Identifying Small-World Connectors across an Academic Web Space – A Webometric Study

Lennart Björneborn

lb@db.dk
Department of Information Studies, Royal School of Library and Information Science,
DK-2300 Copenhagen (Denmark)

Abstract
This webometric study identifies web links, pages, and sites that function as small-world connectors affecting short link distances across topics in an academic web space. A five-step methodology is developed to sample and identify small-world properties by zooming stepwise into more and more fine-grained web node levels and link structures among 7669 subsites harvested from 109 UK universities. The methodology includes shortest path nets functioning as investigable small-world link structures, *'mini small worlds'*, generated by deliberate juxtaposition of topically dissimilar subsites. The network analysis tool *Pajek* identified all shortest link paths within the data set between 10 pairs of subsites. The study includes a novel corona-shaped model of reachability structures in a web subgraph. Indicative findings suggest that personal link creators and computer science subsites may be important small-world connectors across sites and topics in an academic web space. Such small-world connectors are important as they counteract balkanization of the Web into insularities of disconnected and unreachable subpopulations. The study also suggests how the Web is a *web of genres* with richly diversified genre connectivity and with *genre drift*, i.e. changes in page genres along link paths that may affect small-world properties.

Introduction

An intriguing dimension of the vast hypertext networks on the Web deals with small-world properties in the shape of short distances along link paths traversing intermediate web pages, web sites, and web clusters. Small-world networks are characterized by a combination of highly clustered network nodes and short average path lengths between pairs of network nodes (Watts & Strogatz, 1998). In recent years, an avalanche of research has revealed small-world properties in a large variety of networks, including biochemical, neural, ecological, physical, technical, social, economical, and informational networks. For instance, scientific collaboration networks, citation networks and semantic networks may show small-world features (Newman, 2001; Steyvers & Tenenbaum, 2001). Containing both high local clusterization and short global separation, small-world networks simultaneously have small local *and* small global distances, which facilitate high efficiency in disseminating information, ideas, contacts, signals, energy, viruses, etc., both on a local and global scale in the concerned networks.

On the Web, small-world link structures are concerned with core library and information science issues such as navigability and reachability of information across vast document networks. For instance, short link distances along link paths affect the speed and exhaustivity with which web crawlers can reach and retrieve web pages when following links from web page to web page. Further, small-world link structures may reflect cross-social connections between different interest communities and cross-disciplinary contacts in scientific networks. Moreover, small-world web topologies may have implications for the ways users explore the Web and the ease with which they gather information.

So far, research on small-world phenomena in complex networks including the Web has yielded important results regarding overall structural factors such as graph components, clustering coefficients, characteristic path lengths and scale-free properties including power-law frequency distribution of network connections (cf. reviews in, e.g., Albert & Barabási, 2002; Scharnhorst, 2003). However, there is a need to reveal more details on what *micro-level* web activities contribute to the emergence of small-world phenomena across macro-level web structures. As most links within and between web sites connect web pages containing similar topics (e.g., Davison, 2000) leading to topically clustered aggregations of web pages and sites, it would thus be interesting to identify *cross-*

Proceedings of ISSI 2005, pages 56–66.
Edited by P. Ingwersen and B. Larsen.

topic connections in order to reveal how small-world properties emerge in a web space. On this background, the main research question in the present webometric study (Björneborn, 2004) was:

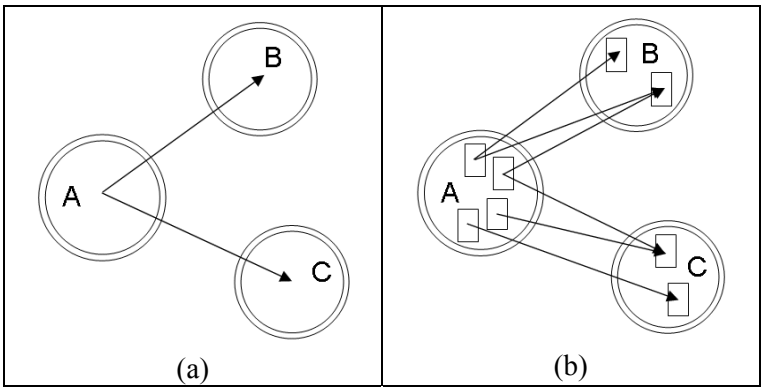
> *What types of web links, web pages, and web sites function as cross-topic connectors in small-world link structures across an academic web space?*

Data set

The data set in the study was harvested from 109 UK university web sites in June/July 2001 by a special web crawler (Thelwall, 2001).[1] The data set comprised all identified 7669 university subsites, i.e. departments, research groups, etc., with derivative university URLs, for instance, *www.geog.plym.ac.uk* (Geography Department, University of Plymouth). Data files contained source URLs of harvested web pages listed with target URLs of all outlinks from each web page. Because many universities have variant domain names, including old versions, canonical domain name forms were used to ensure data comparability. For example, *.doc.edinburgh.ac.uk* was converted to *.doc.ed.ac.uk*.

