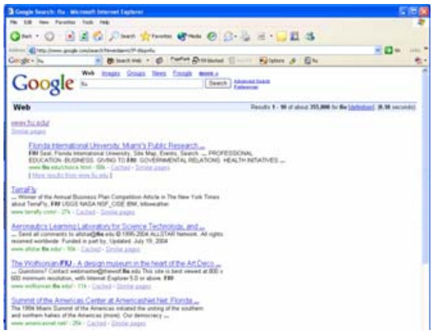
Link-based Web Search

Vagelis Hristidis
School of Computer Science
Florida International University
COP 6727

Roadmap

- Web Search
- PageRank
- HITS
- Stability Issues
- Current Research

Search the Web


Google Search: fiu - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Media
Google
Search Web
Options
Web Images Groups News Froogle more »
Search
Advanced Search
Preferences

Web Results 1 - 10 of about 355,000 for fiu [definition]. (0.18 seconds)

www.fiu.edu/
Similar pages

Florida International University: Miami's Public Research ...
FIU Seal, Florida International University, Site Map, Events, Search. ... PROFESSIONAL EDUCATION. BUSINESS. GIVING TO FIU. GOVERNMENTAL RELATIONS. HEALTH INITIATIVES ...
www.fiu.edu/choice.html - 59k - Cached - Similar pages
[More results from www.fiu.edu]

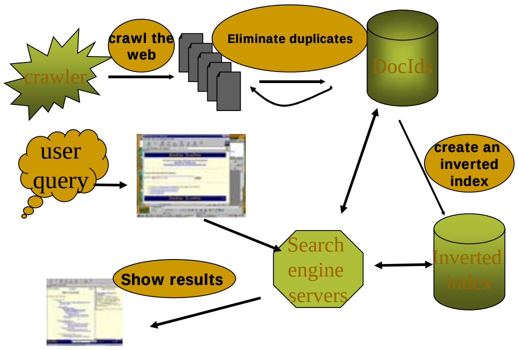
TerraFly
... Winner of the Annual Business Plan Competition Article in The New York Times about TerraFly. FIU USGS NASA NSF_CISE IBM, Infoweather.
www.terrafly.com/ - 27k - Cached - Similar pages

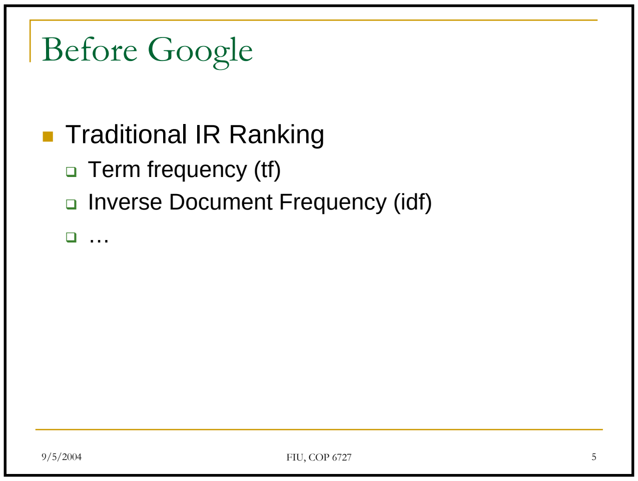
Aeronautics Learning Laboratory for Science Technology, and ...
... Send all comments to allstar@fiu.edu © 1995-2004 ALLSTAR Network. All rights reserved worldwide. Funded in part by. Updated: July 19, 2004
www.allstar.fiu.edu/ - 16k - Cached - Similar pages

The Wolfsonian-FIU - A design museum in the heart of the Art Deco ...
... Questions? Contact webmaster@thewolf.fiu.edu This site is best viewed at 800 x 600 minimum resolution, with Internet Explorer 5.0 or above. FIU
www.wolfsonian.fiu.edu/ - 11k - Cached - Similar pages

Summit of the Americas Center at AmericasNet.Net, Florida ...
The 1994 Miami Summit of the Americas initiated the uniting of the southern and northern halves of the Americas (more). Our democracy ...
www.americasnet.net/ - 25k - Cached - Similar pages

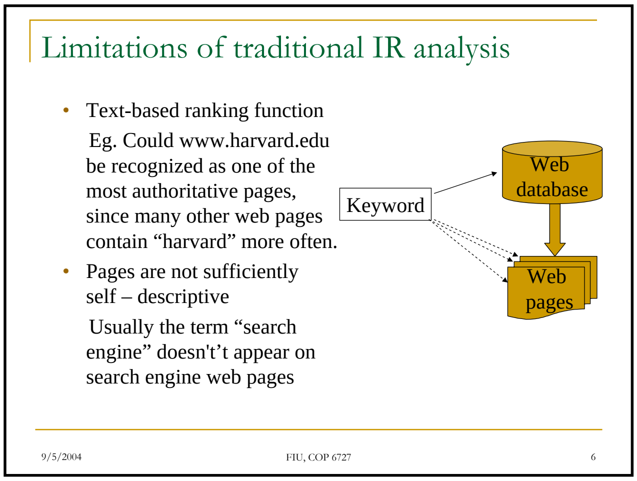
Standard Web Search Engine Architecture





Before Google

- Traditional IR Ranking
  - Term frequency (tf)
  - Inverse Document Frequency (idf)
  - …



