IN2P3
Lyon, France.
KVI
Kernfysisch Versneller Instituut (nuclear physics accelerator institute), Groningen, Netherlands. VMS server.

- **In Germany:** LEO, "Link Everything Online" at TU Munich

Brief History of Web Search

Detour

1993–1998:

- Many new search engines, most popular:
  Lycos, AltaVista, Excite, Inktomi, HotBot, Ask Jeeves
- All of them mainly rely on **classical IR techniques** and focus on the **problem of scaling**

1998:

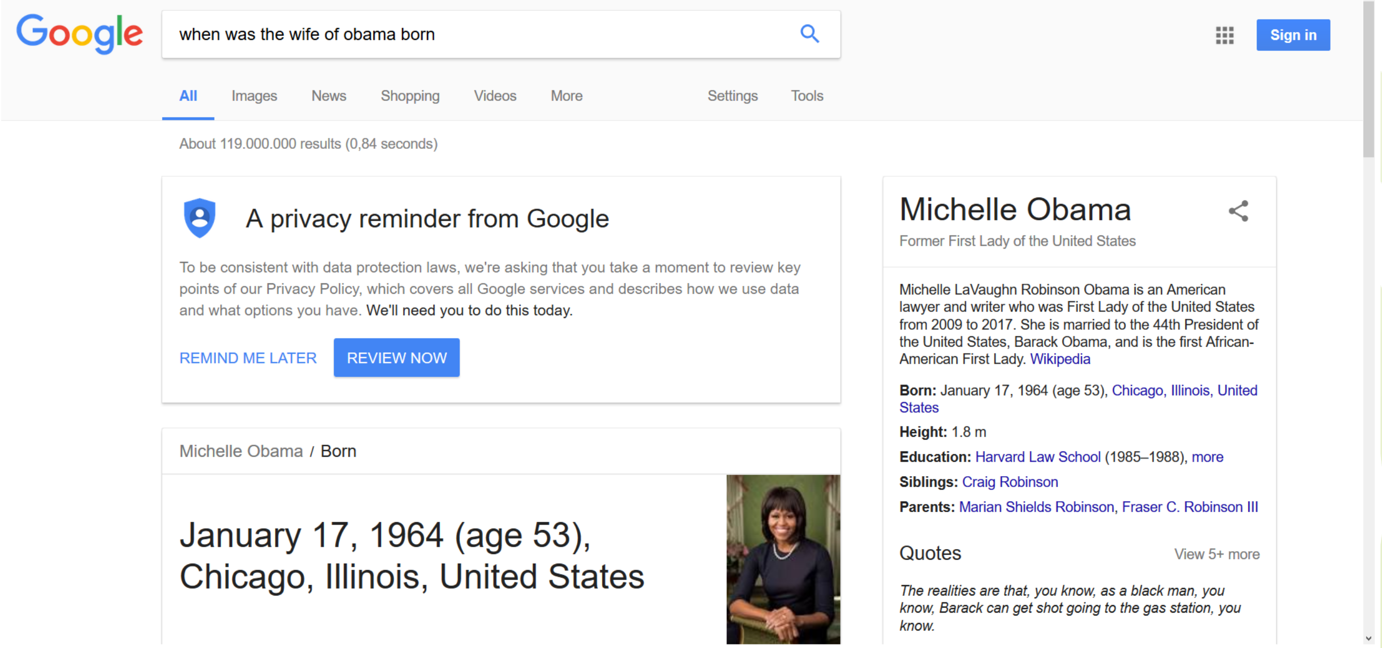
- **Google** is founded
- The first engine that heavily exploits the Web's **link structure**
- Google's success has a name: **PageRank**

1998–Today:

- Large companies try to **keep up with Google**
- Most noteworthy: Yahoo and Microsoft

The next big thing in Web search?

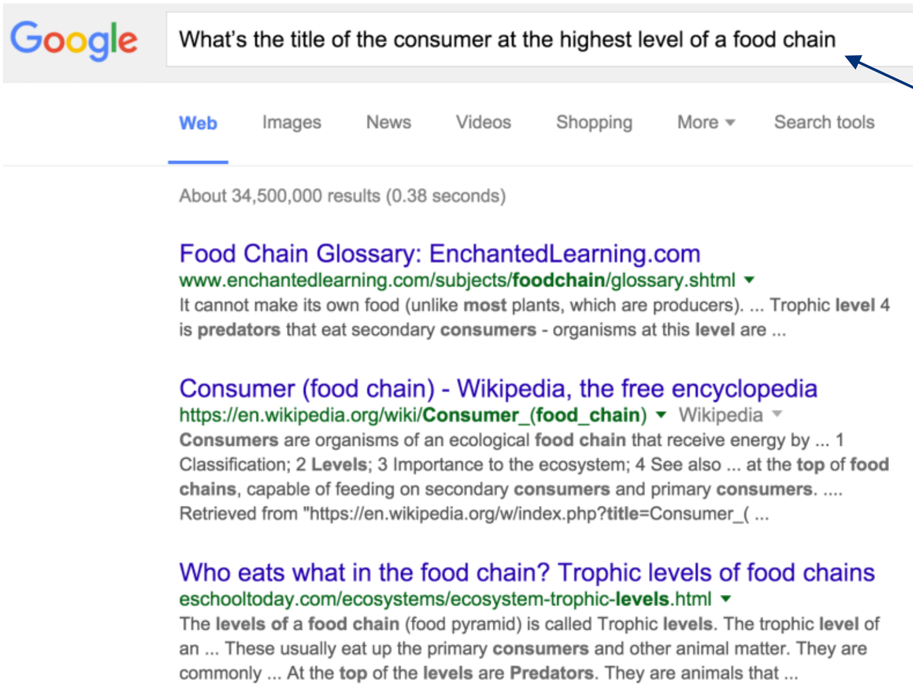
- Clustering?
- Natural language query processing?
- The "Semantic Web"? e.g. Knowledge Graph



