Introduction to link analysis & Temporal/Trend extensions of Pagerank

M. Vazirgiannis (mvazirg@aueb.gr)
http://db-net.aueb.gr/michalis

Introduction - Link Analysis

Based on slides from Mark Levene

Why link analysis?

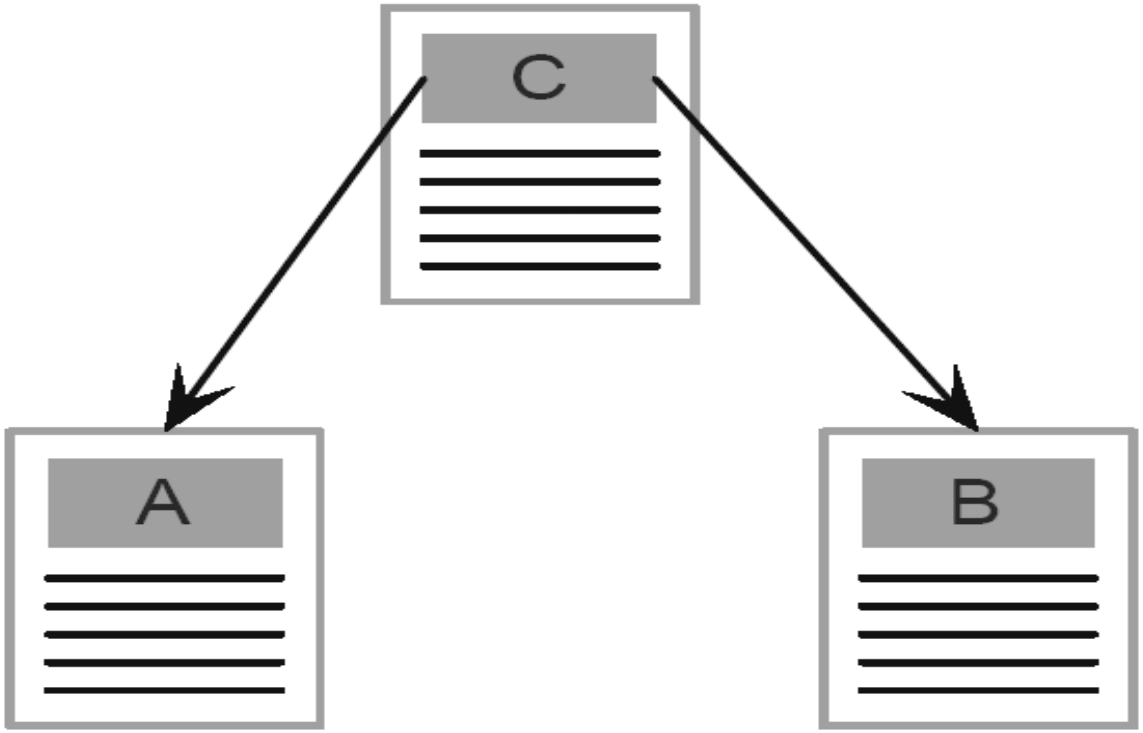
- The web is **not** just a collection of documents – its hyperlinks are important!
- A link from page *A* to page *B* may indicate:
  - *A* is related to *B*, or
  - *A* is recommending, citing or endorsing *B*
- Links are either
  - referential – *click here and get back home*, or
  - Informational – *click here to get more detail*

Citation Analysis

- The **impact factor** of a journal = *A/B*
  - *A* is the number of current year citations to articles appearing in the journal during previous two years.
  - *B* is the number of articles published in the journal during previous two years.

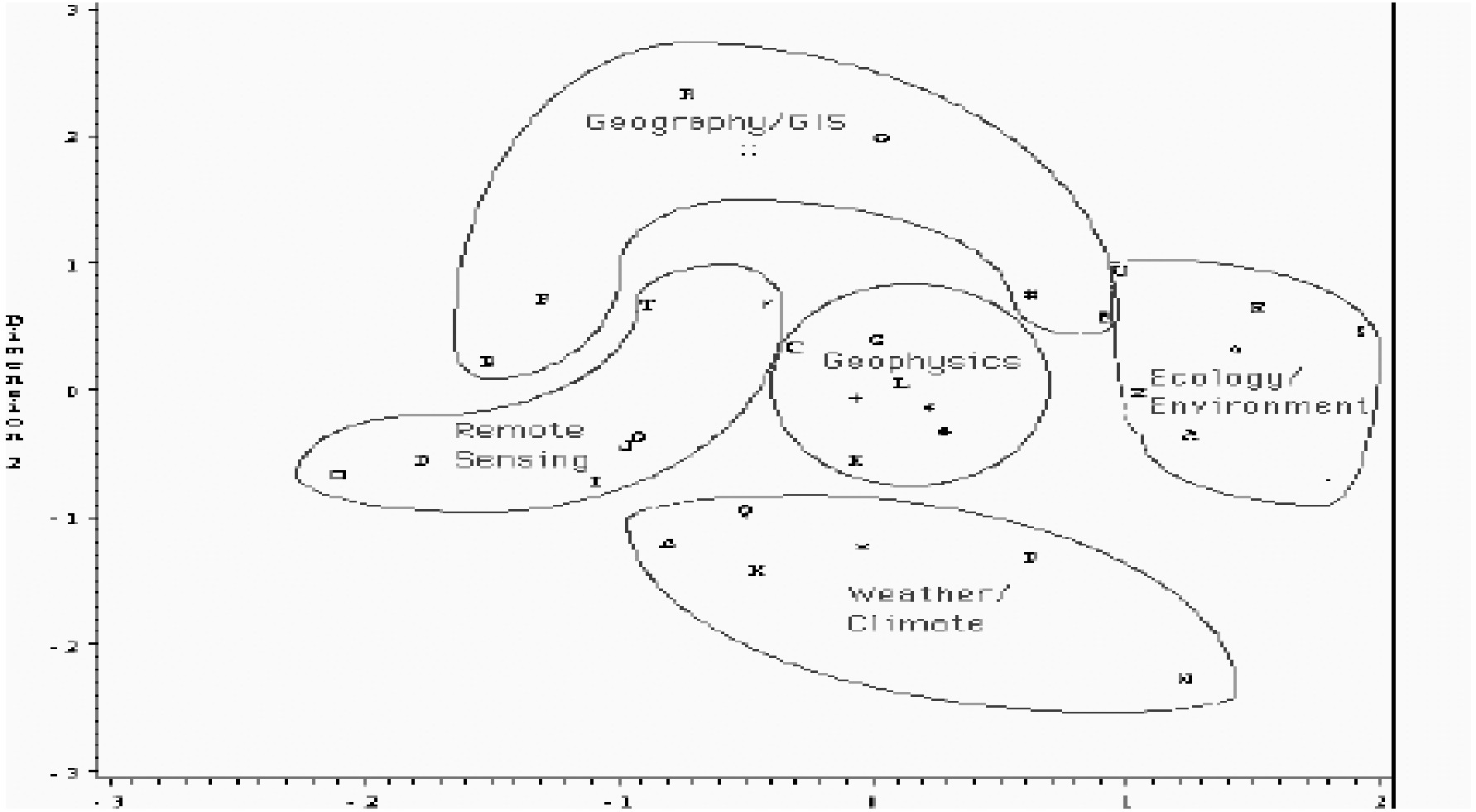
Journal Title	**Impact Factor (2002)**
J. Mach. Learn. Res.	3.818
IEEE T. Pattern Anal.	2.923
Mach. Learn.	1.944
IEEE Intell. Syst.	1.905
Artif. Intell.	1.703

Co-Citation



- *A* and *B* are co-cited by *C*, implying that
  - they are related or associated.
- The strength of co-citation between *A* and *B* is the number of times they are co-cited.