The data set contained almost 3.4 million web pages with outlinks. Of the 39.3 million page outlinks, 34.4 million were site selflinks pointing to a web page at the same university. An average outlinking university web page in the data set thus had 11.6 outlinks comprising 10.1 site selflinks and 1.5 site outlinks.[2]

In the study, links to and from the 109 university main sites were excluded from the data set because multi-disciplinary contents allocated in sub-directories would impede identification of cross-topic links across sites. University site selflinks were also excluded because topically different subsites belonging to the same university could be connected by embedded university navigational links. The exclusion of site selflinks and links to and from university main sites logically created delimited link structures. However, the included data subset still reflected real-world connectivity structures in the investigated academic web space. Only links connecting subsites belonging to different universities were thus included in the study. There were 207,865 such links located at 105,817 web pages. However, a data set this size was not tractable for the employed network analysis tool *Pajek*[3]. Instead, the 7669 university subsites were selected as units of analysis in the study. Data runs in *Pajek* were based on an adjacency matrix of the 7669 subsites interconnected by 48,902 *subsite level* connections comprising the above 207,865 *page level* links, cf. Figures 1a & b.



Figures 1a & b: Two *subsite level* links (a) between subsites A, B, C comprise six *page level* links (b).

Methodology

A five-step methodology was developed in the study in order to sample and identify small-world properties by zooming stepwise into more and more fine-grained web node levels and link structures among the harvested 7669 UK academic subsites.

[1] The 2001 data set is a subset of a freely accessible database at *http://cybermetrics.wlv.ac.uk/database* (cf. Thelwall, 2002/2003)

[2] Cf. link typology in Björneborn (2004) and Björneborn & Ingwersen (2004).

[3] Available at *http://vlado.fmf.uni-lj.si/pub/networks/pajek/*.

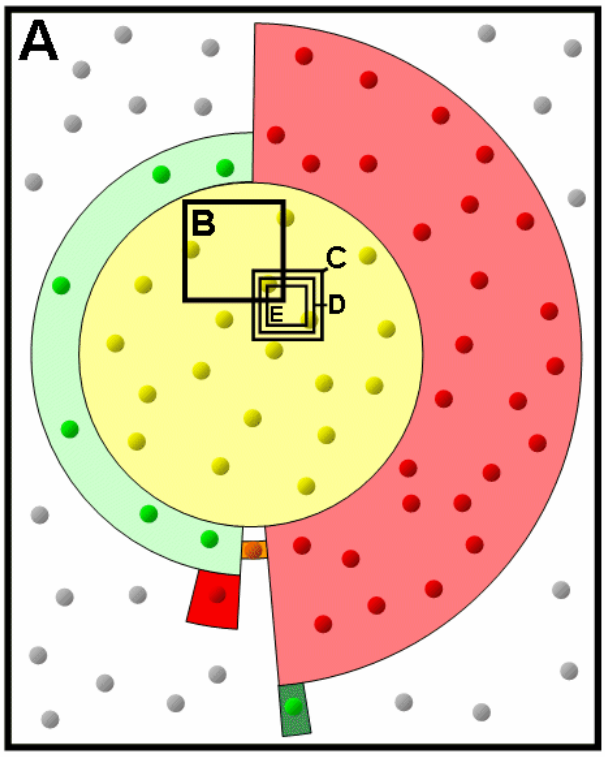


Figure 2: Five-step methodology.

The five steps will be described in more details further below. In summary, cf. Figure 2, step A identified overall web graph components and connectivity patterns among the 7669 subsites. From the strongly connected component (SCC) identified in step A, a random sample of 189 subsites was examined in step B in order to classify overall subsite topics. Within the SCC, all subsites can reach each other through directed link paths. This enabled a stratified sample to be extracted in step C resulting in 10 path nets comprising all shortest link paths between pairs of SCC subsites belonging to dissimilar topics. In step D, page genres and topics of source and target pages along followed link paths in the 10 path nets were classified. The objective of all these steps was to lead up to the final step E concerned with identifying what types of web links, pages, and sites function as cross-topic connectors in small-world link structures across an academic web space, the main research question in the study.

Step A: Corona model

The first step in the five-step methodology was to establish a graph model of connectivity patterns among the 7669 UK university subsites in order to enable the selection of a suitable sample for further investigation. Figure 3 shows a *Corona*[4] web graph model (Björneborn, 2004) of graph components and reachability structures among the 7669 subsites extracted by a special program. Reachability structures are graph connectivity patterns that reflect whether and how web nodes can reach each other along intermediate paths of directed links. The Corona model is a modification of the Bow-tie model by Broder *et al.* (2000). The Corona model depicts actual inter-component reachability structures in the investigated web subgraph not evident in the Bow-tie model. For example, direct links from the IN component to the OUT component indicated at the top of the Corona model are not clearly revealed in the Bow-tie model.

[4] The term *corona* denotes the figure's rudimentary resemblance to a solar corona with protuberances.

