Web Structure Mining

Web Structure Mining

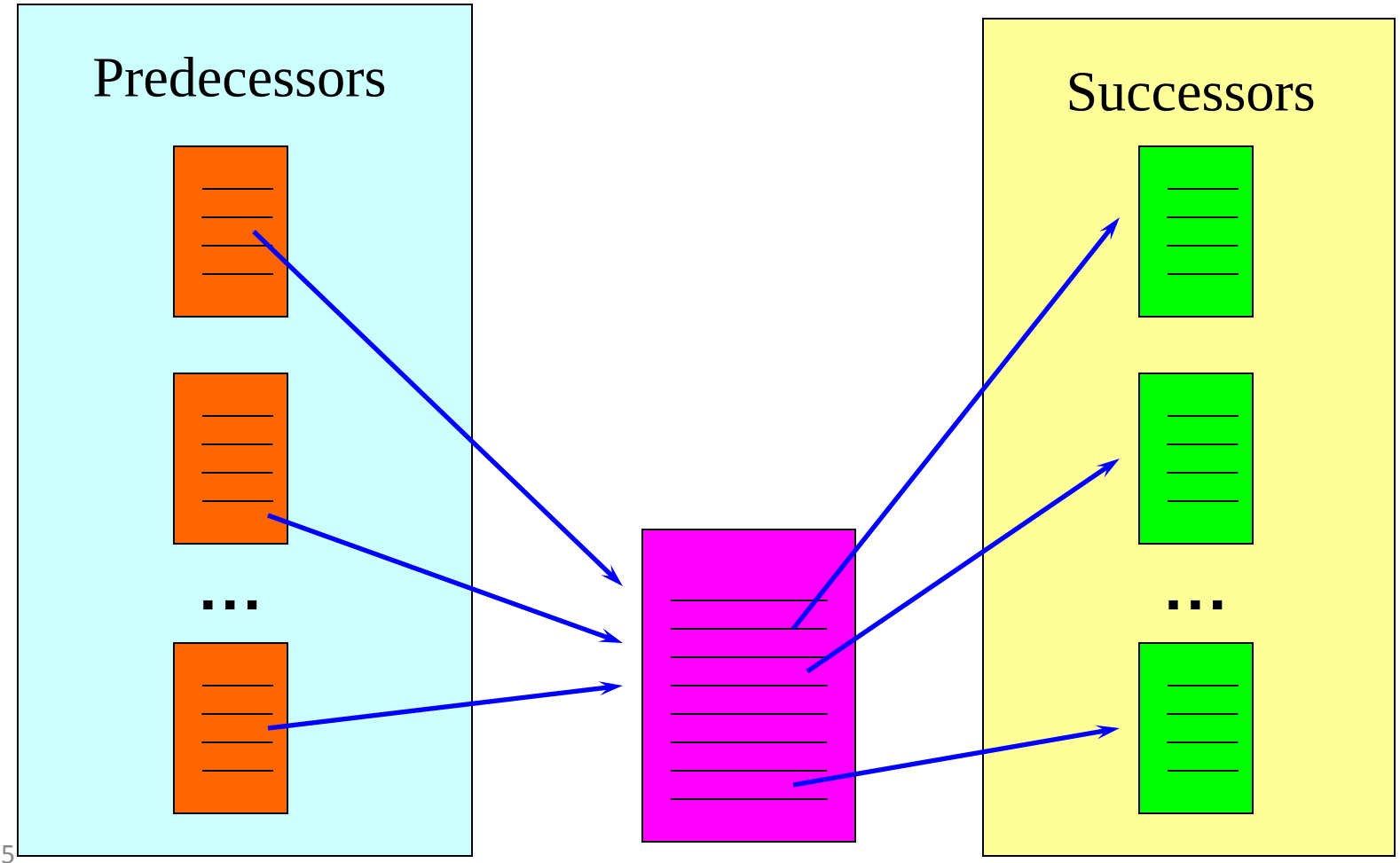
- (1970) Researchers proposed methods of using citations among journal articles to evaluate the quality of reserach papers.
- Customer behavior – evaluate a quality of a product based on the opinions of other customers (instead of product's description or advertisement)
- Unlike journal citations, the Web linkage has some unique features:
  - not every hiperlink represents the endorsement we seek
  - one authority page will seldom have its Web page point to its competitive authorities (CocaCola → Pepsi)
  - authoritative pages are seldom descriptive (Yahoo! may not contain the description „Web search engine")

