Bibliometrics of the Web

Ray R. Larson

Professor

School of Information Management and Systems

University of California, Berkeley

Overview

- Web Characteristics
- Bibliometrics/Webometrics
- Uses of Bibliometrics (link analysis)
- New Approaches and New Uses

Overview

- Web Characteristics
- Bibliometrics/Webometrics
- Uses of Bibliometrics (link analysis)
- New Approaches and New Uses

Web Characteristics

- The WWW differs radically from all previous information systems, lacking any central imposition of control and the ability for anyone or any group to develop sites (or a ”web presence”)
- The WWW is an ”’ecological’ information system” with ”
  - self-organized clusters and communities
  - islands of order + oceans of topical diversity
  - local actions → global consequences (e.g. small-world phenomena)
  - ’3D’ = distributed + diversive + dynamic”

» Bjorneborn *Small-world link structures on the Web* 2002

Web Size and Links

- Estimated 3.5-5 Billion ”visible” web pages
  - Not including over 8Tb in the ‘Deep Web’
- Estimated 35-50 Billion links
  - Dynamic link structures
    - Links are created, modified and destroyed by millions of individuals, institutions, companies, programs, etc.
    - No real typology for web links -- there is an infinity of possible reasons for linking behaviour including similarity and other topical relations, personal preferences or interests, navigatioal means
  - Dynamic adaptation and removal of links
    - may reflect changes in web constructors’ knowledge, ideas, interests, contacts, etc.
    - may reflect cultural and social currents, e.g., diffusion of scientific ideas

Overview

- Web Characteristics
- Bibliometrics/Webometrics
- Uses of Bibliometrics (link analysis)
- New Approaches and New Uses

Bibliometrics

- Bibliometrics is concerned with analysis of the characteristics of published literature including:
  - Studies of dispersion of literature on various topics
  - Statistical analyses of content types, references, etc.
  - Citation and co-citation studies within and across particular disciplines

Bibliometrics and Webometrics

- Larson (1996) applied co-citation analysis to the WWW
- Webometrics (Almind and Ingwersen 1997) is the application of bibliometric methods to the WWW
- Webometrics is similar to traditional informetric and scientometric studies, applying common bibliometric methods.
  - counts and content analysis of web pages are like traditional publication analysis
  - counts and analyses of outgoing links from web pages (*outlinks*) and links pointing to web pages (*inlinks*) are very like reference and citation analyses
    - However circular links (where a page references itself or a part of itself) may not fulfil the same role in pre-web published literature

Citation Analysis

- **Citation analysis** was developed in information science as a tool to identify core sets of articles, authors, or journals of particular fields of study.
- **Cocitation analysis** has been used to map the topical relatedness of clusters of authors, journals or articles (see H. White 1981)
- It can provide a mapping of the intellectual structure of a disciple, showing significant clustering of topically related authors
  - Shown to be effective in a broad range of disciplines, ranging from author cocitation analysis of scientific subfields to journal cocitation analysis of Economics and Marine Sciences

- Some researchers suggest that analyses of the full Web are no longer possible:
  - Such studies, attempting to characterize the "whole Web" are not feasible any more. …the current estimated size of the indexable Web is around 2,200 million pages *(Judit Bar-Ilan,Scientometrics 50(1) 2001)*
- This doesn't take into account use of new resources (such as Grids to harvest and process the data)

Overview

- Web Characteristics
- Bibliometrics/Webometrics
- Uses of Bibliometrics (link analysis)
- New Approaches and New Uses

Webometric analysis

- Link analysis
  - Use in search engines
- Co-citation analysis
  - Discovering web structures

Search Engine Ranking: Link Analysis

- 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
  - References: Page et al. 98, Kleinberg 98

Ranking: Link Analysis

- Why does this work?
  - The official Toyota site will be linked to by lots of other official (or high-quality) sites
  - The best Toyota fan-club site probably also has many links pointing to it
  - Less high-quality sites do not have as many high-quality sites linking to them

Ranking: PageRank

- Google uses the PageRank
- We assume page A has pages T1...Tn which point to it (i.e., are citations). The parameter d is a damping factor which can be set between 0 and 1. d is usually set to 0.85. C(A) is defined as the number of links going out of page A. The PageRank of a page A is given as follows:
- PR(A) = (1-d) + d (PR(T1)/C(T1) + ... + PR(Tn)/C(Tn))
- Note that the PageRanks form a probability distribution over web pages, so the sum of all web pages' PageRanks will be one

PageRank

- Similar to calculations used in scientific citation analysis (e.g., Garfield et al.) and social network analysis (e.g., Waserman et al.)
- Similar to other work on ranking (e.g., the hubs and authorities of Kleinberg et al.)
- Computation is an iterative algorithm and converges to the principle eigenvector of the link matrix

Early Work in Webometrics

- Larson (1996) used an adaptation of traditional co-citation analysis to examine the “intellectual structure of Cyberspace”
- This involved application of traditional co-citation analysis methods to the Web environment
- The assumption was that citation practice in the WWW environment served some of the same intellectual purposes as citation in the print world

Traditional (Drexel-style) ACA

- Selection of authors,
- retrieval of co-citation frequencies,
- compilation of a raw co-citation matrix,
- conversion to a correlation matrix,
- multivariate analysis of correlation matrix (using principle components analysis, cluster analysis, and multidimensional scaling),
- interpretation: author's names on the maps can be translated into subject terms naming the clusters. Clusters and the dimensions (axes) on which they are placed reveal the intellectual structure of a field. *McCain (1990)*

Example -- Geographic Sites

- Methodology
  - Search submitted to AltaVista (using the advanced search mode) was to find ``link:pubweb.parc.xerox.com/map AND link:xtreme.gsfc.nasa.gov", that is, to find a set of WWW documents containing links to both the Xerox Map browser, and the home page for NASA's AVHRR (Advanced Very High Resolution Radiometer) remote sensing projects.
  - Resulted (after throwing out irrelevant pages) in 43 sites that *co-cited* these sites

Example Analysis

- The 43 sites contained 7209 outlinks to other sites, of these 332 existed in at least 3 of the 43 sites. These were examined and 34 “major” sites were selected as the core set for analysis

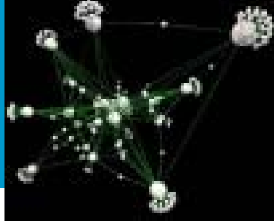
Selected Sites

S1	The American Physical Society (http://aps.org/)
S2	NSF Geosciences Unidata Integrated Earth Information Server (IEIS) (http://atm.geo.nsf.gov/)
S3	Earth Observing System Home Page (http://eos.nasa.gov/)
S4	WeatherNet (http://cirrus.sprl.umich.edu/wxnet/)
S5	NASA-Goddard Climate and Radiation Branch (http://climate.gsfc.nasa.gov/)
S6	Center for Remote Sensing and Spatial Analysis (http://deathstar.rutgers.edu/welcome.html)
S7	EcoWeb (http://ecosys.drdr.virginia.edu/EcoWeb.html)
S8	EROS (Earth Resources Observation Systems) Data Center (http://edcwww.cr.usgs.gov/eros-home.html)
S9	The EnviroWeb– A Project of the EnviroLink Network (http://envirolink.org/)
S10	Live Access to Climate Data (http://ferret.wrc.noaa.gov/ferret/main-menu.html)
S11	Global Change Master Directory (http://gcmd.gsfc.nasa.gov/)
S12	Earth Science Division-Home Page (http://geo.arc.nasa.gov/esd)
S13	Information Center for the Environment (ICE), UC Davis (http://ice.ucdavis.edu/)
S14	Climate Prediction Center (http://nic.fb4.noaa.gov/)
S15	Xerox PARC Map Viewer (http://pubweb.parc.xerox.com/map)
S16	Public Use of Remote Sensing Data (http://rsd.gsfc.nasa.gov/rsd/)
S17	SeaWiFS Project - Homepage (http://seawifs.gsfc.nasa.gov/)
S18	USGS Geo Data (http://sun1.cr.usgs.gov/doc/edchome/ndcdb/ndcdb.html)
S19	Planet Earth Home Page (http://www.nosc.mil/planet_earth/info.html)
S20	GeoWeb Home Page (http://wings.buffalo.edu/geoweb/)
S21	Earthquake Information (http://www.civeng.carleton.ca/cgi-bin/quakes)
S22	ESRG (Earth Space Research Group) Gateway (http://www.crseo.ucsb.edu/esrg.html)
S23	EOS Volcanology (http://www.geo.mtu.edu/eos/)
S24	Mission to Planet Earth (http://www.hq.nasa.gov/office/mtpe/)
S25	Global Warming Update (http://www.ncdc.noaa.gov/gblwrmupd/global.html)
S26	The National Center for Geographic Information and Analysis (http://www.ncgia.ucsb.edu/)
S27	National Geophysical Data Center (http://www.ngdc.noaa.gov/ngdc.html)
S28	NOAA National Oceanographic Data Center (NODC) Home Page (http://www.nodc.noaa.gov/)
S29	National Operational Hydrologic Remote Sensing Center (http://www.nohrsc.nws.gov/)
S30	NOAA/PMEL/TAO El Nino Theme Page (http://www.pmel.noaa.gov/toga-tao/el-nino/home.html)
S31	NCAR Home Page (http://www.ucar.edu/)
S32	United States Geological Survey Home Page (http://www.usgs.gov/)
S33	Current Weather Maps/Movies (http://wxweb.msu.edu/weather/)
S34	AVHRR Pathfinder Home Page (http://xtreme.gsfc.nasa.gov/)

