Link Analysis

Why link analysis?

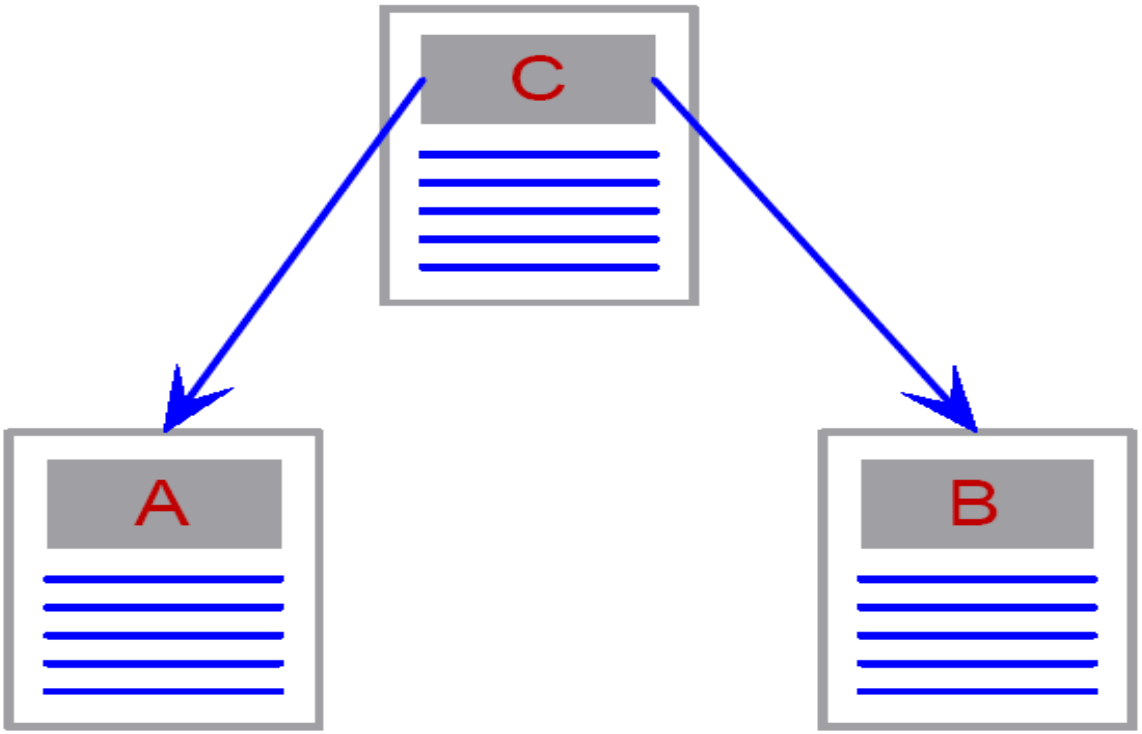
- The web is **not** just a collection of documents – its hyperlinks are important!
- A link from page *A* to page *B* may indicate:
  - *A* is related to *B*, or
  - *A* is recommending, citing, voting for or endorsing *B*
- Links are either
  - referential – *click here and get back home*, or
  - Informational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

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.