Clusters from Co-Citation Graph (Larson 96)



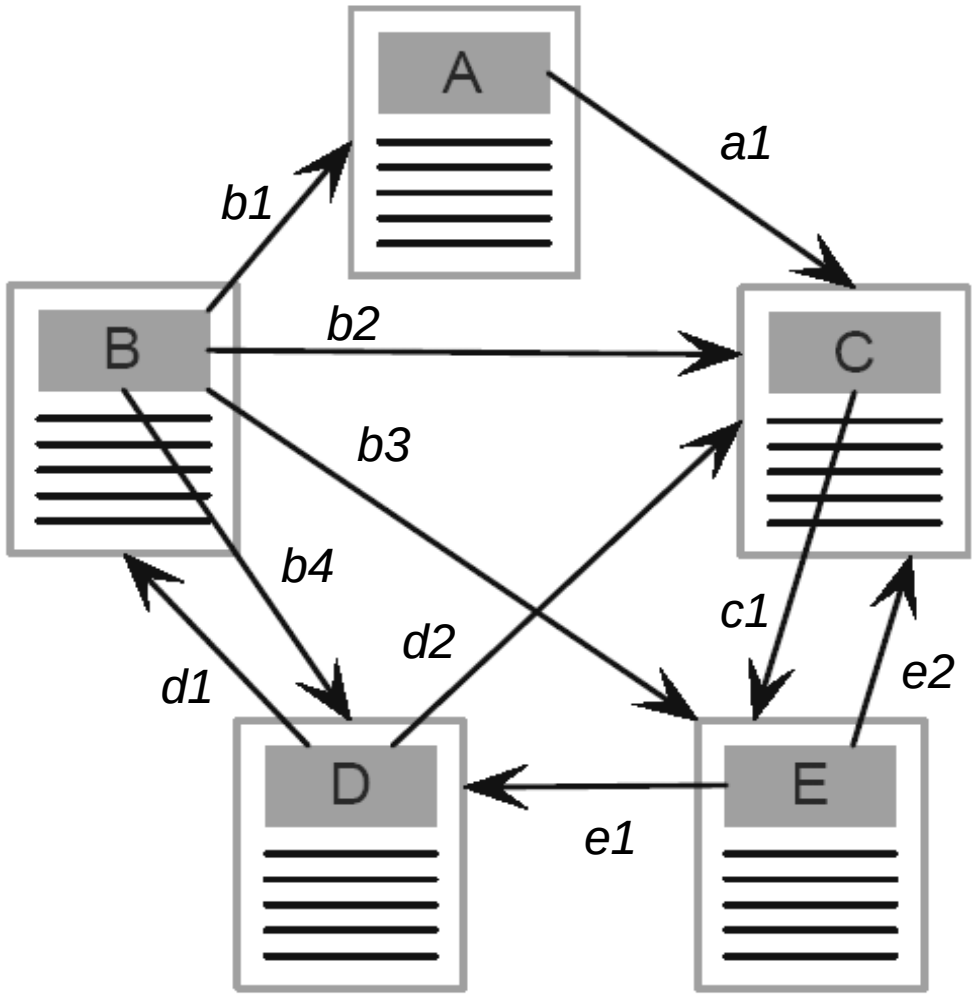
What is a Markov Chain?

- A Markov chain has two components:

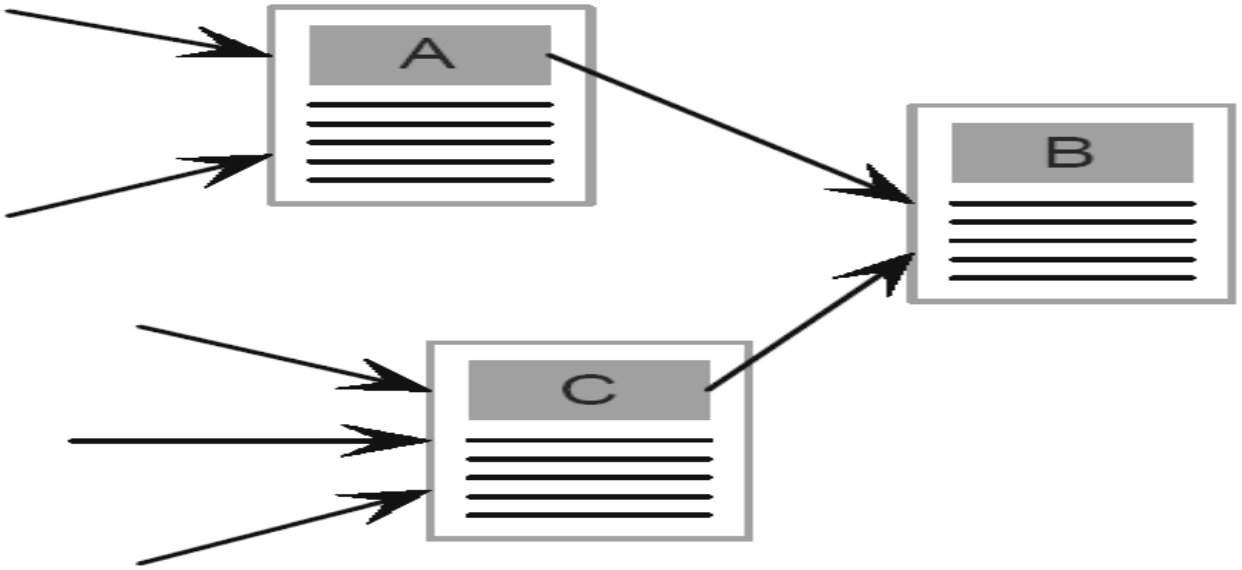
1) A network structure much like a web site, where each node is called a state.
2) A transition probability of traversing a link given that the chain is in a state.
   - For each state the sum of outgoing probabilities is one.

- A sequence of steps through the chain is called a *random walk*.

