Extracting Information from the Links in Academic Webs

An overview of methods and results

Mike Thelwall
Statistical Cybermetrics Research Group
University of Wolverhampton, UK

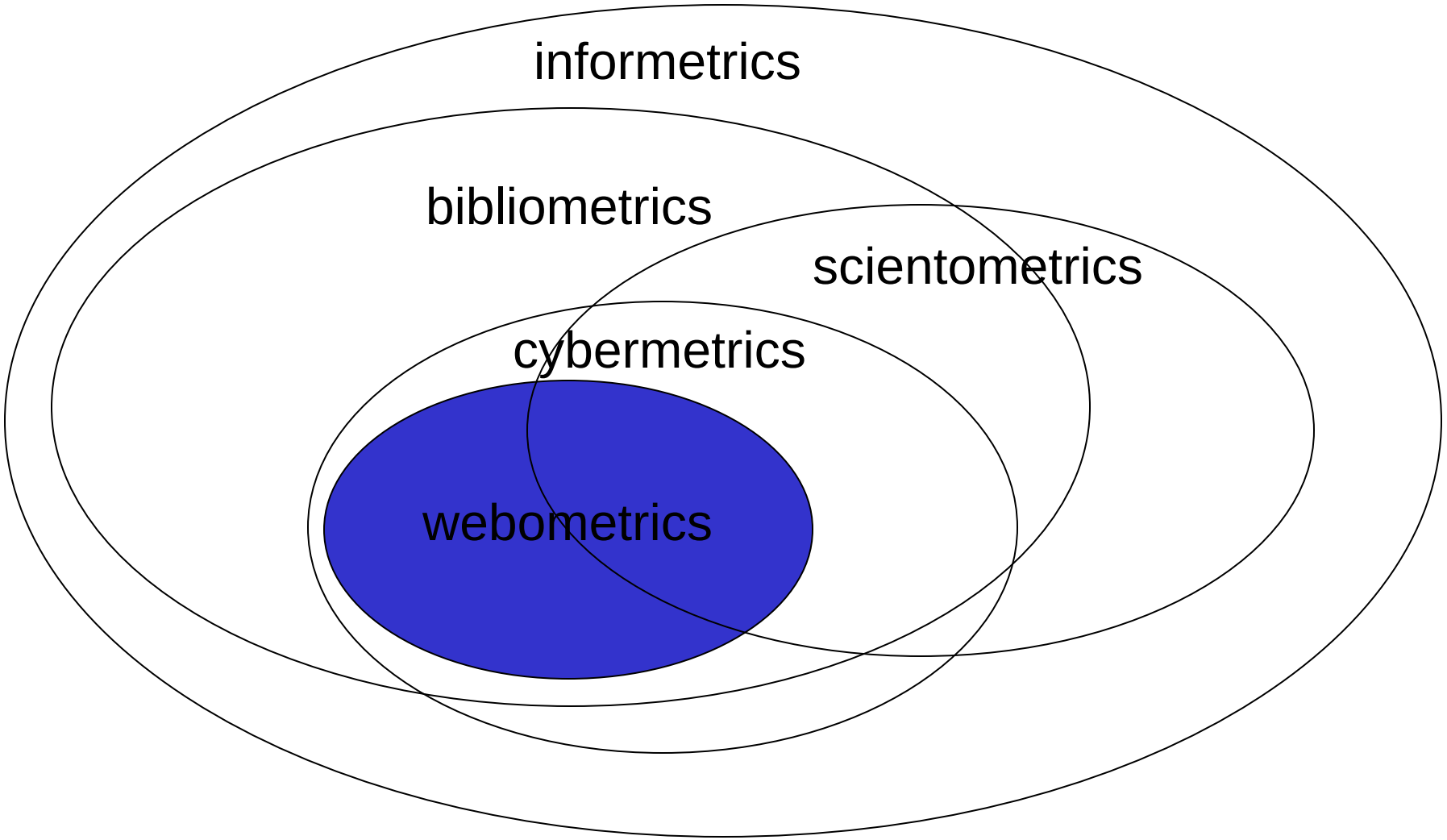
Contents

- Introduction to Webometrics
- Computer Science uses for Web links
- Main talk: analysing university Web links
  - Data collection
  - Data processing
  - Analysis
  - Results

Part 1:
Introduction to Webometrics

A new area of Information Science

infor-/biblio-/sciento-/cyber-/webo-/metrics



© Lennart Björneborn 2001-2002

Webometrics

- the study of quantitative aspects of the construction and use of info. resources, structures and technologies on the Web, drawing on bibliometric and informetric methods – LB def.
- four main research areas of Webometric concern:
  - Web page contents
  - link structures (e.g., Web Impact Factors, cohesion of link topologies, etc.)
  - search engine performance
  - users' information behavior (searching, browsing, encountering, etc.)
- cybermetrics = quantitative studies of the whole Internet
  - i.e. chat, mailing lists, news groups, MUDs, etc. - and Web

© Lennart Björneborn 2001-2002

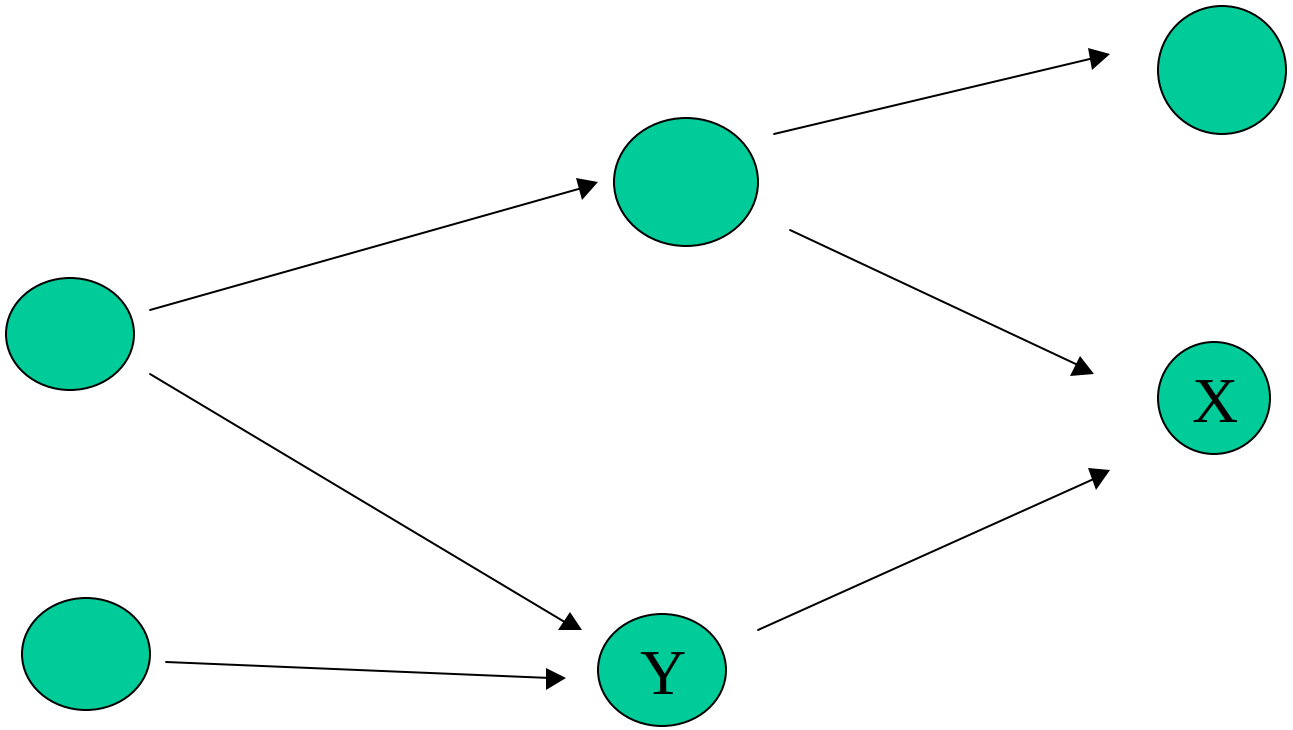
Part 2:
Computer Science uses for Web links