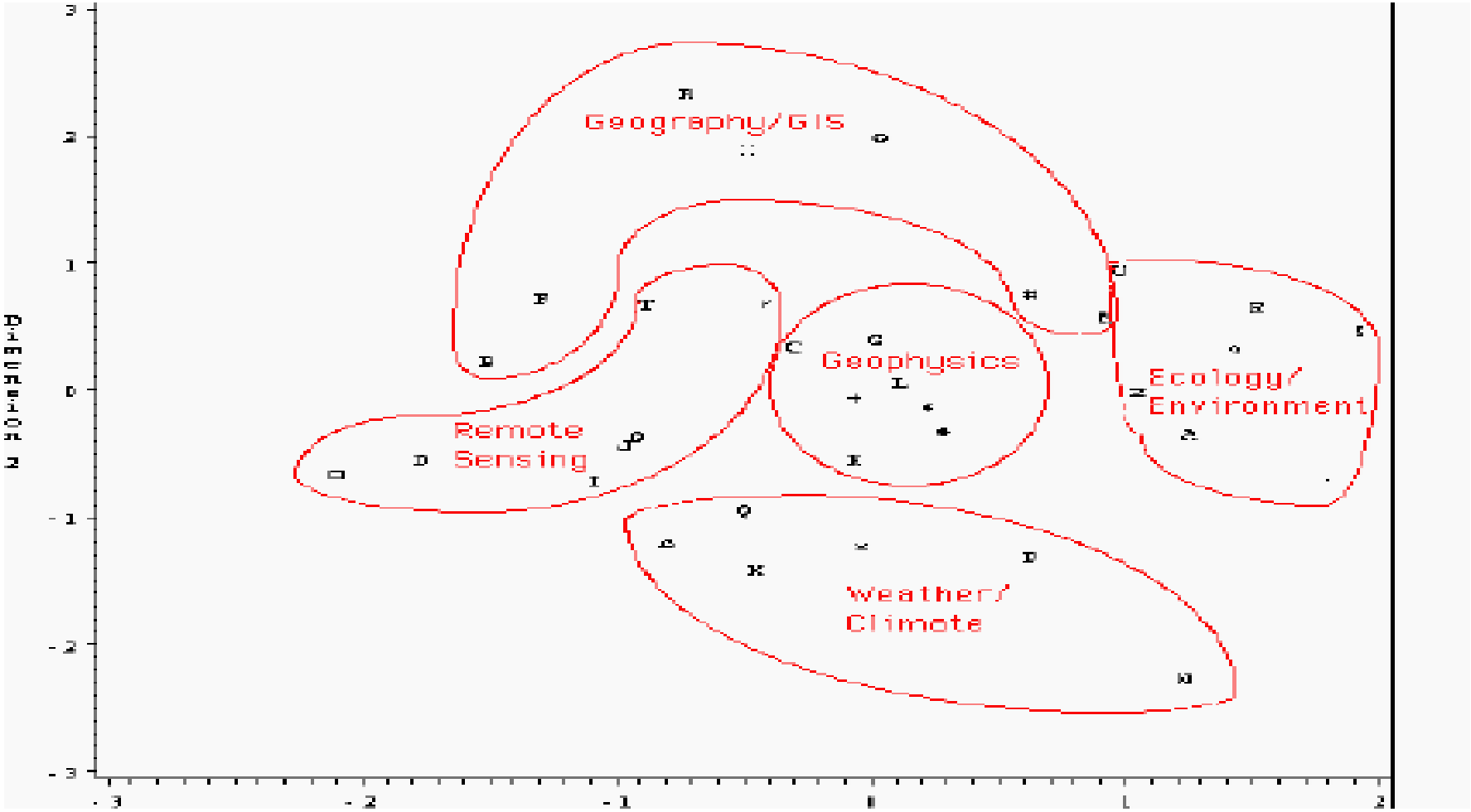
Journal Title (AI)	**Impact Factor (2004)**
J. Mach. Learn. Res.	5.952
IEEE T. Pattern Anal.	4.352
IEEE T. Evolut. Comp.	3.688
Artif. Intell.	3.570
Mach. Learn.	3.258

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*.