The next big thing…

Detour

– Artificial Intelligence (AI) e.g. RankBrain from Google

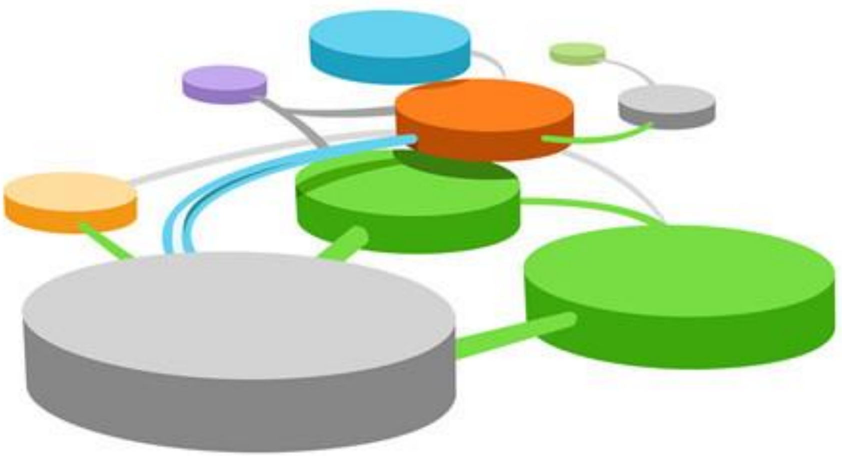


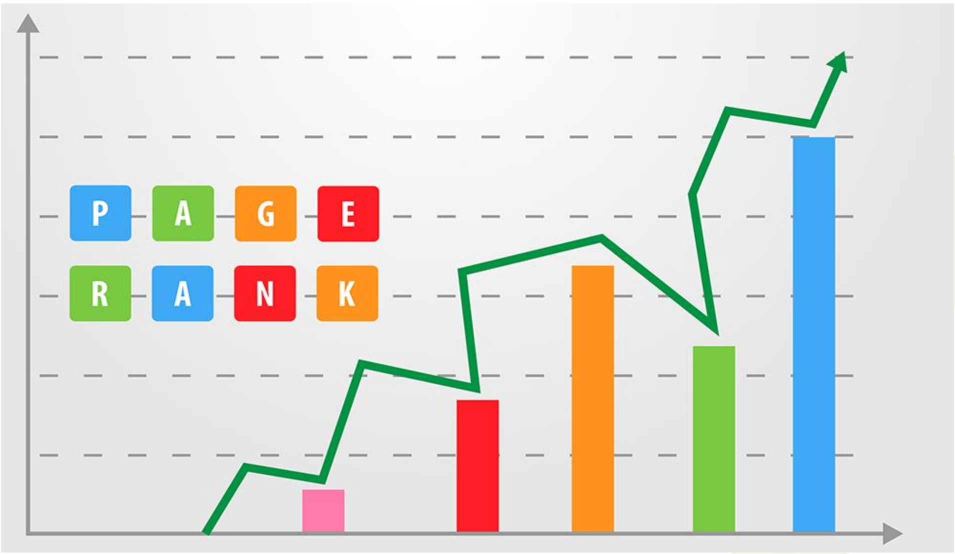
"top level of the food chain"

– Something else?

Link Analysis

1. Link Structures
2. **PageRank**
3. HITS





PageRank

- **Problem:**
  - How to assign a **query-independent** measure of **prestige** to each Web resource?
- **A good but infeasible solution:**
  - Rank Web resources by their **popularity** (measured by traffic?)
- **The PageRank solution:**
  - Apply **John R. Seeley's model** of prestige to the Web graph!
  - The number of in-links is correlated to a resource's prestige
  - Links from good resources should count more than links from bad ones

$$p(u) = \alpha \cdot \sum_{v \longrightarrow u} p(v)$$

The Random Surfer Model

Imagine a Web surfer doing a **random walk** on the Web:

- 90% of the time, the surfer clicks a **random hyperlink**
- 10% of the time, the surfer types in a **random URI**
- **PageRank = The long-term visit rate of each node**

This is **a crude, but useful, Web surfing model**

- No one chooses links with equal probability, surfing usually is topic-driven
- How to surf to a random page?
- What about the back button or bookmarks?

The Random Surfer Model

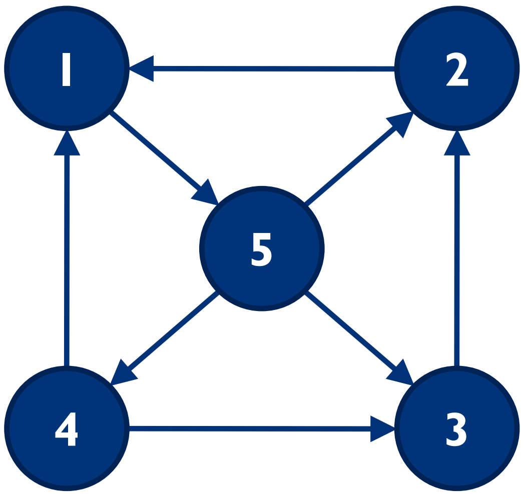
A more detailed version of the model:

1. Start at a random page, chosen uniformly
2. Flip a coin that shows "tails" with probability λ
3. If the coin shows "heads"
   AND the current page has a positive out-degree:
   - Randomly follow one of the pages out-links
   - Continue at (2)

   If the coin shows "tails"
   OR the current page has no out-links:
   - Surf to a random Web page, chosen uniformly
   - Continue at (2)

The Random Surfer Model

Example:



Adjacency matrix:

A	1	2	3	4	5
1					1
2	1				
3		1			
4	1		1		
5		1	1	1	

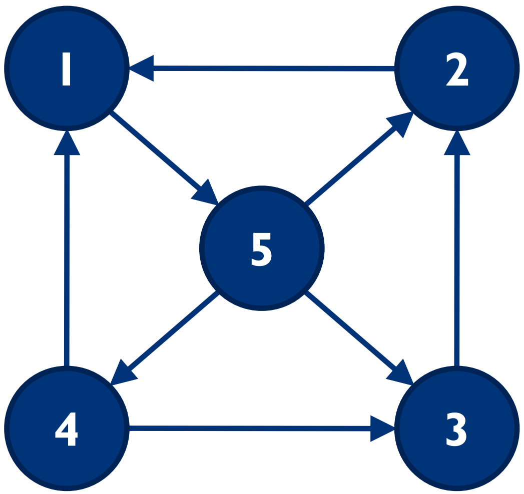
Set λ = 0.25

Transition matrix:

T	1	2	3	4	5
1	0.05	0.05	0.05	0.05	**0.75 + 0.05**
2	**0.75 + 0.05**	0.05	0.05	0.05	0.05
3	0.05	**0.75 + 0.05**	0.05	0.05	0.05
4	**0.375 + 0.05**	0.05	**0.375 + 0.05**	0.05	0.05
5	0.05	**0.25 + 0.05**	**0.25 + 0.05**	**0.25 + 0.05**	0.05

The Random Surfer Model

Example (continued):



Transition matrix:

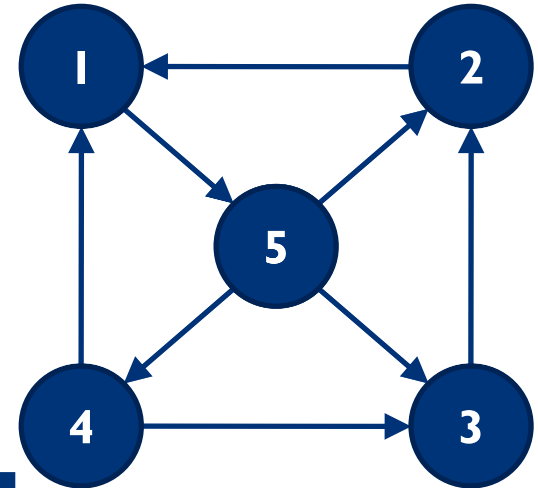
T	1	2	3	4	5
1	0.05	0.05	0.05	0.05	**0.8**
2	**0.8**	0.05	0.05	0.05	0.05
3	0.05	**0.8**	0.05	0.05	0.05
4	**0.425**	0.05	**0.425**	0.05	0.05
5	0.05	**0.3**	**0.3**	**0.3**	0.05

- If the surfer is at page 3 in step *t*
  - He/she will be at page 1 in step $t + 1$ with a probability of 5%
  - He/she will be at page 2 in step $t + 1$ with a probability of 80%
  - He/she will be at page 3 in step $t + 1$ with a probability of 5%
  - He/she will be at page 4 in step $t + 1$ with a probability of 5%
  - He/she will be at page 5 in step $t + 1$ with a probability of 5%

The Random Surfer Model

Example (continued):

- Let's do a **simulation**
- If we start in state 1,
  what's the **probability** of
  being in **state *i* after *t* steps?**



	1	2	3	4	5
$t = 0$	1	0	0	0	0
$t = 1$	0.05	0.05	0.05	0.05	0.8
$t = 2$	0.11	0.29	0.27	0.25	0.09
$t = 3$	0.36	0.27	0.17	0.07	0.13
$t = 4$	0.28	0.21	0.11	0.08	0.32
$t = 5$	0.24	0.21	0.16	0.13	0.26
$t = 6$	0.26	0.24	0.16	0.12	0.23
$t = 7$	0.27	0.23	0.15	0.11	0.24
$t = 8$	0.26	0.22	0.15	0.11	0.25
$t = 9$	0.26	0.23	0.15	0.11	0.25

The probability vector seems to converge…

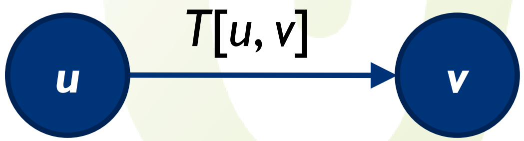
Convergence

- And indeed, **the probability vector converges** as *t* goes to infinity, for any initial probability vector
- To make this point clear, we need some **linear algebra** and some **theory of stochastic processes**
- **Definitions:**
  - Let *n* denote the number of nodes
  - A **probability vector** is an *n*-dimensional vector such that
    (a) all entries are **non-negative** and
    (b) the **sum of entries is 1**
  - A **stochastic matrix** is an $n \times n$ matrix such that
    (a) all entries are **non-negative** and
    (b) the **sum of each row is 1**

Convergence

- Stochastic matrices are closely related to **Markov chains:**
  - A Markov chain consists of **n states** and an **$n \times n$ stochastic matrix T**
  - Each row and column of T corresponds to a state, respectively
  - At any point in time, the Markov chain is in exactly one of these states
  - **Time is discrete,** i.e. it runs in discrete steps: $t = 0, 1, 2, \ldots$
  - From time step to time step, the chain's current state changes according to the stochastic matrix T:

$$\Pr(\text{state } v \text{ at time } t + 1 \mid \text{state } u \text{ at time } t) = T[u, v]$$



Convergence

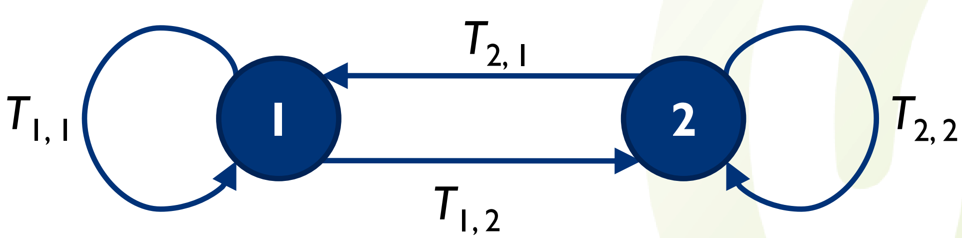
- In essence, a Markov chain is a probabilistic finite state machine
- Knowledge about the current state of a Markov chain can be expressed by **probability vectors** of length *n*
- Remember our example:
  - Knowing for sure that the current state of the chain is state u, can be expressed by a probability vector that is 1 at *u*'s place
  - For example, (0.2, 0.5, 0.3) means that the chain's probability to be in the first, second, and third state is 20%, 50%, and 30%, respectively