Search engine page ranking, topic identification and similarity matching

PageRank

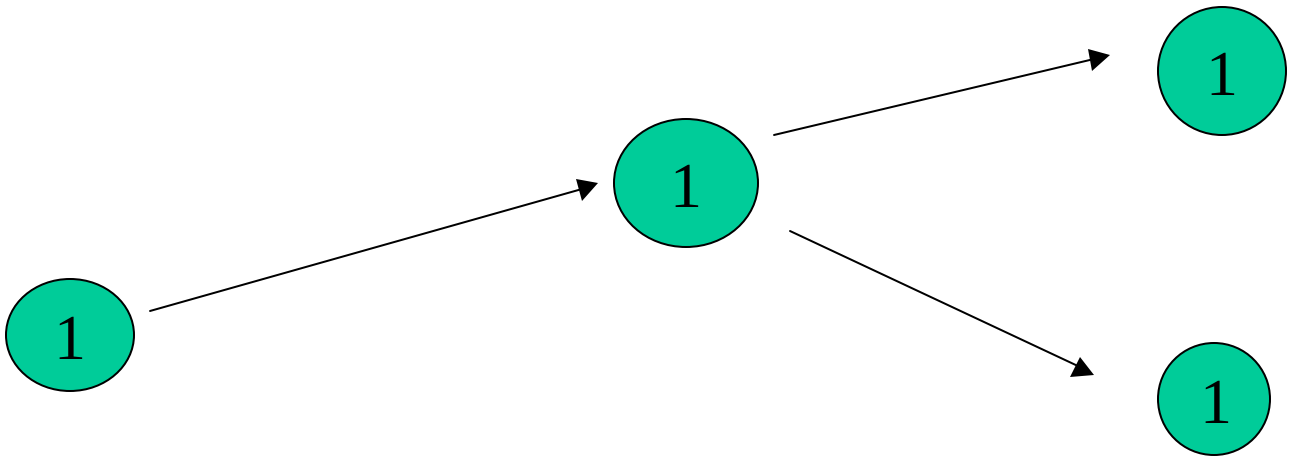
- Assumptions:
  - A page with many links to it is more likely to be useful than one with few links to it
  - The links from a page that itself is the target of many links are likely to be particularly important