Markov Chain Example



PageRank - Motivation

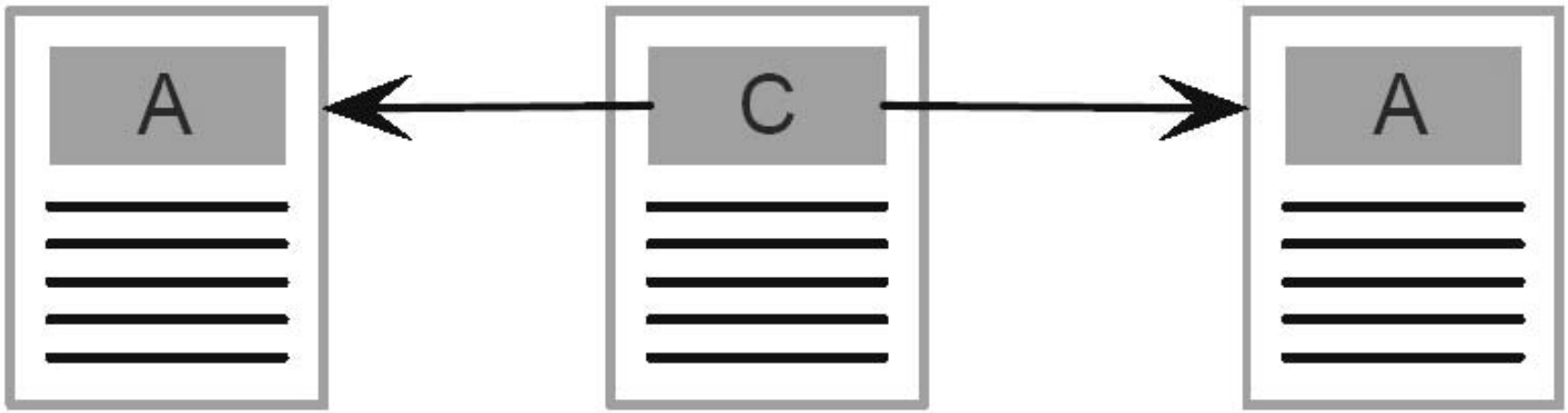


- A link from page *A* to page *B* is a **vote** of the author of *A* for *B,* or a **recommendation** of the page.
- The number incoming links to a page is a measure of importance and authority of the page.
- Also take into account the quality of recommendation, so a page is more important if the sources of its incoming links are important.

The Random Surfer

- Assume the web is a Markov chain.
- Surfers randomly click on links, where the probability of an outlink from page A is *1/m*, where *m* is the number of outlinks from *A.*
- The surfer occasionally gets *bored* and is *teleported* to another web page, say *B,* where *B* is equally likely to be any page.
- Using the theory of Markov chains it can be shown that if the surfer follows links for long enough, *the PageRank of a web page is the probability that the surfer will visit that page*.

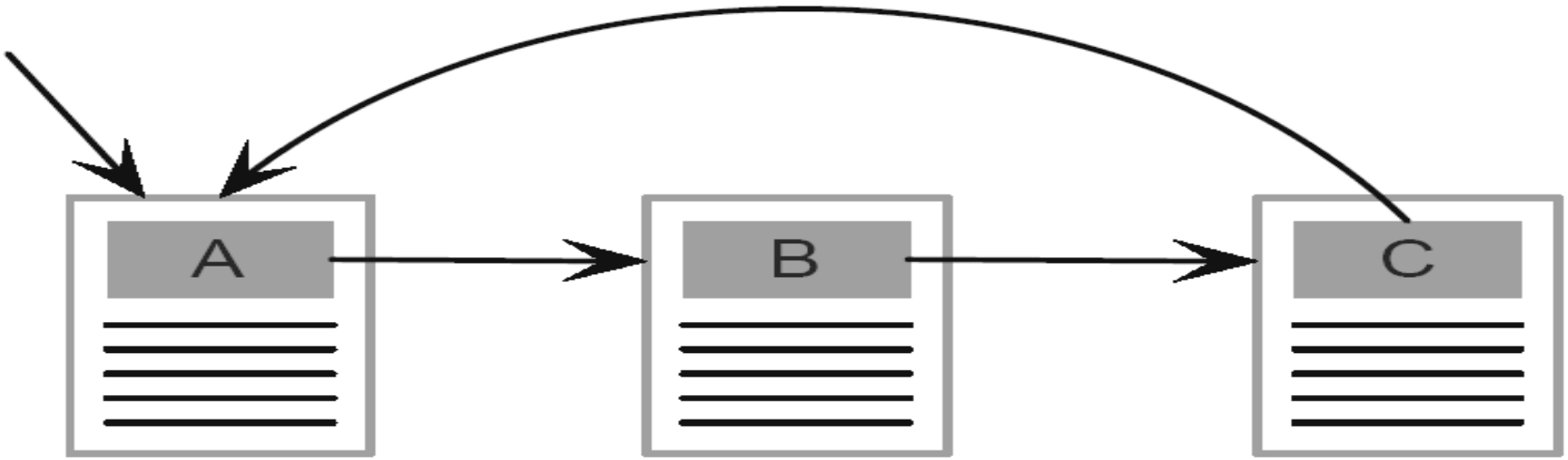
Dangling Pages



- Problem: *A* and *B* have no outlinks.

Solution: Assume *A* and *B* have links to all web pages with equal probability.

Rank Sink



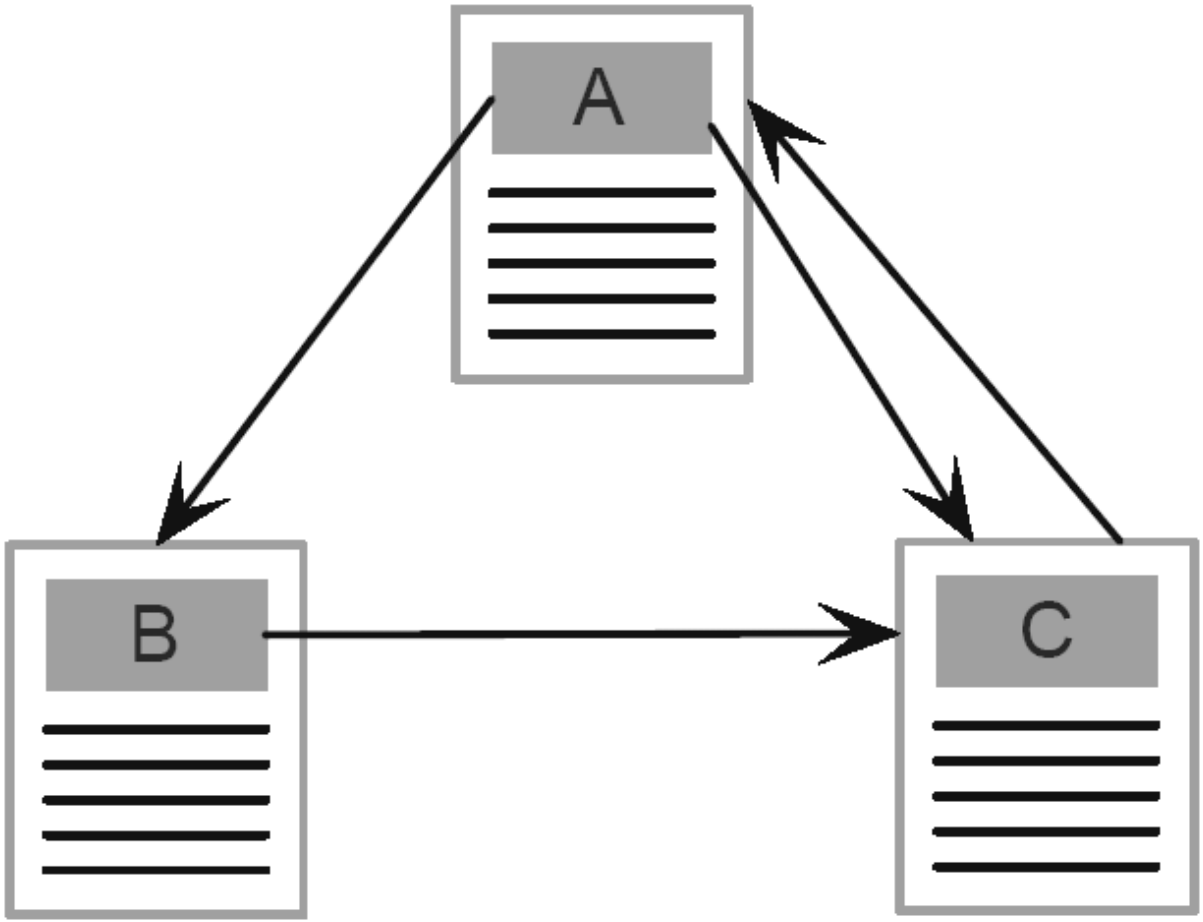
- Problem: Pages in a loop accumulate rank but do not distribute it.
- Solution: Teleportation, i.e. with a certain probability the surfer can jump to any other web page to get out of the loop.