Convergence

- State transitions can be formalized using matrix–vector multiplication
- Let T be a transition matrix and p a probability vector that models the chain's state probabilities at time t
- What are the **state probabilities p' at time t + 1?**

$$p' = T^{\mathrm{T}} \cdot p \qquad p'_i = \sum_{k=1}^{n} T_{k,i} \cdot p_k$$

- **Example (n = 2):**



$p = (p_1, p_2)$
$p' = (p_1', p_2')$

$p_1' = T_{1,1} \cdot p_1 + T_{2,1} \cdot p_2$
$p_2' = T_{1,2} \cdot p_1 + T_{2,2} \cdot p_2$

Convergence

- Now we have everything we need to talk about **convergence properties** of the Markov chain

- Let $\boldsymbol{p_0}$ be some **initial probability state vector**
- Let $\boldsymbol{p_t}$ denote the **probability state vector at time $\boldsymbol{t}$**
- Then, for any t, we have $p_{t+1} = T^{\mathrm{T}} \cdot p_t$

- Clearly, convergence of p_t as $t \to \infty$ means that **$\boldsymbol{p_t}$ converges to a vector $\boldsymbol{p}$** such that

$$\boldsymbol{p} = \boldsymbol{T}^{\mathrm{T}} \cdot \boldsymbol{p}$$

- Well, what we are looking for is an **eigenvector of $\boldsymbol{T}^{\mathrm{T}}$** corresponding to the **eigenvalue 1**

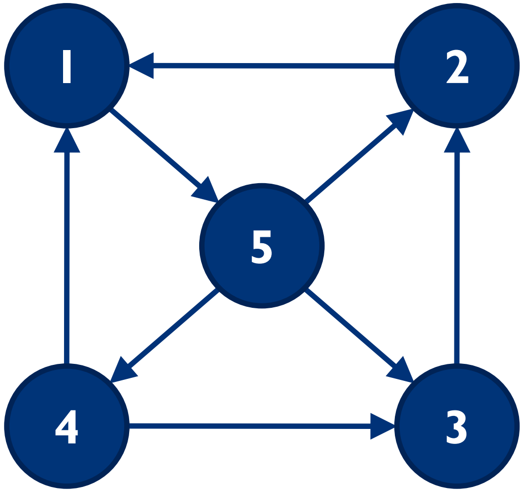
Convergence

- According to the **Perron–Frobenius theorem** from linear algebra the following is true:
  - Every stochastic matrix containing **only positive entries** has **1 as one of its eigenvalues**
  - Furthermore, 1 is the **largest eigenvalue** of the matrix
  - There is **only one eigenvector** having the eigenvalue 1
- Since we do a **random teleport** with probability $\lambda > 0$ in the random surfer model, the theorem applies
- Therefore, we can be sure that there is a probability vector p satisfying $p = T^{\mathrm{T}} \cdot p$
- Such a vector p is called the Markov chain's **stationary probability vector**

PageRank

- In the **random surfer model** there is a **unique stationary probability vector *p***
- Node *u*'s **PageRank** is its stationary probability $p[u]$

	1	2	3	4	5
$t = 0$	1	0	0	0	0
$t = 1$	0.05	0.05	0.05	0.05	0.8
$t = 2$	0.11	0.29	0.27	0.25	0.09
$t = 3$	0.36	0.27	0.17	0.07	0.13
…					
$t \to \infty$	0.26	0.23	0.15	0.11	0.25



- This fits **Seeley's notion of prestige:**

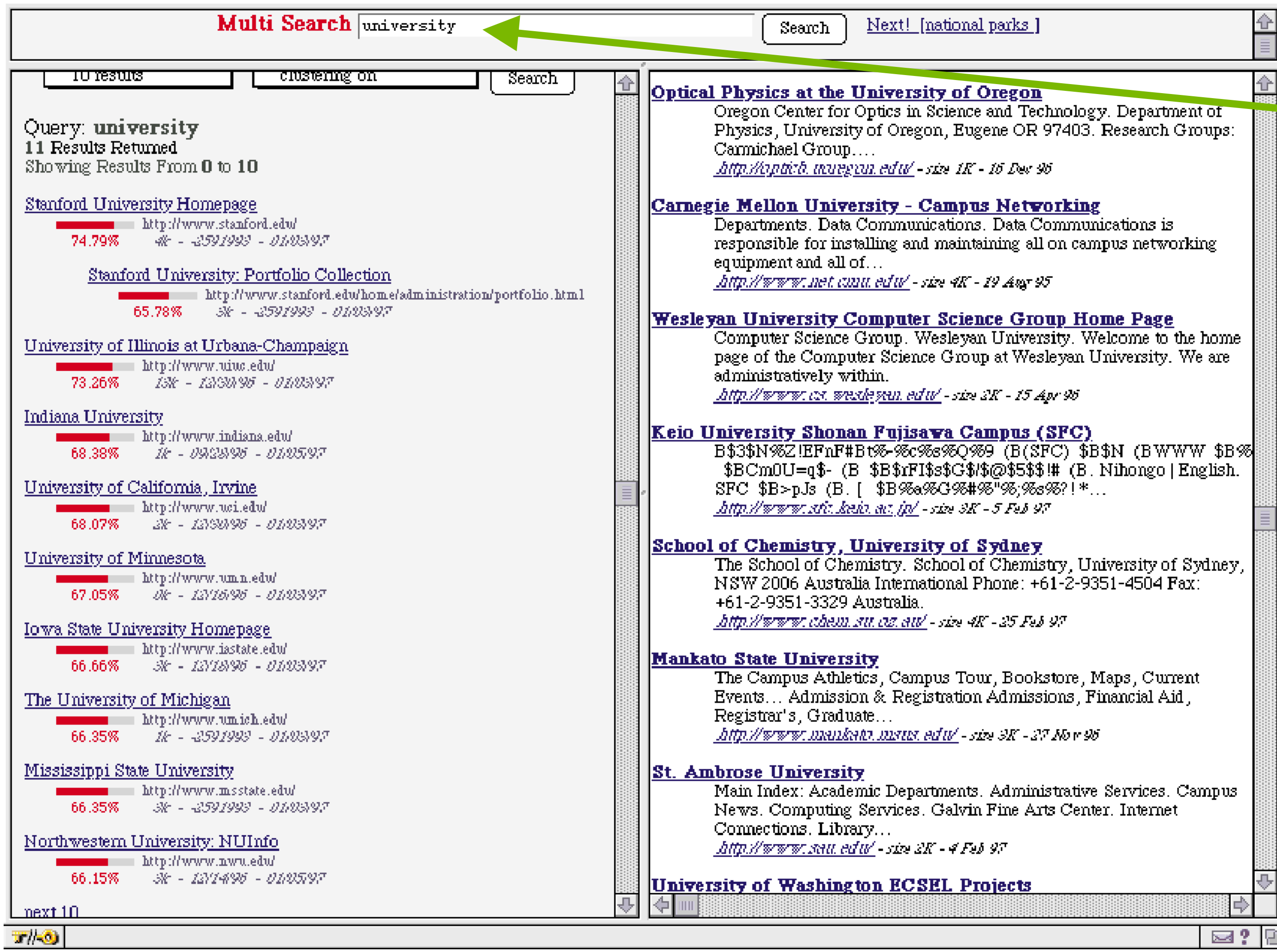
$$p(u) = \alpha \cdot \sum_{v \to u} p(v)$$