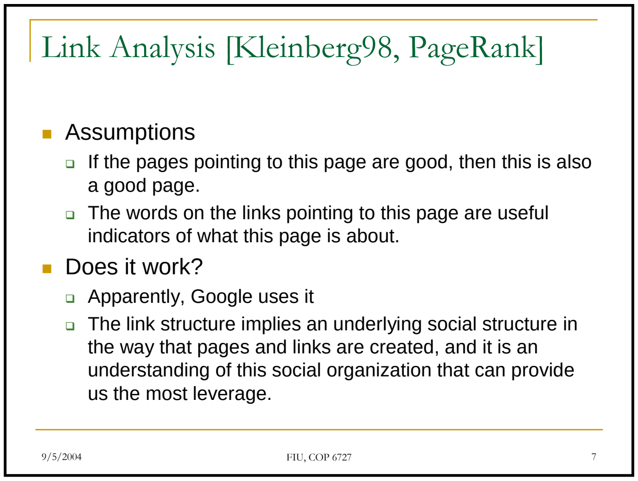
Limitations of traditional IR analysis

- Text-based ranking function

  Eg. Could www.harvard.edu be recognized as one of the most authoritative pages, since many other web pages contain “harvard” more often.
- Pages are not sufficiently self – descriptive

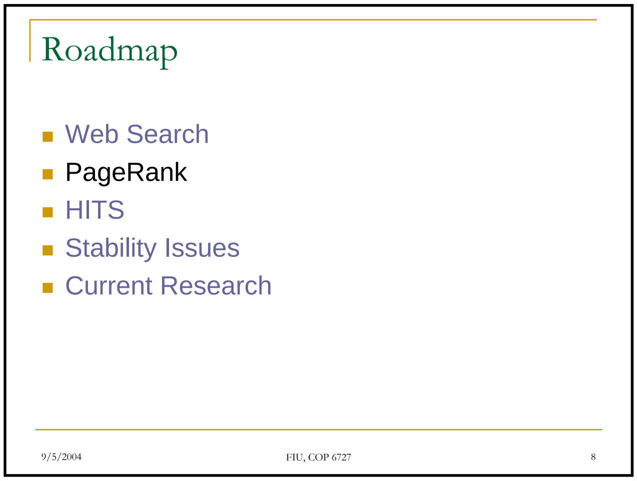
  Usually the term “search engine” doesn't't appear on search engine web pages

Keyword
Web database
Web pages



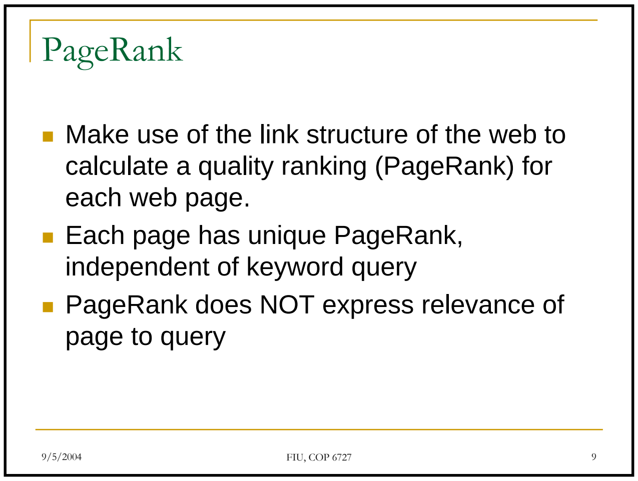
Link Analysis [Kleinberg98, PageRank]

- Assumptions
  - If the pages pointing to this page are good, then this is also a good page.
  - The words on the links pointing to this page are useful indicators of what this page is about.
- Does it work?
  - Apparently, Google uses it
  - The link structure implies an underlying social structure in the way that pages and links are created, and it is an understanding of this social organization that can provide us the most leverage.



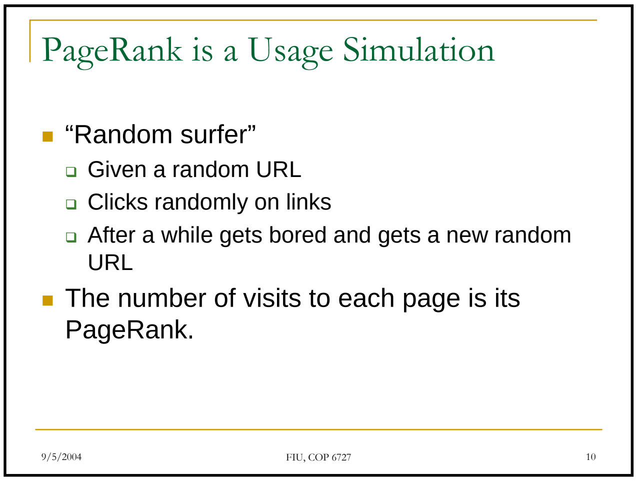
Roadmap

- Web Search
- PageRank
- HITS
- Stability Issues
- Current Research



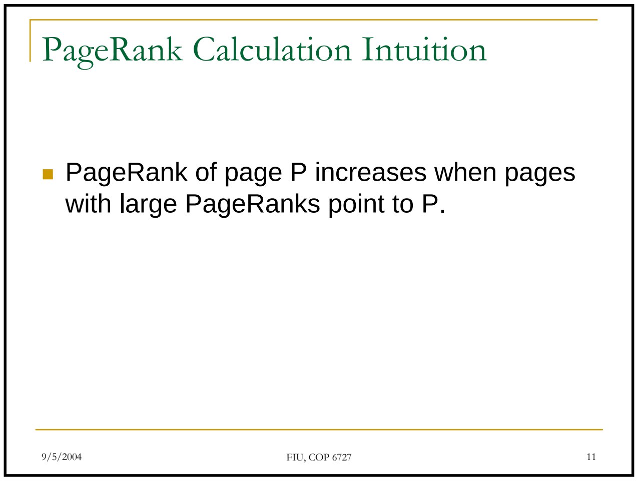
PageRank

- Make use of the link structure of the web to calculate a quality ranking (PageRank) for each web page.
- Each page has unique PageRank, independent of keyword query
- PageRank does NOT express relevance of page to query



