Small-World Link Structures across an Academic Web Space: A Library and Information Science Approach

Lennart Björneborn

**Ph.D. thesis from the Department of Information Studies,
Royal School of Library and Information Science, Denmark, 2004**

‘Small-world’ linkstrukturer på tværs af en akademisk web: En biblioteks- og informationsvidenskabelig tilgang

Lennart Björneborn

Ph.d.-afhandling fra Institut for Informationsstudier,
Danmarks Biblioteksskole, 2004

CIP – Cataloguing in Publication

Björneborn, Lennart
Small-world link structures across an academic web space :
a library and information science approach. PhD dissertation.
Copenhagen: Department of Information Studies, Royal School
of Library and Information Science, 2004. xxxvi, 399 p.
ISBN 87-7415-276-9

Acknowledgements

First of all, I want to commemorate late Tomas Crone Almind whose creative ideas on applying bibliometric approaches to the Web contributed to the birth of a new research field, *webometrics*, in collaboration with Peter Ingwersen.

Early on, Peter understood the potentials – and pitfalls – of Web studies. While writing my master's thesis, Peter placed confidence in my explorative and tentative web ideas revolving around small-world phenomena, knowledge discovery and serendipity. He gave me the decisive push to apply for a PhD scholarship in 2000 pursuing these web ideas. Having Peter as a main supervisor on the PhD project has been a privilege. His great intuition, playful creativity and visionary mind have been very stimulating. He has given me room to pursue my ideas and has been there every time I needed advice. His warm support, concern and confidence have been invaluable, especially when tricky computing problems and time problems became burdensome.

Furthermore, I want to express my great gratitude to Dr. Mike Thelwall who has been a superb second supervisor on the PhD project. Hundreds of e-mails have provided encouragement and constructive feedback. Crucial was Mike's initial generous offer to give me access to the UK link data set as well as to patiently help me with his great programming skills to filter, extract and run the necessary data to be analyzed – so-called *'washomatic webometrics'* in our internal lingo. Spending time together with Mike and his research group at the University of Wolverhampton has been very inspiring and great fun. It has shown me how creatively stimulating, hospitable and enjoyable international scientific collaboration can be.

I am very grateful for the positive and constructive feedback from the review committee: Professor Ronald Rousseau, Professor Olle Persson and Associate Professor Niels Ole Pors. I would also like to express my gratitude to the Royal School of Library and Information Science for granting me the three-year scholarship that enabled this PhD project. Especially, I owe thanks to the Head of Department, Mona Madsen, and to the Rector, Leif Lørring, for their great and warm support. Also many thanks to all my good colleagues and friends at the Royal School, including my fellow 'tapirs' in the TAPIR research group.

My office mate and good friend Birger Larsen has been an invaluable help and support during these three years thanks to his great kindness and humor, as well as his large professional and technical skills. Also many thanks to Jacob Andresen for his programming assistance with tricky data validations.

I am also grateful to Michael Kristiansson for many inspiring and supportive discussions since my undergraduate years. Michael has always encouraged me to follow my curiosity and intuition, and has to a high degree stimulated my wish to become a researcher and explore new frontiers in library and information science.

Many warm hugs to my family and friends for their large indulgence and patience during this almost endless and, alas, frequently quite asocial *"Monk's-Cell Marathon"...* Last, but not least, I want to thank Eva, my beloved life and dance partner. Without her love and great patience I would never have been able to accomplish this dissertation.

Abstract

The Web constitutes an obvious research area for library and information science (LIS), being a document network with documents in the shape of web pages interconnected by billions of links into complex hypertext structures. The Web is constructed through *distributed knowledge organization* by millions of local page and link creators. This *self-organization* of hypertextual link structures on the Web may be conceived as macro-level aggregations of micro-level interactions; as *'collaborative weaving'* of an evolving global document network conducted by a multitude of link creators.

An intriguing dimension of this giant document network deals with so-called *small-world* properties in the shape of short link paths between web pages or web sites. Small-world link structures are concerned with core LIS issues such as navigability and accessibility 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.

The overall objective of the dissertation is to develop a conceptual framework and empirical methods concerning the identification and characterization of whether and how small-world phenomena emerge in link structures across an *academic* web space. The main research question is concerned with what types of web links, web pages and web sites function as connectors across dissimilar topical domains in an academic web space. The dissertation is thus a LIS approach to answering what micro-structure web activities and elements contribute to cohesive macro structures across an academic web space. The UK academic web space *ac.uk* was chosen as a setting for the empirical investigation because a link data set that covered 109 UK universities was available and had a suitable size and coverage for studying small-world link structures.

The dissertation is situated within the new research field of *webometrics*. The dissertation defines webometrics within the LIS framework of informetric studies and bibliometrics, concerned with the study of the quantitative aspects of the construction and use of information resources, structures and technologies on the Web, drawing on bibliometric and informetric approaches. The dissertation incorporates approaches from graph theory and social network analysis into this framework.

The dissertation proposes a consistent and detailed link terminology as well as a novel web node diagram notation in order to fully appreciate and investigate link structures and different web node levels on the Web.

A wide range of *graph measures* (including characteristic path length; clustering coefficient; distribution of in-neighbors/out-neighbors, inlinks/outlinks, and in-distance/out-distance; assortative mixing; betweenness centrality; cores; hubs and authorities; co-linkage) were applied to investigate the *macro-level* (UK academic web) as well as *meso-level* (10 path nets) connectivity patterns in the investigated web space.

The UK academic subsite web showed small-world properties with a high *clustering coefficient* and a low *characteristic path length* of 3.5 between reachable subsites. These measures meet the requirements for a small-world network as introduced by Watts & Strogatz (1998). Power-law-like distributions of in-neighbors/out-neighbors and inlinks/outlinks were found in the UK academic web space as well as within the 10 path nets. These findings are in line with the concept of a *fractal 'self-similar'* Web with

subsets of the Web displaying the same graph properties as the Web at large. Further, there was an indication of a close relation between Kleinberg's (1999a) concepts of *hubs* and *authorities* on the Web and the social network analytic measure of *betweenness centrality*. No literature has been found discussing such a relation.

A five-step methodology was developed in order to sample, identify and characterize small-world properties by 'zooming' stepwise into more and more fine-grained web node levels in the investigated UK academic web space.

The first step A was concerned with identifying the so-called web graph components among 7669 subsites harvested by a special web crawler at 109 UK universities as of June-July 2001. Only links connecting subsites located at different universities were included, thus excluding links to or from the 109 multi-disciplinary university main sites. The dissertation presents a novel *'corona'* graph model that illustrates inter-component adjacencies in the UK web graph. About 25% of the investigated subsites belonged to the *strongly connected component* (*SCC*) of the graph model. In the SCC, all pairs of subsites can reach each other through link paths.

In step B, a large random sample of 189 subsites from the SCC component was examined in order to classify overall subsite topics and genres. This step enabled a stratified sample to be extracted in the next step C resulting in 10 so-called *path nets* comprising all shortest link paths in both directions between five pairs of seed node subsites belonging to dissimilar topics in natural sciences and technology on the one hand, and in humanities and social sciences on the other. The network analysis program *Pajek* was used to extract all the shortest link paths between the seed nodes. The path nets were constructed to function as investigable and illustrative small-world link structures – *'mini small worlds'* – by the juxtaposition of pairs of dissimilar topical seed nodes. In step D, genres and topics of the visited source and target pages along the followed link paths in the 10 path nets were classified by the author. The objective of the developed methodological steps was to lead up to the final step E concerned with identifying what types of web links, web pages and web sites function as *transversal* (cross-topic) connectors in the 10 path nets.