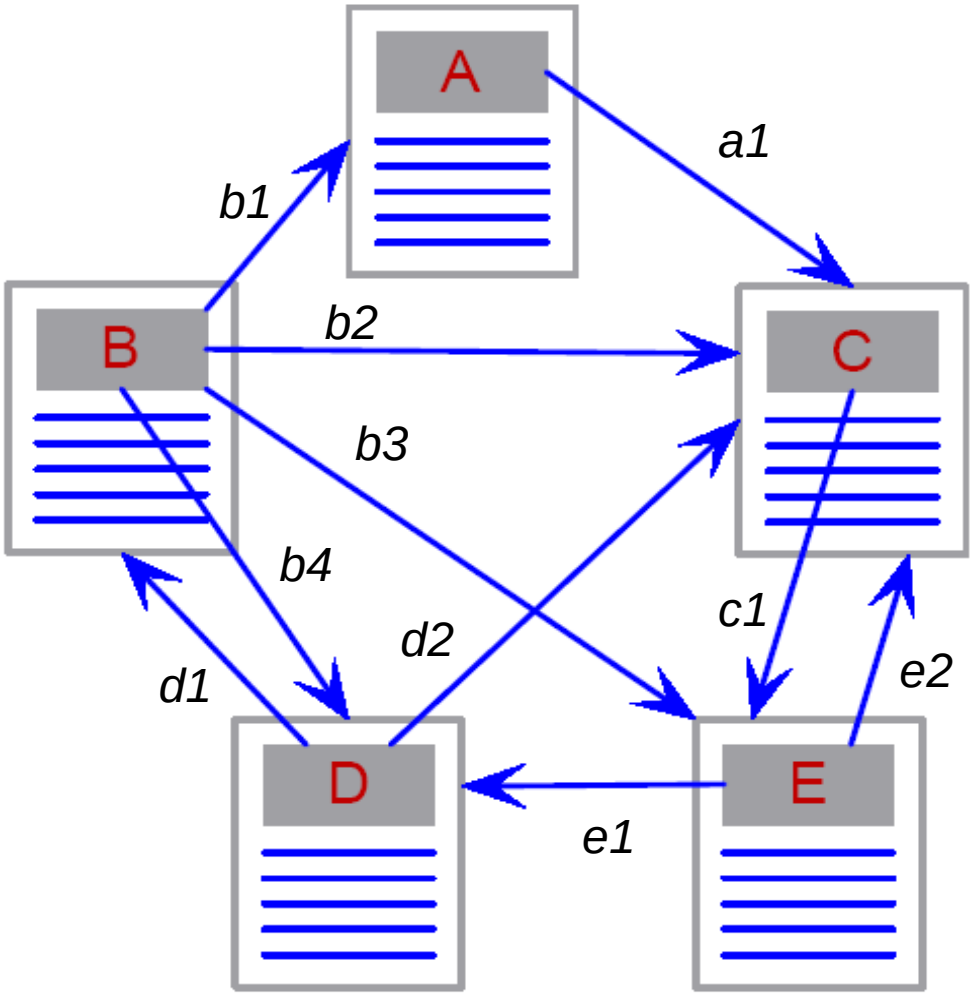
Markov chains

- Markov chains have been extensively studied by statisticians and have been applied in a wide variety of areas.

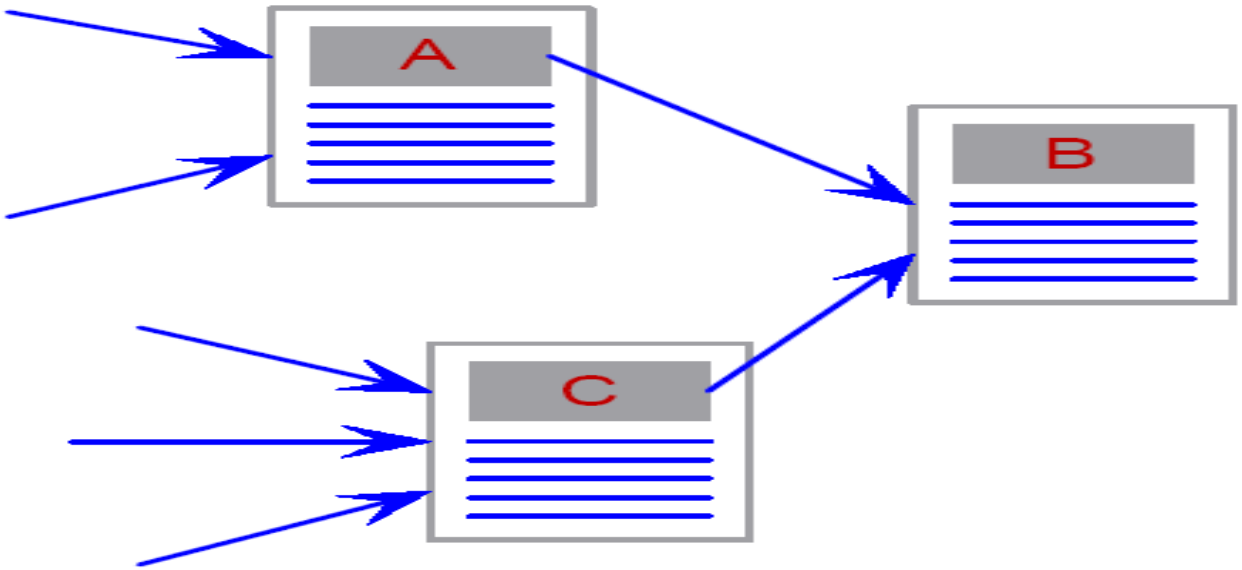
Markov Property:

$$P(S_t \mid S_{t-1}, S_{t-2}, \ldots, S_2, S_1) = P(S_t \mid S_{t-1}) \qquad (3.3)$$