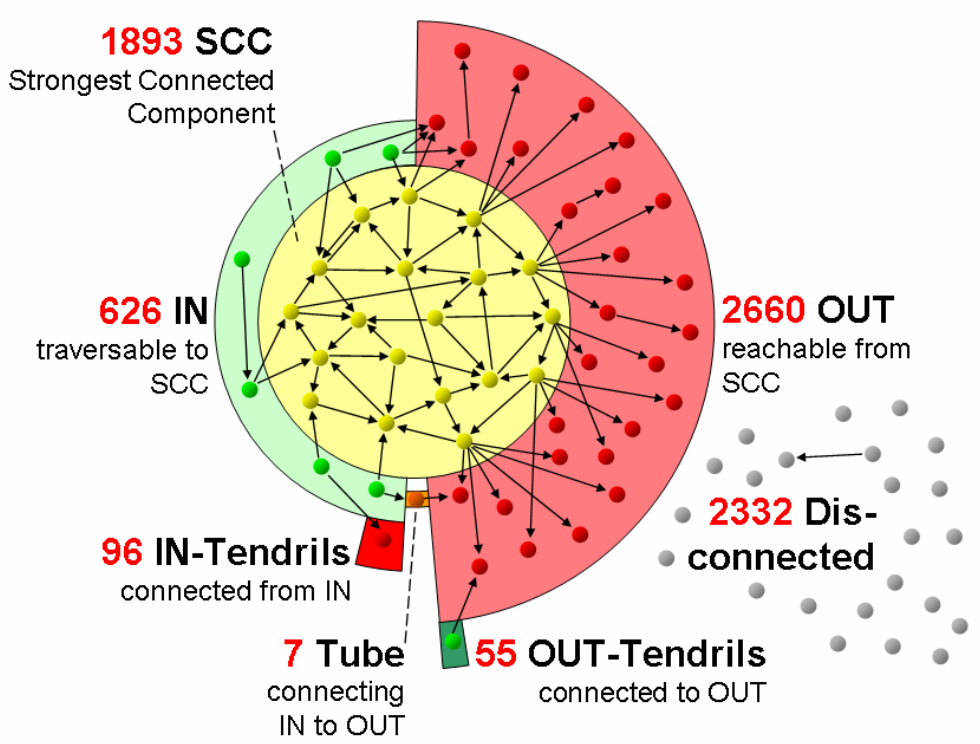


Figure 3: Corona model of web graph components and reachability structures among 7669 UK university subsites with numbers of subsites in each component (Björneborn, 2004).

In the strongly connected component (SCC) any subsite can reach any other along intermediate link paths. The IN component contains subsites that can reach the SCC along link paths – but cannot be reached back to. The OUT component contains subsites that can be reached from the SCC – but cannot reach back. See Björneborn (2004) for more details of the Corona model and its different components.

The Corona model of the UK academic subweb supports the notion of a fractal 'self-similar' Web (Dill *et al.*, 2001) with subsets of the Web displaying the same graph properties as the Web at large.

Step B: SCC subsite topics

The study focused on the strongly connected component because only within the SCC there can be link paths between all pairs of subsites irrespective of topical dissimilarity thus allowing easier identification of cross-topic connections.

A random sample of 189 (10%) subsites was extracted from the 1893 SCC subsites. Overall topics of the sampled subsites were classified by the author by visiting each subsite in the Internet Archive (*www.archive.org*) indexed as closely as possible to the original web crawl in June/July 2001 where no other contents than source and target URLs had been extracted. Internet Archive was used because subsite topics might have changed on the current Web. Only one of the 189 subsites was not available in the Archive. Subsite topics were classified and grouped with related topics in a pragmatic way to form topical groups as a basis for stratified sampling of diversified topical pairs of subsites in the subsequent step C.

Step C: path nets

From the topical groups formed in step B, a stratified sample was extracted consisting of five subsites from different categories in humanities and social sciences (hum/soc) and five subsites from natural sciences and technology (nat/tech). The ten seed subsites were randomly paired, one hum/soc subsite with one nat/tech subsite, giving the five pairs listed in Table 1.

Table 1. Stratified sample of five pairs of SCC subsites (prefix *www* removed from URLs).

path net	hum/soc node	affiliation	nat/tech node	affiliation
HN01	hum.port.ac.uk	Faculty of Humanities and Social Sciences, Portsmouth	atm.ox.ac.uk	Atmospheric, Oceanic and Planetary Physics, Physics Dept, Oxford
HN02	economics.soton.ac.uk	Economics Dept, Southampton	chem.gla.ac.uk	Chemistry Dept, Glasgow
HN03	psy.man.ac.uk	Psychology Dept, Manchester	maths.gcal.ac.uk	Mathematics Dept, Glasgow Caledonian
HN04	speech.essex.ac.uk	Speech Research Group, Language and Linguistics Dept, Essex	palaeo.gly.bris.ac.uk	Palaeontology Research Group, Earth Sciences Dept, Bristol
HN05	geog.plym.ac.uk	Geography Dept, Plymouth	eye.ox.ac.uk	Ophthalmology Dept, Oxford

Subsequently, the order of start and end nodes was reversed, resulting in a total list of five plus five diversified topical pairs of subsites, cf. Table 2 further below. *Pajek* extracted all shortest link paths between the ten subsite pairs based on the adjacency matrix of how the 7669 subsites were interconnected. The objective of this step was to construct investigable small-world link structures, *'mini small worlds'*, generated by the deliberate juxtaposition of topically dissimilar seed nodes.