PageRank

- PageRank was invented by Larry Page at Stanford
- PageRank is **patented** as US patent 6,285,999
  - "Method for node ranking in a linked database"
    - The method for computing the PageRank and related stuff are patented!
  - Patent was assigned to Stanford University (not to Google)
  - **Google has exclusive license rights**
  - Stanford received **1.8 million shares in Google** in exchange for use of the patent
  - These shares were sold in 2005 for **336 million dollars**

PageRank

Result list sorted by PageRank   Result list sorted using IR methods

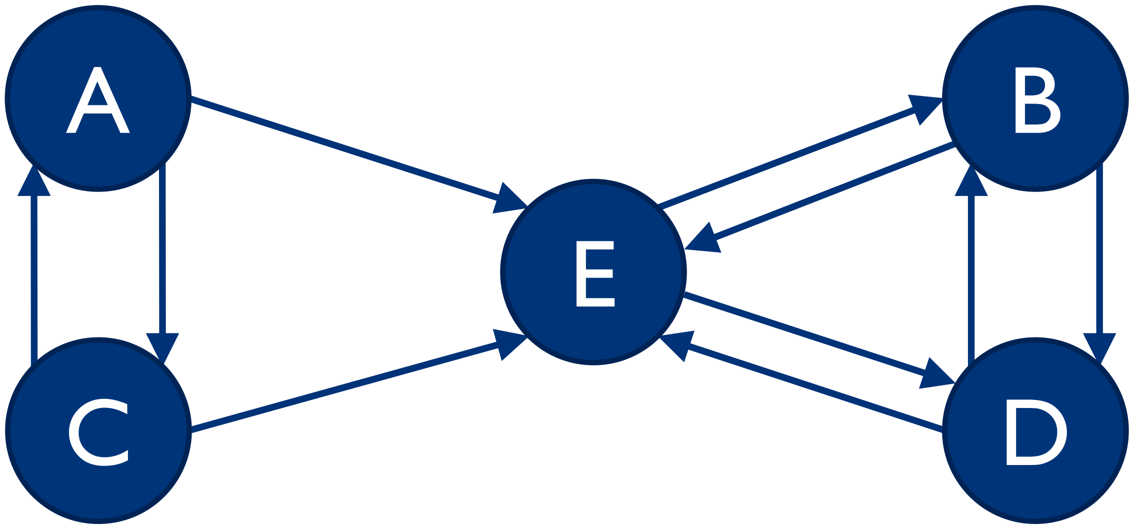


Query: "university"

PageRank Quiz

Quiz

A Web graph:



Which of the following node lists is ordered by PageRank?

a) E > B = D > A = C

b) B = E = D > A = C

c) E > D > B = A > C

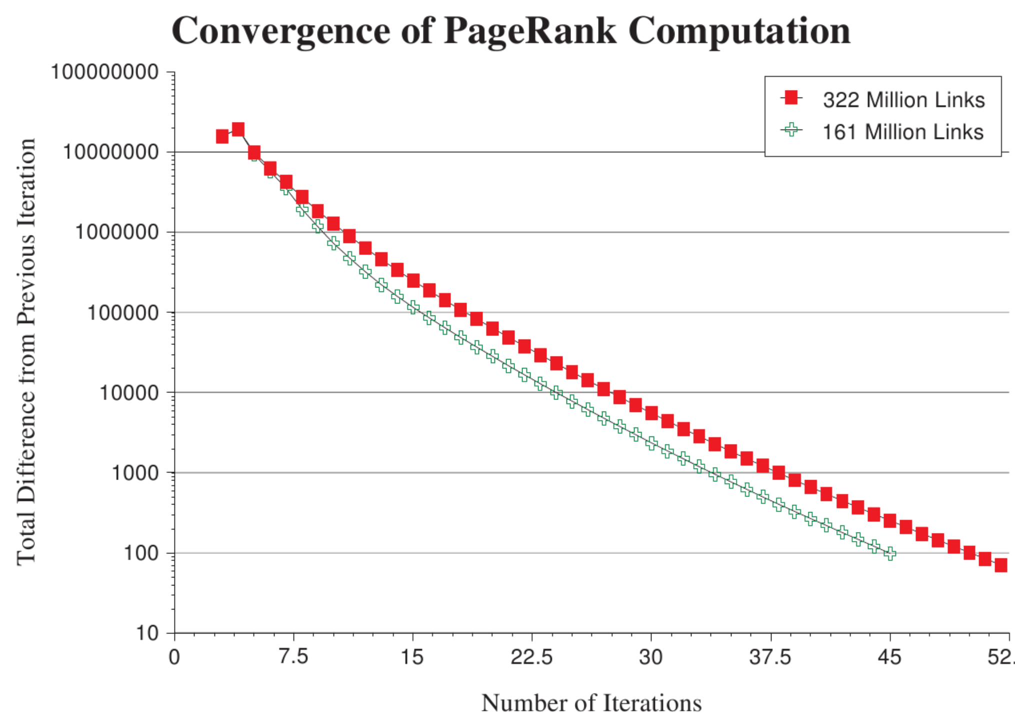
d) D > E > A = C > B

PageRank Computation

- How to compute the PageRank?
- A very simple method for eigenvalue and eigenvector computation is the so-called **power iteration,** which can be applied to any quadratic matrix *A*:
  1. Start with an arbitrary initial vector b_0
  2. Set $i = 0$
  3. Set $b_{i+1} = A \cdot b_i$
  4. Set $b_{i+1} = b_{i+1} / |b_{i+1}|$, i.e. normalize b_{i+1} to unit length
  5. Set $i = i + 1$
  6. GOTO 3