Evaluation of Web pages

Web Search

- There are two approches:
  - **page rank**: for discovering the most important pages on the Web (as used in Google)
  - **hubs and authorities**: a more detailed evaluation of the importance of Web pages
- Basic definition of importance:
  - A page is important if important pages link to it

Predecessors and Successors of a Web Page



Page Rank (1)

Simple solution: create a stochastic matrix of the Web:

- Each page i corresponds to row i and column i of the matrix
- If page j has n successors (links) then the ij^{th} cell of the matrix is equal to 1/n if page i is one of these n succesors of page j, and 0 otherwise.

Page Rank (2)

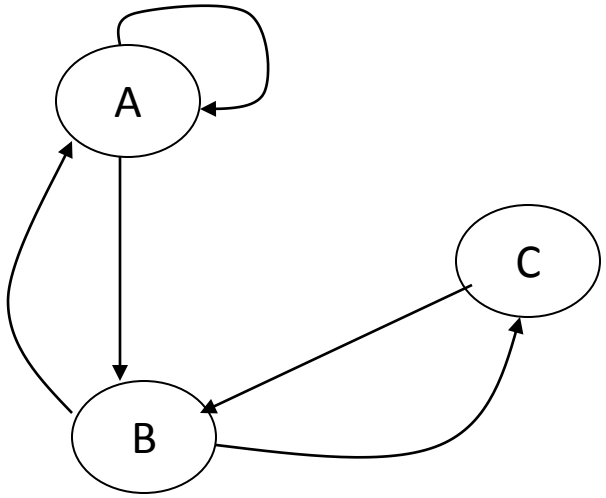
The intuition behind this matrix:

- initially each page has 1 unit of importance. At each round, each page shares importance it has among its successors, and receives new importance from its predecessors.
- The importance of each page reaches a limit after some steps
- That importance is also the probability that a Web surfer, starting at a random page, and following random links from each page will be at the page in question after a long series of links.

Page Rank (3) – Example 1

- Assume that the Web consists of only three pages - A, B, and C. The links among these pages are shown below.

Let [a, b, c] be the vector of importances for these three pages



	A	B	C
A	1/2	1/2	0
B	1/2	0	1
C	0	1/2	0

Page Rank – Example 1 (cont.)

- The equation describing the asymptotic values of these three variables is:

$$\begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 1 \\ 0 & 1/2 & 0 \end{bmatrix} \times \begin{bmatrix} a \\ b \\ c \end{bmatrix}$$

We can solve the equations like this one by starting with the assumption a = b = c = 1, and applying the matrix to the current estimate of these values repeatedly. The first four iterations give the following estimates:

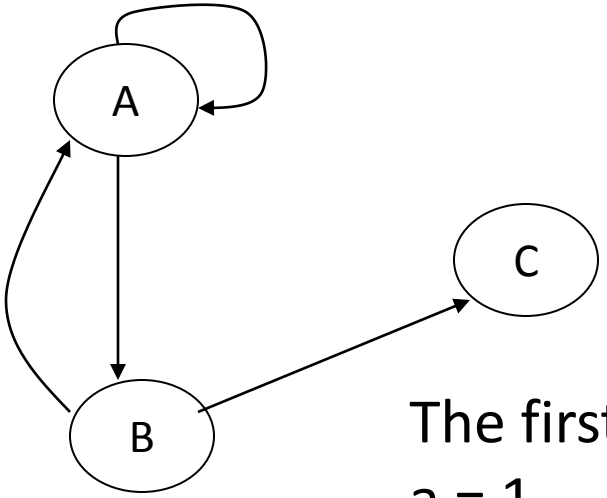
a =	1	1	5/4	9/8	5/4	…	**6/5**
b =	1	3/2	1	11/8	17/16	…	**6/5**
c =	1	1/2	3/4	1/2	11/16	…	**3/5**

Problems with Real Web Graphs

- In the limit, the solution is a=b=6/5, c=3/5. That is, *a* and *b* each have the same importance, and twice of *c*.
- Problems with Real Web Graphs
  - dead ends: a page that has no succesors has nowhere to send its importance.
  - spider traps: a group of one or more pages that have no links out.