Figure 4 below shows one of the resulting 10 subgraphs consisting of all shortest link paths within the data set (see all 10 subgraphs in Björneborn, 2004, Appendix 10). In the study, such an 'all shortest paths' subgraph was called a *path net*.[5] The path net (HN01, cf. Table 1) in the figure consists of all shortest link paths between node 2099 (*hum.port.ac.uk*), the Faculty of Humanities and Social Sciences in Portsmouth, and node 1904 (*atm.ox.ac.uk*), the atmospheric physics subsite in Oxford. All shortest link paths between the two seed nodes have the same path length 3. As shown in Table 2 below, the shortest path lengths were 3 or 4 in the 10 path nets. Due to the reachability structures in the strongly connected component, cf. Figure 3, all intermediary nodes in a path net belong to the SCC when the start and end node in the path net belong to the SCC.

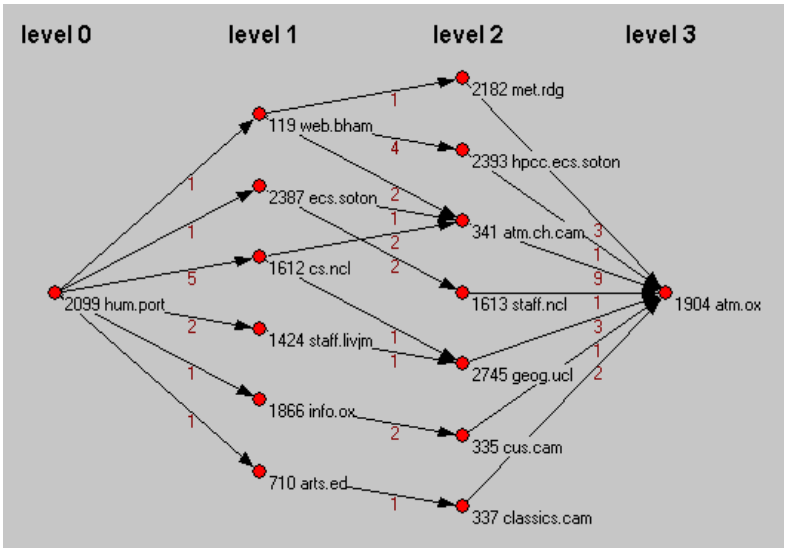


Figure 4: Path net with all shortest link paths (path length 3) between node 2099 (hum.port.ac.uk) and node 1904 (atm.ox.ac.uk) with counts of page level links (cf. Fig. 1b). Path net levels denote degrees of separation from start node. See affiliations in Björneborn (2004, Appendix 10).

[5] No literature has been found with similar investigations of subgraphs of all shortest paths between two network nodes.

Table 2. Summary of 10 path nets. Headings are explained in Step D below. *See Table 1 for affiliations of start and end nodes.

path net	start node*	end node*	path length	# sub-sites in path net	# visited subsites	# subsite level links	# page level links	# followed page links	# source pages	# visited source pages	# target pages	# visited target pages
HN01	hum	atm	3	15	15	23	48	48	41	41	28	28
HN02	econ	chem	3	6	6	7	7	7	7	7	7	7
HN03	psych	math	4	8	8	12	38	38	29	29	28	28
HN04	speech	palaeo	3	4	4	3	4	4	4	4	3	3
HN05	geogr	eye	3	6	6	7	12	12	7	7	12	12
NH01	atm	hum	3	15	8	25	87	33	68	26	47	23
NH02	chem	econ	4	28	12	53	183	53	134	37	114	37
NH03	math	psych	3	5	5	5	8	8	7	7	7	7
NH04	palaeo	speech	4	41	26	99	232	111	167	90	152	77
NH05	eye	geogr	4	13	13	18	38	38	33	33	27	27
			3.4	**141**	**103**	**252**	**657**	**352**	**497**	**281**	**425**	**249**

Step D: page topics and genres

This step zoomed in on source pages and target pages with outlinks and inlinks in the 10 path nets. The objective was to identify page genres and topics and thus enable the identification of pages and links that provide cross-topic connections between subsites in the final step E.

Figure 5 shows a node diagram with an excerpt of a path net (NH05, cf. Table 2). The figure gives an impression of how page level links connect source and target pages along shortest link paths, in this case between the subsites of eye research in Oxford and geography in Plymouth.