Example Analysis

- The 34 sites were analysed to construct a cocitation matrix
  - This stage requires the ability to search for “citing documents”, that is, those with links to the items in the core set and also the ability to conduct the many searches required (for any core set of size *N*, there are $N^2/2$ - N searches required -- one for each pair of items in the core set)
  - We developed a “robot” search program to submit these searches to AltaVista and collect the results
- The result was a cocitation matrix…

Cocitation Matrix

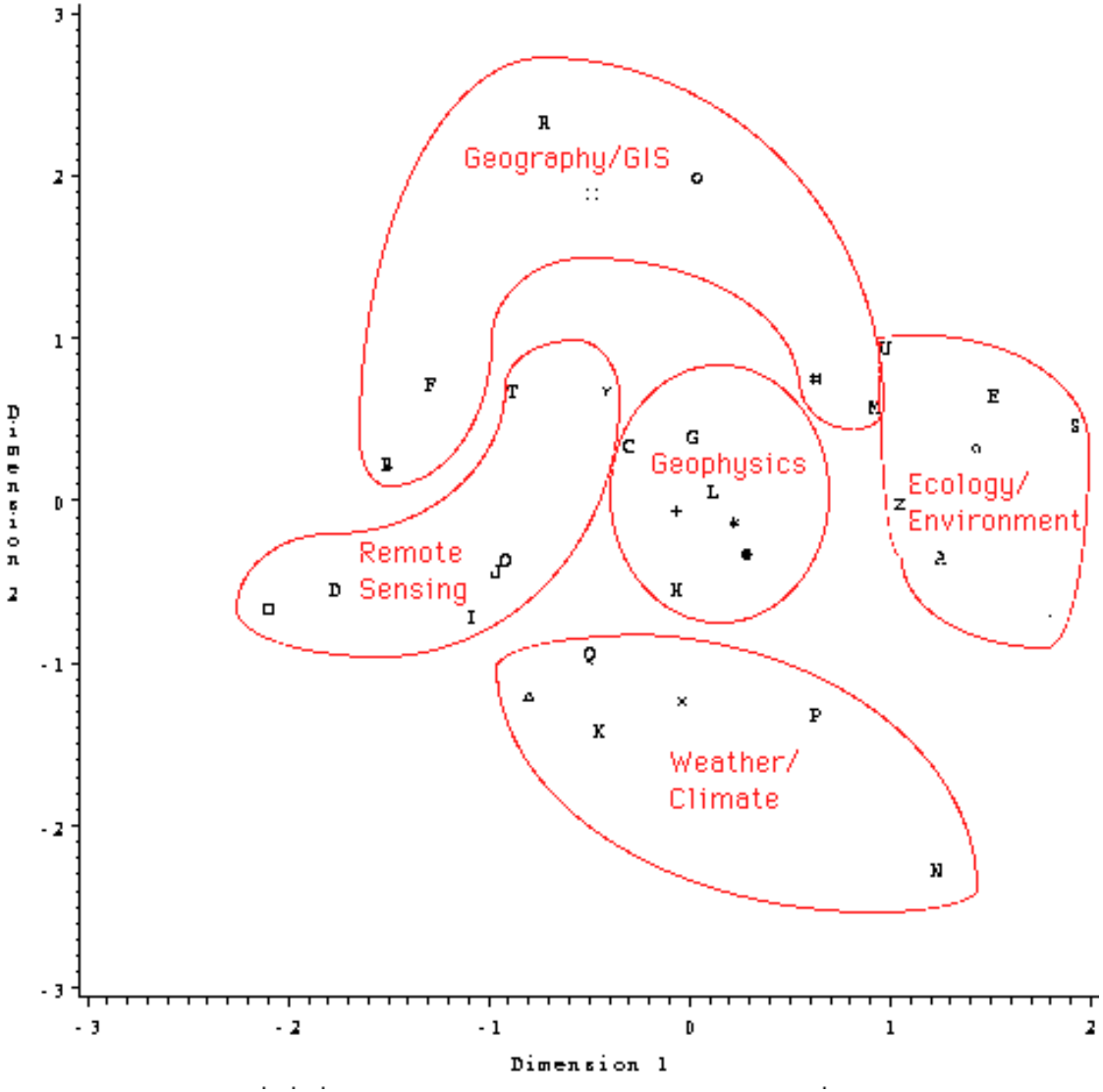


ID	Raw Cocitation Frequency																																
	J1																																
J2	96	J2																															
J3	74	128	J3																														
J4	27	161	25	J4																													
J5	41	182	91	123	J5																												
J6	4	13	24	1	5	J6																											
J7	9	63	19	10	6	8	J7																										
J8	6	16	14	12	4	5	5	J8																									
J9	127	400	113	120	138	20	700	15	J9																								
J10	13	83	20	119	31	1	19	17	75	J10																							
J11	68	200	200	71	151	25	130	26	400	97	J11																						
J12	3	3	8	2	17	2	0	3	1	2	17	J12																					
J13	72	182	91	29	121	19	200	6	1000	23	300	8	J13																				
J14	63	185	82	106	109	6	18	11	109	90	173	3	85	J14																			
J15	132	181	47	164	25	40	58	26	600	50	154	5	200	36	J15																		
J16	5	20	35	15	17	20	33	6	44	16	92	8	45	8	48	J16																	
J17	75	180	150	168	110	14	119	32	700	165	200	24	184	85	300	106	J17																
J18	1	26	7	12	6	12	0	4	13	5	22	0	8	4	48	6	11	J18															
J19	27	13	5	66	2	5	8	3	86	4	11	0	16	45	75	5	51	2	J19														
J20	6	44	22	14	5	51	20	10	47	12	40	0	78	10	142	12	23	39	7	J20													
J21	20	53	37	101	15	2	95	3	200	38	43	1	42	81	170	20	200	3	65	19	J21												
J22	1	6	4	0	2	3	1	1	2	4	7	1	4	5	3	4	4	0	0	0	1	J22											
J23	11	69	76	12	15	16	31	21	86	17	75	2	45	23	91	33	200	3	6	23	63	3	J23										
J24	5	14	74	17	31	4	3	6	15	20	62	11	15	15	16	23	73	0	9	5	6	2	10	J24									
J25	5	38	19	6	11	5	128	0	167	27	124	1	130	35	42	31	65	1	1	3	23	2	29	4	J25								
J26	55	85	75	3	36	42	2	5	59	2	106	2	88	60	58	11	18	11	4	70	2	2	6	6	2	J26							
J27	48	131	59	39	31	33	58	9	159	31	132	2	146	46	200	52	123	29	5	51	35	5	59	20	34	37	J27						
J28	80	175	137	33	84	9	54	22	300	72	200	3	101	192	62	11	400	7	6	27	111	5	37	22	24	64	106	J28					
J29	49	112	81	56	66	16	7	10	66	26	116	2	68	117	32	13	28	6	2	23	4	2	9	10	6	58	53	106	J29				
J30	6	39	4	44	28	0	1	4	33	43	54	2	8	60	16	3	78	2	2	3	11	2	6	7	17	1	12	36	13	J30			
J31	136	300	148	185	126	20	50	23	300	80	300	6	200	187	400	48	194	29	8	32	43	5	56	30	47	92	300	200	120	67	J31		
J32	70	200	99	126	63	46	52	56	600	58	200	11	200	93	600	42	300	73	44	132	108	6	100	40	27	100	157	153	96	29	145	J32	
J33	6	40	2	117	24	2	2	2	26	10	12	0	11	18	43	4	21	1	10	4	16	0	10	3	3	3	4	5	4	9	27	21	J33
J34	88	151	152	25	128	17	13	24	109	59	200	16	95	119	115	28	157	16	6	27	14	2	70	52	18	105	86	127	86	37	175	81	9

Example Analysis

- The raw cocitation matrix was converted to a correlation matrix
- Next the correlation matrix was analysed using Multidimensional Scaling
  - MDS is a class of techniques for uncovering “hidden structure” in data.
- The result of applying MDS to a set of proximities results in the output of a spatial representation of the data consisting of a geometric configuration of points, in effect a mapping of the information onto a two or three dimensional plane.