Page Rank – Example 2

- Assume now that the structure of the Web has changed. The new matrix describing transitions is:



$$\begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} ½ & ½ & 0 \\ ½ & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \times \begin{bmatrix} a \\ b \\ c \end{bmatrix}$$

The first four steps of the iterative solution are:

a = 1 1 3/4 5/8 1/2

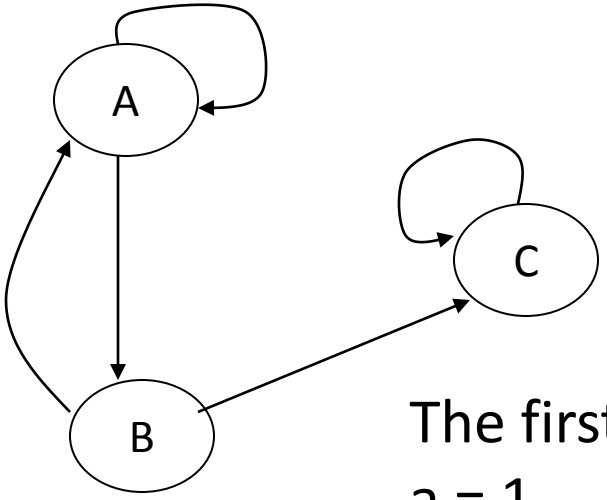
b = 1 1/2 1/2 3/8 5/16

c = 1 1/2 1/4 1/4 3/16

Eventually, each of a, b, and c become 0.

Page Rank – Example 3

- Assume now once more that the structure of the Web has changed. The new matrix describing transitions is:



$$\begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} ½ & ½ & 0 \\ ½ & 0 & 0 \\ 0 & 1 & 1/2 \end{bmatrix} \times \begin{bmatrix} a \\ b \\ c \end{bmatrix}$$

The first four steps of the iterative solution are:

a = 1 1 3/4 5/8 1/2

b = 1 1/2 1/2 3/8 5/16

c = 1 3/2 7/4 2 35/16

c converges to 3, and a=b=0.

Google Solution

- Instead of applying the matrix directly, „tax" each page some fraction of its current importance, and distribute the taxed importance equally among all pages.
- Example: if we use 20% tax, the equation of the previous example becomes:

  a = 0.8 * (½*a + ½ *b +0*c)

  b = 0.8 * (½*a + 0*b + 0*c)

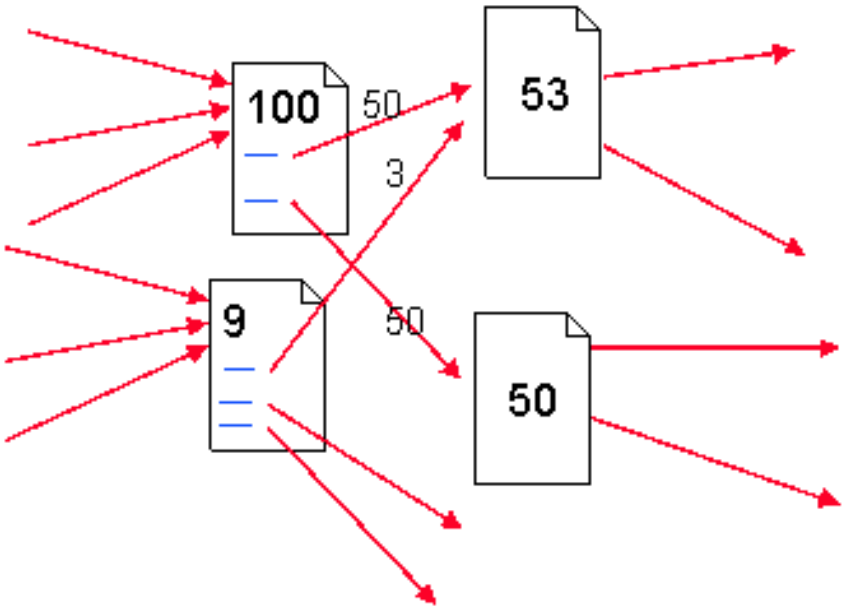
  c = 0.8 * (0*a + ½*b + 1*c)