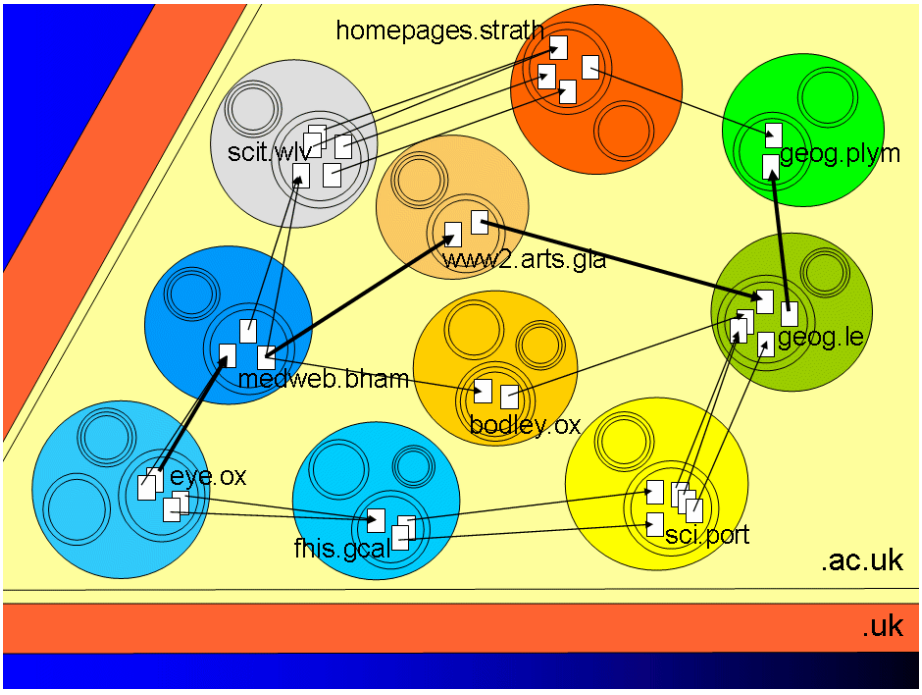


Figure 5: Node diagram with excerpt of a path net (NH05, cf. Table 2) with source and target pages, and page level links belonging to shortest link paths (path length 4) between start node *eye.ox.ac.uk* and end node *geog.plym.ac.uk*. Bold links show one of the link paths. Double border circles denote subsites, and single borders main sites, not reflecting actual sizes or numbers. See affiliations in Björneborn (2004, Appendix 10). Stacked pages are from the same subsite directories.

All source pages and target pages in the smaller path nets were visited in the study (cf. Table 2). In the large path nets (NH01, NH02, NH04, cf. Table 2), only pages belonging to link paths not passing

service-type (with no specific topic) subsites were visited. A total of 530 pages comprising 281 unique source pages and 249 unique target pages were visited on 103 subsites connected by 352 page level links, cf. Table 2. Internet Archive was used to retrieve source and target pages indexed as closely as possible to June/July 2001 when the original link data set was collected. 45 source pages (16.0%) and 30 target pages (12.0%) were not available in the Archive but were visited on the current Web. The higher availability of target pages in the Archive may reflect that more *inlink-prone* page genres like institutional homepages may have larger probability of being reached and retrieved by the Archive's web crawlers.

Page genres were classified by the author for all visited 530 source and target pages in the 10 path nets. The term *genre* is here used in a broad sense in accordance with contemporary web terminology for describing types of functions and purposes of web sites and web pages (e.g., Koehler, 1999). The visited web pages represented a rich diversity of genres reflecting a multitude of creators and purposes. By an iterative process of comparing and grouping similar pages, 17 broader *meta genres* crystallized, divided into two main categories, *personal* and *institutional* pages as listed in Table 3. A prioritized categorization order (Björneborn, 2004, p.156) was used to classify pages in a consistent way. The classification was not exhaustive, as a larger sample of academic web pages would probably yield additional genres. Nor are the genres mutually exclusive, but overlap with fluid borders.

Table 3. Meta genres of academic web pages: 9 institutional genres (*i.*) and 8 personal (*p.*).

meta genre	description
i.archive	collection of data in institutional archive or database
i.conf	conferences, workshops, seminars, etc.
i.generic	non-research-related, campus-wide service web pages
i.hp	homepages of academic institutions
i.list	institutional pages with link lists as main content
i.proj	joint and single research groups, including specific projects
i.publ	institutional publications, e.g., reports
i.soft	institutional software programs, software documentation, tutorials, etc.
i.teach	institutional teaching-focused web pages
p.archive	collection of data in personally maintained archive or database, e.g., discussion group archive
p.hobby	personal hobby webpages not related with the person's main academic activity
p.hp	personal homepages
p.list	personal pages with link lists as main content
p.proj	personal research project pages
p.publ	personal publications, e.g., papers, posters
p.soft	personal software programs, software manuals, etc.
p.teach	personal teaching-focused web pages

Contrary to scientific literature where references interconnect a small number of different document genres such as papers, monographies, etc., academic web spaces contain a much richer diversity of interconnected document genres. There were 83 different interlinked genre pairs among the visited pages in the 10 path nets. The most frequent genre pairs were institutional link lists linking to institutional homepages (9.4%), personal link lists linking to personal publications (4.8%) and personal link lists linking to institutional homepages (4.3%). The study gives support to how the Web literally is a *web of genres* with *genre drift*, that is, changes in genres along link paths, that may affect small-world phenomena (cf. Björneborn, 2004, Section 7.2.3).