The Internet Archive (*www.archive.org*) was a useful *'web archaeological'* tool to retrieve and examine source pages and target pages with interconnecting links from the 10 path nets.

Due to the small and non-random sample of 10 path nets, there are no generalizable findings. However, the close examination in the case studies of the 10 path nets yielded interesting indicative findings. Personal link creators, such as researchers and students, may be important connectors across sites and topics in the investigated academic web space. Personal web pages thus provide about 53% of all followed site outlinks and 35% of site inlinks in the 10 path nets. Further, personal web pages provide about 57% of transversal outlinks and 42% of transversal inlinks in the 10 path nets. Personal link lists was the largest cross-topic page genre providing about 40% of transversal outlinks in the 10 path nets. Over 80% of the identified transversal links in the 10 path nets were related to academic activities such as research or teaching.

Another important indicative finding was that computer science-related (CS) subsites may be important cross-topic connectors in an academic web space. About 46% of subsites providing or receiving transversal links in the 10 path nets thus were CS-related, whereas only 11% CS-related in the population of subsites in the strongest connected component. The indicated connective role of CS-related subsites in academic

link structures probably reflects the auxiliary function of computer science in many scientific disciplines in natural sciences, technology, humanities, and social sciences. This auxiliary function may be combined with a more experienced and unconstrained web presence by CS-related persons and institutions.

The close examination of pages and links in the 10 path nets gave an impression of the rich *genre connectivity* on the Web. This finding gives an intuitive support to how the Web may be conceived as a *web of genres* with a rich diversity of page genres linked to other genres and with *genre drift*, that is, changes in genres of pages along link paths.

An extensive analytical discussion and perspectivation is brought in the dissertation based on the findings of the empirical investigation. The discussion is concerned with (1) the role of personal and institutional link creators for the emergence of small-world link structures across an academic web space; (2) hypothesized complementarities of topical uniformity and diversity in the formation of small-world link structures – including hypothesized complementarities of so-called *topic drift* and *genre drift*; and (3) possible implications for exploratory capabilities including serendipity in *'crumpled-up'* small-world web spaces. Further, possible implications of non-engineered and small-world knowledge organization for LIS frameworks are discussed. It is argued there is a need for redefining the overall aim and explanatory framework of LIS research so it encompasses both *convergent* (goal-directed) and *divergent* (serendipitous) information behavior conducted by users in both 'top-down'-constructed information systems (e.g., traditional libraries and bibliographic databases) and in 'bottom-up'-constructed distributed information systems such as the Web, in order to cope with issues concerned with, for instance, distributed knowledge organization, selforganization in information systems, small-world phenomena, topical diversity, genre connectivity, serendipity, knowledge discovery and creativity stimulation in information systems, as discussed in the dissertation.

Dansk resumé

Webben, *World-Wide Web*, udgør et oplagt forskningsområde for biblioteks- og informationsvidenskab, da den er et dokumentnetværk med dokumenter i form af websider sammenvævede af milliarder af hyperlinks i komplekse hypertekststrukturer. Webben opbygges nedefra og op gennem *distribueret* (dvs. decentraliseret) *vidensorganisation* via millioner af lokale aktører, der tilføjer og fjerner websider og links. Der opstår derved en *selvorganisering* af hypertekstuelle linkstrukturer på webben i form af aggregering på makro-niveau af interaktioner udfoldet på mikro-niveau: som en fortløbende *'fælles-vævning'* ('collaborative weaving') af et globalt dynamisk dokumentnetværk udført af en mangfoldighed af lokale linkskabere.

En spændende dimension ved dette gigantiske dokumentnetværk omhandler såkaldte *'small-world'* egenskaber i form af korte afstande på tværs af webben via stier af links ('link paths'), der går fra webside til webside, og fra websted til websted. 'Small-world' linkstrukturer vedrører centrale områder for biblioteks- og informationsvidenskab, såsom navigationsmuligheder og informationstilgængelighed på tværs af meget omfattende dokumentnetværk. Eksempelvis påvirker korte afstande langs linkstier hvor hurtigt og dækkende, at søgemaskiners web-robotter kan indhøste websider, når de følger links fra webside til webside.

Det overordnede formål med afhandlingen er at udvikle en konceptuel ramme og empiriske metoder til at identificere og karakterisere, hvorvidt og hvorledes 'small-world' fænomener opstår i linkstrukturer på tværs af et akademisk 'web space', dvs. segment af webben. Det overordnede forskningsspørgsmål i afhandlingen vedrører, hvilke typer links, websider og websteder, der fungerer som konnektorer eller bindeled på tværs af uligeartede emneområder i et 'small-world' akademisk web-segment. Afhandlingen anlægger således en biblioteks- og informationsvidenskabelig tilgang til at afdække, hvilke mikro-strukturelle web-aktiviteter og -elementer, der bidrager til sammenhængende makro-strukturer på tværs af et akademisk web-segment. Det britiske akademiske web-segment *ac.uk* blev valgt som setting for den empiriske undersøgelse i afhandlingen, idet et linkdatasæt fra 109 britiske universiteter var tilgængeligt og med passende størrelse og indhold til at kunne undersøge 'small-world' linkstrukturer.

Afhandlingen indskriver sig i det ny forskningsfelt *webometri*, der tilhører de biblioteks- og informationsvidenskabelige domæner informetri og bibliometri. I afhandlingen defineres webometri som studiet af kvantitative aspekter vedrørende konstruktion og brug af informationsressourcer, -strukturer og -teknologier på WWW, baseret på bibliometriske og informetriske tilgange. Afhandlingen inddrager i denne sammenhæng også grafteori og social netværksanalyse.

Afhandlingen foreslår en konsistent og detaljeret linkterminologi samt et nyt notationssystem i form af webnode-diagrammer til brug for præcise beskrivelser af linkstrukturer og forskellige webnode-niveauer (fx websider, websteder, akademiske sub-domæner, lande-domæner).

Et bredt spektrum af statistiske mål fra grafteori og social netværksanalyse blev benyttet for at undersøge konnektivitetsmønstre i såvel den britiske akademiske web som i 10 såkaldte 'path nets', jf. nedenfor. De statistiske mål inkluderer såkaldt 'characteristic path length'; 'clustering coefficient'; distribution af 'in-neighbors/out-

neighbors', indlinks/udlinks, og 'in-distance/out-distance'; 'assortative mixing'; 'betweenness centrality'; 'cores'; 'hubs/authorities'; og 'co-linkage'.

Undersøgelsen viste, at den britiske akademiske web havde 'small-world' egenskaber i form af en høj såkaldt *'clustering coefficient'* og lav *'characteristic path length'*. I gennemsnit var en link-sti på kun 3,5 links tilstrækkelig for at forbinde to vilkårlige underwebsteder ('subsites'), der var indenfor rækkevidde fra hinanden. Disse statistiske mål opfylder specifikationer for et 'small-world' netværk angivet af Watts & Strogatz (1998). Statistiske 'power-law'-lignende skæve fordelinger af 'in-neighbors/out-neighbors' og indlinks/udlinks blev fundet i den britiske akademiske web samt i de 10 'path nets'. Disse fund er i overensstemmelse med relateret forskning omkring såkaldt *fraktale 'self-similar'* aspekter på webben, hvor mindre web-segmenter udviser samme grafmæssige egenskaber som hele webben. Afhandlingen indikerer også en relation mellem Kleinberg's (1999a) begreber *'hubs'* og *'authorities'* på webben og begrebet *'betweenness centrality'* fra social netværksanalyse. Der er ikke fundet litteratur, der diskuterer en sådan relation.