Example

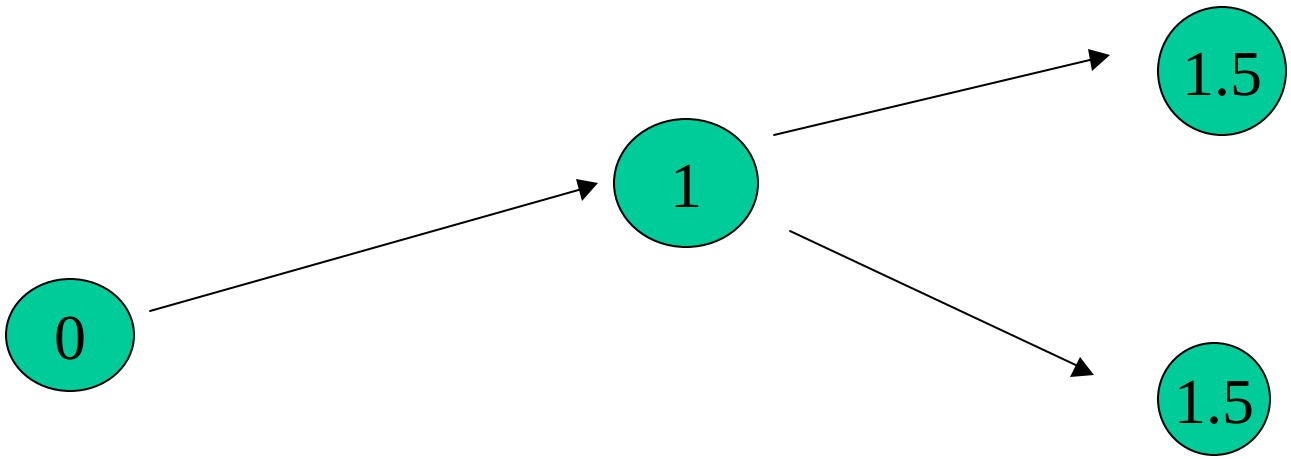


X seems to be the most important page since 2 important pages link to it

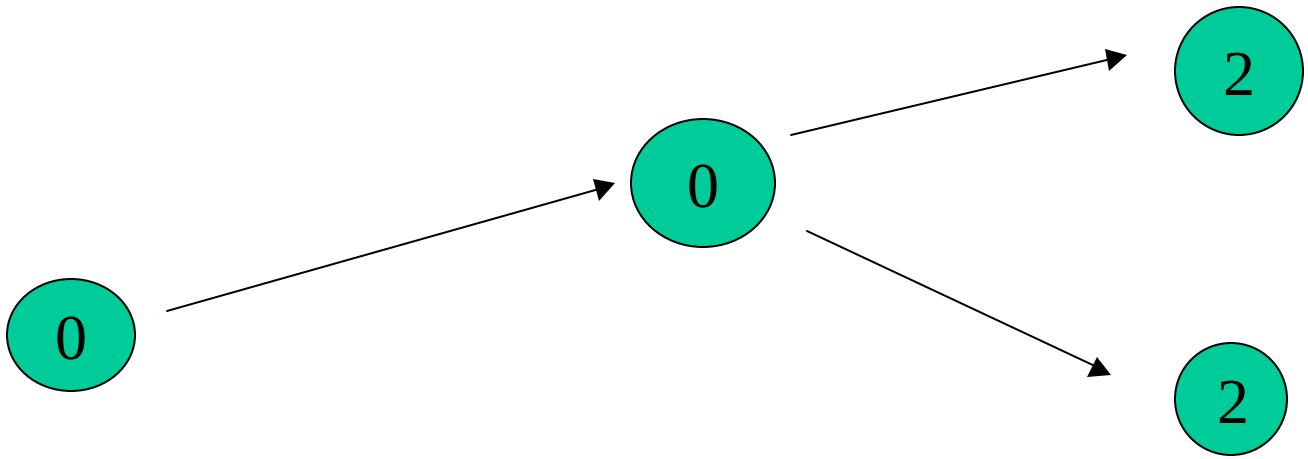
Simple voting model: round 1



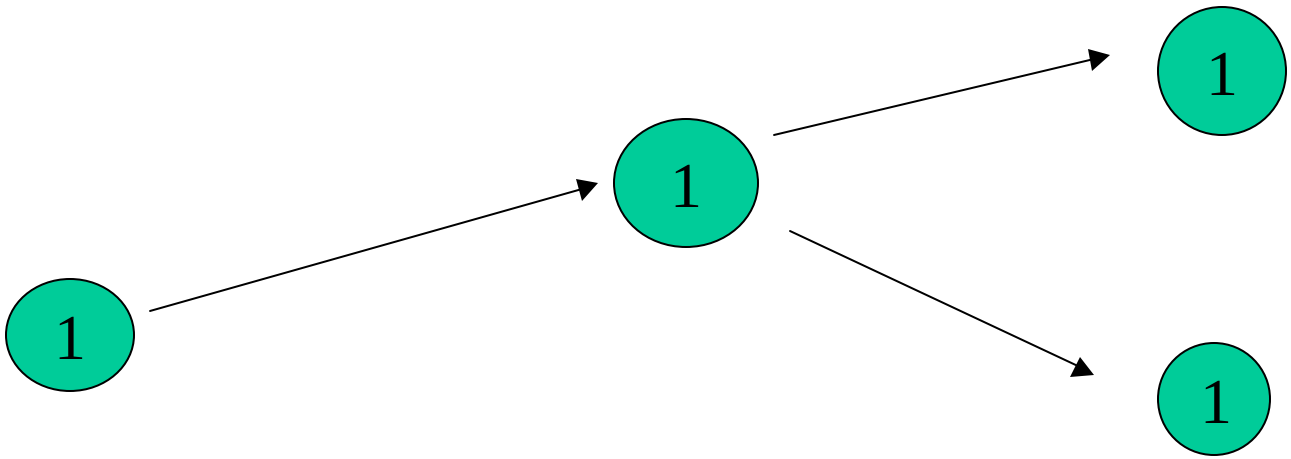
Simple voting model: round 2



Simple voting model: round 3

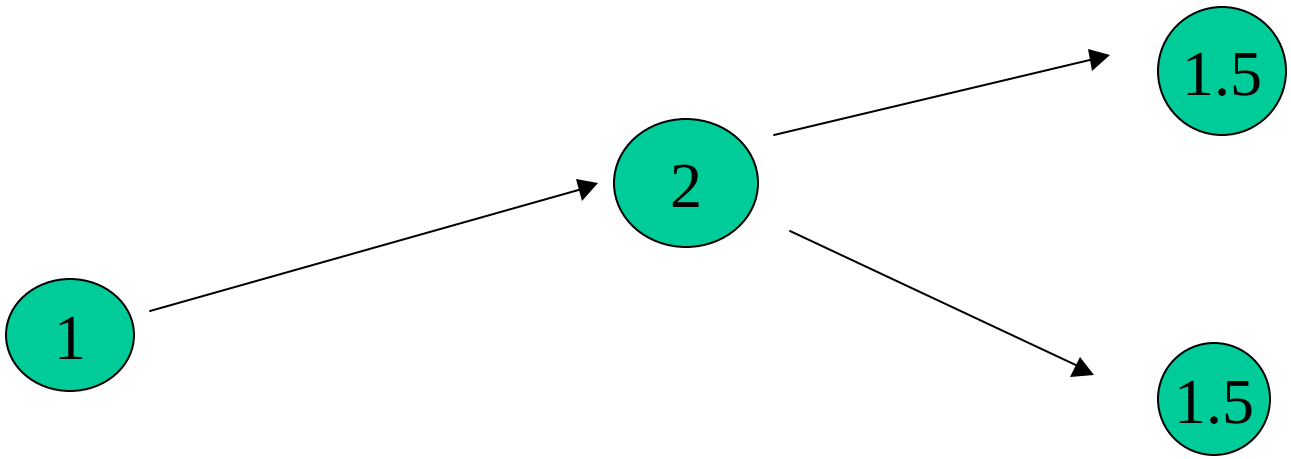


Revised voting model: round 1

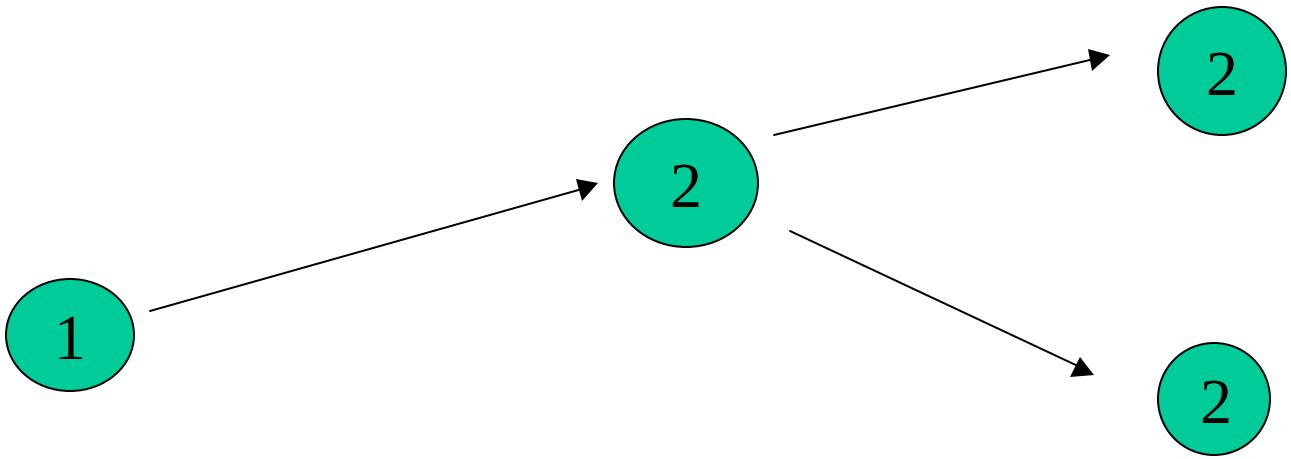


- Allocate 1 vote to each node after each voting round
- Remove votes from ‘leaf’ nodes

Revised voting model: round 2

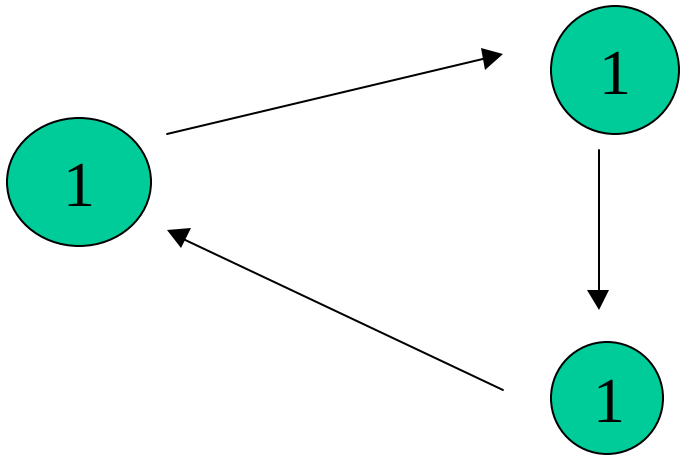


Revised voting model: round 3



The middle node only has one link to it, but this does not share its votes with other nodes