Step E: cross-topic connectors

In the final step E were identified 112 cross-topic links (called *transversal* links in the study) between dissimilar subsite topics among the 352 followed page level links in the 10 path nets. The question of dissimilarity between subsites was determined by the author based on affiliations and topical descriptions given by the visited subsites and web pages. This was not trivial because of many interdisciplinary and multidisciplinary subunits at the UK universities. For example, environmental studies and meteorology posed interdisciplinary overlaps with earth sciences making it inconvenient to designate links between them as cross-topic. This illustrates how topic drift may appear as sliding

transitions along overlapping interdisciplinary topics. A pragmatic view on cross-topic links was employed in the study, focusing on links crossing more clear-cut topical borders, for instance, between computer science and atmospheric physics as in Figure 6.

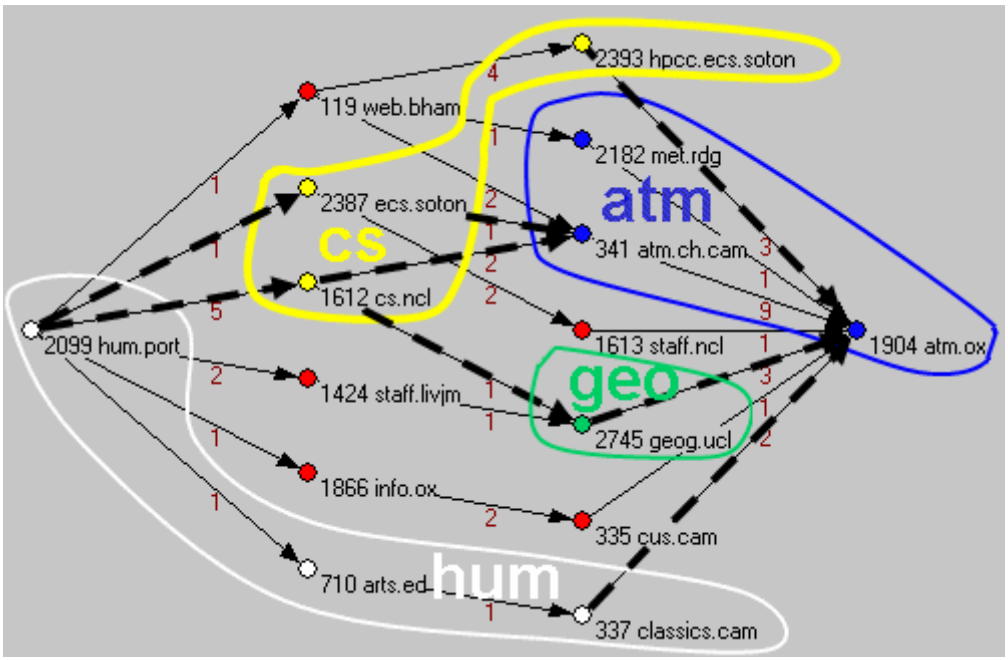


Figure 6: Path net (cf. Fig. 4) with enclosed topical areas in humanities (hum), computer science (cs), geography (geo) and atmospheric sciences (atm). Non-enclosed nodes are service-type subsites. Cross-topic links are marked with dashed bold. See Björneborn (2004, Appendix 10) for affiliations.

Figure 6 shows all the shortest link paths (path length 3) between the Faculty of Humanities and Social Sciences in Portsmouth, and the atmospheric physics subsite in Oxford (cf. Figure 4). One link path passes a humanities area (*hum*). Another topical area (*atm*) consists of three atmospheric science subsites. The computer-science area (*cs*) functions as connector on many of the shortest link paths between the start node and end node. Also one geography subsite (*geo*) is a cross-topic connector. The non-enclosed nodes are service-type subsites, typically providing services for staff and students at the whole university. The dashed links denote cross-topic links connecting one topical area with another in the path net, for example, a link from Faculty of Classics in Cambridge (node 337) to an oceanographer at the atmospheric subsite in Oxford (node 1904) having a hobby web page devoted to an ancient Greek warship.

Selected findings

The findings in the study cannot be generalized to the whole Web or to all academic web spaces. The findings are indicative only as there were delimitations in the investigated data set: *national* (only UK sites), *sectoral* (only academic sites), *institutional* (only university subsites), and *temporal* (snapshot data set from June/July 2001 not covering dynamic changes on the Web). Further, a small stratified sample of 10 path nets was used.

However, the indicative findings and identified phenomena may be fruitful for future large-scale investigations. For example, longitudinal time series of snapshots of the same population of academic web subsites could reveal how graph components, reachability structures, topical clusters, and cross-topic connectors change over the years on the dynamic Web. In order to enable large-scale investigations, automatic classification of overall topics both on the subsite and page levels would be necessary, as well as more objective heuristics for determining dissimilarity between topics, for instance, by including low *co-inlink* or *co-outlink* measures (analogous to co-citations and bibliographic couplings, respectively, cf. Björneborn & Ingwersen, 2004) of web sites combined with low co-word occurrences.