MDS Output


Geography/GIS
Geophysics
Ecology/
Environment
Remote
Sensing
Weather/
Climate
Dimension 2
Dimension 1
3
2
1
0
-1
-2
-3

Overview

- Web Characteristics
- Bibliometrics/Webometrics
- Uses of Bibliometrics (link analysis)
- New Approaches and New Uses

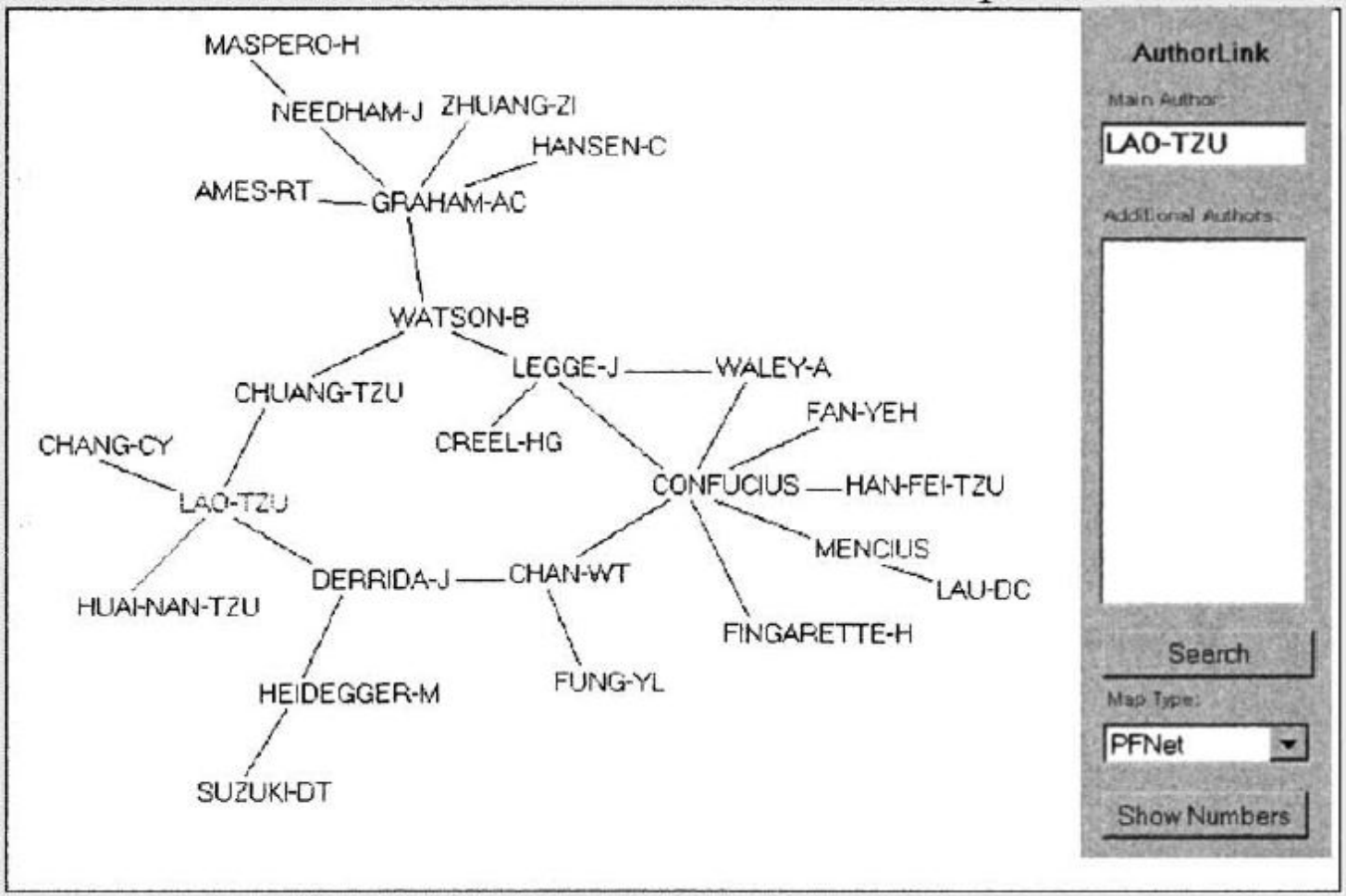
Lin, White AuthorLink

- Lin and White have developed a system for dynamic visualization of cocitation data called AuthorLink
- Provides support for analysis and visualization using the “traditional” tool, ISI’s Science Citation Index

Lin, White AuthorLink Display

Author Search: LAO-TZU Submit

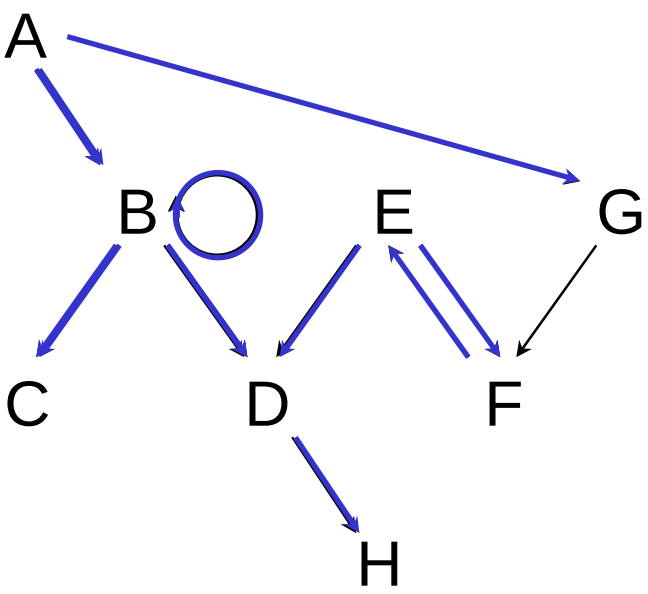
Instant Author Co-citation Map



Complex Link Analysis

- Ingwersen and his research group at the Royal Danish School of Library and Information Science are probably the central research group on Webometrics today. The following pages are derived from a presentation by Lennart Björneborn a Doctoral Student in this research group

Basic Concepts



- B has an outlink to C : ~ reference
- B has an inlink from A : ~ citation
- B has a selflink : ~ self-citation