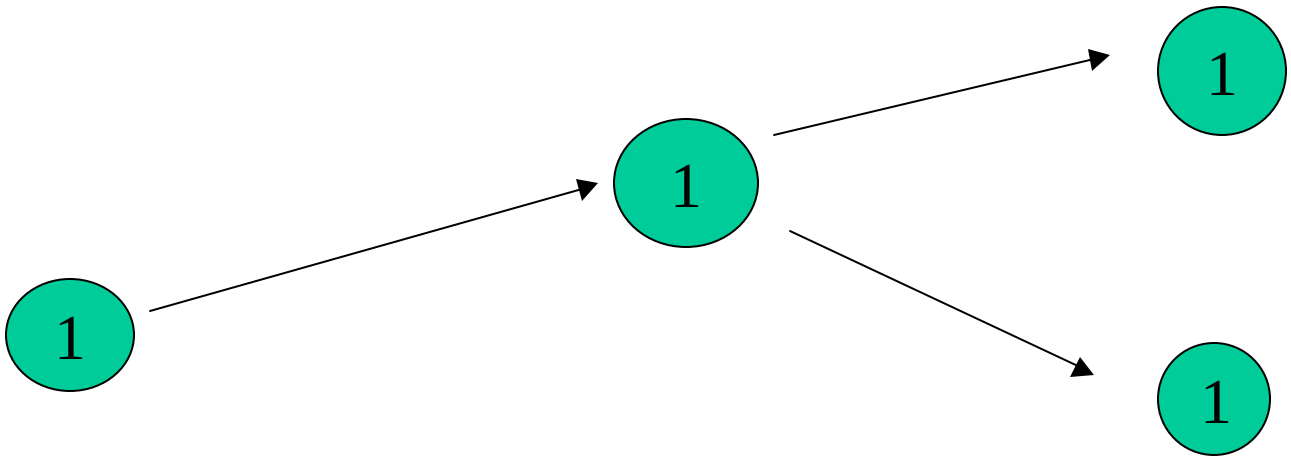
Revised voting model cycling problem



PageRank

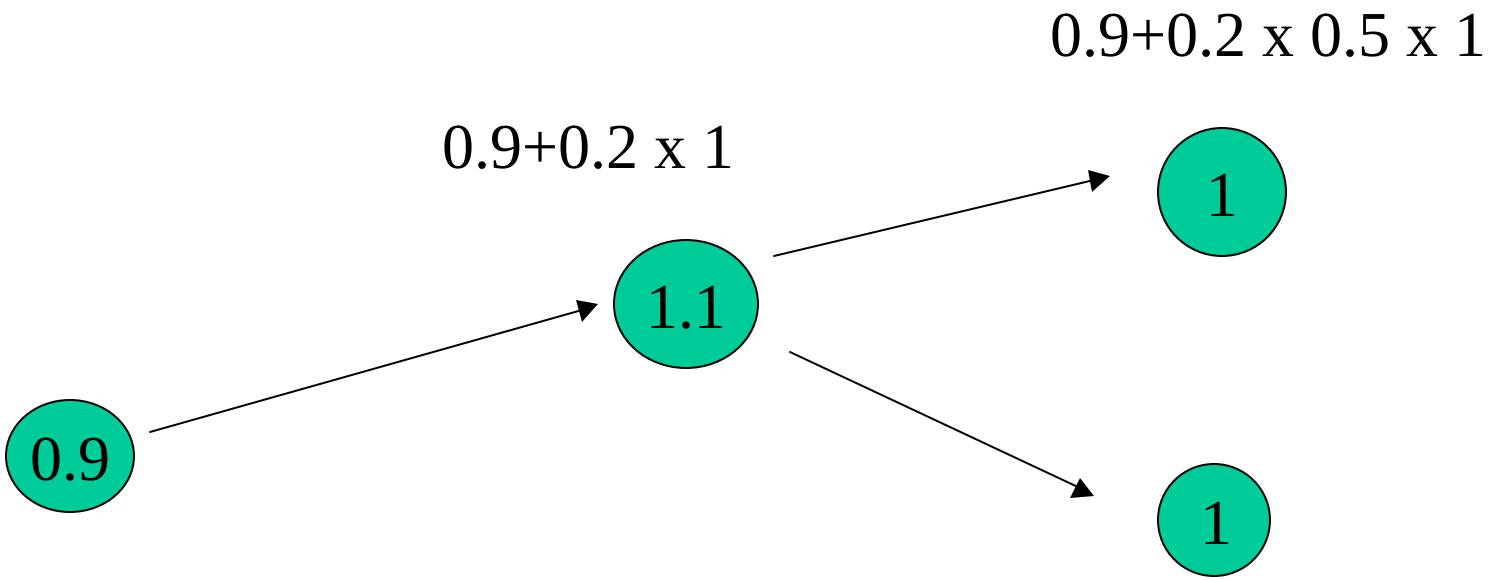
- Use a proportion of vote, redistribute the rest
- If proportion is < 1 then no cycling will occur
- Voting can also be performed by a matrix
- Find votes from principle left eigenvector of matrix

PageRank: round 1

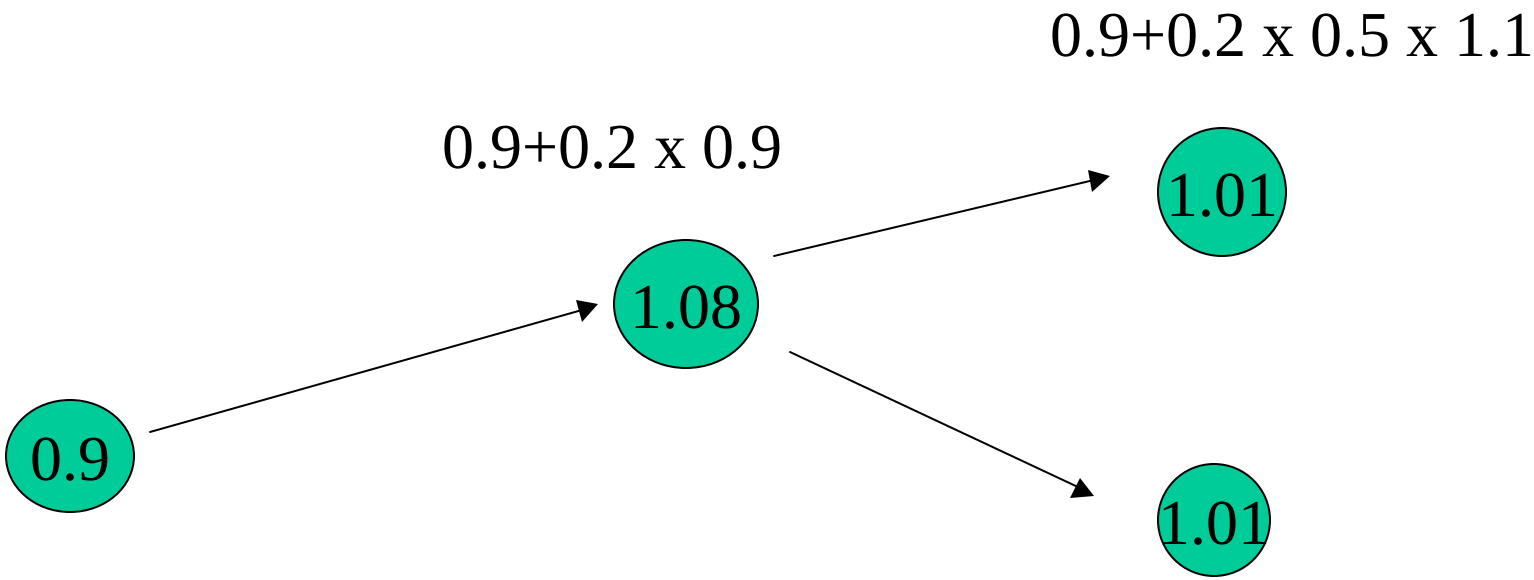


•4 votes in system: allocate 20% of vote, redistribute 80% of each, plus the lost votes from leaf nodes = 3.6 votes

PageRank: round 2



PageRank: round 3



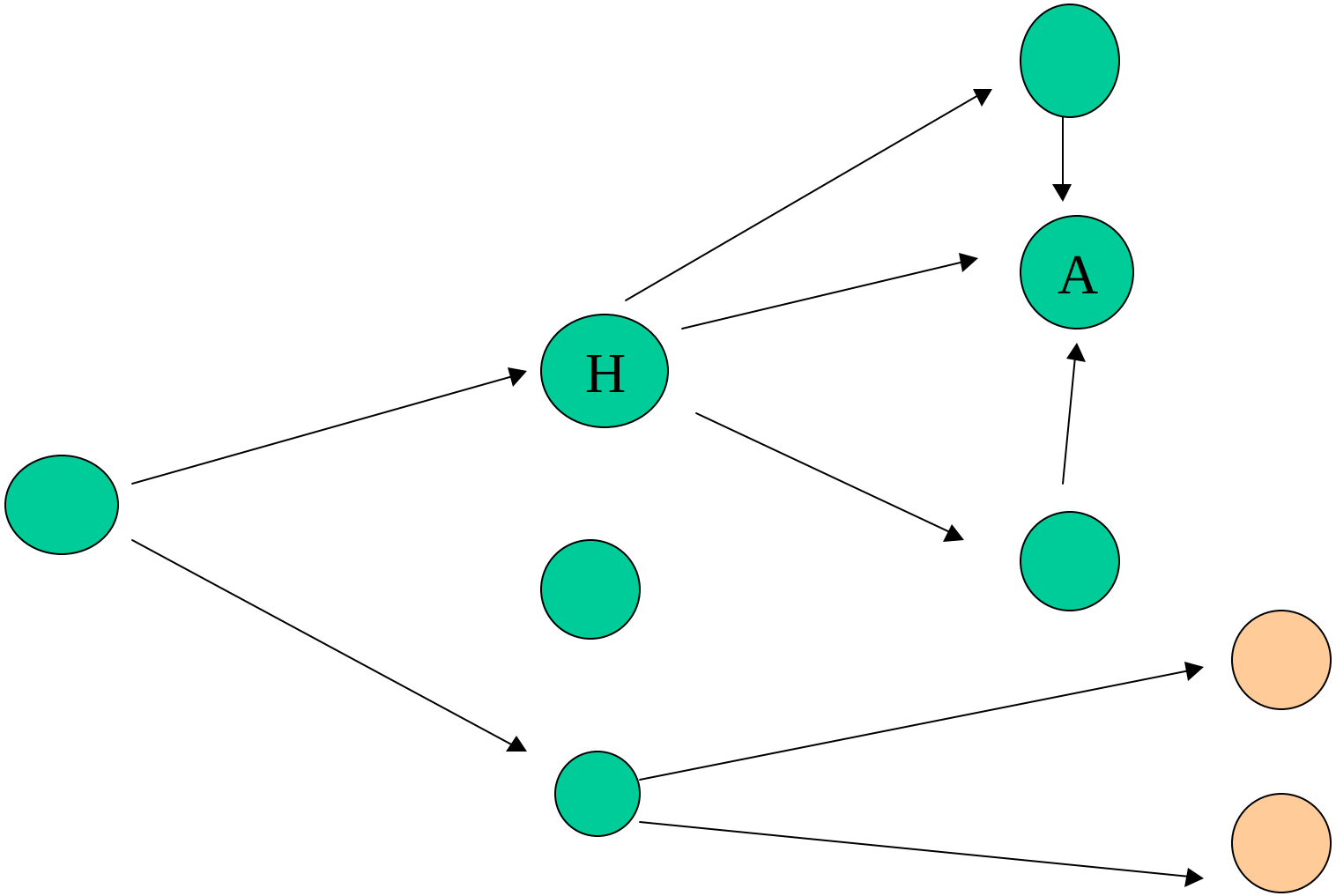
PageRank summary

- The pages that get the highest PageRank are those that are linked to by many pages or by important pages
- Spammers try to exploit this by creating dummy sites to link to their main sites