Below, some selected findings are outlined regarding small-world properties and small-world connectors in the investigated UK academic web space. See Björneborn (2004) for a more comprehensive presentation and discussion.

Small-world properties

The investigated UK academic web space was found to contain small-world properties. Figure 7 shows the distribution of lengths of shortest link paths between pairs of subsites in the data set of 7669 UK academic subsites. Only 19.5% of all pairs of subsites could be connected by a directed link path within the data set. The low percentage is affected by the high share of subsites belonging to the OUT and Disconnected graph components, cf. Figure 3. The characteristic path length, that is, the average path length in the UK academic subweb was 3.46 among subsites that could reach each other along link paths. The diameter of the UK academic subweb was 10, the longest of the existing shortest paths between subsites.

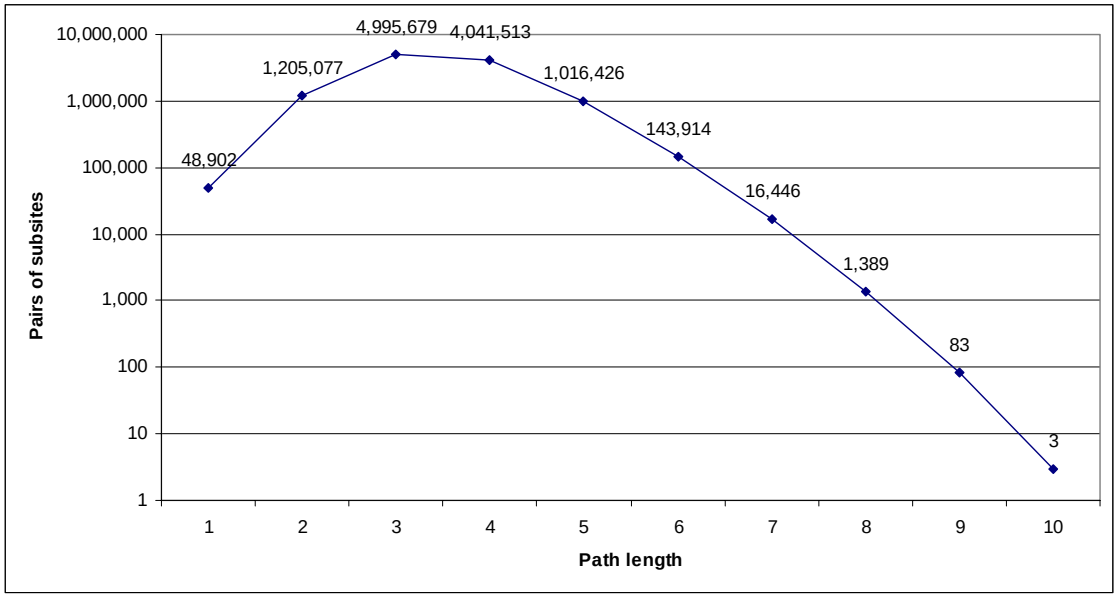


Figure 7: Distribution of shortest path lengths between 7669 UK subsites. Semi-log scale.

Pajek was used to construct a corresponding random graph with 7669 nodes and 48,902 edges as in the UK academic subweb. The characteristic path length of the random graph was 5.04 compared with 3.46 for the UK graph. Pajek also computed all local clustering coefficients for the 7669 subsites in the UK graph giving an average overall clustering coefficient of 0.09038. In other words, if a subsite node v_1 was connected with two subsites nodes v_2 and v_3, there was 9.0% probability that v_2 and v_3 were also connected. The clustering coefficient for the corresponding random graph was 0.00084, that is, over 100 times smaller. The small characteristic path length of the UK subweb and the large clustering coefficient compared with the corresponding random graph meet the requirements for a small-world network as defined by Watts & Strogatz (1998). However, as also noted by Broder *et al.* (2000), true small-world properties were only present within the SCC component, because as stated above, only 19.5% of all pairs of subsite nodes in the investigated UK graph could be connected by link paths.

Small-world connectors

Below are some selected indicative findings answering the main research question in the study: what types of web links, web pages, and web sites function as cross-topic connectors in small-world link structures across an academic web space?

Of the 112 cross-topic links identified in the 10 path nets, 64 (57.1%) originated from a personal web page, whereas 48 (42.9%) were from an institutional page. Over 80% of the identified cross-topic links were related to academic activities such as research and teaching. There was a higher percentage of personal link lists among the cross-topic links (40%) and cross-topic source pages (36%) than

among all the followed links (32%) and visited source pages (30%) in the 10 path nets. This finding may indicate a special impact of personal web creators and their link lists for the emergence of small-world phenomena across dissimilar topical web domains.

About 46% of subsites providing or receiving cross-topic links in the 10 path nets were related to computer science. In the random sample of 189 SCC subsites, only about 11% were related to computer science. Even if the sample of 10 path nets was small, this finding may indicate a special role of computer science subsites as cross-topic connectors on shortest link paths in an academic web space. This probably reflects the auxiliary function of computer technology in many scientific disciplines. Furthermore, a more experienced and extrovert web presence may exist among persons and institutions in computer science.