PageRank (*PR*) - Definition

$$PR(P) = \frac{d}{N} + (1-d)\left(\frac{PR(P_1)}{O(P_1)} + \frac{PR(P_2)}{O(P_2)} + \ldots + \frac{PR(P_n)}{O(P_n)}\right)$$

- *P* is a web page
- *Pi* are the web pages that have a link *to P*
- *O*(*Pi*) is the number of outlinks from *Pi*
- *d* is the teleportation probability
- *N* is the size of the web

Example Web Graph



Iteratively Computing PageRank

- Replace *d/N* in the def. of *PR*(*P*) by *d, so PR* will take values between 1 and *N*.
- d is normally set to 0.15, but for simplicity lets set it to 0.5
- Set initial *PR* values to 1
- *Solve the following equations iteratively:*

$$PR(A) = 0.15/3 + 0.85PR(C)$$

$$PR(B) = 0.15/3 + 0.85(PR(A)/2)$$

$$PR(C) = 0.15/3 + 0.85(PR(A)/2 + PR(B))$$

Example Computation of PR

Iteration	PR(A)	PR(B)	PR(C)
0	1	1	1
1	1	0.75	1.125
2	1.0625	0.765625	1.1484375
3	1.07421875	0.76855469	1.15283203
4	1.07641602	0.76910400	1.15365601
5	1.07682800	0.76920700	1.15381050
…	…	…	…
12	1.07692308	0.76923077	1.15384615

The Largest Matrix Computation in the World

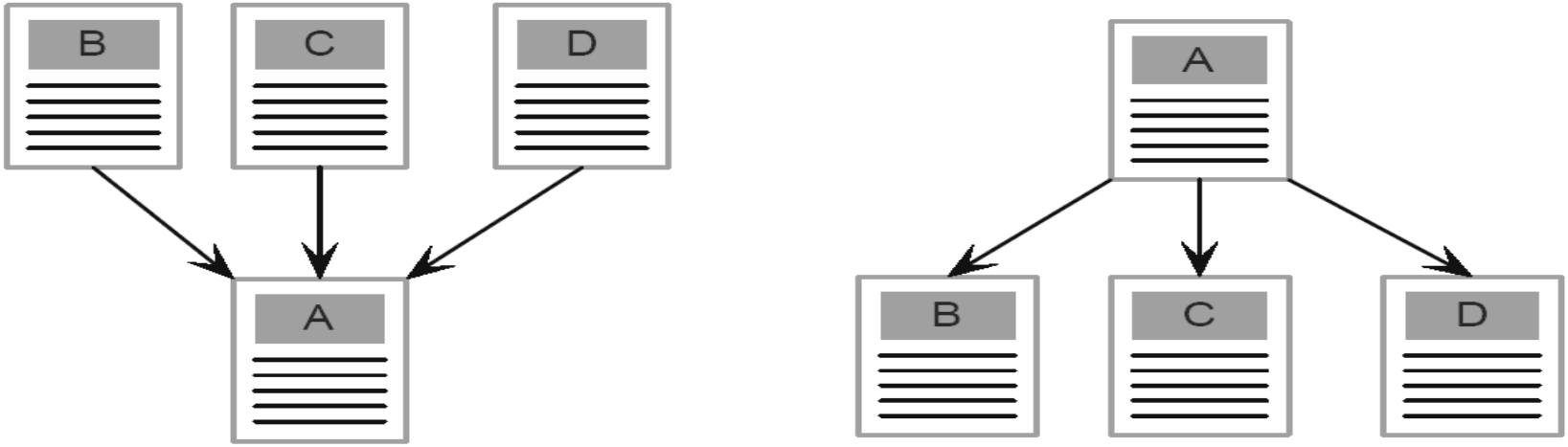
- Computing PageRank can be done via matrix multiplication, where the matrix has 3 billion rows and columns.
- The matrix is sparse as average number of outlinks is between 7 and 8.
- Setting d = 0.85 or below requires at most 100 iterations to convergence.
- Researchers still trying to speed-up the computation.

Personalised PageRank

$$PR(P) = dv + (1-d)\left(\frac{PR(P_1)}{O(P_1)} + \frac{PR(P_2)}{O(P_2)} + \ldots + \frac{PR(P_n)}{O(P_n)}\right)$$

- Change *d/N* with *dv*
- Instead of teleporting uniformly to any page we *bias* the jump prefer some pages over others.
  - E.g. *v* has 1 for your home page and 0 otherwise.
  - E.g. *v* prefers the topics you are interested in.