Kleinberg's HITS

- Also uses link structures, but also uses page content to identify pages that are useful for a coherent topic on the web
- An *Authority* is a page that is linked to by many other pages from the same topic
- A *Hub* is a page that links to many pages from the same topic

Hubs and authorities



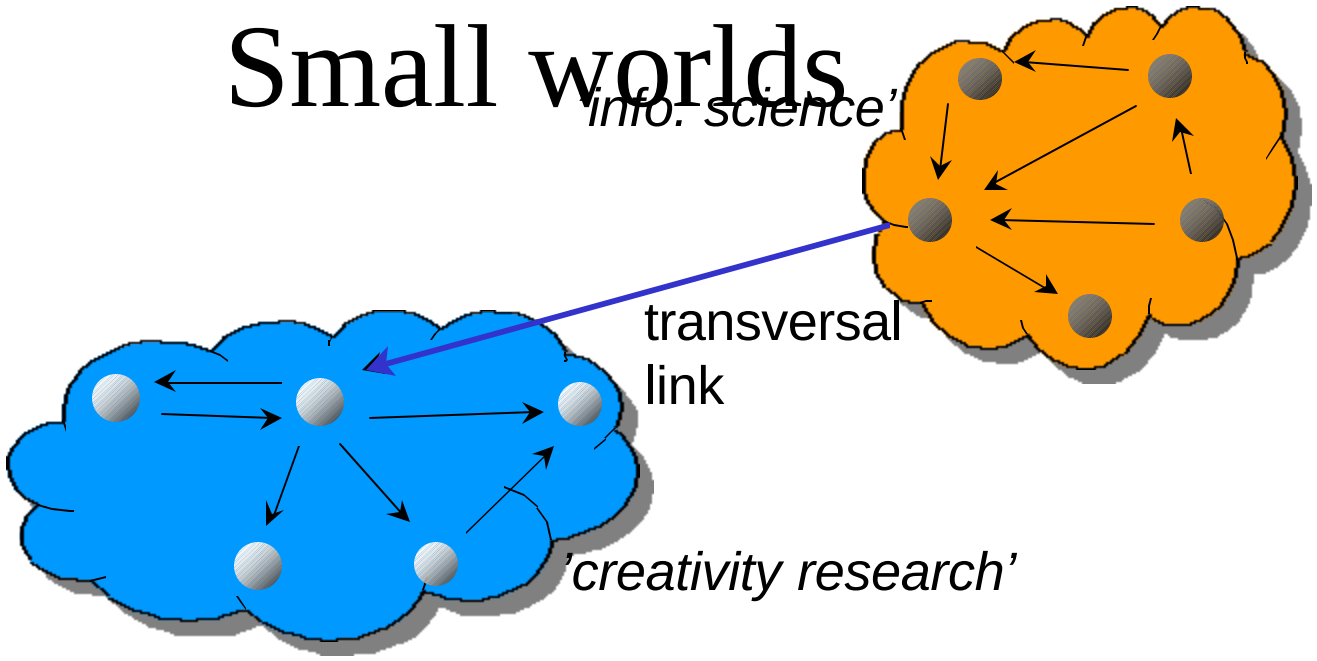
The HITS algorithm

- Another iterative algorithm
- Each page has a hub value and an authority value
- Unlike PageRank, is topic specific, and potentially needs to be recomputed for each user query

Link Algorithms - Overview

- The success of HITS and PageRank indicates the importance of links as a new information source
- More needs to be known about patterns of linking
- But there is still no hard evidence that link approaches work – academic paper report unscientific experiments or inconclusive results

Small worlds



- short cuts or 'weak ties' between otherwise 'distant' web clusters (e.g., subject domains, interest communities)

© Lennart Björneborn 2001-2002

Part 3:
Analysing University Link Structures

Information science approaches

Why analyse university link structures?

- Analogies with citation studies
- Ensure that the Web is efficiently used for research communication
- Identify trends in informal scholarly communication
- Suggest improvements in search tools
- Exploratory research: the Web is important and a valid object for scientific study

Methodologies: Data collection

- Web crawler
- AltaVista advanced queries

```
host:wlv.ac.uk AND link:albany.edu
```

- AllTheWeb advanced queries
- Google
  - Does not support same level of Boolean querying

AltaVista found 2 results About

Using the Net

Links. Site Map University Learning and Teaching in Higher Education Teaching Humanities and Social Sciences Teaching Politics Higher Education in Europe Feminist Pedagogy. University...

pers-www.wlv.ac.uk/~le1810/netuse.htm ▪ Translate

University of Wolverhampton, Learning Centres - School of computing and Information Studies Internet Resources Menu

University of Wolverhampton, Learning Centre Internet Resources page for SCIT

www.wlv.ac.uk/lib/systems/math-ir-reference.htm ▪ Translate

REFERENCE SOURCES

Dave's Math Tables
Includes a search engine on topics e.g. theore

Favourite Mathematical Constants

Mathematical Quotations

Newton's Method
Finds roots of equations

The Prime Page - information and additional s

Probability Theorythe logic of Science.
On-line book

Visual Dictionary of Special Plane Curves

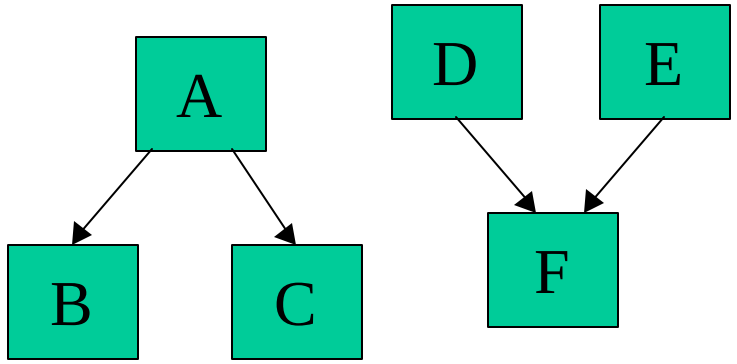
PROBABILITY THEORY: THE LOGIC OF SCIENCE

by
E. T. Jaynes
Wayman Crow Professor of Physics
Washington University
St. Louis, MO 63130, U.S.A.