PageRank Computation

- One can prove that the **power iteration converges** to the eigenvector of *A* having the **largest eigenvalue**
- In our case, the largest eigenvalue is 1
  - The power iteration finds the stationary probability vector *p*
- How many iterations are needed?
  - Actually, the number is quite low since we don't need a perfect result anyway…



PageRank Computation

- How to compute the PageRank for a Web graph containing 60 billion nodes?
  - Use a highly scalable distributed algorithm
  - Actually, this is one of Google's secrets…

Importance of PageRank

- **A search engine myth:**
  "PageRank is the most important component of ranking"

- **The reality:**
  - There are several components that are at least as important: Anchor text, phrases, proximity, …
  - Google uses **hundreds of different features** for ranking
  - There are rumors that PageRank in its original form (as presented here) has a negligible effect on ranking
  - However, variants of PageRank are still an essential part of ranking
  - Addressing **link spam** is difficult and crucial!

Topic-Sensitive PageRank

- A disadvantage of PageRank is that it computes only a single overall score for each web resource
  - A web resource might be unimportant from a global view but highly important for a specific topic
- **Topic-sensitive PageRank** tries to address this issue:
  - Define a set of popular **topics** (e.g. football, Windows, Obama)
  - Use **classification** algorithms to assign each Web resource to one (or more) of these topics
  - For each topic, compute a **topic-sensitive PageRank** by **limiting the random teleports** to pages of the current topic
  - At query time, **detect the query's topics** and **use the corresponding PageRank scores…**

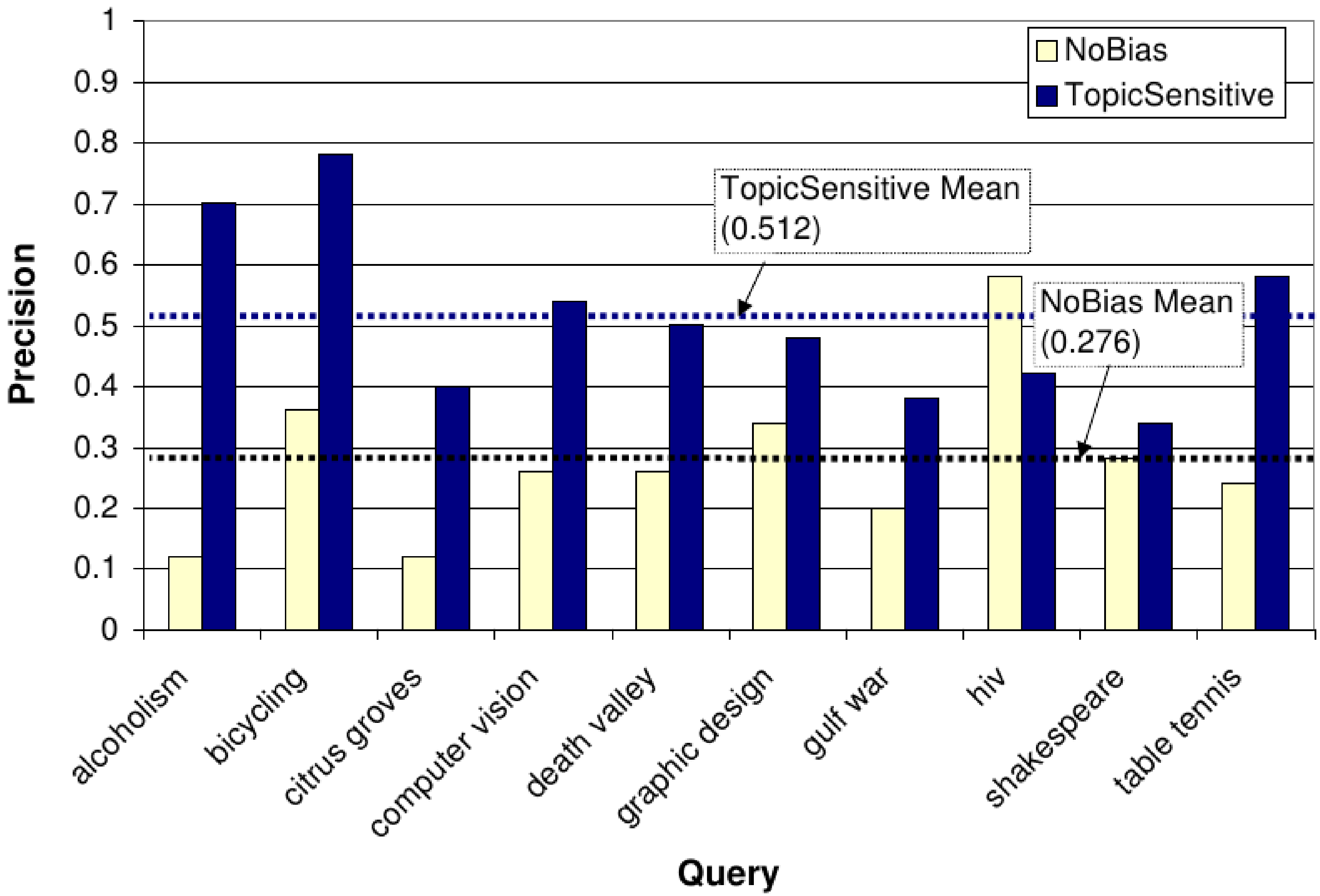
Topic-Sensitive PageRank

Example (query: bicycling):