I afhandlingen introduceres en 5-trins metodologi for at identificere og karakterisere 'small-world' aspekter ved at 'zoome' trinvis ind i mere og mere detaljerede webnode-niveauer blandt de britiske akademiske websteder. I det første trin A identificeredes overordnede såkaldte *grafkomponenter* i linkstrukturerne mellem 7669 underwebsteder ('subsites') indhøstet fra 109 britiske universiteter og højere læreanstalter i juni/juli 2001. Kun links mellem underwebsteder tilhørende forskellige universiteter blev inkluderet. Indlinks og udlinks til og fra hovedwebstedet ved de 109 universiteter blev således ikke inkluderet. Afhandlingen introducerer en *'corona'* grafmodel, der viser strukturelle sammenhængsforhold mellem grafkomponenterne. 25% af underwebstederne i data-materialet tilhørte den såkaldt *'strongly connected component'* (*SCC*) i grafmodellen. I SCC grafkomponenten kan ethvert underwebsted nå ethvert andet via stier af mellemliggende links.

I trin B blev en randomiseret stikprøve af 189 underwebsteder fra SCC graf-komponenten undersøgt med henblik på at klassificere overordnede emner og genrer for underwebstederne. Dette trin muliggjorde et stratificeret stikprøve i næste trin C, der resulterede i 10 såkaldte *'path nets'*. Disse bestod af alle de korteste link-stier i begge retninger mellem fem par af underwebsteder, der tilhørte forskelligartede emner i naturvidenskab og teknologi på den ene side, samt humaniora og socialvidenskaber på den anden. Netværksanalyseprogrammet *Pajek* blev brugt til at identificere alle de korteste link-stier mellem de udvalgte underwebsteder. De 10 'path nets' blev konstrueret for at fungere som afgrænsede og illustrative 'small-world' linkstrukturer – *'mini small worlds'* – ved at finde de korteste link-stier mellem par af emnemæssigt forskelligartede underwebsteder. I trin D, klassificerede forfatteren genrer og emner for kilde- og destinations-websider med henholdsvis udlinks og indlinks langs de undersøgte link-stier i de 10 'path nets'. Formålet med rækken af metodologiske trin var at lede frem til det sidste trin E, der omfattede identifikation af, hvilke typer links, websider og websteder der fungerer som *transversale* ('cross-topic', dvs. på tværs af emneområder) konnektorer i de 10 'path nets'. Internet Archive (*www.archive.org*) viste sig at være et udmærket *web-'arkæologisk'* værktøj til at finde og undersøge websider i de 10 'path nets' som de så ud så tæt som muligt på den oprindelige *'web crawl'* i 2001.

Grundet den lille og ikke-randomiserede stikprøve bestående af de 10 'path nets', er der ingen fund, der giver mulighed for generalisering. De grundige undersøgelser i

casestudierne af de 10 'path nets' gav dog interessante indikative fund. Personlige linkskabere, såsom forskere og studerende, kan således være vigtige konnektorer og bindeled på tværs af emner og websteder i det undersøgte akademiske web-segment. Personlige websider bidrog således med ca. 53% af alle fulgte websteds-udlinks ('site outlinks') og 35% af websteds-indlinks ('site inlinks') i de 10 'path nets'. Ydermere havde personlige websider ca. 57% af transversale udlinks og 42% af transversale indlinks på tværs af forskelligartede emneområder i de 10 'path nets'. Personlige linklister var den største webside-genre, og bidrog med ca. 40% af transversale udlinks i de 10 'path nets'. Over 80% af de identificerede transversale links i de 10 'path nets' var relaterede til akademiske aktiviteter såsom forskning eller undervisning.

En andet interessant indikativt resultat var, at computer science-relaterede (CS) underwebsteder kan være vigtige konnektorer på tværs af emneområder i en akademisk web. Ca. 46% af underwebstederne med transversale udlinks eller indlinks i de 10 'path nets' var således CS-relaterede, hvorimod kun 11% var CS-relaterede i den store stikprøve af underwebsteder i SCC grafkomponenten. Denne mulige konnektive rolle for CS-relaterede underwebsteder i akademiske linkstrukturer afspejler formodentlig, at computer science fungerer som 'hjælpevidenskab' for mange viden-skabelige discipliner i natur-videnskab, teknologi, humaniora og socialvidenskab. Denne funktion er muligvis kombineret med et mere erfarent og afslappet 'web-nærvær' ('web presence') hos CS-relaterede personer og institutioner.

Den detaljerede undersøgelse af websider og links i de 10 'path nets' gav et indtryk af omfattende *genre-konnektivitet* på webben. I afhandlingen opfattes webben derfor som en *'web of genres'* med stor diversitet af webside-genrer linket sammen med hinanden – og med *'genre drift'*, dvs. skift i webside-genrer, når links følges fra webside til webside langs link-stier.

Afhandlingen indeholder en analytisk diskussion og perspektivering der følger tre hovedspor: (1) personlige og institutionelle linkskaberes rolle for fremkomsten af 'small-world' linkstrukturer på tværs af en akademisk web; (2) hypotetiseret komplementaritet mellem emnemæssig uniformitet og diversitet i formationen af 'small-world' linkstrukturer – inkl. hypotetiseret komplementaritet mellem såkaldt *'topic drift'* og *'genre drift'*; og (3) mulige implikationer for eksplorative muligheder herunder serendipitet (dvs. uventede informationsfund) i 'sammenkrøllede' 'small-world' web-rum. I denne forbindelse diskuteres også mulige implikationer for biblioteks- og informationsvidenskab på baggrund af distribueret vidensorganisation. Der argumenteres for nødvendigheden af at udvide biblioteks- og informationsvidenskabs traditionelle forskningsfoki og forklaringsmodeller, så de omfatter både *konvergent* (målrettet) og *divergent* (serendipitiv) brugeradfærd i både 'top-down'-konstruerede informationssystemer (fx traditionelle bibliotek og bibliografiske databaser) og i 'bottom-up'-konstruerede distribuerede informationssystemer såsom webben. Dette foreslås med henblik på at kunne håndtere spørgsmål vedrørende fx distribueret vidensorganisation, selvorganisation i informationssystemer, 'small-world' fænomener, emne-diversitet, genre-konnektivitet, serendipitet, 'knowledge discovery' og kreativitetsstimulering i informationssystemer.

Prelude

This dissertation is concerned with how decentralization and individualization of control in the construction and use of an information system such as the Web may affect overall structures and functionalities of this information system.