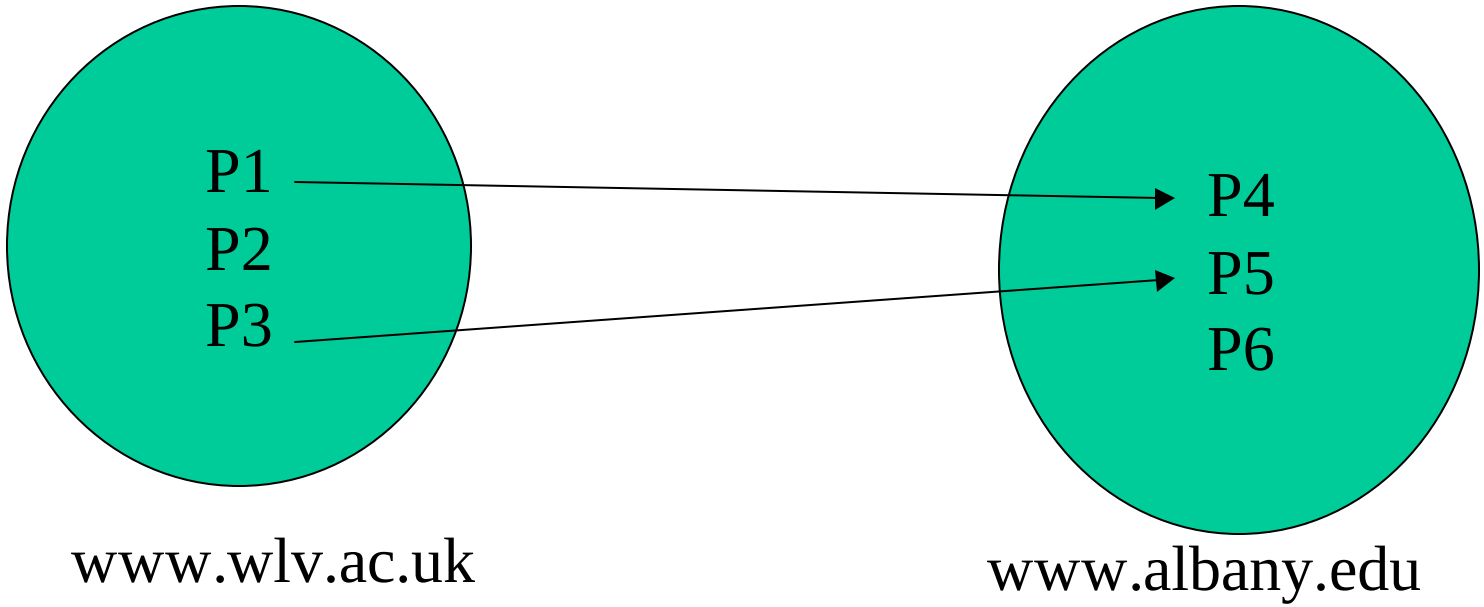
Methodologies: Data processing 1

- Link counts to target universities
  - Inter-site links only
- Colink counts
  - B and C are colinked
- Couplings
  - D and E are coupled



Methodologies: Data processing 2

- Alternative Document Models
  - E.g. count links between domains (ignoring multiple links) instead of pages



Methodologies: Data analysis

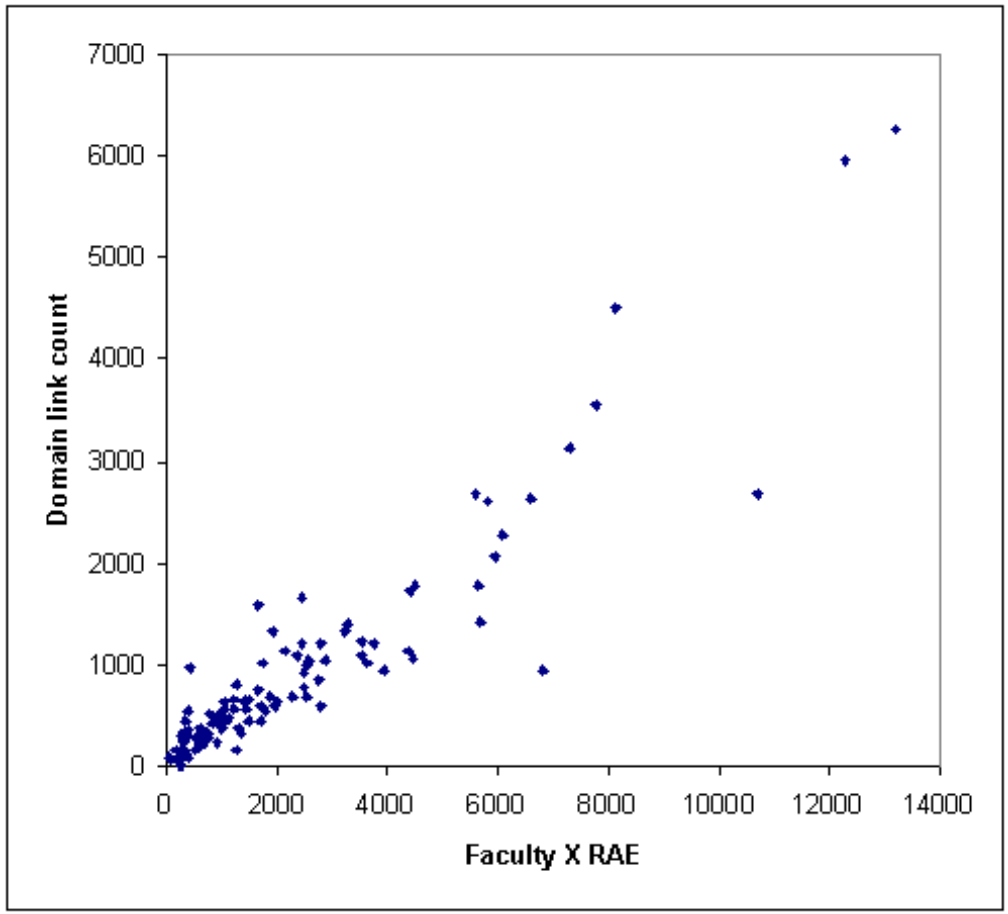
- Statistical techniques for evaluating results
  - Correlation with known research performance measures
  - Factor analysis, Multi-Dimensional Scaling, Cluster analysis for patterns
- Simple graphical techniques
- Techniques from Communication Networks research / Geography

Results section 1 – Patterns of links between university Web sites

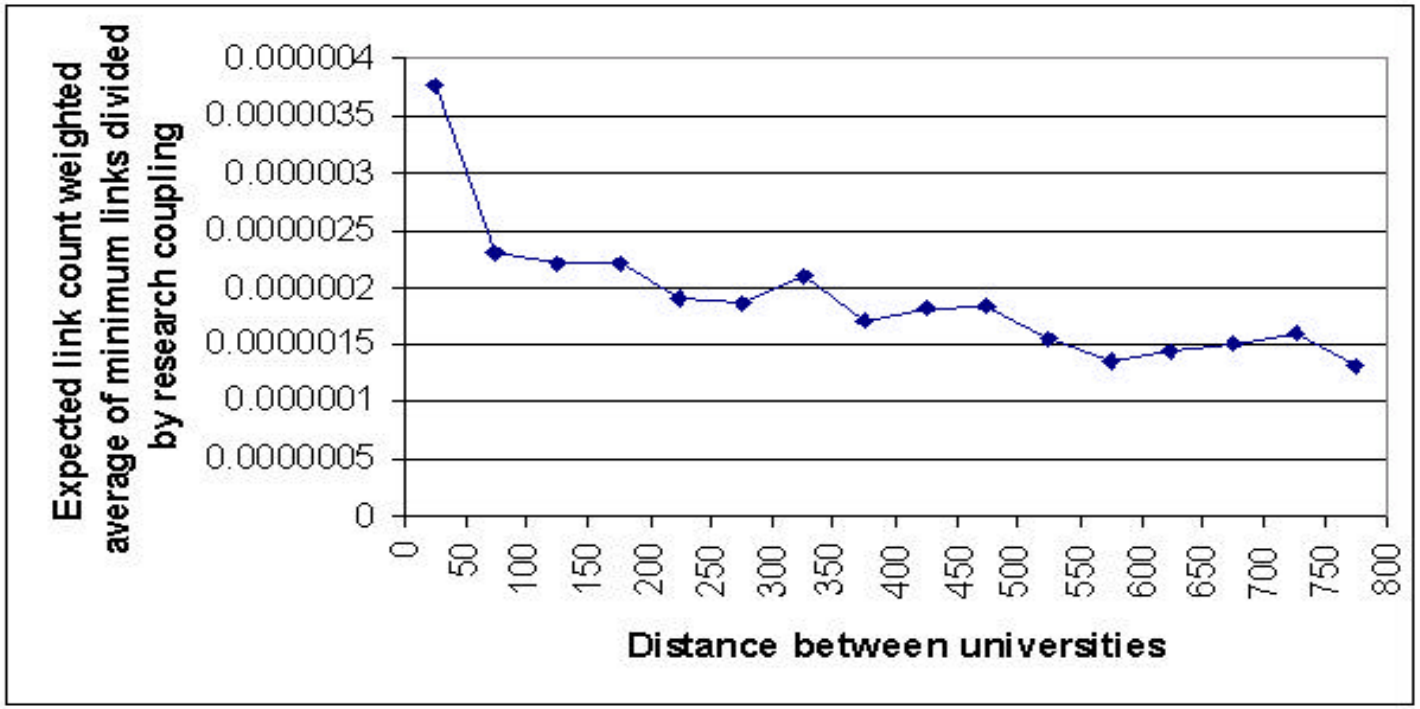
Results 1: Links associate with research