The solution to this equation is a=7/11, b=5/11, and c=21/11

Google Anti-Spam Solution

- „Spamming” is the attept by many Web sites to appear to be about a subject that will attract surfers, without truly being about that subject.
- Solutions:
  - Google tries to match words in your query to the words on the Web pages. Unlike other search engines, Google tends to belive what others say about you in their anchor text, making it harder from you to appear to be about something you are not.
  - The use of Page Rank to measure importance also protects against spammers. The naive measure (number of links into the page) can easily be fooled by the spammers who creates 1000 pages that mutually link to one another, while Page Rank recognizes that none of the pages have any real importance.

PageRank Calculation



HITS Algorithm

--Topic Distillation on WWW

- Proposed by Jon M. Kleinberg
- Hyperlink-Induced Topic Search

Key Definitions

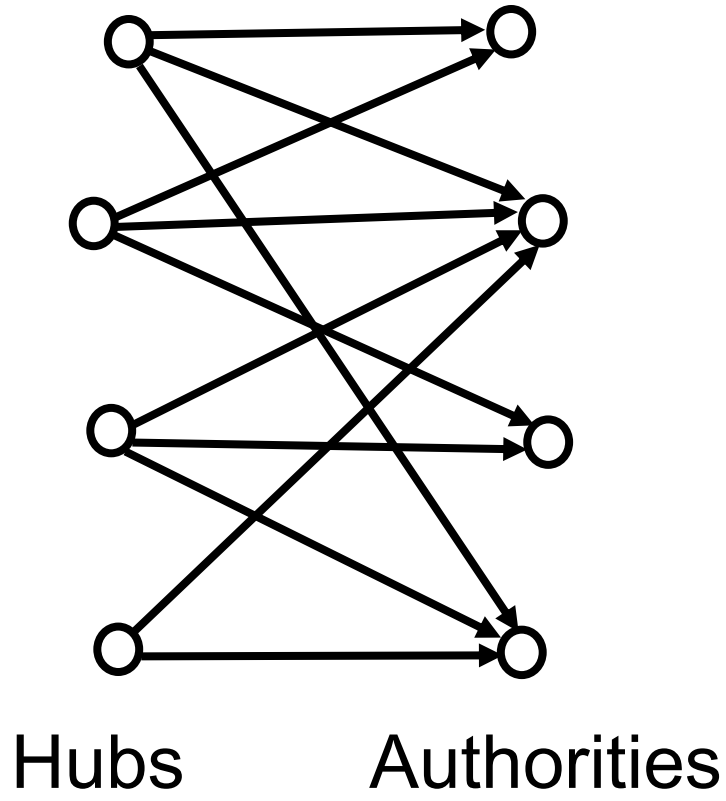
- **Authorities**

  Relevant pages of the highest quality on a broad topic

- **Hubs**

  Pages that link to a collection of authoritative pages on a broad topic

Hub-Authority Relations



Hyperlink-Induced Topic Search (HITS)

The approach consists of two phases:

- It uses the query terms to collect a starting set of pages (200 pages) from an index-based search engine – root set of pages.
- The root set is expanded into a base set by including all the pages that the root set pages link to, and all the pages that link to a page in the root set, up to a designed size cutoff, such as 2000-5000.
- A weight-propagation phase is initiated. This is an iterative process that determines numerical estimates of hub and authority weights