Markov Chain Example



PageRank - Motivation

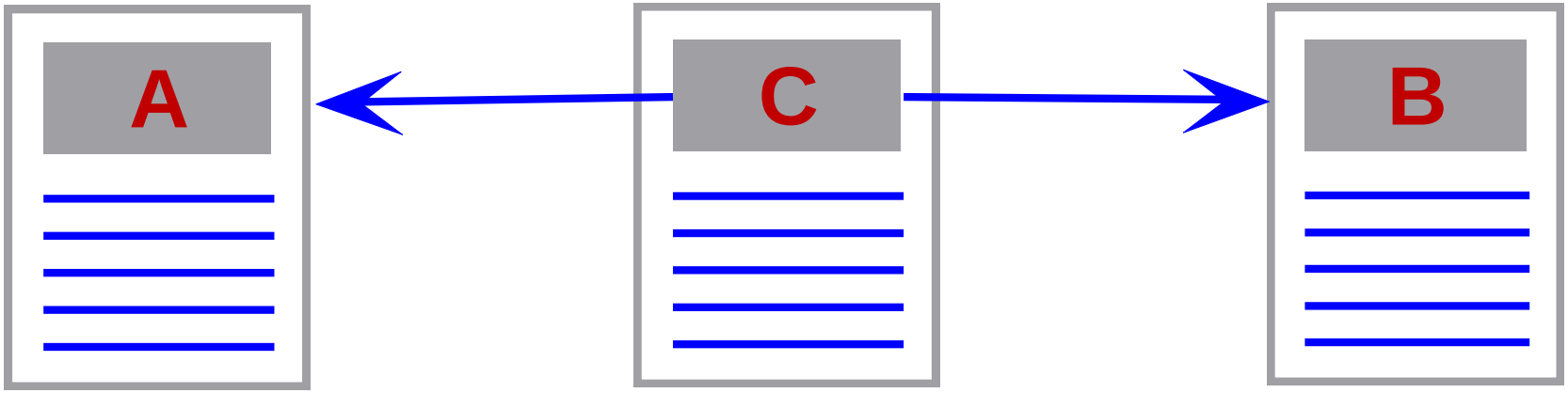


- A link from page *A* to page *B* is a **vote** of the author of *A* for *B,* or a **recommendation** of the page.
- The number incoming links to a page is a measure of importance and authority of the page.
- Also take into account the quality of recommendation, so a page is more important if the sources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

The Random Surfer

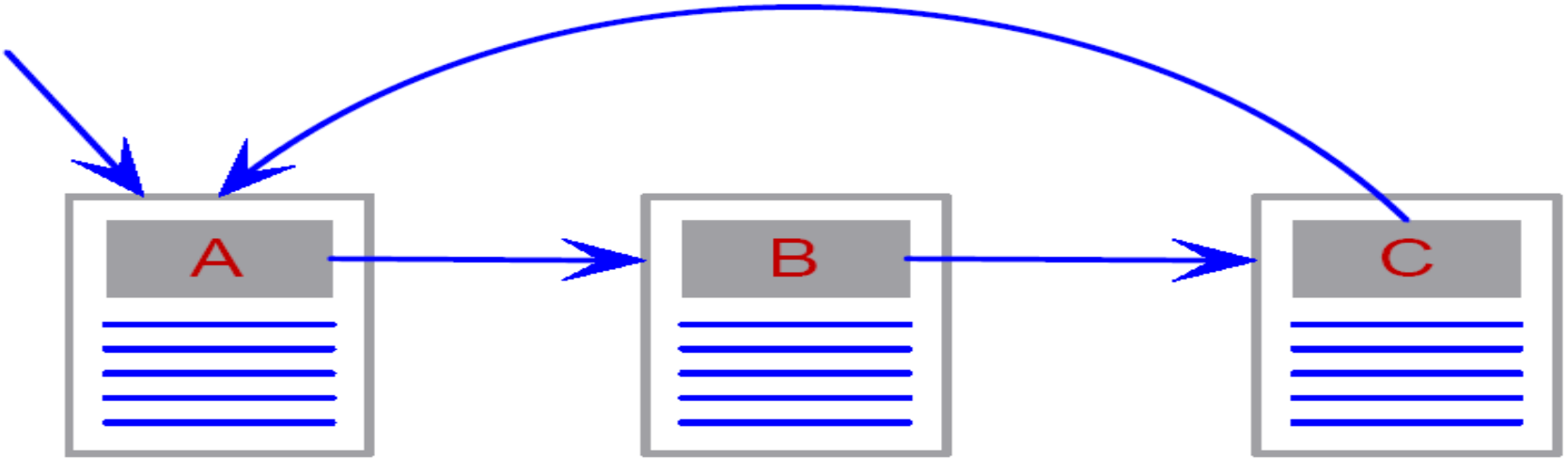
- Assume the web is a Markov chain.
- Surfers randomly click on links, where the probability of an outlink from page A is *1/m*, where *m* is the number of outlinks from *A.*
- The surfer occasionally gets *bored* and is *teleported* to another web page, say *B,* where *B* is equally likely to be any 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