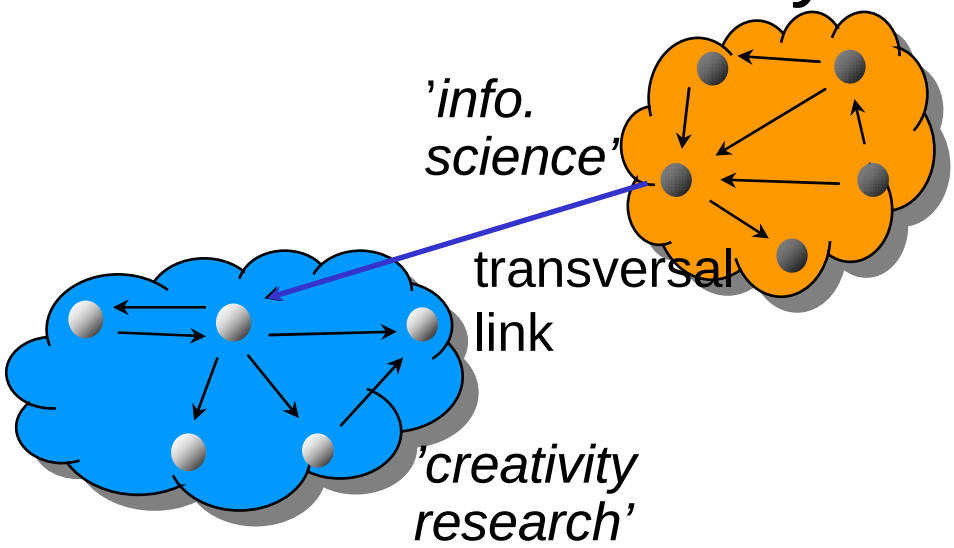
- E and F are reciprocally linked
- A is transitively linked with H via B - D
- A has a transversal link to G : short cut

co-links

- C and D are co-linked from B, i.e. shared inlinks: co-citation
- B and E are co-linking to D, i.e. shared outlinks: bibliog.coupling

transversal links

- transversal links = short cuts or ‘weak ties’ between otherwise ‘distant’ web clusters (e.g., subject domains, interest communities)
- example 1: info.sci. researcher with transversal link to creativity research



- transversal links may create small-world phenomena
  - in the shape of short distances between nodes in the Web graph
  - making the Web more strongly connected and ‘crumpled up’

’crumpled-up’ Web

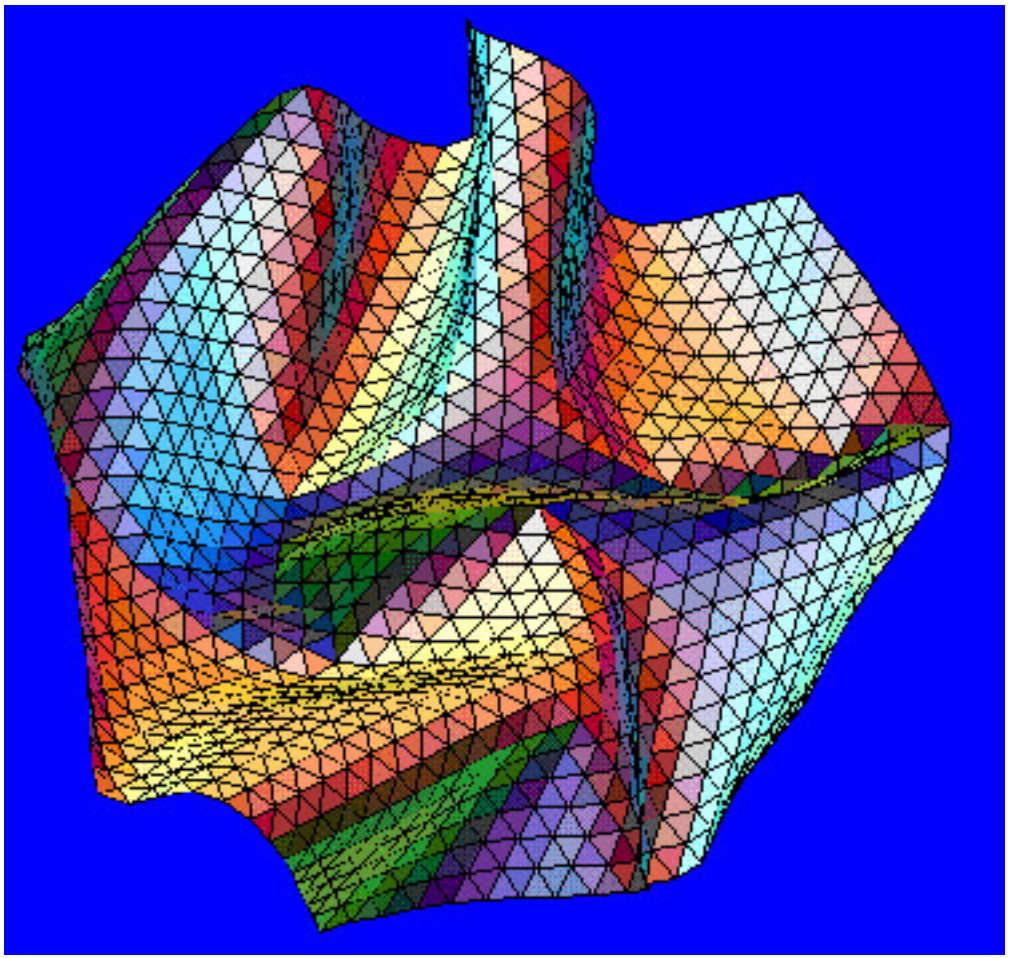
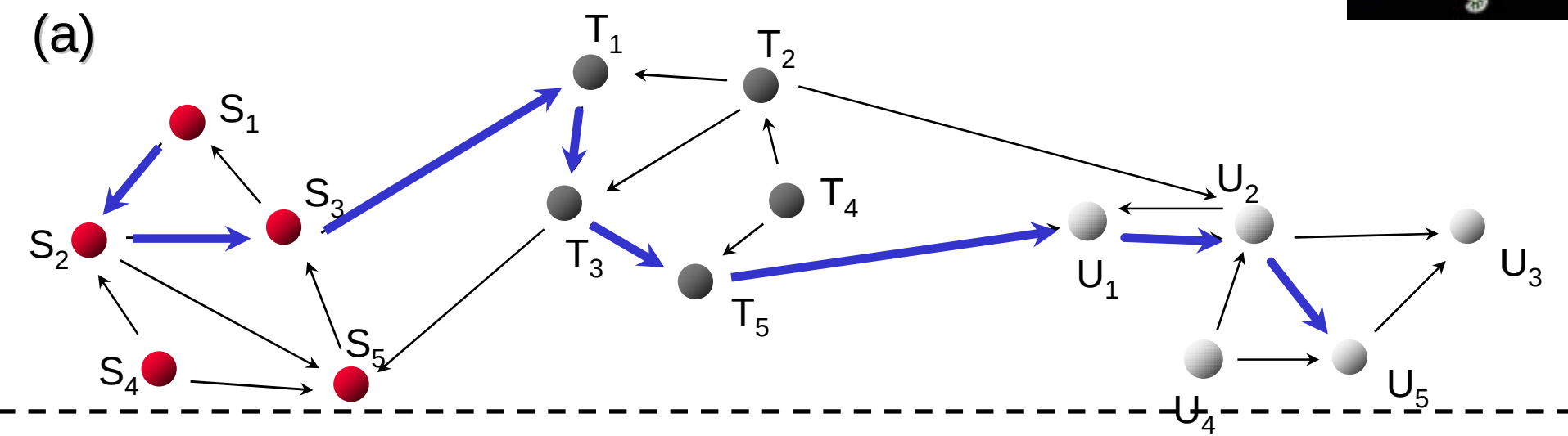
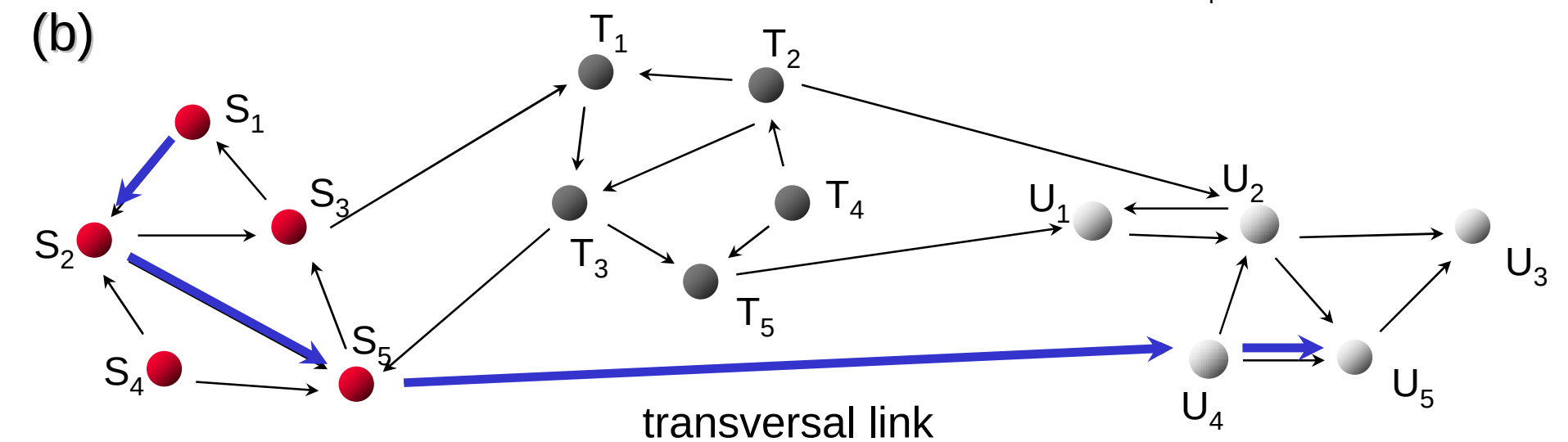