An early propagator of decentralization and individualization of control in the construction and use of information systems was the 'father of hypertext' Vannevar Bush in the 1940s with his visions of a hypertext-like association-based and personalized information system. Bush (1890-1974) was originally an electrical engineer and professor at the Massachusetts Institute of Technology (MIT) developing analogue computers and microfilm selectors in the 1930s. During the Second World War he coordinated the research activities of over 6000 scientists in the American scientific warfare program (e.g., the atomic bomb and the radar). Bush's seminal article 'As we may think' published in *The Atlantic Monthly*, July 1945 (Bush, 1945) was based on his experiences with the creative synergetic effects deriving from the historically unprecedented collaboration of so many researchers from different scientific domains. According to Bush, methods used by scientific libraries to organize knowledge hindered possibilities for scientific creativity and innovation, because traditional indexing methods did not sufficiently support exploration of relations across rigid classification hierarchies:

> *"Our ineptitude in getting at the record is largely caused by the artificiality of systems of indexing. When data of any sort are placed in storage, they are filed alphabetically or numerically, and information is found (when it is) by tracing it down from subclass to subclass. It can be in only one place, unless duplicates are used; one has to have rules as to which path will locate it, and the rules are cumbersome. Having found one item, moreover, one has to emerge from the system and re-enter on a new path. [...] The human mind does not work that way. It operates by association."* (Bush, 1945, p. 101)

In the article Bush outlines a vision of a writing-table-looking machine, *Memex* ('memory extender'), functioning as a researcher's personal tool for associative indexing. Text paragraphs on microfilm that were separately displayed on the built-in screens of the *Memex* could be connected by so-called *trails* (analogous to paths of links on the Web) inserted by the researcher. Some central quotes from the article may illustrate Bush's ideas:

> *"any item can be joined into numerous trails" ... "gathered together to form a new book" ... "builds a trail of his interest through the maze of materials available to him" ... "new forms of encyclopedias will appear, ready-made with a mesh of associative trails"* (p. 104-105).

Bush envisaged that a researcher could share his personalized view of a 'trail-blazed' interconnected document universe encompassing diverse scientific domains, by

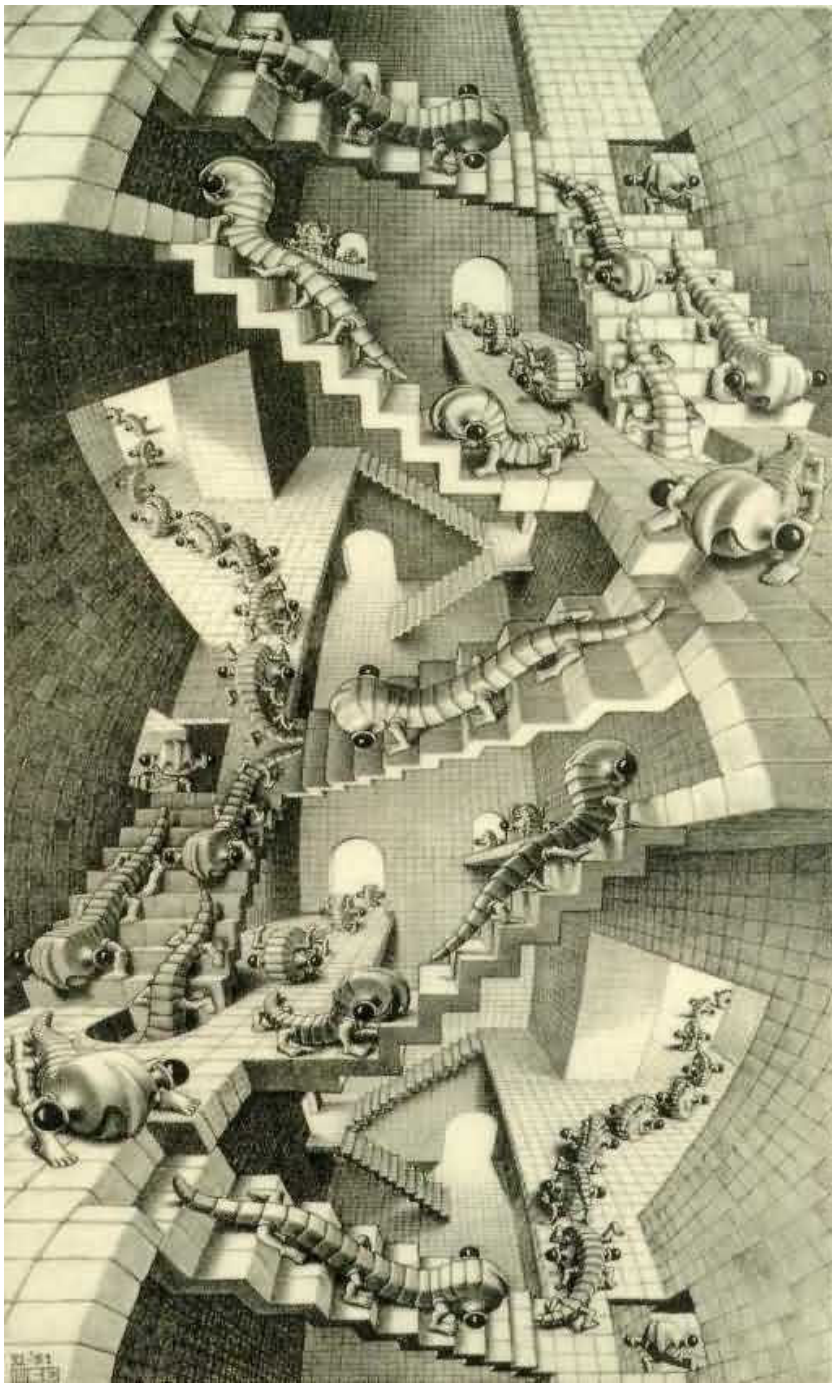
distributing his trails photographed on microfilm to other researchers, each having a personal *Memex*.

Even though Bush's *Memex* was never constructed in real life, his ideas of a personal tool for associative indexing had a strong impact on key persons behind the development of personal computers, hypertext, the Internet and the WWW (cf., e.g., Simpson *et al.*, 1996).

The visionary ideas of Bush also influenced the author of this dissertation, with regard to how scientific serendipity and creativity may be affected by so-called *small-world* phenomena in the shape of short distances on link paths (*'trails'*) between 'distant' topical domains due to decentralized link creations across the Web.

Apropos the art of serendipity and finding one's way through a PhD project – walking long enough along transversal trails:

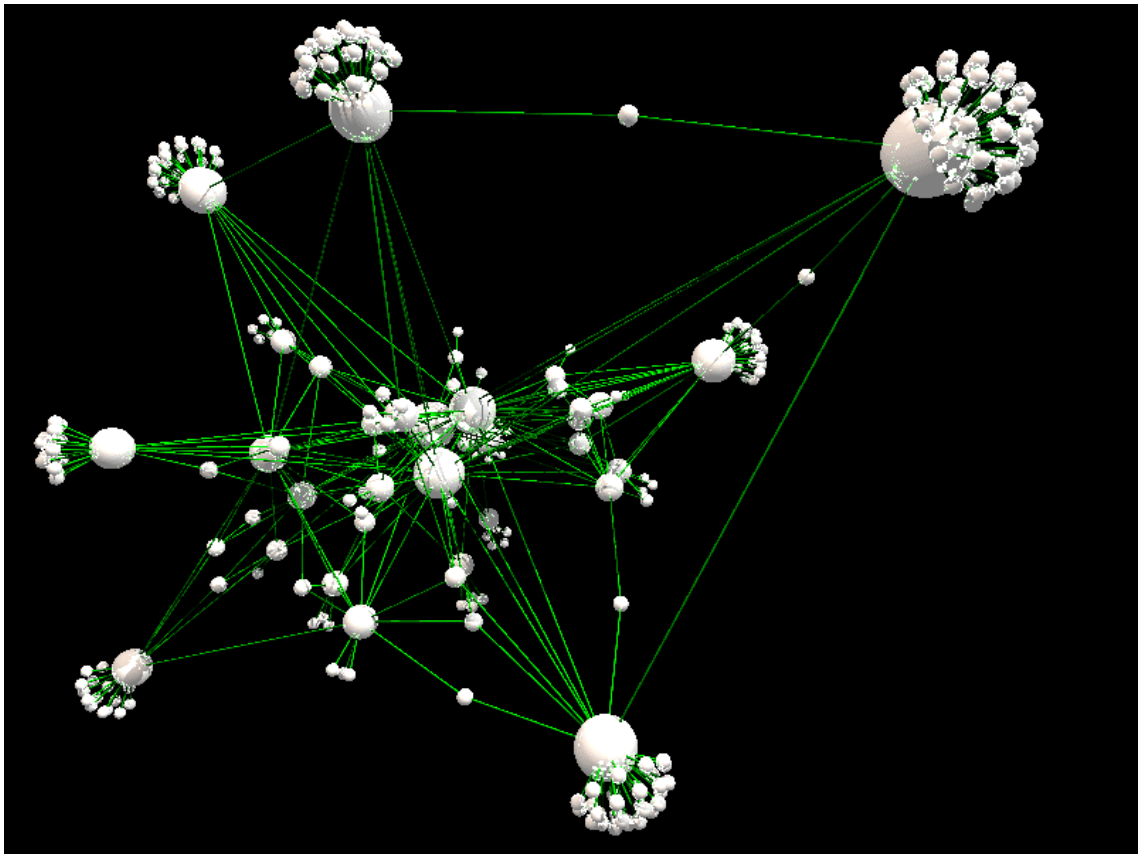
> *"'Cheshire Puss,' she began, rather timidly, as she did not at all know whether it would like the name: however, it only grinned a little wider.*
> *'Come, it's pleased so far,' thought Alice, and she went on.*
> *'Would you tell me, please, which way I ought to go from here?'*
> *'That depends a good deal on where you want to get to,' said the Cat.*
> *'I don't much care where--' said Alice.*
> *'Then it doesn't matter which way you go,' said the Cat.*
> *'--so long as I get somewhere,' Alice added as an explanation.*
> *'Oh, you're sure to do that,' said the Cat, 'if you only walk long enough."*
>
> (Lewis Carroll, *Alice's Adventures in Wonderland,* 1865. Chapter VI)