HITS – Hubs and Authorities - Hyperlink-Induced Topic Search



- **A** on the left is an **authority**
- **A** on the right is a **hub**

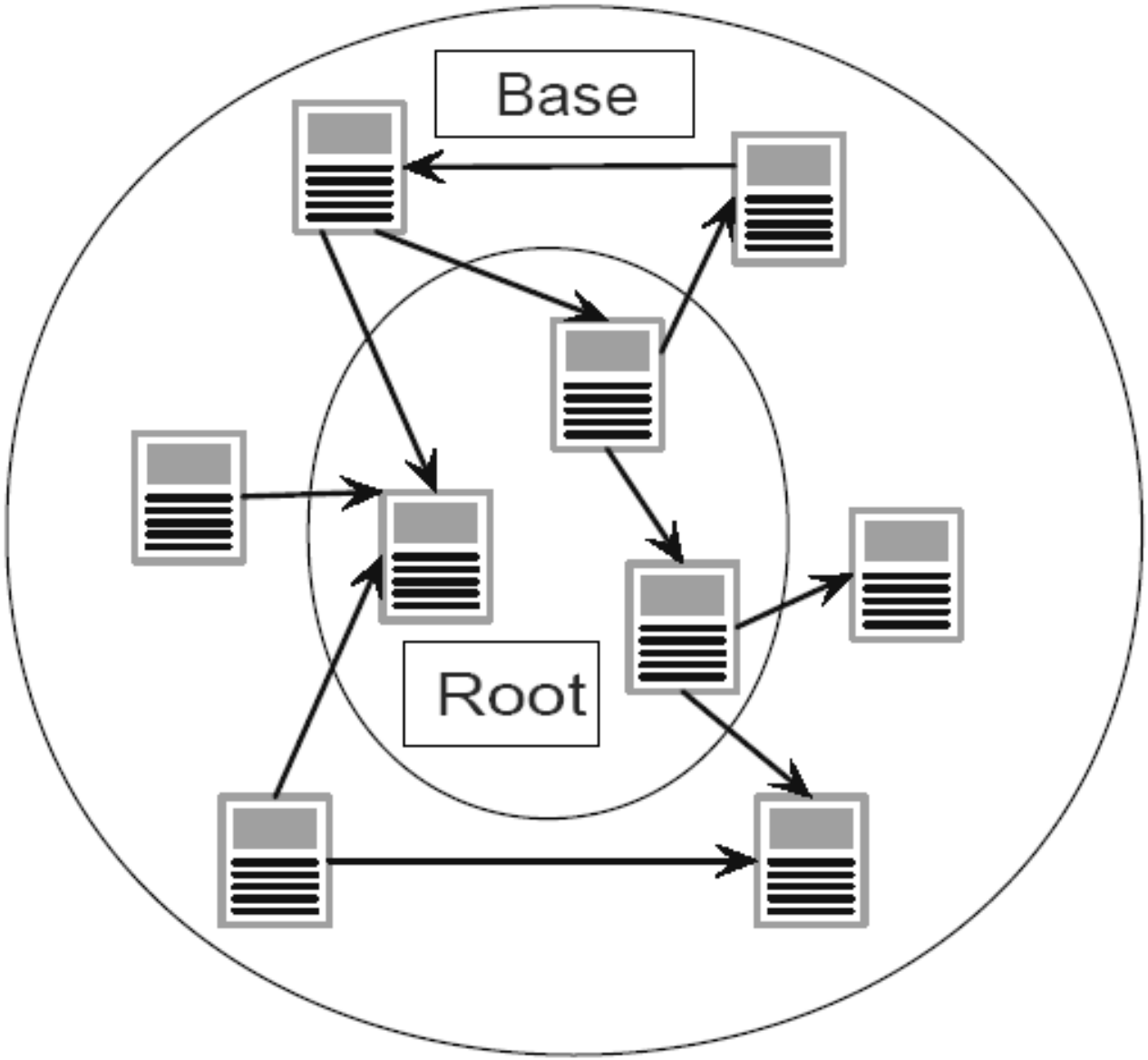
Communities on the Web

- A densely linked focused sub-graph of hubs and authorities is called a *community.*
- Over 100,000 emerging web communities have been discovered from a web crawl (a process called *trawling*).
- Alternatively, a community is a set of web pages *W* having at least as many links to pages in *W* as to pages outside *W*.

Pre-processing for HITS

1) Collect the top *t* pages (say t = 200) based on the input query; call this the **root set**.
2) Extend the root set into a **base set** as follows, for all pages *p* in the root set:
    1) add to the root set all pages that *p* points to, and
    2) add to the root set up-to *q* pages that point to *p* (say *q* = 50).
3) Delete all links within the same web site in the base set resulting in a **focused sub-graph**.

Expanding the Root Set



HITS Algorithm – Iterate until Convergence

$$A(p) = \sum_{q \in B | q \to p} H(q)$$

$$H(p) = \sum_{q \in B | p \to q} A(q)$$

- *B* is the base set
- *q* and *p* are web pages in *B*
- *A*(*p*) is the authority score for *p*
- *H*(*p*) is the hub score for *p*

Applications of HITS

- Search engine querying (speed an issue)
- Finding web communities.
- Finding related pages.
- Populating categories in web directories.
- Citation analysis.

Link Spamming to Improve PageRank

- Spam is the act of trying unfairly to gain a high ranking on a search engine for a web page without improving the user experience.
- *Link farms* - join the farm by copy a hub page which links to all members.
- *Selling links* from sites with high PageRank.

Temporal aspects - Motivation I

- The World Wide Web evolves at a high pace (25% new links, 8% new pages per week), therefore
  - rankings must be frequently recomputed, but still they do not always reflect the current authorities
- Availability of archived web content (e.g., the Internet Archive at www.archive.org)
  - creates a need for rankings with respect to time
  - allows tracing the evolution of pages and their authority
- Link-analysis techniques (e.g., PageRank, HITS) do not take into account the evolution and its associated temporal aspects, although
  - the users' interest has a temporal dimension
  - evolutionary data reflects current trends

Temporal aspects - Motivation II

- First objective: integration of temporal aspects (e.g., freshness, rate of change) into link-analysis techniques, to produce
  - rankings that better reflect the users' demand for recent information
  - rankings that reflect the authorities with respect to a temporal interest
- Second objective: a ranking based on the trends the pages' authority values exhibit with respect to time.
  - Ranking not by absolute authority, but by relative gain or loss of authority with respect to a temporal interest
  - Such a ranking should precisely reflect the importance with respect to a temporal interest taking into account only developments around that time

Temporal aspects - Basics II

- Time represented by **integers** (e.g., 20040701)
- Model of the **evolving graph** G(V,E)
  - Temporal annotations on nodes and edges
  - $TS_{Creation}$ refers to the moment of creation
  - $TS_{Deletion}$ refers to the moment of deletion (set to infinity while node still alive)
  - The set $TS_{Modifications}$ refers to the moments, when the node or edge was modified
  - $TS_{Lastmod}$ as a shortcut to the moment of the last modification (viz. max($TS_{Modifications}$))