Image: http://jfi.uchicago.edu/~tten/rainbow/Crumpling/

- analogy: crumpled-up paper
- paths between selected points get shorter as opposite parts of structure are pulled together
- paper = three-dimensional
- www = billion-dimensional
- every new web outlink may interconnect with any existing and accessible web node
  - creating a new dimension in the ’crumpled-up’ Web space
- every new link = 'hook'
  - stretching and reshaping existing web network

Traversal Links





small-world networks

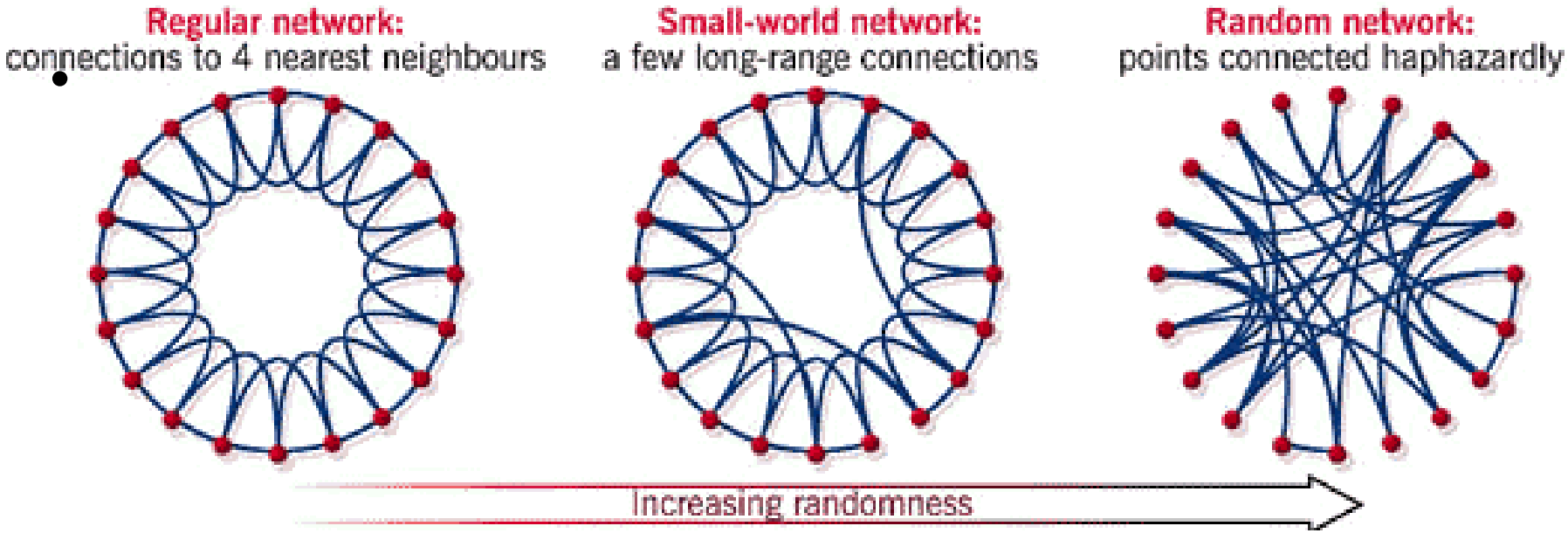
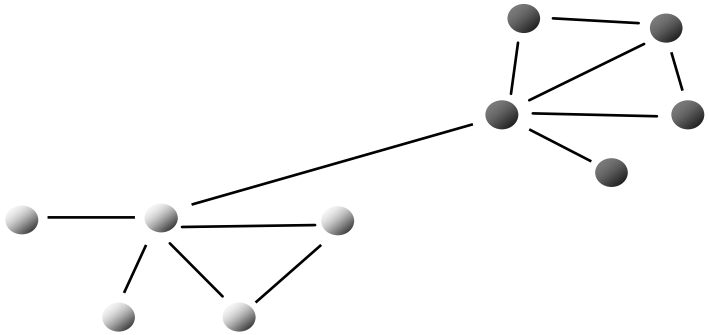


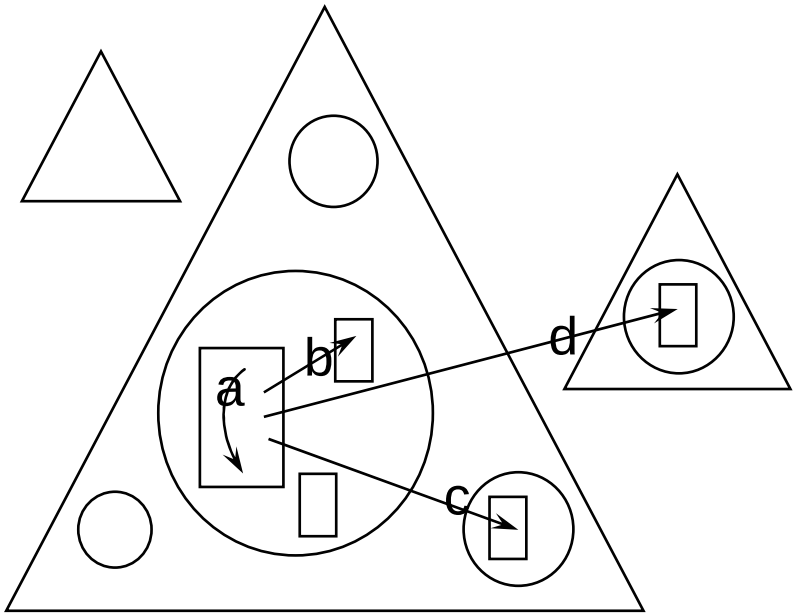
Fig.: Matthews (1998), *New Scientist*

- small-world = highly clustered + short paths [graph theory]
- small-world = order + diversity = complementarities
- short distances through short cuts between nodes in network
- small-world = short local + short global distances

Small-World phenomena



- small-world theories in social network analysis in 1960s
  - small-world phenomena in the shape of short distances between two arbitrary persons through intermediate chains of acquaintances
- ‘six degrees of separation’ (cf. play 1990, movie 1993)
  - cf. ‘Six degrees of Kevin Bacon’ (http://www.cs.virginia.edu/oracle)
- ‘small worlds’ in biological, technological and social networks: neural networks, electrical power grids, diffusion of ideas and epidemics,…
- lack of research on small-world phenomena in informational networks: WWW, citation databases, semantic networks, thesauri, ...



basic levels of web nodes

- 3 basic levels of web nodes: pages □, domains ◯, TLDs △
- different levels of selflinks and outlinks
  - a = page selflink
  - b = page outlink and domain selflink
  - c = domain outlink and TLD selflink
  - d = TLD outlink
- more levels: frames, directories, sub-domains, sub-TLDs ...

top level domains

more web node notation

top-level domain
TLD

sub-TLD

Example: .uk

.ac.uk

domain

.wlv.ac.uk

sub-dom.