NoBias	Arts
"RailRiders Adventure Clothing" www.RailRiders.com www.Waypoint.org/default.html www.Gorp.com/ www.FloridaCycling.com/	"Photo Contest & Gallery (Bicycling)" www.bikescape.com/photogallery/ www.trygve.com/ www.greenway.org/ www.jsc.nasa.gov/Bios/htmlbios/young.html
Business	**Computers**
"Recumbent Bikes and Kit Aircraft" www.rans.com www.BreakawayBooks.com java.oreilly.com/bite-size/ www.carbboom.com	"GPS Pilot" www.gpspilot.com www.wireless.gr/wireless-links.htm www.linkstosales.com www.LiftExperts.com/lifts.html
Games	**Kids and Teens**
"Definition Through Hobbies" www.flick.com/~gretchen/hobbies.html www.BellaOnline.com/sports/ www.npr.org/programs/wesun/puzzle/will.html www.trygve.com/	"Camp Shohola For Boys" www.shohola.com www.EarthForce.org www.WeissmanTours.com www.GrownupCamps.com/homepage.html
Recreation	**Science**
"Adventure travel" www.gorp.com/ www.GrownupCamps.com/homepage.html www.gorp.com/gorp/activity/main.htm www.outdoor-pursuits.org/	"Coast to Coast by Recumbent Bicycle" hypertextbook.com/bent/ www.SiestaSoftware.com/ www.BenWiens.com/benwiens.html www.SusanJeffers.com/jeffbio.htm
Shopping	**Sports**
"Cycling Clothing & Accessories for Women" www.TeamEstrogen.com/ www.ShopOutdoors.com/ www.jub.com.au/books/ www.bike.com/	"Swim, Bike, Run, & Multisport" www.multisports.com/ www.BikeRacing.com/ www.CycleCanada.com/ www.bikescape.com/photogallery/

Topic-Sensitive PageRank

Comparison to PageRank (precision at 10):

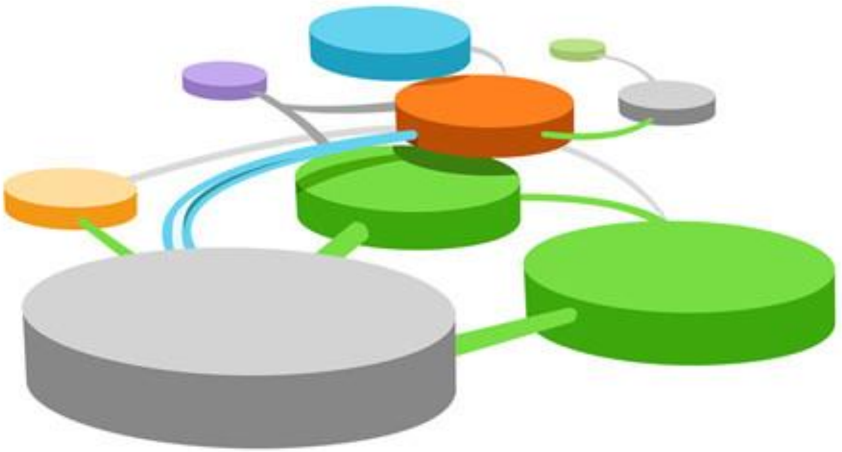


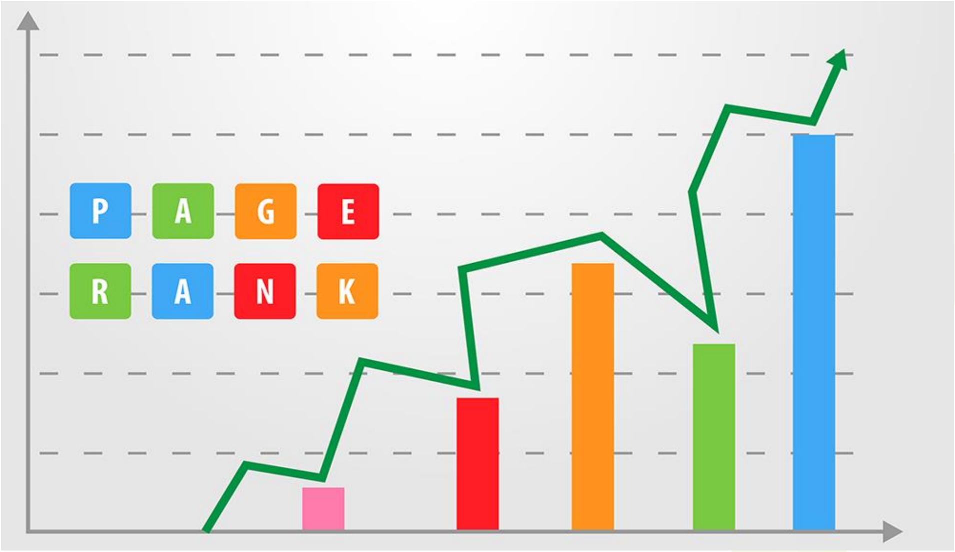
Possible Enhancements

- **Eliminate navigational links:**
  - Most web pages contain **navigational structures**
  - The **quality assumption** does only hold if a hyperlink was created as a result of **editorial judgment**
  - Therefore, navigational links should be removed before computing the PageRank
- **Eliminate nepotistic links:**
  - Nepotism = favoritism based on kinship
  - Links between **pages authored by the same person** also are problematic
  - Again, they should be removed before doing any computations
  - Unfortunately, it's much harder to detect them than detecting navigational links…