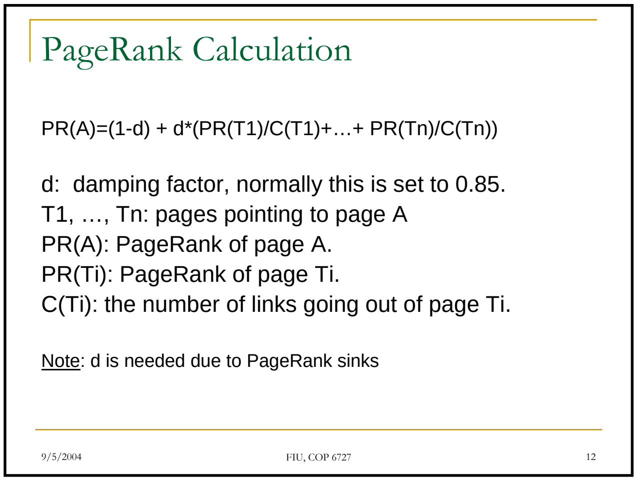
PageRank is a Usage Simulation

- “Random surfer”
  - ❑ Given a random URL
  - ❑ Clicks randomly on links
  - ❑ After a while gets bored and gets a new random URL
- The number of visits to each page is its PageRank.



PageRank Calculation Intuition

- PageRank of page P increases when pages with large PageRanks point to P.

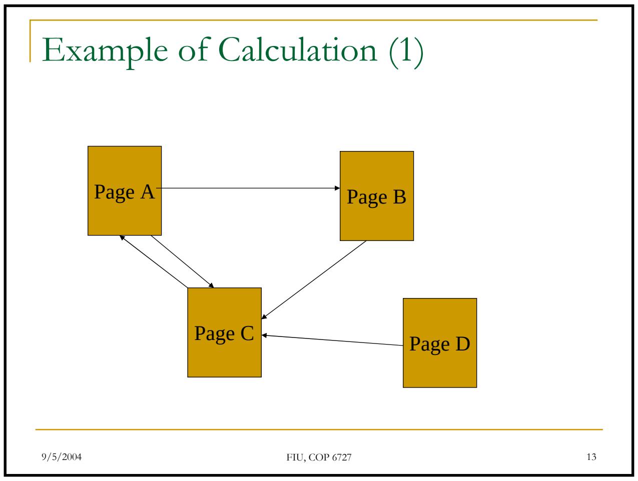


PageRank Calculation

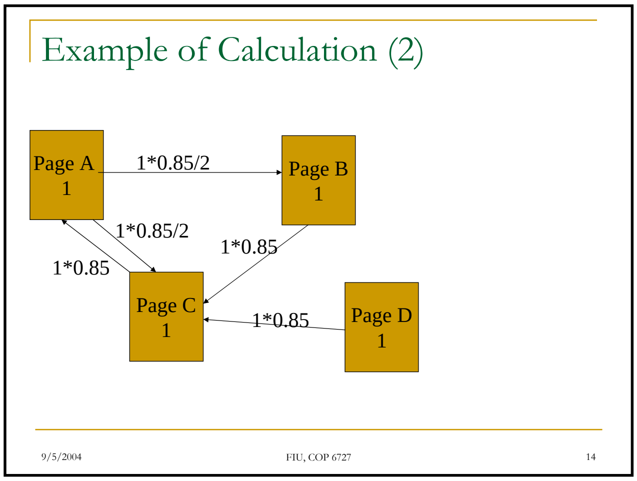
$PR(A)=(1-d) + d*(PR(T1)/C(T1)+\ldots+ PR(Tn)/C(Tn))$

d: damping factor, normally this is set to 0.85.
T1, …, Tn: pages pointing to page A
PR(A): PageRank of page A.
PR(Ti): PageRank of page Ti.
C(Ti): the number of links going out of page Ti.

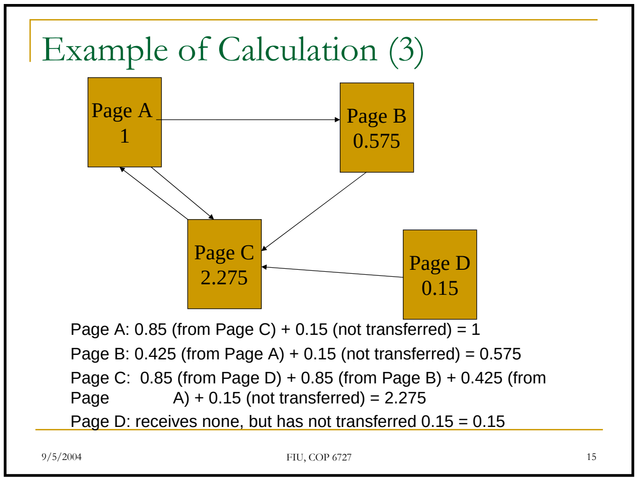
Note: d is needed due to PageRank sinks



Example of Calculation (1)



Example of Calculation (2)