Hub and Authorities

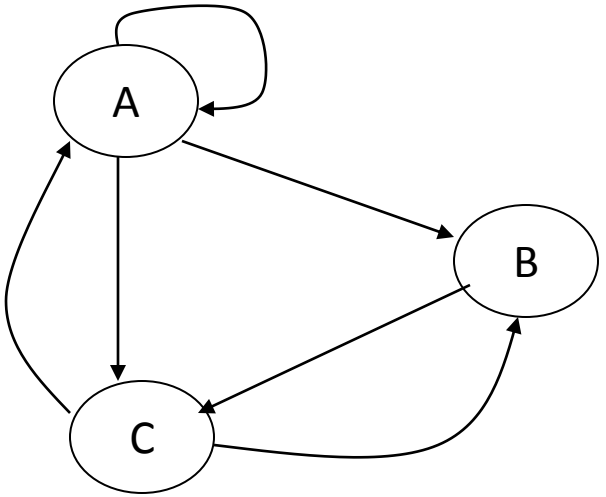
- Define a matrix **A** whose rows and columns correspond to Web pages with entry $A_{ij}=1$ if page `i` links to page `j`, and 0 if not.
- Let **a** and **h** be vectors, whose i^{th} component corresponds to the degrees of authority and hubbiness of the `i`th page. Then:
- $\mathbf{h} = \mathbf{A} \times \mathbf{a}$. That is, the hubbiness of each page is the sum of the authorities of all the pages it links to.
- $\mathbf{a} = \mathbf{A}^T \times \mathbf{h}$. That is, the authority of each page is the sum of the hubbiness of all the pages that link to it ($\mathbf{A}^T$ - transponed matrix).

Then, $\mathbf{a} = \mathbf{A}^T \times \mathbf{A} \times \mathbf{a}$ $\mathbf{h} = \mathbf{A} \times \mathbf{A}^T \times \mathbf{h}$

Hub and Authorities - Example

Consider the Web presented below.

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$$



$$A^T = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$$

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

$$A^TA = \begin{bmatrix} 2 & 2 & 1 \\ 2 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix}$$

Hub and Authorities - Example

If we assume that the vectors

$h = [h_a, h_b, h_c]$ and $a = [a_a, a_b, a_c]$ are each initially [1,1,1], the first three iterations of the equations for a and h are the following:

$a_a = 1 \quad 5 \quad 24 \quad 114$

$a_b = 1 \quad 5 \quad 24 \quad 114$

$a_c = 1 \quad 4 \quad 18 \quad 84$

$h_a = 1 \quad 6 \quad 28 \quad 132$

$h_b = 1 \quad 2 \quad 8 \quad 36$

$h_c = 1 \quad 4 \quad 20 \quad 96$

Discovering cyber-communities on the web

Based on link structure

What is cyber-community

- Defn: a *community on the web* is a group of web pages sharing a common interest
  - Eg. A group of web pages talking about POP Music
  - Eg. A group of web pages interested in data-mining
- Main properties:
  - Pages in the same community should be similar to each other in contents
  - The pages in one community should differ from the pages in another community
  - Similar to cluster

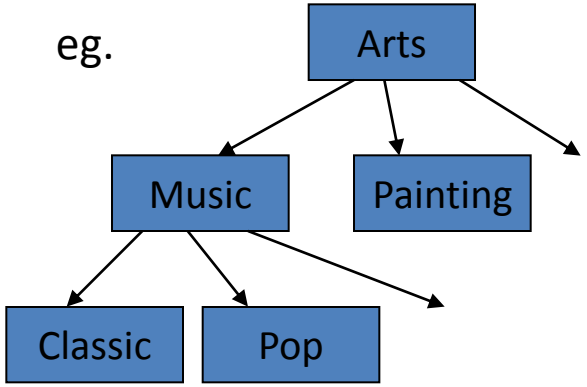
Recursive Web Communities

- **Definition:** A *community* consists of members that have more links within the community than outside of the community.
- Community identification is NP-complete task

Two different types of communities

- Explicitly-defined communities
  - They are well known ones, such as the resource listed by Yahoo!

eg.