.scit.wlv.ac.uk

sub-sub-domain … sub-sub-sub-…

x1.scit.wlv.ac.uk

directory

x1.scit.wlv.ac.uk/yy

sub-sub-directory

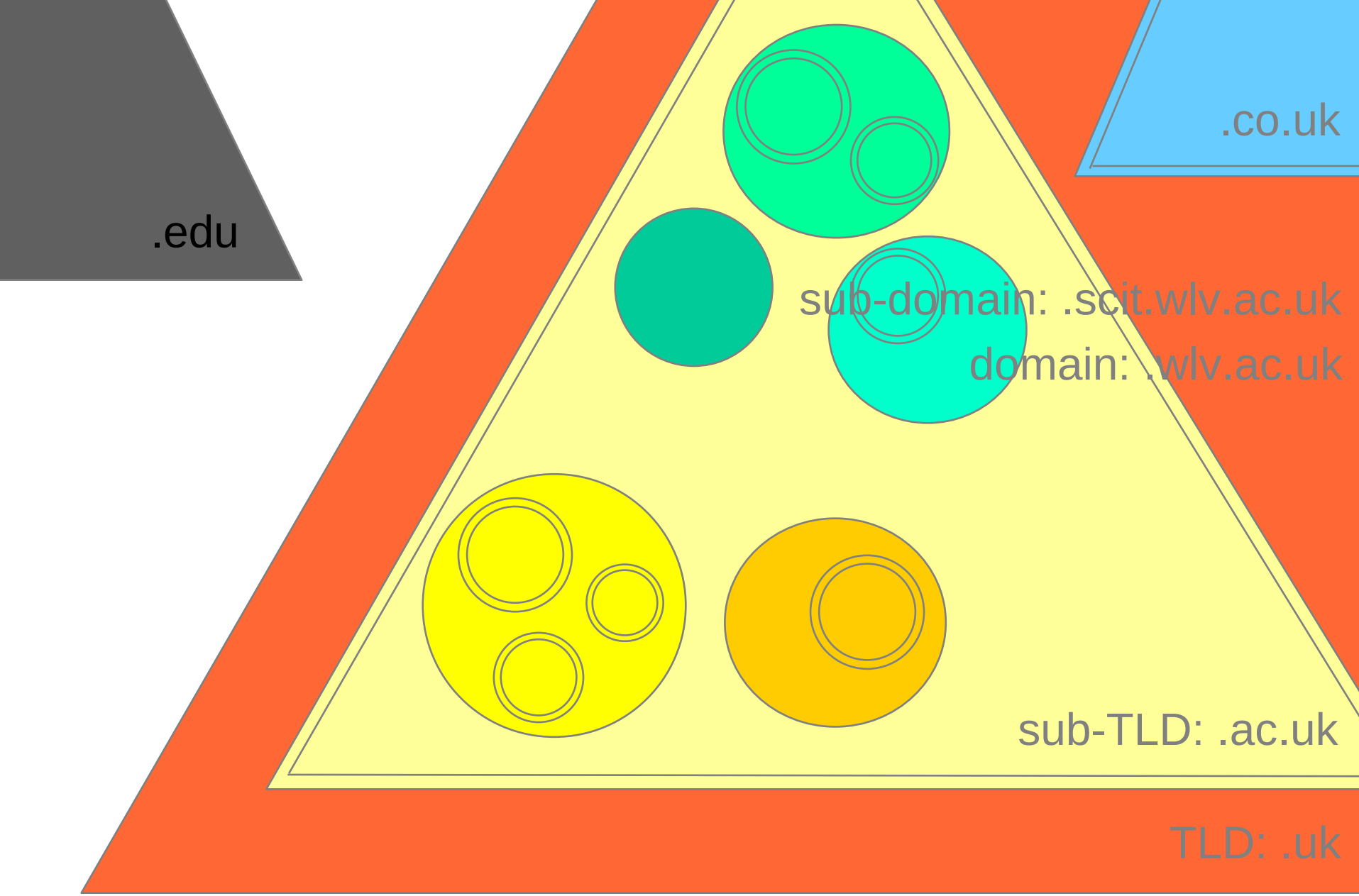
.scit.wlv.ac.uk/zz/aa/bb

.html
.pdf
.gif
. …

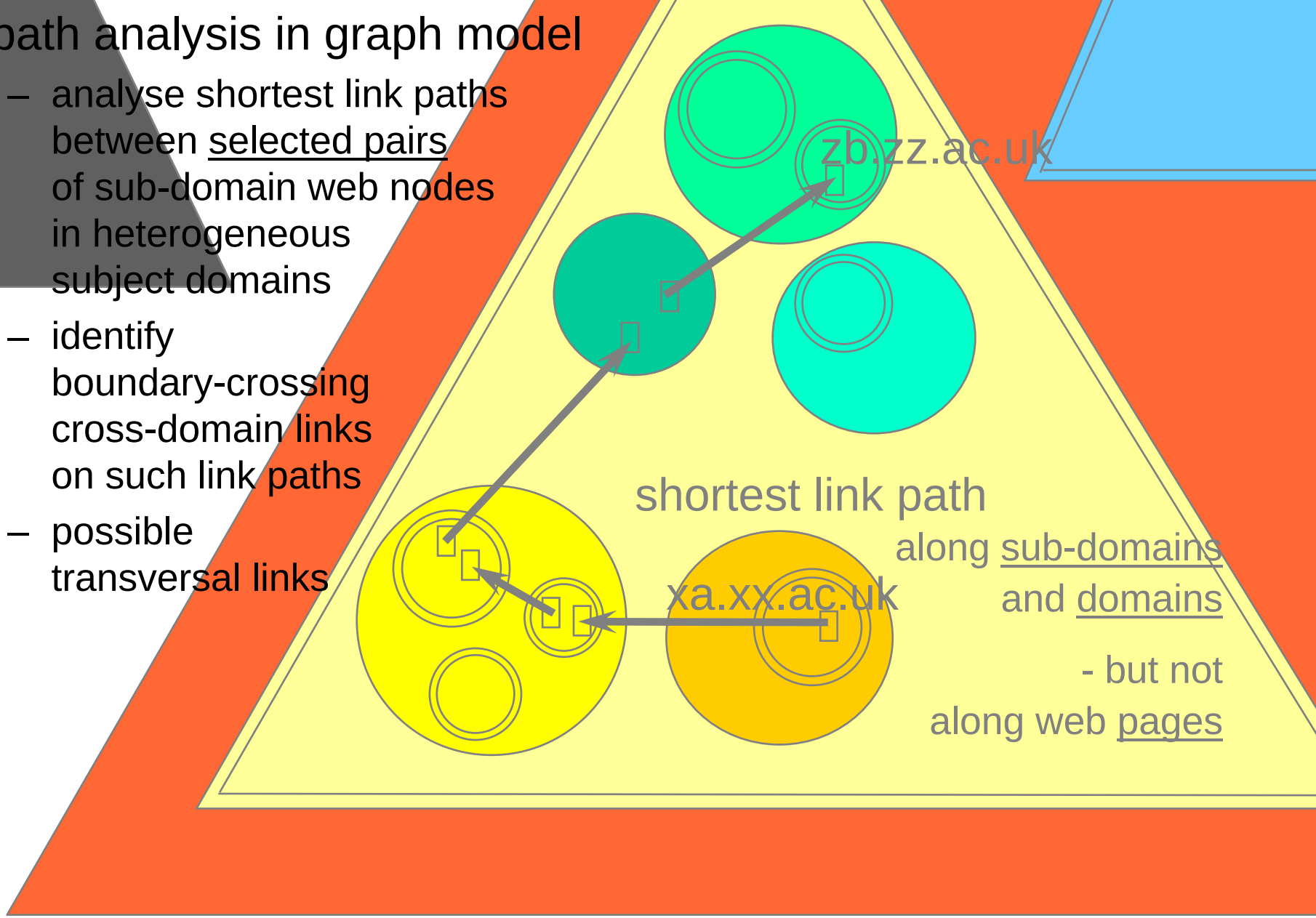
page

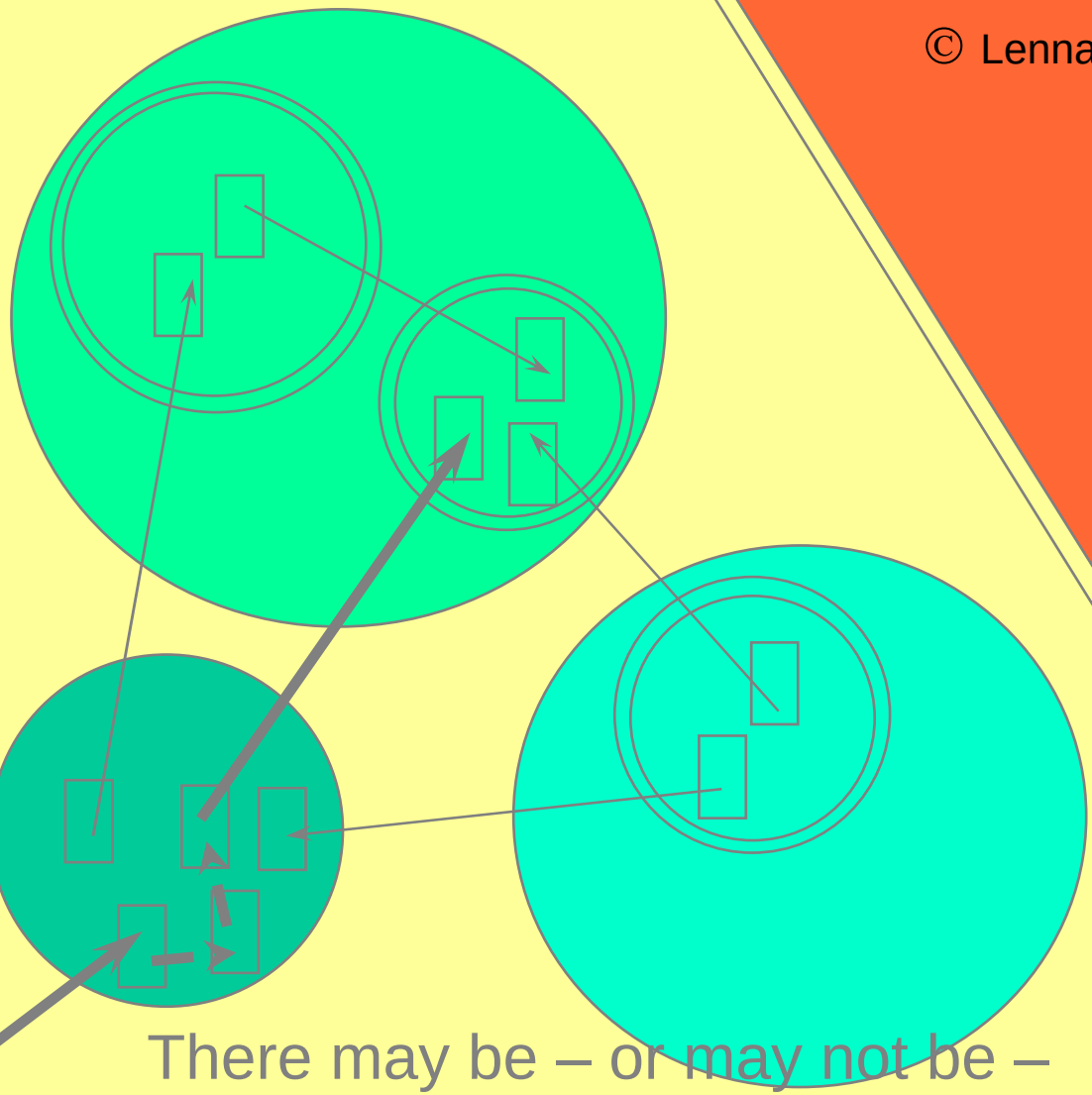
page subsection / frame

example of web node notation

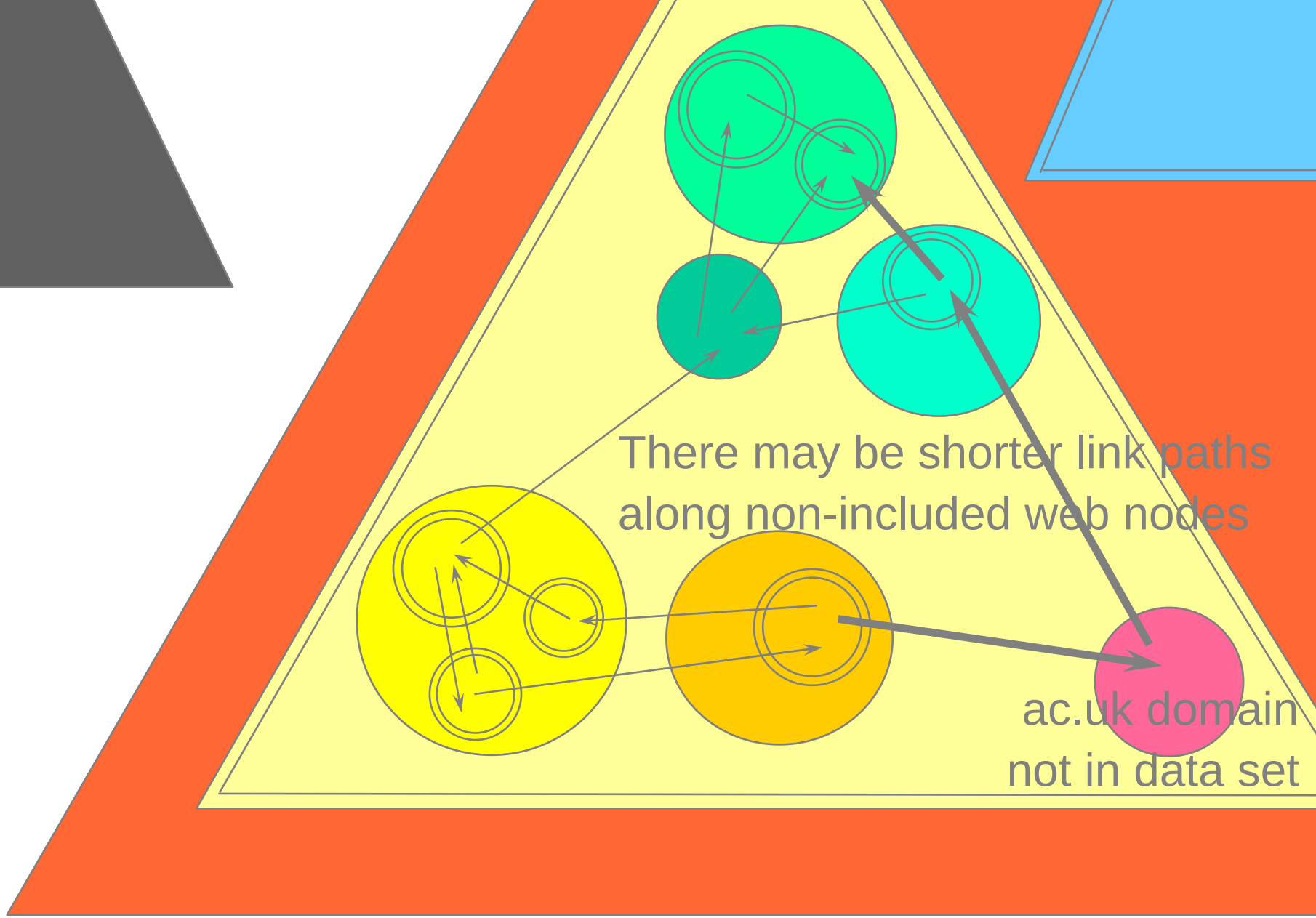


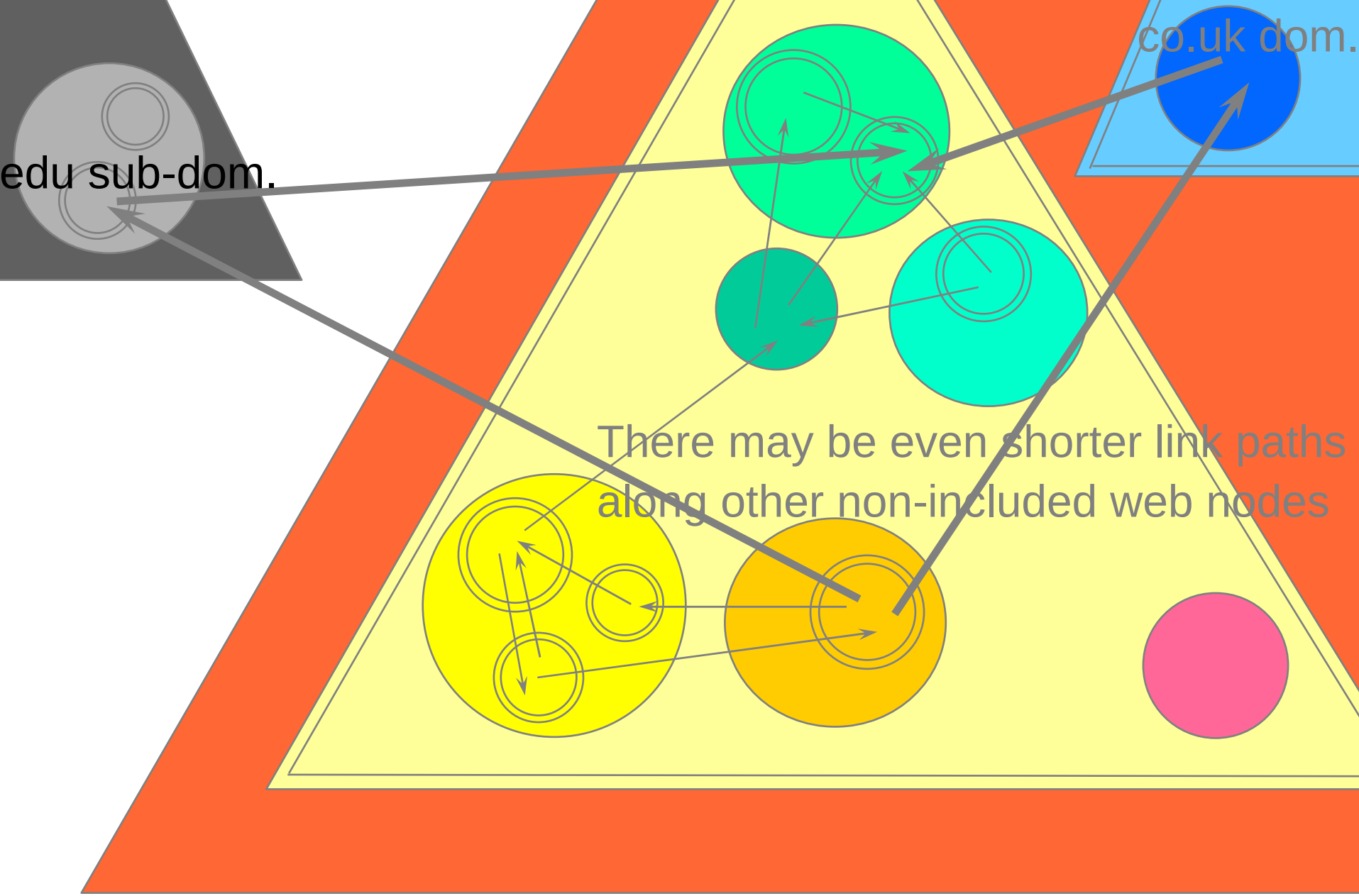
- path analysis in graph model
  - analyse shortest link paths between selected pairs of sub-domain web nodes in heterogeneous subject domains
  - identify boundary-crossing cross-domain links on such link paths
  - possible transversal links





There may be – or may not be –
local link paths connecting web pages
on the overall link path
along web domains and sub-domains

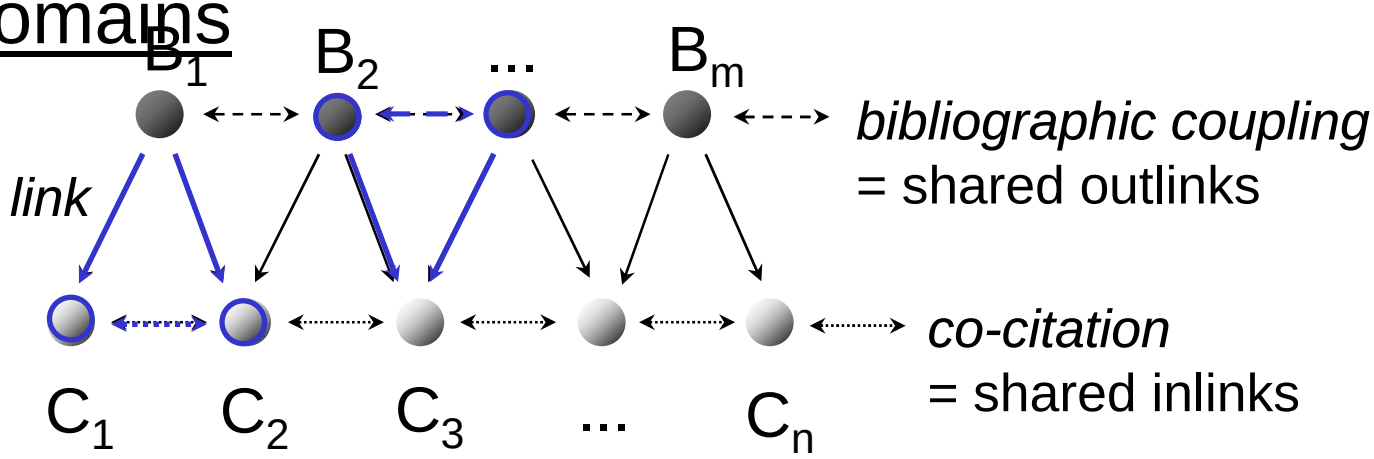




shared outlinks in link data set

Björneborn 2001-2002

- analysing shared outlinks between domains / sub-domains



- shared outlinks between web pages with bookmark lists
  - researchers' diverse interests, preferences and actions on the Web
  - may reflect emerging 'research fronts' or 'invisible colleges'
  - correlation with co-citations/bib.couplings in Science Citation Index?
- web mining 'undiscovered public knowledge' (cf. Swanson 1986)
- possible power law in distribution of shared outlinks

co-linkage chain