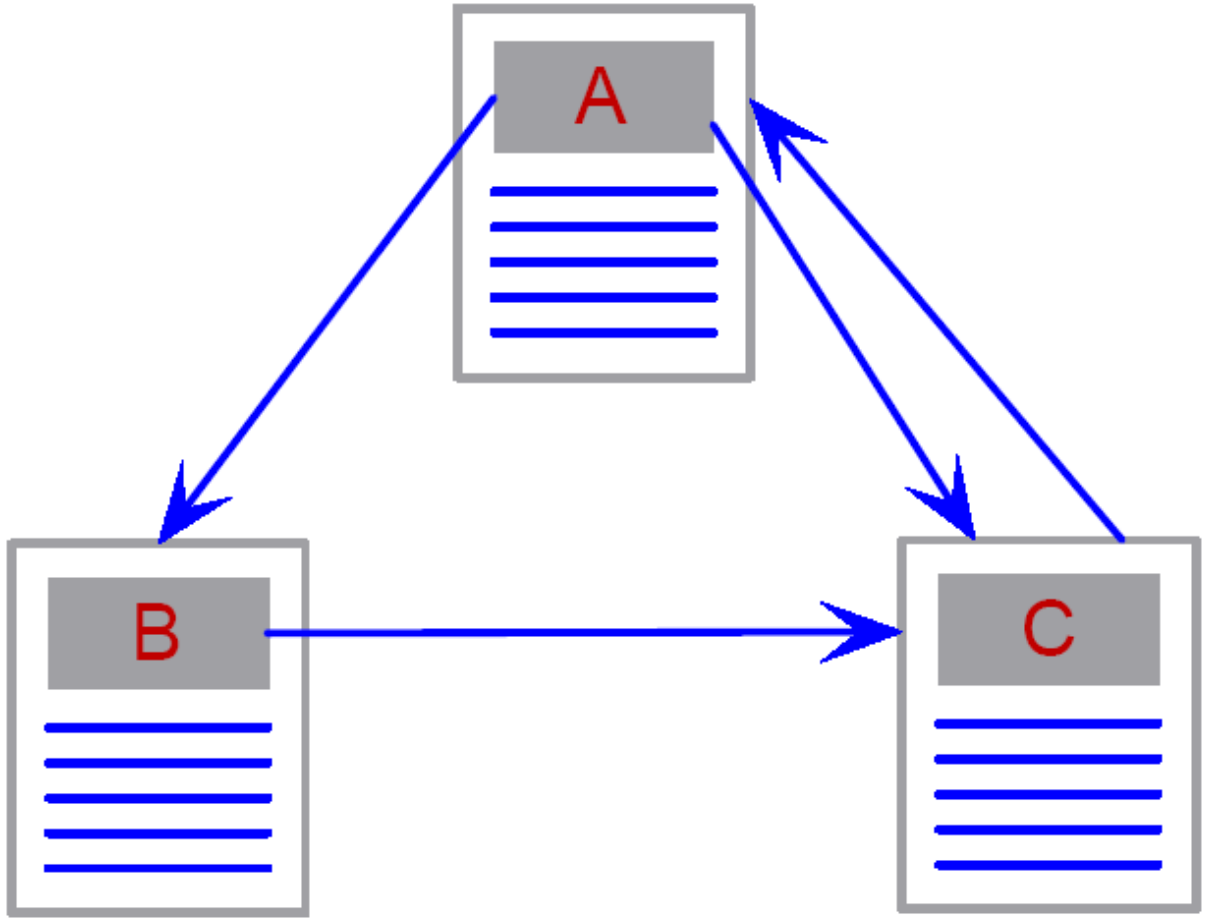
- Problem: Pages in a loop accumulate rank but do not distribute it.
- Solution: Teleportation, i.e. with a certain probability the surfer can jump to any other web page to get out of the loop.