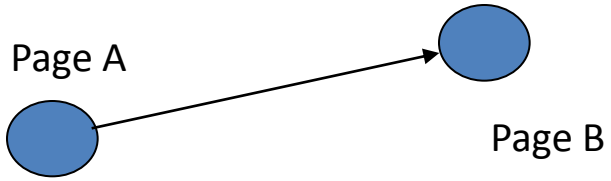


- Implicitly-defined communities
  - They are communities unexpected or invisible to most users

eg. The group of web pages interested in a particular singer

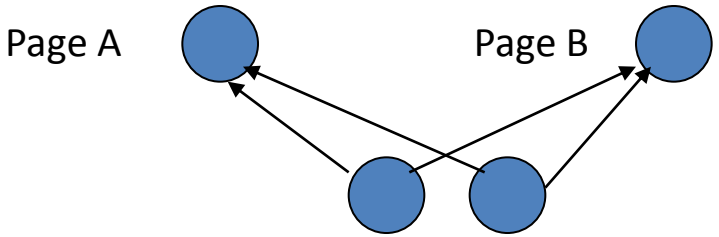
Similarity of web pages

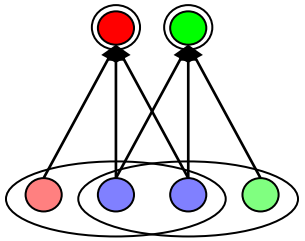
- Discovering web communities is similar to clustering. For clustering, we must define the similarity of two nodes
- A Method I:
  - For page and page B, A is related to B if there is a hyper-link from A to B, or from B to A
  - Not so goc f IBM and Microsoft.



Similarity of web pages

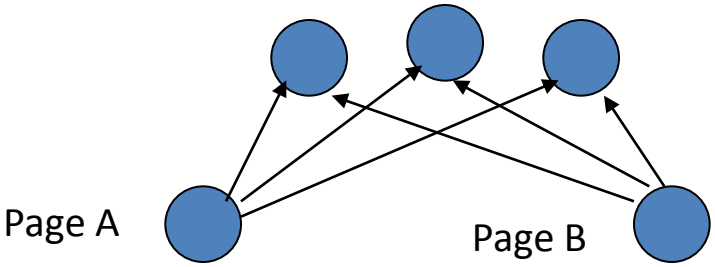
- Method II (from Bibliometrics)
  - **Co-citation**: the similarity of A and B is measured by the number of pages cite both A and B

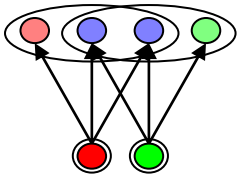




The normalized degree of overlap in inbound links

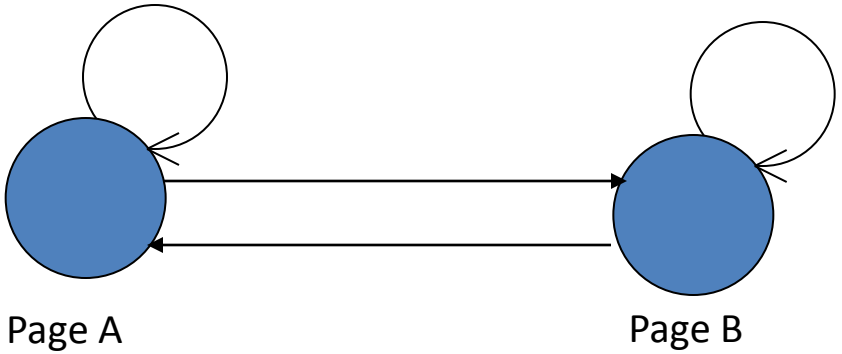
  - **Bibliographic coupling**: the similarity of A and B is measured by the number of pages cited by both A and B.