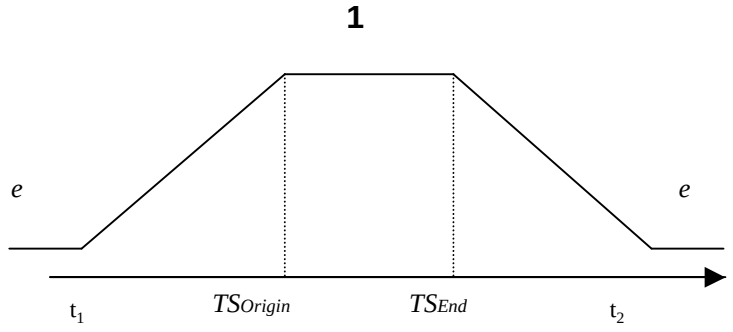
Temporal aspects - Basics III

- Concept of **temporal interest** defined by
  - A **time window** [*tsOrigin*,*tsEnd*]
  - A surrounding **tolerance interval** [*t1*,*t2*]
  - A smoothing parameter *e*
  - For the timestamps t1 <= tsOrigin <= tsEnd <= t2 must hold
- Graph $G_{ti}(V,E)$ contains all nodes and edges that exist at some point in the interval [*t1*,*t2*], that is whose timestamps fulfill:
  $TS_{Deletion} > t1 \wedge TS_{Creation} < t2$

Temporal aspects - Basics IV

- **Freshness** *f* measures the relevance of a timestamp *ts* with respect to a temporal interest

$$f(ts) = \begin{cases} if\ \ TS_{Origin} \le ts \le TS_{End}: & 1 \\ if\ \ t_1 \le ts < TS_{Origin}: & \dfrac{1}{(TS_{Origin} - ts) + 1} \\ if\ TS_{End} < ts \le t_2: & \dfrac{1}{(ts - TS_{End}) + 1} \\ otherwise: & e \end{cases}$$



- **Freshness** of **node *x***: *f(x) = f(TS$_{Lastmod}$ (x))*
- **Freshness** of **edge *x,y***: *f(x,y) = f(TS$_{Lastmod}$ (x,y))*

Temporal aspects - Basics V

- **Activity** *a* measures the frequency of change expressed by a set of timestamps *TS* with respect to a temporal interest

$$a(TS) = \begin{cases} if\ \ \mathrm{TS} \neq \varnothing: & \sum_{t1}^{t2} \{ f(ts) | ts \in TS \} \\ otherwise: & e \end{cases}$$

- **Activity** of **node *x***: $a(x) = a(TS_{Modifications}(x))$
- **Activity** of **edge *x,y***: $a(x,y) = a(TS_{Modifications}(x,y))$

T-Rank I

- **Objective** is a ranking of nodes according to the authority with *respect to the temporal interest*
- **Modified PageRank** on graph $G_{ti}(V,E)$
  - Transition probabilities *t(x,y)* depend on
    - Freshness of the node *y*
    - Freshness of the edge *x,y*
    - Freshness of the incoming edges of *y*
  - Random jump probabilities *s(y)* depend on
    - Freshness of node *y*
    - Activity of node *y*
    - Freshness of the incoming edges of *y*
    - Activity of the incoming edges of *y*

T-Rank II

$$t(x,y) = w_{t1} \cdot \frac{f(y)}{\sum_{(x,z)\in E} f(z)} + w_{t2} \cdot \frac{f(x,y)}{\sum_{(x,z)\in E} f(x,z)} + w_{t3} \cdot \frac{avg\{f(v,y \mid (v,y) \in E\}}{\sum_{(x,w)\in E} avg\{f(v,w \mid (v,w) \in E\}}$$

- **Transition probabilities** as weighted sum with coefficients wt_i that must add up to 1
- Before making a transition, the random surfer rolls a three-sided dice with probability distribution according to the wt_i. Seeing the
  - 1st side the edge *x,y* is followed with probability proportional to the freshness of the node *y*
  - 2nd side the edge *x,y* is followed with probability proportional to the freshness of the edge *x,y*
  - 3rd side the edge *x,y* is followed with probability proportional to the average freshness of the incoming edges of node *y*

T-Rank III

$$s(y) = w_{s1} \cdot \frac{f(y)}{\sum_{z \in V} f(z)} + w_{s2} \cdot \frac{a(y)}{\sum_{z \in V} a(z)} +$$

$$w_{s3} \cdot \frac{avg\{f(v, y \mid (v, y) \in E\}}{\sum_{z \in V} avg\{f(w, z \mid (w, z) \in E\}} + w_{s4} \cdot \frac{avg\{a(v, y \mid (v, y) \in E\}}{\sum_{z \in V} avg\{a(w, z \mid (w, z) \in E\}}$$

- **Random jump probabilities** with tunable parameters ws_i that must add up to 1
- In case of a random jump, a four-sided dice is rolled with probability distribution according to the ws_i. Seeing the
  - 1st side node *y* chosen with probability proportional to *f(y)*
  - 2nd side node y chosen with probability proportional to *a(y)*
  - 3rd side node *y* chosen with probability proportional to average freshness of the incoming edges of node *y*
  - 4th side node y chosen with probability proportional to average activity of the incoming edges of node *y*

E-Rank I

- **Objective** is a ranking by the *emerging authority*, that is not by the absolute authority but by *the trend the authority of a web page shows with respect to a temporal interest*
- **Idea** is to base the ranking only on things that happened in the period of interest
- With respect to the temporal [*t1,t2*] interval, we distinguish
  - the set *Nt consisting* of links created in the interval
  - the set *Dt consisting* of links deleted in the interval
  - the set *Mt consisting* of links modified in the interval
- Link-analysis techniques on the web commonly assume, that *links embody recommendations and transfer credit*

E-Rank II

- We extend this idea and assume, that
  - links modified within [*t1,t2*] transfer credit
  - links created within [*t1,t2*] transfer credit
  - links deleted within [*t1,t2*] transfer discredit (withdraw formerly given credit)

  with respect to the temporal interest
- Random walk based on these ideas:
  - new and modified links followed in regular direction with probability depending on their freshness and inverse indegree of the target page
  - deleted links followed in reversed direction with probability depending on their freshness and inverse outdegree of the soure page
  - still with a probability ε a random jump is performed

E-Rank III

$$t(x,y) = w_{e1} \cdot \frac{Nt(x,y)\, f(TScreation(x,y))}{\ln(indegree(y, TScreation(x,y)) + c)} \cdot \Big(\sum_{(x,z)\in E} \frac{Nt(x,z)\, f(TScreation(x,z))}{\ln(indegree(z, TScreation(x,z)) + c)} \Big)^{-1}$$

$$+ w_{e2} \cdot \frac{Dt(y,x)\, f(TSdeletion(y,x))}{\ln(outdegree(y, TSdeletion(y,x)) + c)} \cdot \Big(\sum_{(z,x)\in E} \frac{Dt(z,x)\, f(TScreation(z,x))}{\ln(outdegree(z, TScreation(z,x)) + c)} \Big)^{-1}$$

$$+ w_{e3} \cdot \frac{Mt(x,y)\, f(TSlastmod(x,y))}{\ln(indegree(y, TSlastmod(x,y)) + c)} \cdot \Big(\sum_{(x,z)\in E} \frac{Mt(x,z)\, f(TSlastmod(x,z))}{\ln(indegree(z, TSlastmod(x,z)) + c)} \Big)^{-1}$$