PageRank (*PR*) - Definition

$$PR(W) = \frac{T}{N} + (1-T)\left(\frac{PR(W_1)}{O(W_1)} + \frac{PR(W_2)}{O(W_2)} + \ldots + \frac{PR(W_n)}{O(W_n)}\right)$$

- *W* is a web page
- *Wi* are the web pages that have a link *to W*
- *O*(W*i*) is the number of outlinks from W*i*
- *T* is the teleportation probability
- *N* is the size of the web

Example Web Graph



Iteratively Computing PageRank

- Replace *T/N* in the def. of *PR*(W) by *T, so PR* will take values between 1 and *N*.
- T 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:*

$$PR(A) = 0.5 + 0.5PR(C)$$

$$PR(B) = 0.5 + 0.5(PR(A)/2)$$

$$PR(C) = 0.5 + 0.5(PR(A)/2 + PR(B))$$

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 T = 0.15 or above requires at most 100 iterations to convergence.
- Researchers still trying to speed-up the computation.

Monte Carlo Methods in Computing PageRank

- Rather than following a single long random walk, the random surfer can follow many sampled random walks.
- Each walk starts at a random page and either teleports with probability *T* or continues choosing a link with uniform probability.
- The PR of a page is the proportion of times a sample random walk ended at that page.
- Rather than starting at a random page, start each walk a fixed number of times from each page.

Inlinks and Outlinks

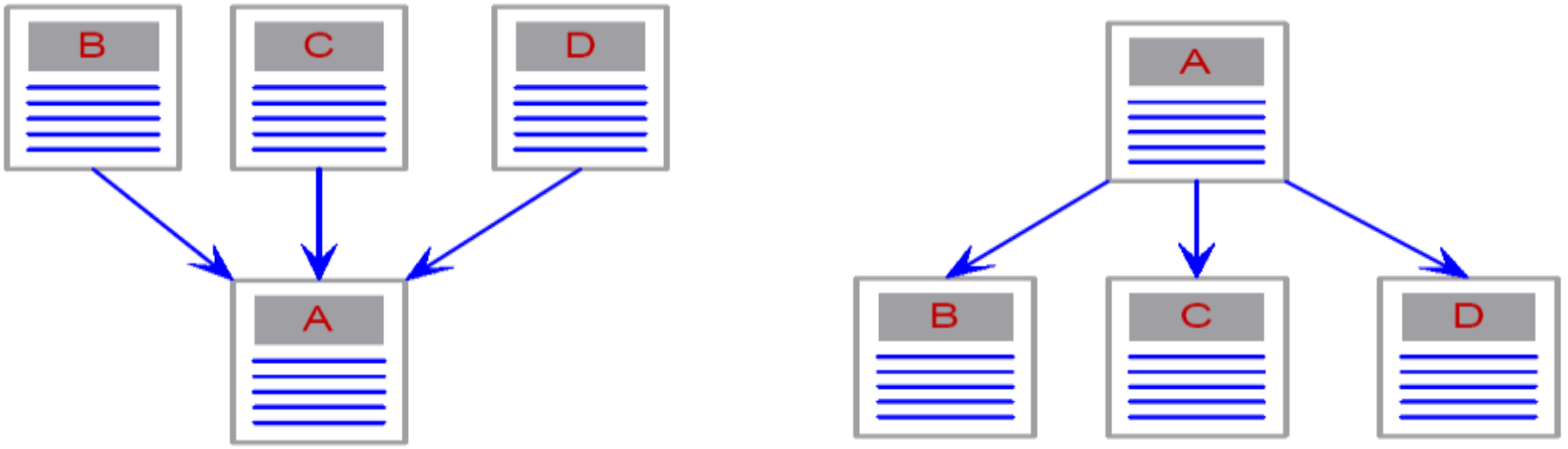
- The number of incoming links to a web page is correlated to the PageRank, but ... this measure is a noisy metric!
- PageRank is biased against new pages (*Googlearch*y), why?
- The long tail of queries, gives less popular web pages some visibililty.
- A single outlink to one in the community minimises the loss of PageRank within a community.