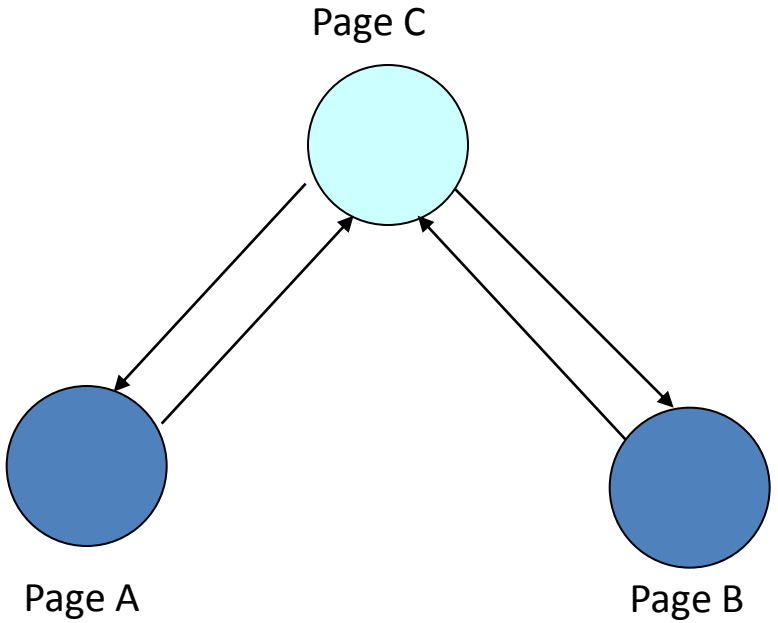


The normalized degree of overlap in outbound links

Simple Cases (co-citations and coupling)



Better not to account self-citations



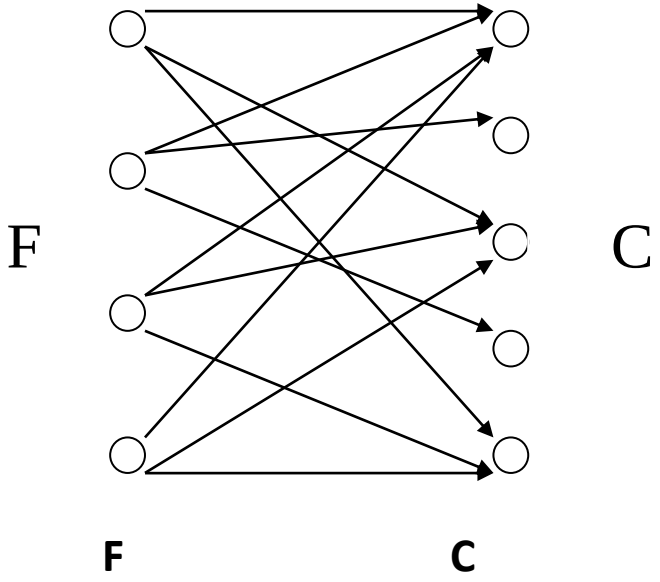
Number of pages for similarity decision should be big enough

Example method of clustering

- The method from R. Kumar, P. Raghavan, S. Rajagopalan, Andrew Tomkins
  - IBM Almaden Research Center
- They call their method ***communities trawling (CT)***
- They implemented it on the graph of 200 millions pages, it worked very well

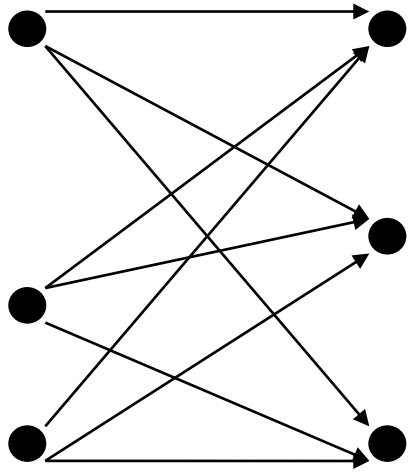
Basic idea of CT

- Bipartite graph: Nodes are partitioned into two sets, **F** and **C**
- Every directed edge in the graph is directed from a node in **F** to a node in **C**



Basic idea of CT

- Definition Bipartite cores
  - a complete bipartite subgraph with at least i nodes from **F** and at least j nodes from **C**
  - i and j are tunable parameters
  - A (i, j) Bipartite core



A (i=3, j=3) bipartite core

- Every community have such a core with a certain i and j

Basic idea of CT

- A bipartite core is the identity of a community
- To extract all the communities is to enumerate all the bipartite cores on the web

Web Communities