- Counts of links to universities within a country can correlate significantly with measures of research productivity

Links to UK universities counted by domain



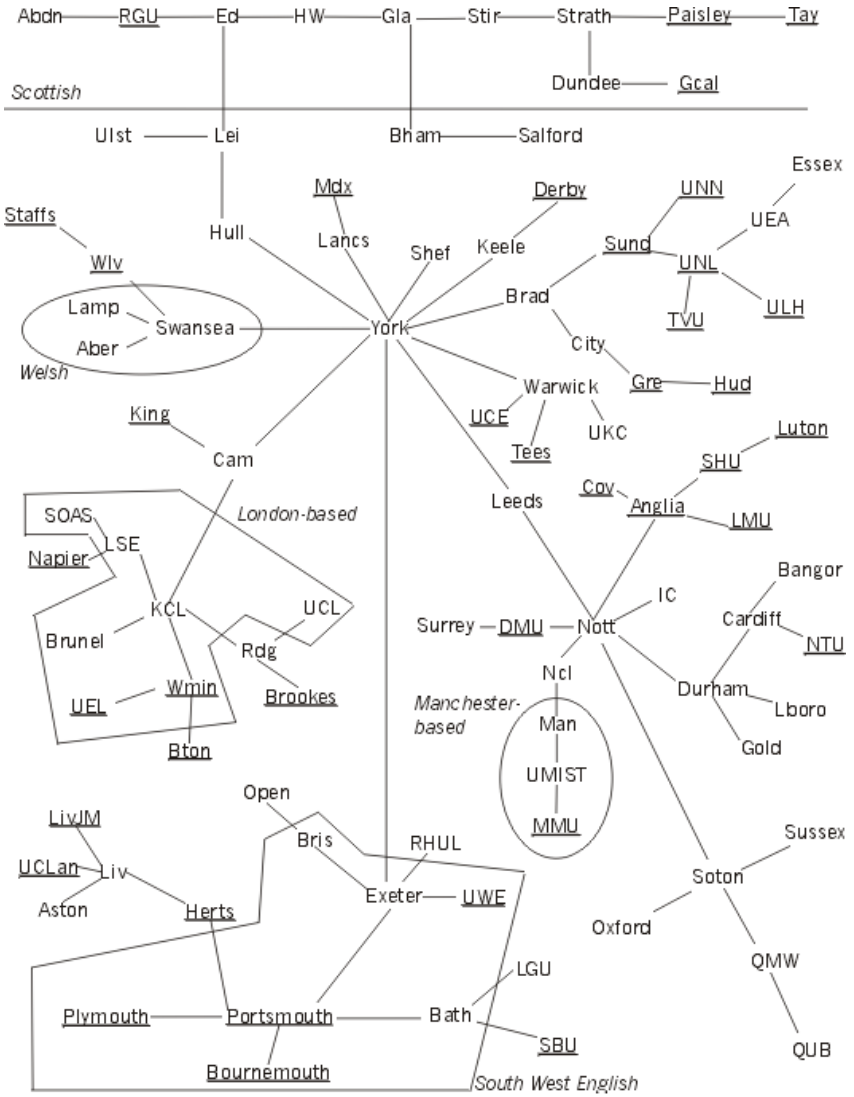
Results 2: Links between universities in a country can be



Results 3: Universities cluster by geographic region

- This is clearest for Scotland but also for other groupings, including Manchester-based universities
- Coherent clusters are difficult to extract because of overlapping trends

A pathfinder network of UK university interlinking with geographic clusters indicated



Results section 2: Links and subject areas

Results 4: Links to departments associate with research

- In the US, links to chemistry and psychology departments from other departments associate with total research impact
- No evidence of a significant geographic trend
- Disciplinary differences in the extent of interlinking: history Web use is very low

{Research with Rong Tang}

Results 5: Links for precision, colinks and couplings for recall

- For the UK academic Web, about 42% of domains connected by links alone are similar, and about 43% connected by links, colinks and couplings
- But over 100 times more domains are colinked or coupled than are directly linked
- Colinks and couplings can help the task of finding additional subject-based pages

Results 6: Most links are only loosely related to research

- A random sample of links between UK university sites revealed over 90% had some connection with scholarly activity, including teaching and research.
- Less than 1% were equivalent to citations

Results section 3: International academic links

Results 7: Linguistic factors in EU communication

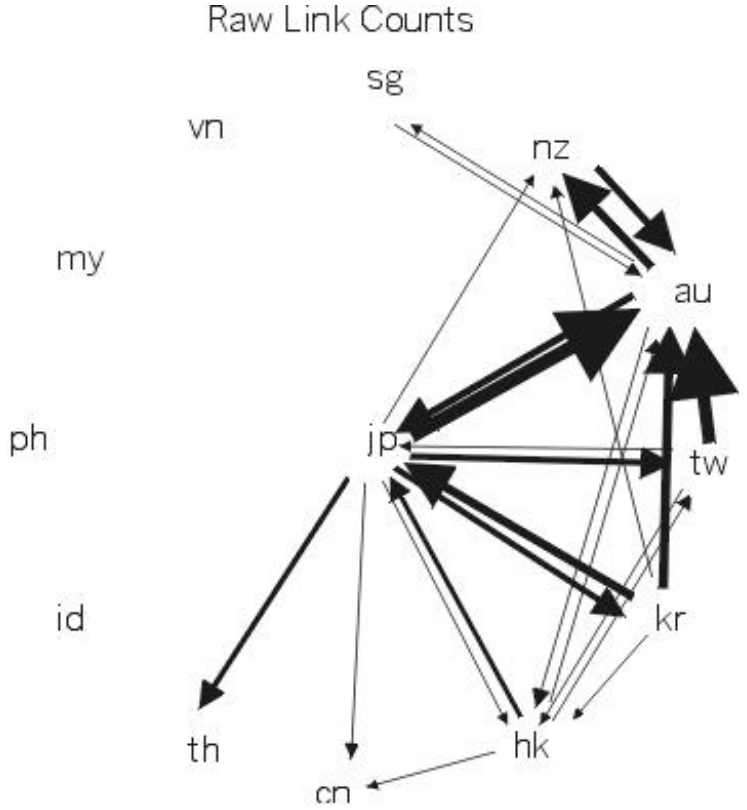
- English the dominant language for Web sites in the Western EU
- In a typical country, 50% of pages are in the national language(s) and 50% in English
- Non-English speaking extensively interlink in English

{Research with Rong Tang}

Results 8: Can map patterns of international communication

Counts of links between Asia-Pacific universities are represented by arrow thickness.