The special role of computer science subsites as connectors across an academic web space is supported by the fact that 8 of the 10 subsites with the highest *betweenness centrality* in the whole data set were related to computer science. The social network analytic measure of betweenness centrality reflects the probability that a node occurs on a shortest link path between two arbitrary nodes (Freeman, 1977). The present study (Björneborn, 2004, Section 6.3.2.4) indicated a close relation between Kleinberg's (1999) concepts of *hubs* and *authorities* on the Web and the measure of betweenness centrality, as the subsites with the highest betweenness centrality also were among the strongest hubs and authorities as computed by *Pajek*. No other literature has been found discussing such a relation.

Conclusion

This webometric study has developed a five-step methodology to identify web links, pages, and sites that function as small-world connectors across an academic web space. The methodology includes a novel Corona model of reachability structures in a web graph, as well as shortest path nets functioning as investigable small-world link structures, '*mini small worlds'*, generated by deliberate juxtaposition of topically dissimilar subsites.

Of importance to web sociology and science studies, small-world link structures may reflect cross-social connections between different interest communities including cross-domain contacts in scientific networks. Indicative findings in the present study suggest that personal link creators, such as researchers and students, as well as computer science-related subsites may be important small-world connectors across sites and topics in an academic web space. Such small-world connectors are important as they counteract *balkanization* of the Web into insularities of disconnected subpopulations (cf. Figure 3 of the Corona model) not reachable by search engine crawlers following paths of links. Further, such small-world connectors 'crumple up' the Web by pulling web sites and their neighborhoods close together across the Web. Possibilities of 'crumpling up' web spaces are unlimited because the virtual space of the Web is billion-dimensional, as every new link literally adds a new dimension.

As academic web spaces increasingly include extensive scholarly self-presentations and link creations, science studies may employ small-world approaches including social network analytic concepts like the abovementioned *betweenness centrality* for automatic tracking of central gatekeepers and interdisciplinary boundary crossings in academic web spaces. This might identify fertile areas for interdisciplinary exploration and cross-pollination.

Special decentralized algorithms have been developed that utilize local connectivity information for identifying short traversal paths through a network where link data for the whole network are not available (e.g., Adamic *et al.*, 2001). In particular, well-connected hub-like web nodes may be exploited in such decentralized algorithms. The findings in the present study suggest that the rich diversity of inlinks and outlinks to and from computer-science web sites and personal link lists may be utilized for such computer-aided navigation along small-world shortcuts, which also may be exploited as transit points for more exhaustive web coverage by search engine crawlers.

References

Adamic, L.A., Lukose, R.M., Puniyani, A.R. & Huberman, B.A. (2001). Search in power-law networks. *Physical Review E*, 64, 46135.

Albert, R. & Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1), 47-97.

Björneborn, L. (2004). *Small-world link structures across an academic web space : a library and information science approach*. PhD thesis. Royal School of Library and Information Science. Retrieved January 14, 2005 from: http://www.db.dk/lb/phd

Björneborn, L. & Ingwersen, P. (2004). Toward a basic framework for webometrics. *Journal of the American Society for Information Science and Technology*, 55(14), 1216-1227.

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A. & Wiener, J. (2000). Graph structure in the Web. *Computer Networks*, 33(1-6), 309-320.

Davison, B.D. (2000). Topical locality in the Web. *Proceedings of the 23rd International ACM SIGIR Conference*. ACM Press. pp. 272-279.

Dill, S., Kumar, S.R., McCurley, K., Rajagopalan, S., Sivakumar, D. & Tomkins, A. (2001). Self-similarity in the Web. *Proceedings of the 27th International Conference on Very Large Data Bases*, pp. 69-78.

Freeman, L.C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, 40(1), 35-41.

Kleinberg, J.M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5), 604-632.

Koehler, W. (1999). Classifying web sites and web pages: the use of metrics and URL characteristics as markers. *Journal of Librarianship and Information Science*, 31(1), 297-307.

Newman, M.E.J. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, 16 Jan. 2001, 98(2), 404-409.

Scharnhorst, A. (2003). Complex networks and the Web: insights from nonlinear physics. *Journal of Computer-Mediated Communication*, 8(4). Retrieved January 14, 2005 from: http://www.ascusc.org/jcmc/vol8/issue4/scharnhorst.html

Steyvers, M. & Tenenbaum, J.B. (2001). The large-scale structure of semantic networks: statistical analyses and a model for semantic growth. Retrieved January 14, 2005 from: http://arxiv.org/pdf/cond-mat/0110012

Thelwall, M. (2001). A web crawler design for data mining. *Journal of Information Science*, 27(5), 319-326.

Thelwall, M. (2002/2003). A free database of university web links: data collection issues. *Cybermetrics*, 6/7(1): paper 2. Retrieved January 14, 2005 from: http://www.cindoc.csic.es/cybermetrics/articles/v6i1p2.html

Watts, D.J. & Strogatz, S.H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393(June 4), 440-442.