M.C. Escher, 1951. House of Stairs (Trappenhuis).[1]

[1] Anticipating the art of crawling the Web along transversal trails.

"Connecto ergo sum." (Björneborn, 1998)



(Wood *et al.,* 1995)[2]

[2] Visualizing link structures in a web space, with web pages as nodes.

Contents

List of Figures

List of Tables

1 Introduction

> *"New technologies alter the structure of our interests: the things we think about.*
> *They alter the character of our symbols: the things we think with.*
> *And they alter the nature of community: the arena in which thoughts develop."*
> (Postman, 1993, p. 20)

Library and information science (LIS) is concerned with how different information resources and information structures (interrelations of information resources) are generated, organized, distributed and utilized by different users in different contexts. Core research areas in LIS are concerned with *documents* (broadly defined as information carriers containing text, graphics, audio, video, etc.), *document representations* (for example, bibliographic data and metadata), and *relations* between the documents or document representations (for instance, links, cross-references, citations, co-citations, and bibliographic couplings). The World-Wide Web thus constitutes an obvious research area for LIS, being a *document network* with documents in the shape of web pages interconnected by links into complex hypertext structures.

The Internet-based hypertext system *World-Wide Web*, WWW, was launched in 1991, initially just as an internal 'intranet' for researchers affiliated to the European research centre for nuclear physics, CERN, and was proposed by Berners-Lee (1989/1990) and Berners-Lee & Cailliau (1990) as a tool to facilitate geographically dispersed CERN researchers' information sharing through easy access to online publishing and browsing.

In their WWW project, Berners-Lee and his colleagues at CERN employed Internet technologies developed since the launch of the first inter-university computer networks in 1969 (cf., e.g., Guice, 1998; Abbate, 1999) and merged them with hypertext technologies drawing on research since the mid-1960s where Ted Nelson (cf. Nelson, 1967) coined the term *hypertext* for electronically supported transitions or 'jumps' between text units, using the Greek term *hyper* for *over*, *beyond*, *transcendent*.

The WWW technology was made freely available for non-CERN individuals, institutions, and companies world-wide in 1993 (Cailliau, 1995). Subsequently, in the span of less than a decade, the World-Wide Web (hereafter: the Web) evolved from the small initial hypertext project into the largest repository and richest source of information – and misinformation – ever known to man (cf. Weare & Lin, 2000; Lyman & Varian, 2000). The Web is the fastest growing medium ever, strongly spurred by the fact that broader social interests vastly extended the technology's originally intended function (Granic & Lamey, 2000). The Web has thus proliferated into all spheres of human enterprise as a medium for social, cultural, political, economic, and scientific interaction (cf. e.g., Castells, 1996; 2001). Today, this ever-evolving global document network probably contains well over five billion web pages – not including database

request-generated document formats, etc., in the so-called *Deep Web* (Bergman, 2001) – interconnected by over 50 billion links.[3]

1.1 Brief characteristics of the self-organizing Web

The Web has become a highly complex and dynamic conglomerate of a diversity of information carriers constructed and utilized by a diversity of actors for a diversity of purposes. The Web thus increasingly integrates and influences all types of knowledge organization and information finding[4] – core research areas in library and information science as mentioned above. In this context, the Web could be characterized by being *'3D'*: *distributed*, *diversified* and *dynamic* (Björneborn & Ingwersen, 2001). Hypermedial documents containing text, graphics, audio, video, etc., are *distributively* located on millions of web servers. The distributiveness of the Web reflects the 'bottom-up' decentralized construction of the Web; as local individual and institutional inputs in a global collective medium. The content and link targets of web pages reflect the immense *diversity* of human enterprise and interests ranging from commercial commodities, pornographic material, and Nazi propaganda, to extensive scholarly work available on the Web. The last 'D' for *dynamic* is concerned with the continuous changes and mutual adaptations of content and links across the Web.

The Web is thus a new type of information system without central control, without centrally coordinated acquisition and indexing of contents. Contrary to traditional information systems such as libraries and bibliographic databases, the Web is constructed in a distributed and self-organizing way like an ecological system (cf. Pirolli *et al.*, 1996; Pitkow, 1997; Bøgh Andersen, 1998; Huberman *et al.*, 1998; Huberman & Adamic, 1999; Kleinberg & Lawrence, 2001) by millions of individuals, institutions, companies, etc., that dynamically create, adapt and remove web pages and links. As stated by Huberman *et al.* (1998),

> *"... the sheer reach and structural complexity of the Web makes it an ecology of knowledge, with relationships, information 'food' chains, and dynamic interactions that could soon become as rich, if not richer, than many natural ecosystems."* (p. 97).

However, the lack of central control and coordination does not imply that a distributed information system as the Web necessarily is totally chaotic and unordered. On the contrary, analyses of the Web reveal a remarkable degree of *self-organization* in the shape of aggregated link structures that reflect topic-focused web clusters, e.g., in the shape of interest communities related to work and leisure (cf. Clever Project, 1999; Kumar *et al.*, 1999; Kleinberg & Lawrence, 2001; Girvan & Newman, 2002; Flake *et al.*, 2002). Such link aggregations can be research-related, for instance, web clusters of interlinked web pages and web sites of researchers and their projects, papers and

[3] Based on conservative extrapolation of estimates from Lyman & Varian (2000) and Broder *et al.* (2000).

[4] The term *information finding* is here used as a generic term for information searching, browsing, and serendipitous information encountering.

institutions within a scientific domain. Other web clusters comprise topic-specific web portals, subject gateways and resource guides.