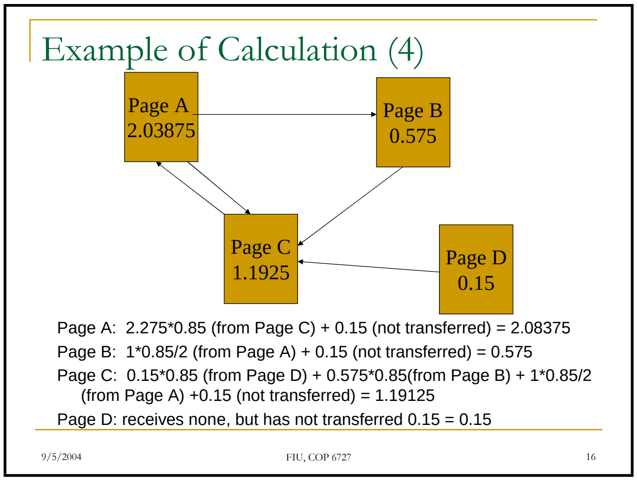
Example of Calculation (3)

Page A: 0.85 (from Page C) + 0.15 (not transferred) = 1

Page B: 0.425 (from Page A) + 0.15 (not transferred) = 0.575

Page C: 0.85 (from Page D) + 0.85 (from Page B) + 0.425 (from Page A) + 0.15 (not transferred) = 2.275

Page D: receives none, but has not transferred 0.15 = 0.15



Example of Calculation (4)

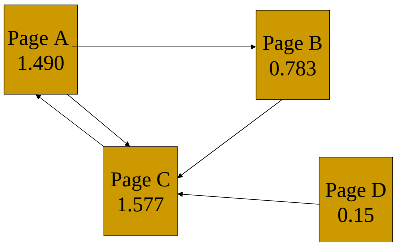
Page A: 2.275*0.85 (from Page C) + 0.15 (not transferred) = 2.08375

Page B: 1*0.85/2 (from Page A) + 0.15 (not transferred) = 0.575

Page C: 0.15*0.85 (from Page D) + 0.575*0.85(from Page B) + 1*0.85/2 (from Page A) +0.15 (not transferred) = 1.19125

Page D: receives none, but has not transferred 0.15 = 0.15

Example of Calculation (5)



- After 20 iterations it converges
- Converges because Web data graph irreducible (strongly connected) and aperiodic

Google

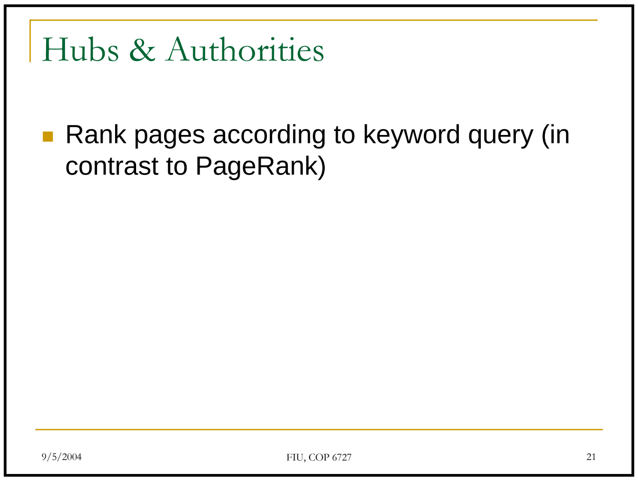
- Uses PageRank as one of the criteria to rank keyword query results.
- Other criteria (may) include:
  - Term frequencies
  - Term proximities
  - Term position (title, top of page, etc)
  - Term characteristics (boldface, capitalized, etc)
  - Link analysis information
  - Category information
  - Popularity information

Roadmap

- Web Search
- PageRank
- HITS
- Stability Issues
- Current Research

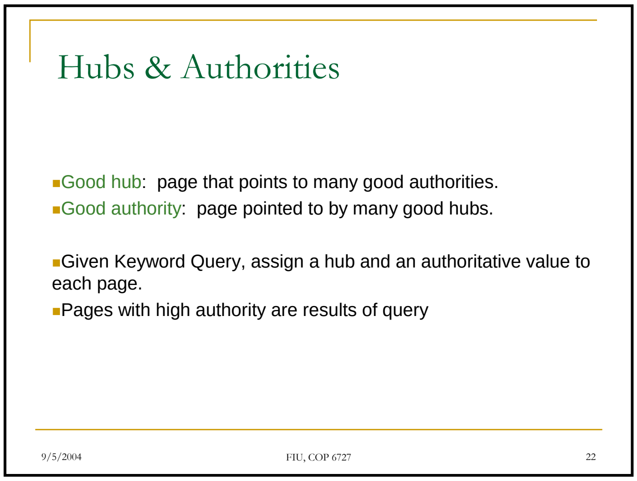
HITS [Kleinberg98]
Hubs & Authorities

- Jon M. Kleinberg: **Authoritative Sources in a Hyperlinked Environment**. JACM 46(5): 604-632 (1999)
- HITS (Hypertext-Induced Topic Search) developed by Jon Kleinberg, while visiting IBM Almaden.
- IBM expanded HITS into Clever.
- IBM doesn't see Clever as real-time search engine. But create constantly refreshed lists of relevant pages for categories