- **Transition probabilities** *t(x,y)* with tunable parameters we_i
- Indicator function *Nt(x,y)*, *Dt(x,y)* and *Mt(x,y)* represent the sets of *Nt, Dt* and *Mt*
- Natural logarithm dampens in-/outdegree values
- Constant *c* needed to guarantee non-zero denominator
- **Random jump probabilities** *s(y)* are defined uniformly

Implementation

- **Java** Implementation (J2SE 1.4.3)
- **Oracle 9i** used for storage of data
- **Bingo!** focused crawler collects the web data
- Evolving graph stored in **database relations** that do neither depend on Bingo! nor on application on the web graph
- **Multi-threaded** implementationof the **Power Method** based on a Compressed Row Storage (**CRS**) datastructure tailored to the problem

Experiments – DBLP I

- **Digital Bibliography & Library Project** (DBLP) freely available bibliographic dataset (as XML)
- **Evolving graph** derived from DBLP
  - Authors as nodes, citations as edges
  - ~350K (~16K) nodes, ~350K edges
- ***T-Rank*** and ***PageRank*** applied for temporal interests on **decades** (70s to 00s)

Experiments – DBLP II

	***PageRank** 2000s*	***T-Rank** 2000s*
1	E. F. Codd	Jim Gray
2	Michael Stonebraker	Michael Stonebraker
3	Jim Gray	Jeffrey D. Ullman
4	Donald D. Chamberlin	Philip A. Bernstein
5	Jeffrey D. Ullman	Hector Garcia-Molina
6	Philip A. Bernstein	Jeffrey F. Naughton
7	Raymond A. Lorie	Donald D. Chamberlin
8	Morton M. Astrahan	David J. DeWitt
9	Kapali P. Eswaran	Jennifer Widom
10	John Miles Smith	Rakesh Agrawal

Experiments – DBLP III

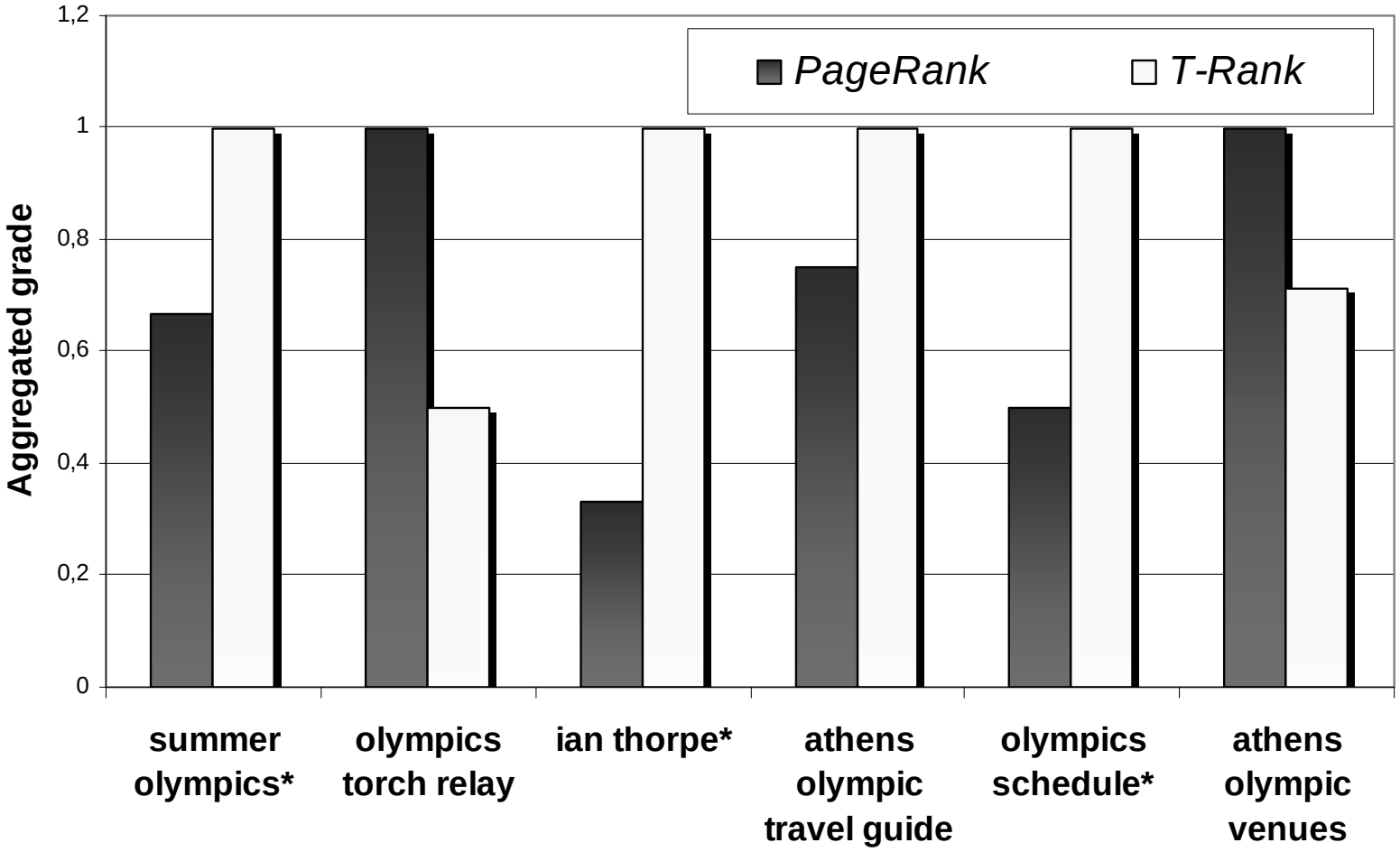
	1970s		1980s	
	PageRank	E-Rank	PageRank	E-Rank
1	E. Codd	E. Codd	E. Codd	E. Codd
2	D. Chamberlin	D. Chamberlin	M. Stonebraker	M. Stonebraker
3	J. Smith	P. Chen	D. Chamberlin	J. Ullman
4	P. Bernstein	J. Smith	R. Lorie	R. Lorie
5	M. Stonebraker	R. Boyce	P. Bernstein	P. Chen

	1990s		2000s	
	PageRank	E-Rank	PageRank	E-Rank
1	E. Codd	M. Stonebraker	E. Codd	A. Eisenberg
2	M. Stonebraker	J. Gray	M. Stonebraker	H. Garcia-Molina
3	J. Gray	J. Ullman	J. Gray	J. Gray
4	D. Chamberlin	D. DeWitt	D. Chamberlin	J. Ullman
5	J. Ullman	H. Garcia-Molina	J. Ullman	M. Stonebraker