In this context, the self-organization of link structures on the Web may be conceived as *macro-level* aggregations of *micro-level* interactions; as *'collaborative weaving'* of an evolving global document network conducted by a multitude of local link creators.[5] In that respect, the Web is thus similar to complex social networks that do not have an engineered architecture but are self-organized by the local interactions of a large number of individuals and groups. From their micro-level positions, local link creators on the Web cannot overview how their links fit into the complex and dynamic macro-level link structures – as there exists no global registry and mapping of the Web. One may thus say that the self-organizing link topologies of the Web emerge through collectively *non-engineered* and *non-intentional* link aggregations. The term *topology* is here used in a network analytic and graph theoretic sense as the geometrical/spatial arrangement of connections in a network.

1.2 Small-world phenomena in the Web graph

In recent years there has been a strongly increasing research interest in investigating the dynamics and intricate structures of web link topologies. Especially researchers from computer science, physics and mathematics but also information scientists have applied methods from graph theory and social network analysis to treat the Web as a so-called *directed graph* consisting of *nodes* (or *vertices*) in the shape of web pages or web sites connected by directed *edges* (or *arcs*) in the shape of directed hyperlinks. Such approaches have been used, for example, for identifying web graph components (Broder *et al.*, 2000), inferring web communities (Gibson, Kleinberg & Raghavan, 1998; Clever Project, 1999), identifying authoritative web pages (Kleinberg, 1999a; Cui, 1999), topic distillation (Bharat & Henzinger, 1998), or improving search engine ranking algorithms as in Google (Brin & Page, 1998).

Furthermore, this research – which will be more outlined in Chapter 3 – has shown that the Web contains structural properties in the shape of so-called *small-world phenomena* and *scale-free* features typically including skewed *power-law* distributions of connections resembling those found in other dynamic and complex networks such as social networks, ecological food webs, neural networks, etc. (cf., e.g., Albert, Jeong & Barabási, 1999; 2000; Adamic, 1999; Barabási, 2001; 2002; Albert & Barabási, 2002).

In current research on complex networks, the Web plays an important role, because the Web is the largest complex network for which topological information presently is available (Albert & Barabási, 2002). The Web has thus become a *testing ground* (ibid.) and *model system* (Barabási, 2002, p. 178) for many current research efforts to build models of the emergence and dynamics of complex networks. This

[5] In this context, an interesting etymological point is that the term 'text' comes from the Latin *texere* meaning 'weave'.

explains the large interest in Web link topologies also from physicists, mathematicians, and other researchers in complex networks.

The abovementioned *small-world phenomena* are in focus in this dissertation. Such phenomena are concerned with short distances along link paths between nodes in a network graph (cf. Chapter 3). For example, short distances between two arbitrary persons through intermediate chains of acquaintances of acquaintances as studied in social network analysis (e.g., Milgram, 1967), and popularized by the notion of 'six degrees of separation'. In a seminal paper, Watts & Strogatz (1998) introduced a small-world network model characterized by highly clustered nodes as in regular graphs, yet with short characteristic path lengths between pairs of nodes as in random graphs. The revival of small-world theory commenced by Watts & Strogatz (ibid.) catalyzed an avalanche of research in a wide range of scientific domains regarding small-world phenomena occurring in a broad variety of biological, biochemical, physical, technical and social networks, such as, for instance, *brains* (Sporns, 2003; Bohland & Minai, 2001), *electricity power grids* (Watts & Strogatz, 1998), and *Internet router networks* (Watts, 1999c; Yook *et al.*, 2002; Jin & Bestavros, 2002) just to mention a few areas (see Chapter 3 for a more extensive coverage). Small-world network features combining high clustering and short link distances affect the diffusion speed of properties as, for example, data, energy, signals, contacts, ideas, economic values or epidemics across the networks in question.

As noted earlier, small-world properties have also been identified on the Web (cf., e.g., Albert, Jeong & Barabási, 1999; Adamic, 1999; Barabási, 2001; Albert & Barabási, 2002). In this context, it is noteworthy that the originators of the World-Wide Web early on envisaged small-world properties: "Yet a small number of links is usually sufficient for getting from anywhere to anywhere else in a small number of hops" (Berners-Lee & Cailliau, 1990).

In a widespread paper 'Diameter of the World-Wide Web', Albert, Jeong & Barabási (1999) constructed a topological model of the Web indicating that two randomly chosen documents on the Web were on average only "19 clicks" (ibid.) away from each other, that is, a link path consisting of 19 links would, on average, be sufficient to connect any two web pages. Later on, in their so-called *'bow-tie'* model of the Web, Broder *et al.* (2000) showed that such short link distances are only present in specially well-connected areas of the Web graph, in the so-called *Strongly Connected Component* – the 'bow-tie knot' (cf. Sections 3.5 and 5.1).

As indicated above, the coincidence of high local clustering and short global separation (Watts, 1999a) means that small-world networks simultaneously consist of small local *and* global distances, leading to high efficiency in propagating information both on a local and global scale (Marchiori & Latora, 2000). However, web links do not *directly* channel information flows like social networks, neural networks or computer networks. On the other hand, web links *indirectly* reflect information diffusion among link creators, because added or removed links may reflect changes in the link creators' knowledge, ideas, topical interests, social preferences and contacts. On an aggregated macro level, such dynamic link adaptations thus could reflect cognitive, cultural and social currents and formations, including the emergence of scholarly networks and the diffusion of scientific ideas across topical domains in such networks.

1.3 Motivation and objective

From a library and information science perspective, there is still a lack of research on investigating small-world phenomena and their possible usabilities regarding different types of *nodes* and *edges* in *informational networks* such as the Web, bibliographic and citation databases, semantic networks, thesauri, etc. (cf. Chapter 3).

On the Web, small-world phenomena are concerned with core library and information science issues such as *navigability* and *accessibility* of information across vast document networks. For instance, short distances along link paths on the Web affect the speed and exhaustivity with which search engine web crawlers can traverse and harvest the Web when following links from web page to web page. Furthermore, as noted by Adamic (1999), small-world link topologies of the Web may have implications for the way users surf the Web and the ease with which they gather information. The Web's self-organizing link topology – that reflects what here will be called *distributed knowledge organization*[6] – thus determines how information is interconnected on the Web and hence how efficiently information may be located on the Web (cf. Barabási, Albert & Jeong, 1999).

From a library and information science perspective, it would thus be interesting to investigate how the Web, as a new kind of information system – constructed through the abovementioned distributed knowledge organization by millions of web creators – may facilitate information access. However, before it is possible to investigate possible relations between *structure* (topology) and *functionality* (e.g., accessibility and navigability) in an information system; between the system's knowledge organization and the users' information behavior; it is necessary to establish an understanding how the knowledge organization *actually* is constructed across the information system.

The dissertation is concerned with a special aspect of the distributed knowledge organization on the Web; small-world link structures. So far, research on small-world phenomena in complex networks including the Web has yielded important results regarding *overall* structural factors like graph components, clustering coefficients, characteristic path lengths, scale-free properties including power-law frequency distribution of network connections, etc. These factors are further outlined in Chapter 3. However, there is a need to reveal more details on what actually occur on a *micro-level* scale that contribute to the formation of cohesive macro-structures displaying small-world properties in the distributed knowledge organization of the Web's document network. This realization gives rise to the following overall question in this dissertation:

- **What kind of links, web pages and web sites actually contribute to the emergence of small-world phenomena on the Web?**

Motivated by the urge to respond to such a challenging question, this dissertation is thus an attempt to cast more light into small-world properties of the Web from a LIS perspective. Furthermore, the dissertation focuses on the *structural* link creation aspect rather than the *functional* link traversal aspect of the Web – however symbiotic both

[6] This term is also used by, e.g., Bopp & Hampel (2001).

aspects are for truly understanding hypertextuality, since the very essence of hypertext is the ability both to *create* links and to *follow* them across a hypertextual network.