Hubs & Authorities

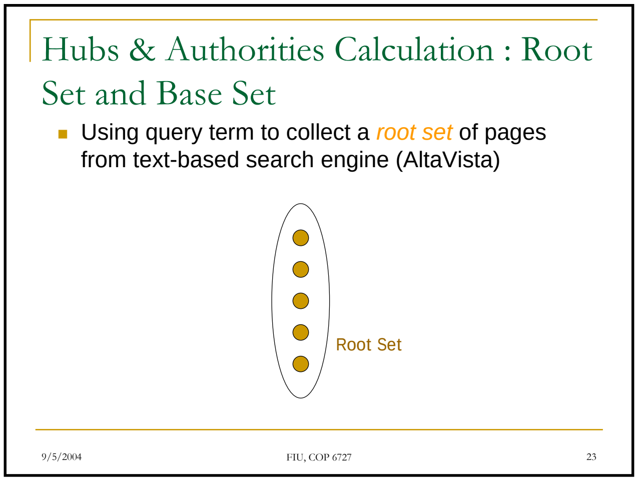
- Rank pages according to keyword query (in contrast to PageRank)



Hubs & Authorities

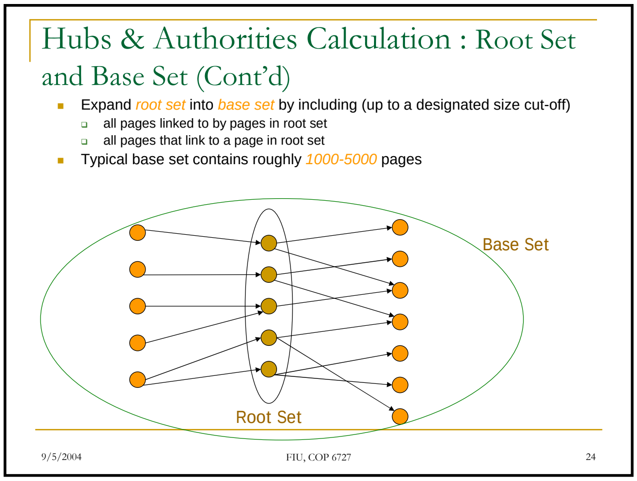
- Good hub: page that points to many good authorities.
- Good authority: page pointed to by many good hubs.

- Given Keyword Query, assign a hub and an authoritative value to each page.
- Pages with high authority are results of query



Hubs & Authorities Calculation : Root Set and Base Set

- Using query term to collect a *root set* of pages from text-based search engine (AltaVista)

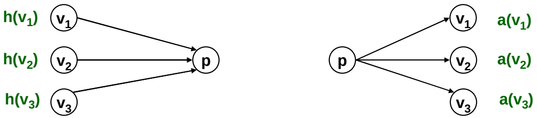


Hubs & Authorities Calculation : Root Set and Base Set (Cont'd)

- Expand *root set* into *base set* by including (up to a designated size cut-off)
  - all pages linked to by pages in root set
  - all pages that link to a page in root set
- Typical base set contains roughly *1000-5000* pages

Hubs & Authorities Calculation

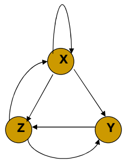
- Iterative algorithm on Base Set: authority weights *a*(p), and hub weights *h*(p).
  - Set authority weights *a*(p) = 1, and hub weights *h*(p) = 1 for all p.
  - Repeat following two operations
    (and then re-normalize *a* and *h* to have unit norm):



$$a(p) = \sum_{q \text{ points to } p} h(q) \qquad h(p) = \sum_{p \text{ points to } q} a(q)$$

Example: Mini Web

$$H = \begin{bmatrix} h_x \\ h_y \\ h_z \end{bmatrix} \quad A = \begin{bmatrix} a_x \\ a_y \\ a_z \end{bmatrix} \quad M = \begin{array}{c|ccc} & X & Y & Z \\ \hline X & 1 & 1 & 1 \\ Y & 0 & 0 & 1 \\ Z & 1 & 1 & 0 \end{array}$$



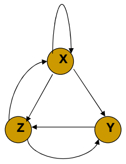
$$H_i = M * A_{i-1} \rightarrow H_i = M * M^T H_{i-1}$$

$$A_i = M^T * H_{i-1} \rightarrow A_i = M^T * M * A_{i-1}$$

Example

$$M = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} \quad M^T = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} \quad M\ M^T = \begin{bmatrix} 2 & 2 & 1 \\ 2 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix} \quad M^T M = \begin{bmatrix} 3 & 1 & 2 \\ 1 & 1 & 0 \\ 2 & 0 & 2 \end{bmatrix}$$

Iteration 0 1 2 3 ... ∞



$$H = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \to \begin{bmatrix} 6 \\ 2 \\ 4 \end{bmatrix} \to \begin{bmatrix} 28 \\ 8 \\ 20 \end{bmatrix} \to \begin{bmatrix} 132 \\ 36 \\ 96 \end{bmatrix} \longrightarrow \begin{bmatrix} \boxed{2+\sqrt{3}} \\ 1 \\ 1+\sqrt{3} \end{bmatrix}$$

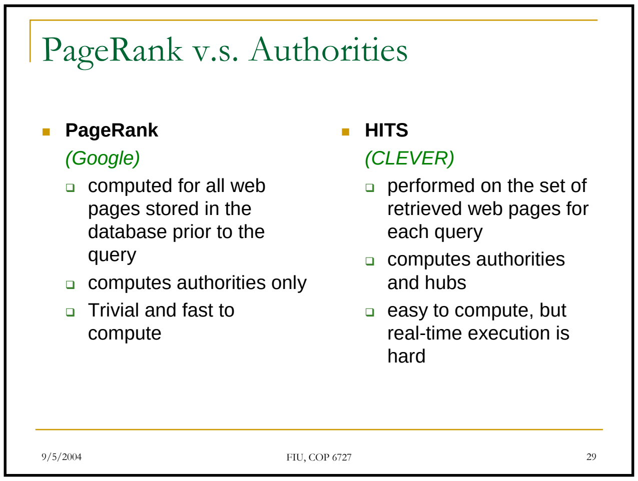
X is the best hub

$$A = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \to \begin{bmatrix} 5 \\ 5 \\ 4 \end{bmatrix} \to \begin{bmatrix} 24 \\ 24 \\ 18 \end{bmatrix} \to \begin{bmatrix} 114 \\ 114 \\ 84 \end{bmatrix} \longrightarrow \begin{bmatrix} 1+\sqrt{3} \\ 1+\sqrt{3} \\ \boxed{2} \end{bmatrix}$$