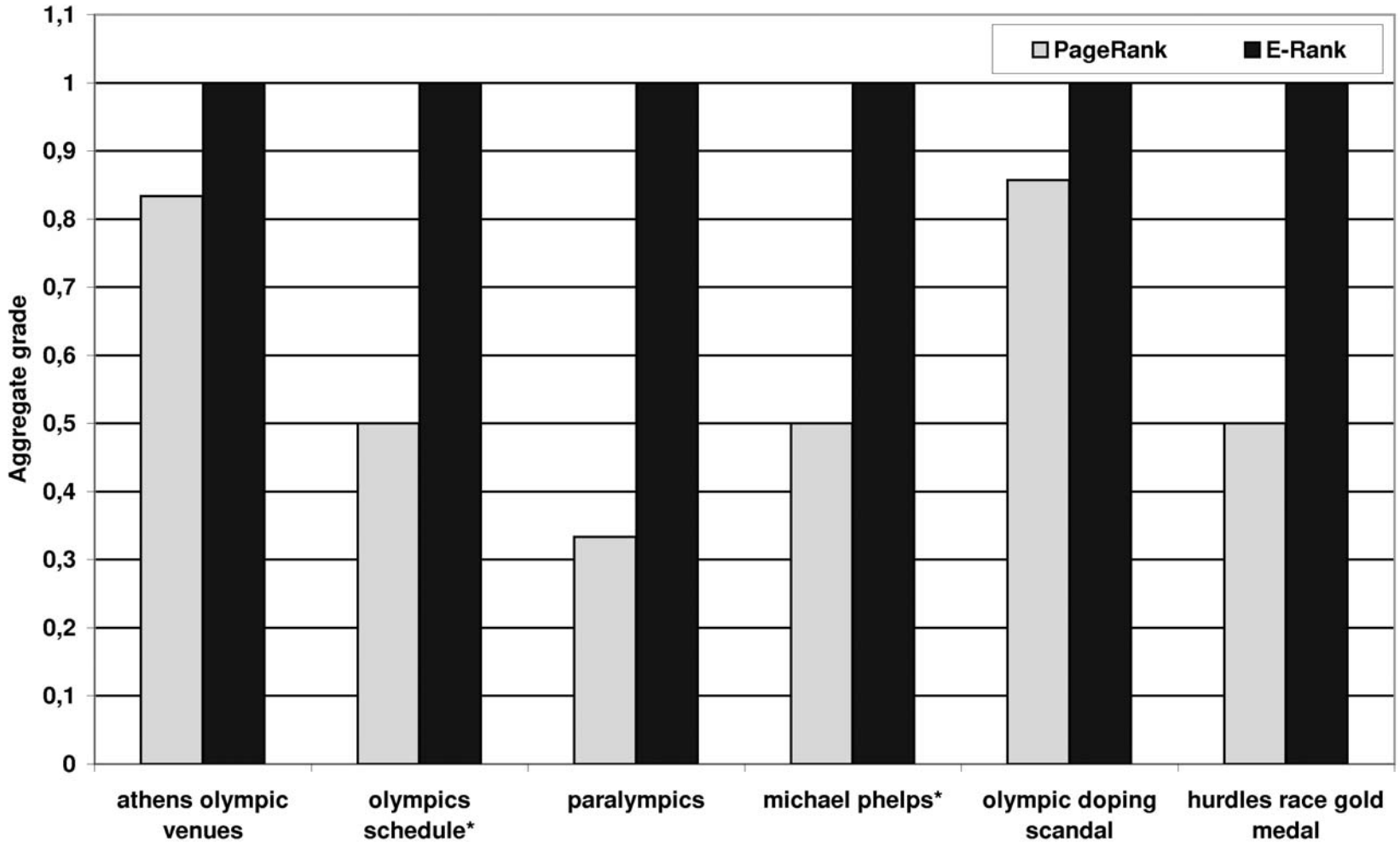
Experiments – Web I

- **Olympic Games 2004**
  - ~200K thematically related Web pages
  - 9 crawls in period July 26th to September 1st
- **Blind test** comparing *PageRank* and *T-Rank*
  - Users asked to **grade quality** of given top-*10* lists
  - Half of the queries drawn from Google Zeitgeist

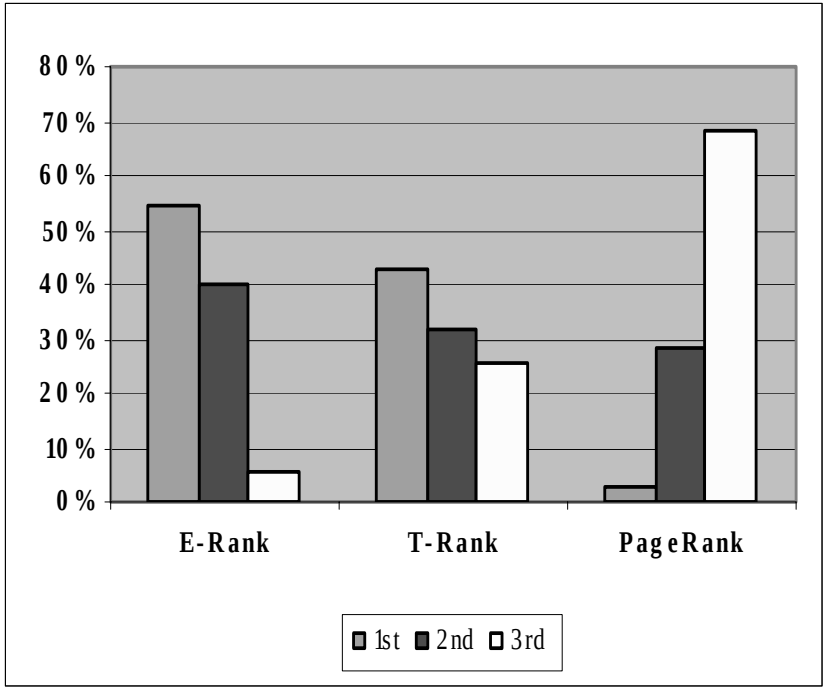
Experiments – Web II



Experiments – Web III



Experiments IV



- E-Rank and T-Rank were chosen in more than 90% as the best ranking
- PageRank was chosen as the worst ranking in >68% of the user assessments

Outlook

- Experiments based on ‘real’ web data
- Approximating variants of E-Rank and T-Rank using only a skewed random jump probabilities
- Closer investigation of E-Rank properties
  - possible advantages due to lower amount of data, that is used (stall edges are neglected)
  - can we approximate emerging authority comparing multiple static authority rankings for different times?

Summary

- Integration of temporal aspects into link-analysis
  - ameliorates rankings
  - gives rankings that do reflect authority with respect to a temporal interest
- Experiments have shown, that promising results can be obtained taking into account the trends exhibited by the link structure

Relevant Publications

• K. Berberich, M. Vazirgiannis, and G. Weikum.T-Rank: Time-aware Authority Ranking. In S. Leonardi, editor, Algorithms and Models for the Web-Graph: Third International Workshop, WAW 2004, pages 131–141. Springer-Verlag, 2004.

• K. Berberich, M. Vazirgiannis, and G. Weikum. “E-Rank: what is new and important on the web”, submitted for publication.