In this context, the *objective* of this dissertation is:

- to develop a conceptual framework and empirical methods concerning the identification and characterization of whether and how small-world phenomena emerge in link structures across an academic web space;
- to identify and characterize phenomena especially in relation to links that cross disciplinary borders in an academic web space
- to yield a better understanding of what factors contribute to the formation of an interconnected topology of link structures across an academic web space.

The focus on an *academic* web space as a setting for the empirical investigation in the dissertation was chosen for three main reasons:

1. The author's research interest in how the Web may be used to stimulate *scientific* creativity across topical domains;
2. A webometric 'tradition' for investigating academic web spaces with lineages to bibliometric and scientometric traditions (cf. Sections 2.4.1 and 2.4.2);
3. A link data set available and suitable for studying small-world link structures was covering the academic web space of 109 UK universities (cf. Chapter 4).

Indeed, academic web spaces are interesting because the Web highly reflects and facilitates scholarly activities. As indicated earlier, the Web was initially developed for scholarly use (Berners-Lee & Cailliau, 1990) and has today become an exceedingly important platform for both formal and informal scholarly communication and collaboration – also across scientific domains (cf. e.g., Cronin *et al.*, 1998; Hurd, 2000; Zhang, 2001; Thelwall & Wilkinson, 2003a; Wilkinson *et al.*, 2003). This aspect is further elaborated in Chapter 2. The new research field of *webometrics* thus offers potentials of tracking and 'mining' aspects of scientific endeavor traditionally more hidden for bibliometric or scientometric studies, for instance, the use of research results in teaching and by the general public (Björneborn & Ingwersen, 2001; Thelwall & Wilkinson, 2003a).

The dissertation is thus situated within this new research field of webometrics – a term introduced by Almind & Ingwersen (1997) – that has grown out of the library and information science-related fields of informetrics, bibliometrics and scientometrics (cf. Bar-Ilan, 2001; Björneborn & Ingwersen, 2001; forthcoming; Thelwall, Vaughan & Björneborn, forthcoming). Webometrics is thus concerned with a wide range of studies of quantitative aspects of link structures, page content, search engine performance, and users' information behavior on the Web (cf. Chapter 2).

In accordance with the reasons for focusing on an academic web space as outlined above, the *original* idea in the PhD project was to investigate how small-world link structures may affect the potential for *serendipity*, computer-supported *knowledge discovery* (*'data mining'*) and *creativity stimulation* in information spaces. The plan was to include so-called small-world *co-linkage chains* (chains of co-linking and co-linked nodes analogous to bibliographic couplings and co-citations, cf. Section 2.3.1) and other

topological factors. The underlying ideas follow approaches from, e.g., de Bono (1967), Swanson (1986), Bawden (1986), O'Connor (1988); Davies (1989); Van Andel (1994), de Jong & Rip (1997), Ford (1999), Björneborn (2001a) and Björneborn & Ingwersen (2001) – inspired by the early visions by Bush (1945) on using an information system for creativity stimulation, as noted earlier in the Prelude.

The intuition behind this initial research idea was that the shorter link distances there are between web nodes belonging to different topical domains on the Web, the larger the probability is to encounter and discover something unexpected when human web surfers and digital web crawlers traverse link structures on the Web. A hypothesis was that so-called *transversal links* (Björneborn, 2001a; Björneborn & Ingwersen, 2001; cf. Section 2.3.1) across dissimilar topical web domains could give useful hints for finding unexpected relations between scientific disciplines in order to identify fertile areas for cross-disciplinary exploration. Possibilities for scientific creativity and discovery could thus be stimulated by small-world phenomena emerging through topical 'inconsistencies' and diversities due to the earlier mentioned multi-purpose and multi-creator construction that also take place in the academic regions of a distributed information system as the Web. In that respect, the original idea was also to include aspects from the information retrieval theory of *poly-representation* (Ingwersen, 1992; 1994). This theory is concerned with how to explore and exploit the diversity of cognitive representations (e.g., title, metadata, bibliographic references, outlinks, inlinks, as well as body text elements) relating and pointing to scientific documents, for example, on the Web.

However, this ambitious approach turned out to be too infeasible to pursue due to difficulties in developing tractable methodologies and objective selection criteria to identify and capture serendipitous information encounters and computer-extracted knowledge discoveries. Nevertheless, some aspects of this original idea will be discussed in Chapter 7, for instance, regarding the *underlying hypothesis* in the dissertation that small-world phenomena emerge through *complementarities* of contrasting elements of topic-focused *uniformity* (clusters) and topic-scattering *diversity* (boundary-crossing shortcuts: transversal links) in link structures on the Web.

1.4 Research questions

As stated above, the overall objective of the dissertation is to develop a conceptual framework and empirical methods concerning the identification and characterization of how small-world phenomena emerge in link structures in an academic web space. Further, such insights are intended to yield a better understanding of what factors contribute to the formation of an interconnected topology of link structures across an academic web space. In this context, the dissertation is concerned with answering the following four research questions:

1. *How cohesively interconnected are link structures in an academic web space?*
2. *In particular, to what extent can so-called small-world properties be identified in this web space?*

3. *If small-world link structures can be identified in this academic web space, which properties can be observed that contribute to such link structures?*
4. ***Especially, what types of web links, web pages and web sites function as cross-topic connectors in small-world academic web spaces?***

The first three research questions regard more general aspects of cohesion, interconnectivity and small-world properties in an academic web space leading up to the fourth and main research question regarding types of links, pages and sites functioning as cross-topic connectors or 'binding elements' across dissimilar topical domains in small-world academic web spaces. The main and fourth research question is thus concerned with a *LIS approach* to answering what micro-structure web activities and elements that contribute to cohesive macro structures on the Web. In other words, the dissertation deals with how basic *'topology generators'* of the Web may be identified and characterized.

1.5 Dissertation structure

The dissertation structure falls into three overall parts: a theoretical background, an empirical investigation, and an analytical discussion and perspectivation.

The *theoretical* background comprises two chapters. Chapter 2 on 'Webometrics' introduces a webometric framework including how webometrics may be placed in a bibliometric context. The chapter also proposes a consistent link terminology and web node diagrams for conducting webometric link structure analysis – as in the dissertation. Furthermore, the chapter contains a literature review on research on link structures in academic web spaces. Chapter 3 on 'Small-world networks' is concerned with describing small-world research drawing on theories and methodologies in graph theory and social network analysis. The chapter includes a literature review on small-world approaches to Web studies.

The *empirical* investigation spans three chapters. Chapter 4 on the 'UK link data' outlines the original collection of link data from a harvest in 2001 of 109 UK academic web sites as well as methodological considerations regarding the extraction of a data subset comprising 7669 university subsites to be analyzed in the present study. Chapter 5 on 'Basic graph measures of the UK academic subweb' implements a wide range of graph measures including a developed *'corona'* graph model of link connectivity structures in the data set of the UK academic web space in order to primarily answer the first two research questions (cf. Section 1.4):

1. *How cohesively interconnected are link structures in an academic web space?*
2. *In particular, to what extent can so-called small-world properties be identified in this web space?*

Chapter 6 on 'Five-step methodology' gives a detailed presentation of a developed methodology for sampling, identification and characterization of small-world properties in an academic web space in order to primarily answer the last two research questions:

3. *If small-world link structures can be identified in this academic web space, which properties can be observed that contribute to such link structures?*

4. *Especially, what types of web links, web pages and web sites function as cross-topic connectors in small-world academic web spaces?*

The developed methodology includes the use of the Internet Archive (*www.archive.org*) as a *'web archaeological'* tool for 'old' web data. Chapter 6 includes detailed case studies of 10 so-called *'path nets'* (subgraphs with all shortest link paths between a pair of nodes) that function as investigable and illustrative small-world link structures – *'mini small worlds'* – constructed by juxtaposition of pairs of dissimilar topical seed nodes. The employed detailed case studies of such small-world link structures enable generation and development of concepts and hypotheses.