- case studies of transversal **co-links**

researchers' homepages

researchers' bookmark lists

Jon Kleinberg: link structures, small-world phenomena

co-citation

Frances Heylighen: cybernetics

David Chalmers: philosophy of mind

Mark Warschauer: language/literacy

Michael Barlow: linguistics

in-link

out-link

Cliff Joslyn: distributed knowledge systems

bibliographic coupling

Liane Gabora: interdiscipl. studies

Don Sorsa: education research

James Wong: computer-assisted language learning

NN

AltaVista's Advanced Search:
(title:bookmark* OR title:link*) AND anchor:"Michael Barlow"

users' information behavior

Björneborn 2001-2002

traditional LIS research areas	complementary LIS research areas
● 'convergent' behavior	● 'divergent' behavior
● goal-directed, rational	● non-goal-directed, intuitive
● e.g. Boolean searching	● e.g. browsing, serendipity
● present, explicit info.needs	● latent, implicit info.needs
● problems, work tasks	● triggered interests, curiosity
● 'information recovery'	● 'information discovery'
● order = topic-focused, 'convergent' link structures	● diversity = topic-scattered, 'divergent' link structures

- *which* combinations of link structures may facilitate *which* combinations of behavior?

exploratory capabilities

- LB proposal: broadened aim of LIS research:
- ”facilitating users to explore and exploit options embedded in information environments”
- small-world link structures may enhance exploratory capabilities in an informational network such as the Web
  - including chances for serendipity - encountering the unexpected
- easier to traverse and explore an information space if both local and global distances are short = ’small world’
- small-world link structures may facilitate combinations of ’convergent’ and ’divergent’ information behavior

possible implications

- webometric studies of small-world link structures may provide a better understanding of complex topologies, functionalities and potentials of the web
- may be utilised in
  - social informatics – exploring emerging cultural/social formations
  - digital libraries – making them as serendipity-prone as physical libraries…
    - facilitating exploratory capabilities!
  - creativity research – networked knowledge creation and diffusion
  - web mining – identifying fertile areas for cross-disciplinary exploration
  - search engines – more exhaustive web traversal + harvesting + ranking
  - browsers – visualisation / navigation facilities stimulating serendipity
  - etc.

Future

- With massively parallel systems (Grid computing) and very-large scale storage (the Data Grid) it *should* be possible to do very-large-scale analyses of the structure and content of the WWW
- Perhaps we finally begin to see the “intellectual structure of Cyberspace”