Z is most authoritative

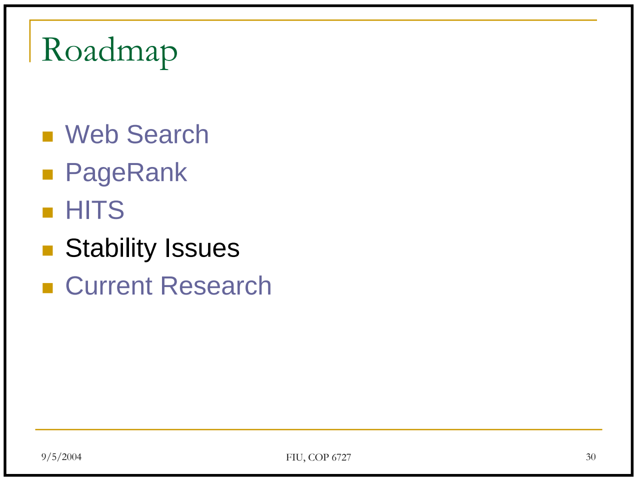
Hubs & Authorities Calculation

- Theorem (Kleinberg, 1998). The iterates a(p) and h(p) converge to the principal eigenvectors of M^TM and MM^T, where M is the adjacency matrix of the (directed) Web subgraph.



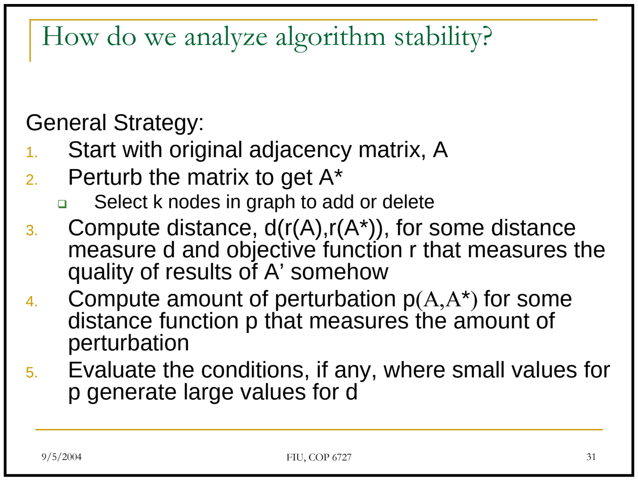
PageRank v.s. Authorities

- **PageRank**
  (Google)
  - computed for all web pages stored in the database prior to the query
  - computes authorities only
  - Trivial and fast to compute
- **HITS**
  (CLEVER)
  - performed on the set of retrieved web pages for each query
  - computes authorities and hubs
  - easy to compute, but real-time execution is hard



Roadmap

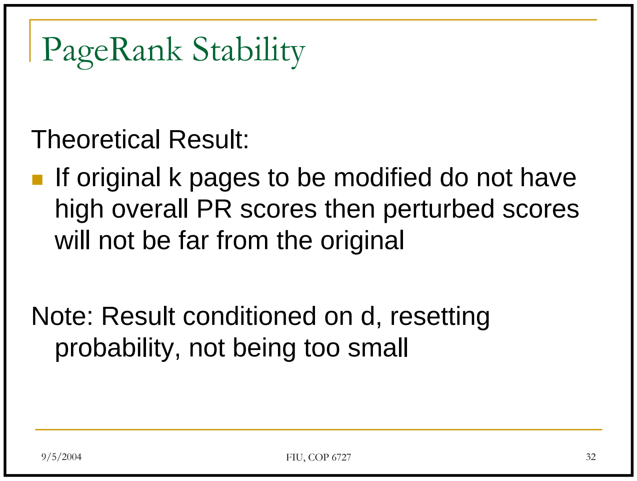
- Web Search
- PageRank
- HITS
- Stability Issues
- Current Research



How do we analyze algorithm stability?

General Strategy:

1. Start with original adjacency matrix, A
2. Perturb the matrix to get A*
   - Select k nodes in graph to add or delete
3. Compute distance, d(r(A),r(A*)), for some distance measure d and objective function r that measures the quality of results of A' somehow
4. Compute amount of perturbation p(A,A*) for some distance function p that measures the amount of perturbation
5. Evaluate the conditions, if any, where small values for p generate large values for d

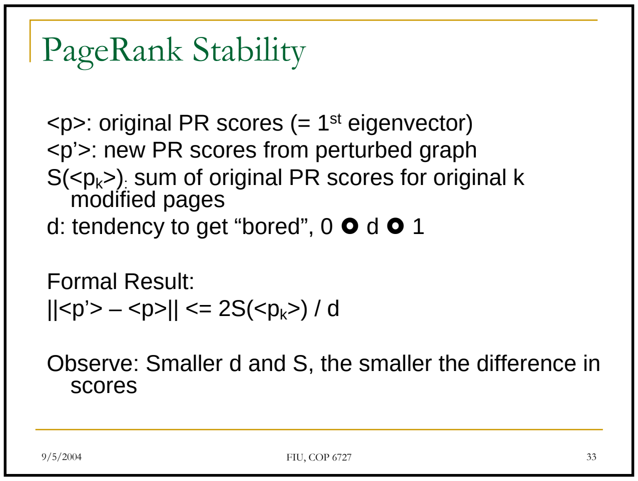


PageRank Stability

Theoretical Result:

- If original k pages to be modified do not have high overall PR scores then perturbed scores will not be far from the original

Note: Result conditioned on d, resetting probability, not being too small

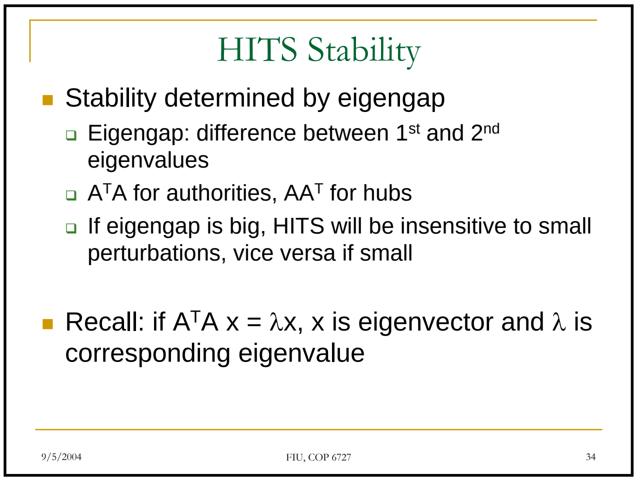


PageRank Stability

<p>: original PR scores (= 1st eigenvector)
<p’>: new PR scores from perturbed graph
S(<p_k>): sum of original PR scores for original k modified pages
d: tendency to get “bored”, 0 ≤ d ≤ 1

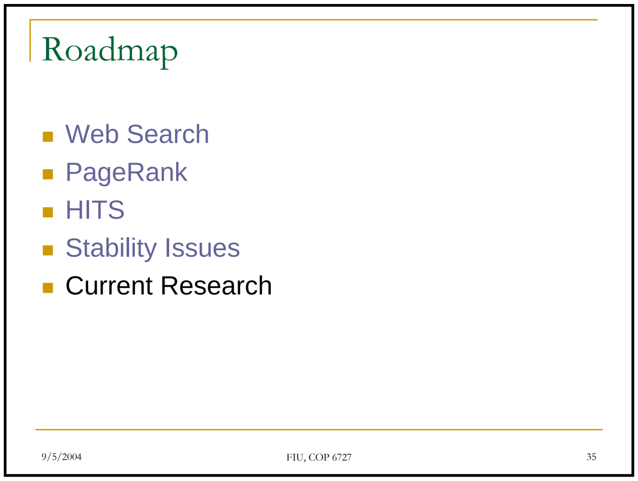
Formal Result:
||<p’> – <p>|| <= 2S(<p_k>) / d

Observe: Smaller d and S, the smaller the difference in scores



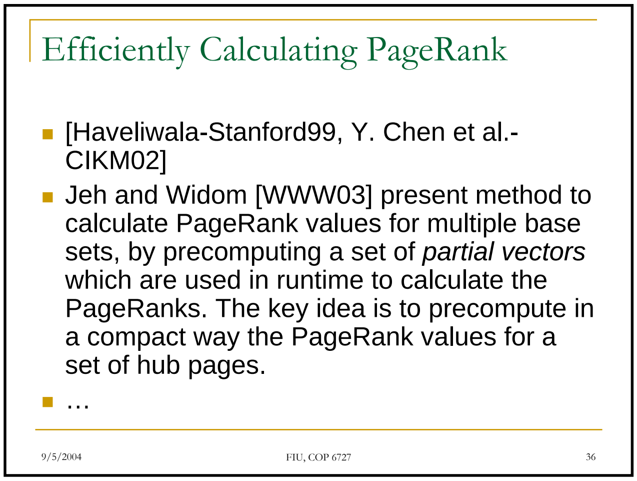
HITS Stability

- Stability determined by eigengap
  - Eigengap: difference between 1st and 2nd eigenvalues
  - A^TA for authorities, AA^T for hubs
  - If eigengap is big, HITS will be insensitive to small perturbations, vice versa if small
- Recall: if $A^TA\ x = \lambda x$, x is eigenvector and λ is corresponding eigenvalue



Roadmap

- Web Search
- PageRank
- HITS
- Stability Issues
- Current Research



Efficiently Calculating PageRank

- [Haveliwala-Stanford99, Y. Chen et al.-CIKM02]
- Jeh and Widom [WWW03] present method to calculate PageRank values for multiple base sets, by precomputing a set of *partial vectors* which are used in runtime to calculate the PageRanks. The key idea is to precompute in a compact way the PageRank values for a set of hub pages.
- …

Topic-Specific PageRank [Haveliwala-WWW02]

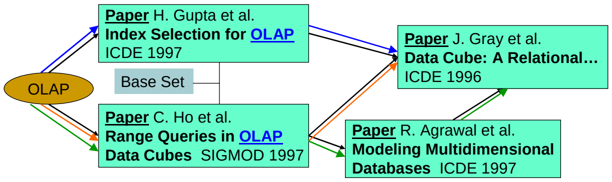
- topic-specific PageRanks for each page precomputed
- PageRank values of the most relevant topics used for each query.
- 16 topics