Link Analysis

1. Link Structures
2. PageRank
3. **HITS**





HITS

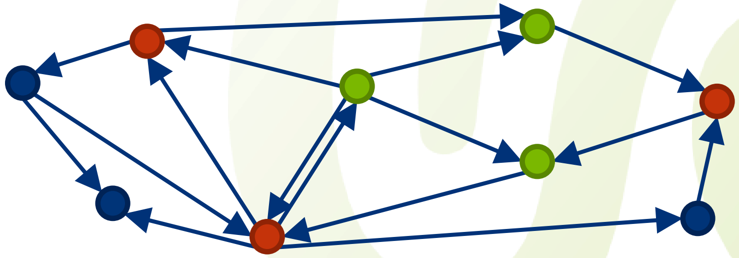
- HITS stands for
  hyperlink induced topic search
- Invented by **Jon Kleinberg**
- **Problem setting:**
  - For any information need,
    there are **hubs** and **authorities**
    - **Authority:** Definitive high-quality information (query-dependent!)
    - **Hub:** Comprehensive lists of links to authorities (query-dependent!)
  - To a certain degree, each page is a hub as well as an authority
- **Task:**
  - Given a query, estimate the degree of
    authority and hubness of each Web page

HITS

- **Obvious:**
  The authority and hubness scores are query-dependent, therefore the computation has to be done at query time

- **Idea:**
  - **Given:** A query q
  - Send q to a **standard IR system** to collect a **root set *R*** of nodes in the Web graph
  - Collect the **base set** V_q of nodes, which includes R as well as all nodes that are connected to R by an in-link or out-link

Root set



HITS

- **Idea (continued):**
  - Finally, **compute hub and authority scores** on the base set
- Hubs and authority scores are defined similar to prestige:
  - Let *A* be the base set's **adjacency matrix**
  - Denote the nodes' hub scores by a vector *h* and their authority scores by a vector *a*
  - **A recursive definition of *h* and *a*:**

$$a = \alpha \cdot A^{T} \cdot h \qquad\qquad h = \beta \cdot A \cdot a$$

  - Again, α **and** β **are proportionality constants**
  - The **authority score** of a page is proportional to the **sum of hub scores** of the pages linking to it
  - The **hub score** of a page is proportional to the **sum of authority scores** of the pages to which it links

HITS

$$a = \alpha \cdot A^T \cdot h \qquad h = \beta \cdot A \cdot a$$

- By **combining** both equations we arrive at:

$$a = \alpha\beta \cdot A^T A \cdot a \qquad h = \alpha\beta \cdot A A^T \cdot h$$

- As we see:
  - The authority vector a is an eigenvector of $A^T A$
  - The hub vector h is an eigenvector of $A A^T$
- Kleinberg decided to take the **principal eigenvectors** in each case, i.e. the eigenvectors corresponding to the eigenvalues with the **highest absolute values**
- Again, they can be computed using the **power iteration**

Example (query: japan elementary schools):

Hubs

- schools
- LINK Page-13
- ˆú–{‚ÌŠw□Z
- □a‰„□¬Šw□Zƒz□[ƒ□ƒy□[ƒW
- 100 Schools Home Pages (English)
- K-12 from Japan 10/...rnet and Education)
- http://www...iglobe.ne.jp/~IKESAN
- ‚l‚f‚j□¬Šw□Z‚U”N‚P‘g•¨Œê
- □ÒŠ—’¬—§□ÒŠ—“Œ□¬Šw□Z
- Koulutus ja oppilaitokset
- TOYODA HOMEPAGE
- Education
- Cay's Homepage(Japanese)
- –y“ì□¬Šw□Z‚Ìƒz□[ƒ□ƒy□[ƒW
- UNIVERSITY
- ‰J—³□¬Šw□Z DRAGON97-TOP
- □Â‰ª□¬Šw□Z‚T”N‚P‘gƒz□[ƒ□ƒy□[ƒW
- ¶µ°é¼ÂÁ© ¥á¥Ë¥å¡¼ ¥á¥Ë¥å¡¼

Authorities

- The American School in Japan
- The Link Page
- ‰ª□è□s—§ˆä“c□¬Šw□Zƒz□[ƒ□ƒy□[ƒW
- Kids' Space
- ˆÀ□é□s—§ˆÀ□é□¼•”□¬Šw□Z
- ‹{□é‹³ˆç‘åŠw•□‘®□¬Šw□Z
- KEIMEI GAKUEN Home Page (Japanese)
- Shiranuma Home Page
- fuzoku-es.fukui-u.ac.jp
- welcome to Miasa E&J school
- □_“Þ□ìŒ§□E‰¡•l□s—§‘†□ì□¼□¬Šw□Z‚Ìƒy
- http://www...p/~m_maru/index.html
- fukui haruyama-es HomePage
- Torisu primary school
- goo
- Yakumo Elementary,Hokkaido,Japan
- FUZOKU Home Page
- Kamishibun Elementary School...

HITS

- As PageRank, **HITS has been patented:**
  - US patent 6,112,202
  - "Method and system for identifying authoritative information resources in an environment with content-based links between information resources"
  - Inventor: Jon Kleinberg
  - **Assignee: IBM**

Connection to LSI/SVD

- There is a direct mapping between finding the **singular value decomposition** of A and finding an eigen-decomposition of A^TA and AA^T
- A short recap from Lecture 4:
  - Let $A = USV$ be the SVD of A
  - **Theorem:**
    U's columns are the **eigenvectors** of AA^T,
    the matrix S^2 contains the corresponding **eigenvalues**
  - Similarly, V's rows are the eigenvectors of A^TA,
    S^2 again contains the eigenvalues
- Therefore, HITS is equivalent to running the SVD on the adjacency matrix of the base set

Extensions

- If the query is ambiguous (e.g. “Java” or “jaguar”) or polarized (e.g. “abortion” or “cold fusion”), the base set will contain a few, almost disconnected, link communities
- Then, the principal eigenvectors found by HITS will reveal hubs and authorities in the largest link community
- One can tease of this structure by computing not only the principal eigenvectors but some more

HITS vs. PageRank

- PageRank can be precomputed,
  HITS has to be computed at query time
  - HITS is very expensive
- Different choices regarding the formal model
  - HITS models hubs and authorities
  - HITS uses a subset of the Web graph
  - But: We could also apply PageRank to a subset and HITS on the whole Web graph…
- On the Web, a good hub usually is also a good authority
- The difference between HITS and PageRank is not that large…

Next Lecture

- Spam detection
- Metasearch
- Privacy issues