Personalised PageRank

$$PR(P) = Tv + (1-T)\left(\frac{PR(W_1)}{O(W_1)} + \frac{PR(W_2)}{O(W_2)} + \ldots + \frac{PR(W_n)}{O(W_n)}\right)$$

- Change *T/N* with *Tv*
- 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

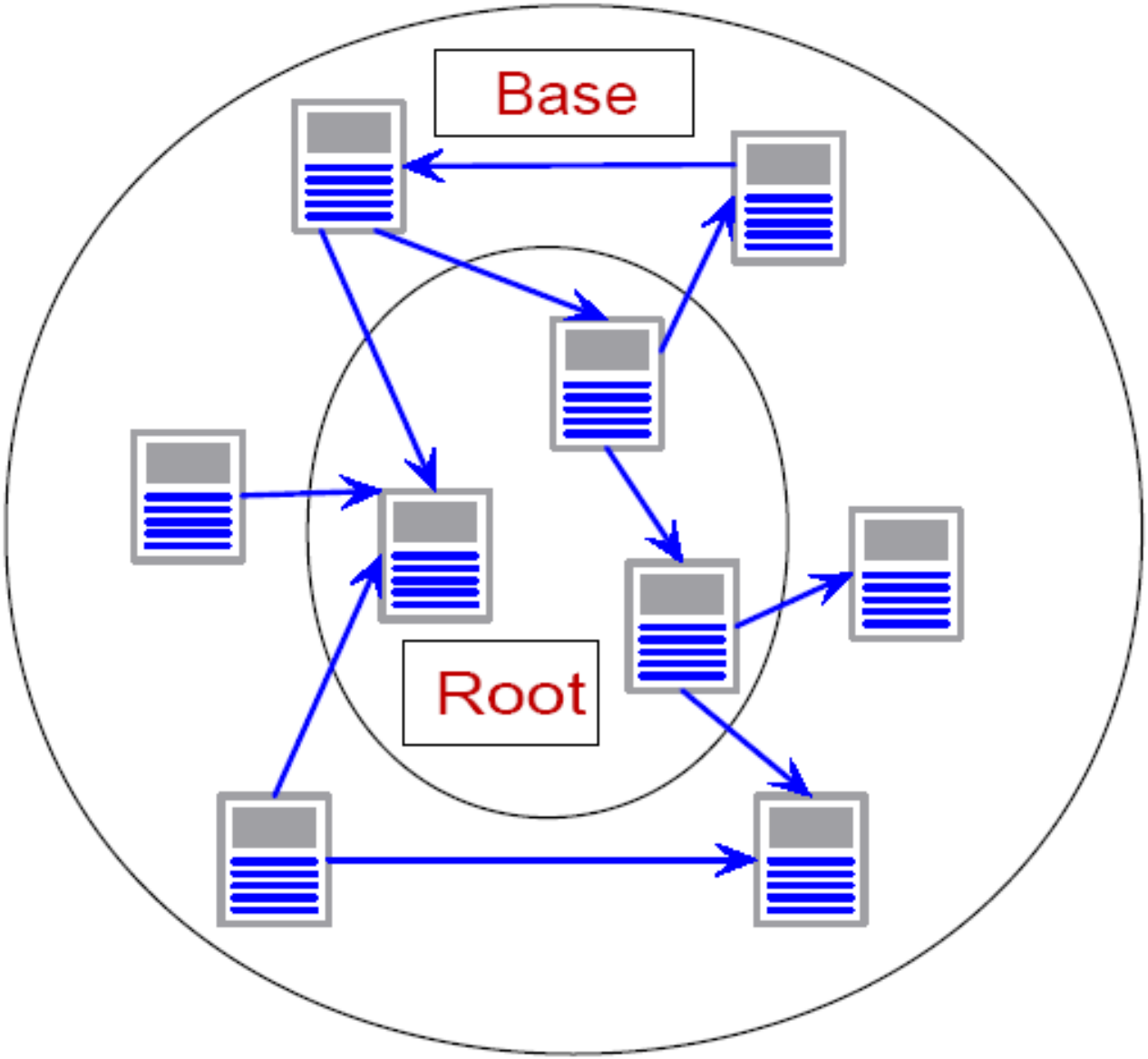


- **A** on the left is an **authority**
- **A** on the right is a **hub**

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.

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*.

Weblogs influence on PageRank

- A weblog (or blog) is a frequently updated web site on a particular topic, made up of entries in reverse chronological order.
- Blogs are a rich source of links, and therfore their links influence PageRank.
- A “google bomb” is an attempt to influence the ranking of a web page for a given phrase by adding links to the page with the phrase as its anchor text.

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 copying a hub page which links to all members.
- *Selling links* from sites with high PageRank.