Personalized PageRank

- Favorites in Base Set
- Too Expensive!
- [WWW03] Linearity theorem

VH1

ObjectRank [VLDB2004]

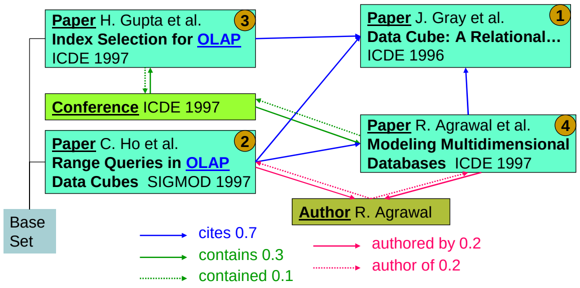

OLAP
Base Set
Paper H. Gupta et al. Index Selection for OLAP ICDE 1997
Paper C. Ho et al. Range Queries in OLAP Data Cubes SIGMOD 1997
Paper J. Gray et al. Data Cube: A Relational… ICDE 1996
Paper R. Agrawal et al. Modeling Multidimensional Databases ICDE 1997

- **ObjectRank** Ranks Objects According to Probability of Reaching Result Starting from Base Set

VH3

ObjectRank - Example

Keyword Query: [OLAP]


Paper H. Gupta et al. Index Selection for OLAP ICDE 1997
3
Paper J. Gray et al. Data Cube: A Relational… ICDE 1996
1
Conference ICDE 1997
Paper C. Ho et al. Range Queries in OLAP Data Cubes SIGMOD 1997
2
Paper R. Agrawal et al. Modeling Multidimensional Databases ICDE 1997
4
Author R. Agrawal
Base Set
cites 0.7
contains 0.3
contained 0.1
authored by 0.2
author of 0.2

Slide 39

VH1 [Proximity98-Goldman] Rank Objects According to Distance from Base Set

Drawback: Ignore Multiple Paths Between Result and Base Set

ObjectRank Ranks..., in a way similar to PageRank for the web, where the base set is... In contrast proximity works rank according to distance from base set.

for some
databases" authority-based and random walk-based search makes sense.
Clearly it is not applicable to all databases. For example for a database of cities with their temperatures there is no authority flow.
Vagelis, 2/22/2004

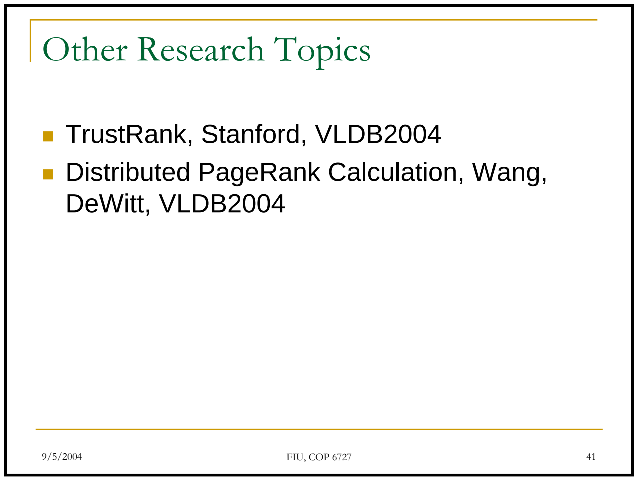
Slide 40

VH3 Database have edges of different types.
Different authority flows through various edges...
The authority transfer rates, which are shown at the bottom, show the maximum ratio of a node's authority transfered over edges of this type.
P->P edge has higher rate than the others because...

Another difference from the way that Web-search engines use PageRank is that we have keyword-specific ObjectRanks

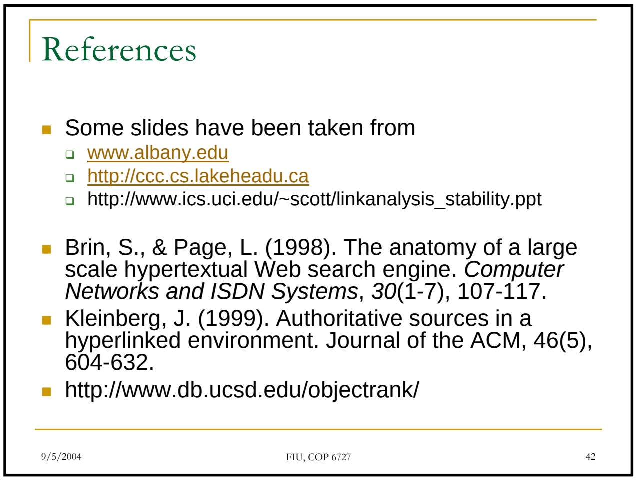
Now assume we have the keyword query OLAP...

In contrast to PageRank on the Web, we can do keyword specific ObjectRanks because (a) smaller size dbs and (b) exploit schema properties to optimize algorithm.
Vagelis, 3/2/2004



Other Research Topics

- TrustRank, Stanford, VLDB2004
- Distributed PageRank Calculation, Wang, DeWitt, VLDB2004



References

- Some slides have been taken from
  - www.albany.edu
  - http://ccc.cs.lakeheadu.ca
  - http://www.ics.uci.edu/~scott/linkanalysis_stability.ppt
- Brin, S., & Page, L. (1998). The anatomy of a large scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, *30*(1-7), 107-117.
- Kleinberg, J. (1999). Authoritative sources in a hyperlinked environment. Journal of the ACM, 46(5), 604-632.
- http://www.db.ucsd.edu/objectrank/