{Research with Alastair Smith, VUW, NZ}

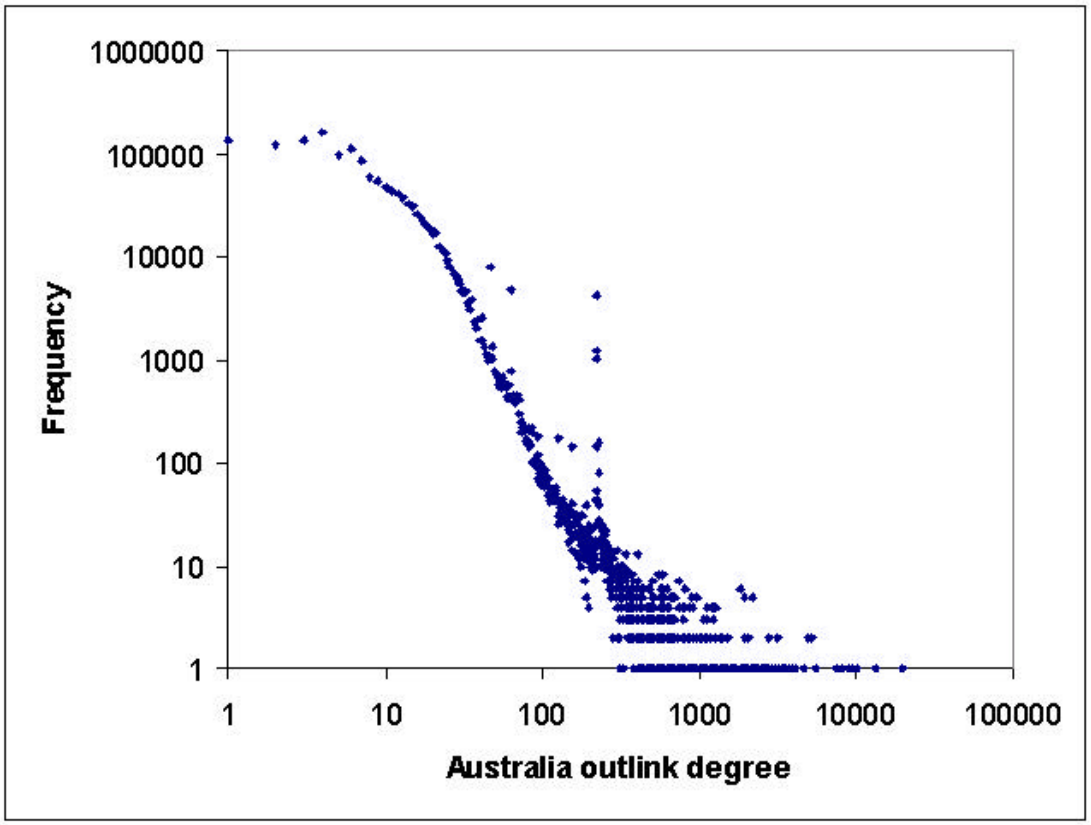


Results section 4: The topology of national academic Webs

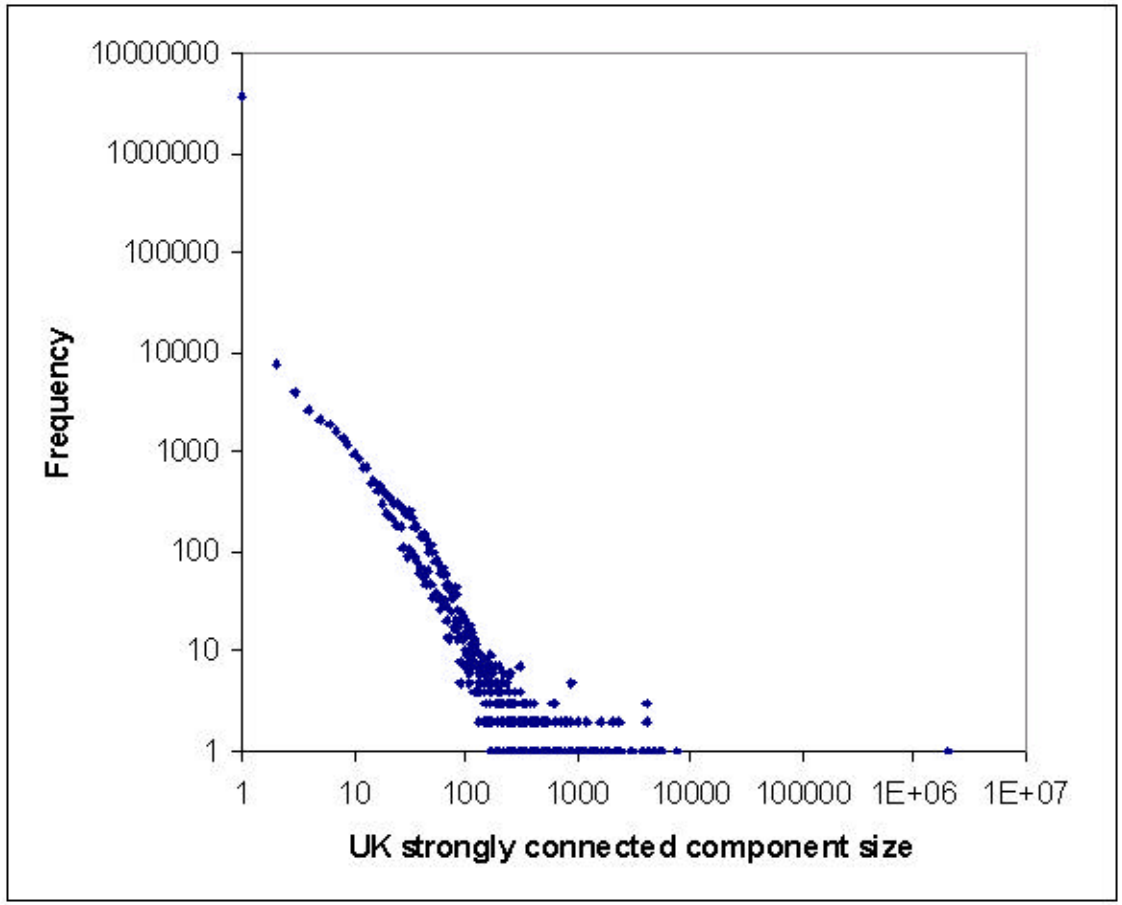
Results 9: “Power laws” in the Web

- Academic Webs have a topology dominated by power laws, including
  - Counts of links to pages (inlink counts)
  - Counts of links to pages (outlink counts)
  - Groups of interconnected pages
    - Directed component sizes
    - Undirected component sizes

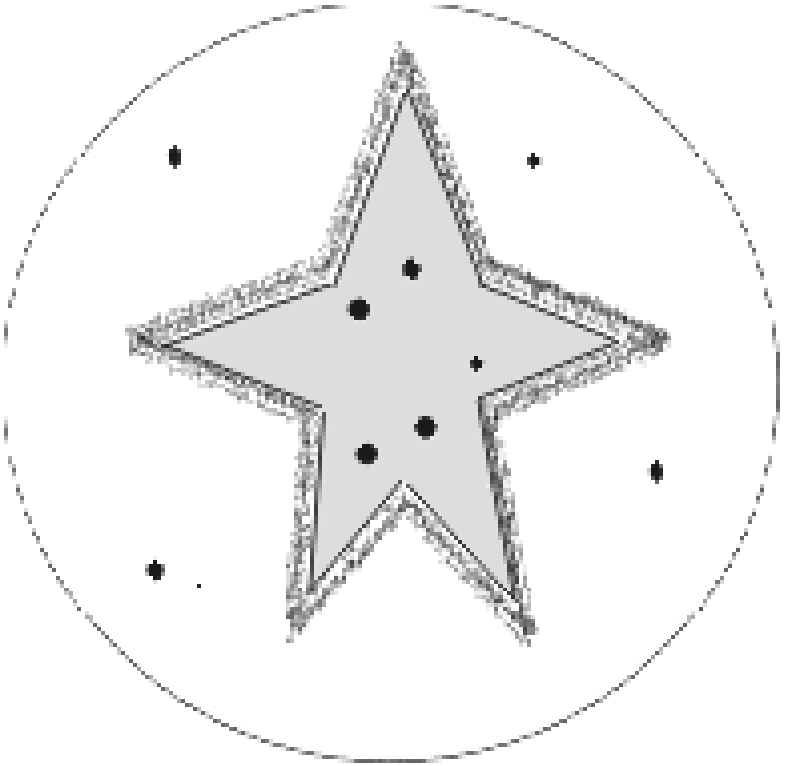
Results 9: “Power laws” in the Web



Results 9: “Power laws” in the Web



Results 10: Academic Web topology



A mess!

The future

- Results of research leading into:
  - Improved Web-related policy making
  - Improved Web information retrieval algorithms
  - Improved understanding of informal scholarly communication on the Web
    - More effective use of the Web by scholars, e.g. via PhD training