An *analytical* discussion and perspectivation is brought in Chapter 7 based on the findings of the empirical investigation. The discussion follows three 'trails' or 'cross-sections': (1) the role of personal and institutional link creators for the emergence of cohesive small-world link structures in a distributed and 'non-engineered' academic document space; (2) hypothesized complementarities of topical uniformity and diversity in the formation of small-world link structures – as well as hypothesized complementarities of so-called *topic drift* and *genre drift*; and (3) possible implications for exploratory capabilities including serendipity in *'crumpled-up'* small-world web spaces. Finally, possible implications for overall library and information science frameworks are discussed. Chapter 8 concludes the dissertation with summaries of how the research questions have been answered, and what contributions the dissertation has made, as well as possible applications regarding the findings in the dissertation.

The dissertation is 'non-mathematical' in the sense that there is limited display of mathematical formula, etc. Instead, the mathematical-rich small-world theory and graph theory are presented in verbalized form. This form of dissertation is in accordance with the employed library and information science approach to small-world link structures across an academic web space.

At the beginning of this PhD project three years ago, this mastodon-sized dissertation was not anticipated. However, the scope of the overall objective, the four research questions, and not least the richness of aspects discovered in the close examination of the UK academic web space and shortest link paths resulted in this behemoth.

2 Webometrics

This chapter gives a contextual background and overview of webometrics as a new LIS research domain, especially regarding academic web spaces as is the focus of this dissertation. Section 2.1 outlines the historical background of webometrics and heritage in bibliometrics. Webometrics is contextualized with other information science metrics like bibliometrics, informetrics, scientometrics and cybermetrics in Section 2.2. Section 2.3 presents a consistent link terminology and web node diagrams for conducting webometric link structure analysis – as in the present dissertation. Finally, Section 2.4 gives a literature review on webometric research in academic web spaces – the setting of the dissertation.

2.1 Historical background

Library and information science and related fields in the sociology of science and science and technology studies have developed a range of theories and methodologies – now including webometrics – concerning quantitative aspects of how different types of information are generated, organized, distributed and utilized by different users in different contexts. Historically, this development arose during the first half of the twentieth century from statistical studies of bibliographies and scientific journals (Hertzel, 1987). These early studies revealed bibliometric power laws like *Lotka's law* on productivity distribution among scientists (Lotka, 1926); *Bradford's law* on the scattering of literature on a particular topic over different journals (Bradford, 1934); and *Zipf's law* of word frequencies in texts (Zipf, 1949). Similar power-law distributions have been identified on the Web, e.g., the distribution of TLDs (top level domains) on a given topic (Rousseau, 1997) or inlinks per web site (Albert, Jeong & Barabási, 1999; Adamic & Huberman, 2000; 2001). Section 3.5 further outlines power-law distributions on the Web.

Decisive for the development of bibliometrics and scientometrics was the arrival of citation indexes of scientific literature introduced by Garfield (1955) that enabled analyses of citation networks in science (e.g., Price, 1965). Access to online citation databases catalyzed a wide range of citation studies, especially mapping scientific domains, including growth, diffusion, specialization, collaboration, impact and obsolescence of literature and concepts (cf. e.g., White & McCain, 1989; Borgman & Furner, 2002).

The breakthrough of online citation analysis parallels the later avalanche of webometric studies enabled by access to large-scale web data. In particular, the apparent yet ambiguous resemblance between citation networks and the hypertextual inter-document structures of the Web triggered much interest from the mid-1990s (e.g., Bossy, 1995; Moulthrop & Kaplan, 1995; McKiernan, 1996; Kuster, 1996; Larson,

1996; Downie, 1996; Rousseau, 1997; Almind & Ingwersen, 1997; Pitkow & Pirolli, 1997; Spertus, 1997; Ingwersen, 1998). Further, the central bibliometric measures of co-citation (Small, 1973) and bibliographic coupling (Kessler, 1963) have been applied to studies of web clustering, web growth and web searching (e.g., Larson, 1996; Weiss *et al.*, 1996; Pitkow & Pirolli, 1997; Efe *et al.*, 2000; Ding *et al.*, 2001; Menczer, 2002).

Since its advent, the Web has been widely used in both formal and informal scholarly communication and collaboration (e.g., Cronin *et al.*, 1998, Harter & Ford, 2000; Hurd, 2000; Zhang, 2001; Thelwall & Wilkinson, 2003a; Wilkinson *et al.*, 2003). As noted earlier, webometrics thus offers potentials for tracking aspects of scientific endeavor traditionally more hidden from bibliometric or scientometric studies, such as the use of research results in teaching and by the general public (Björneborn & Ingwersen, 2001; Cronin, 2001; Thelwall & Wilkinson, 2003a; Thelwall, Vaughan & Björneborn, forthcoming) – but also the actual use of scientific web pages.

A range of new terms for the emerging research field were rapidly proposed from the mid-1990s, for instance, *netometrics* (Bossy, 1995); *webometry* (Abraham, 1996); *internetometrics* (Almind & Ingwersen, 1996); *webometrics* (Almind & Ingwersen, 1997); *cybermetrics* (journal started 1997 by Isidro Aguillo)[7]; *web bibliometry* (Chakrabarti *et al.*, 2002). Webometrics and cybermetrics are currently the two most widely adopted terms, often used as synonyms. In the next section, webometrics and cybermetrics are defined in a bibliometric framework.

2.2 Webometrics and bibliometrics

Being a global document network initially developed for scholarly use as mentioned earlier (Berners-Lee & Cailliau, 1990) and now inhabited by a diversity of users, the Web constitutes an obvious research field for bibliometrics, scientometrics and informetrics. As noted above, webometrics and cybermetrics are currently the two most widely adopted terms for this emerging research field, often used as synonyms. However, the dissertation proposes a differentiated terminology distinguishing between studies of the Web and studies of *all* Internet applications (cf. Björneborn & Ingwersen, forthcoming). In this framework, *webometrics* is defined as:

> *The study of the quantitative aspects of the construction and use of information resources, structures and technologies on the Web, drawing on bibliometric and informetric approaches*.

This definition thus covers quantitative aspects of both the *construction* side and the *usage* side of the Web embracing four main areas of present webometric research:

1. web page *content* analysis;
2. web *link structure* analysis;
3. web *usage* analysis (e.g., log files of users' searching and browsing behavior);
4. web *technology* analysis (including search engine performance).

[7] Available at http://www.cindoc.csic.es/cybermetrics/

This typology includes hybrid forms, for example, Pirolli *et al.* (1996) who explored web analysis techniques for automatic categorization utilizing link graph topology, text content and metadata similarity, as well as usage data. Further, all four main research areas include longitudinal studies of changes on the dynamic Web, for example, of page contents, link structures and usage patterns. So-called *web archaeology* (Björneborn & Ingwersen, 2001) could in this webometric context be important for recovering historical web developments, for instance, by means of the Internet Archive (*www.archive.org*), as also demonstrated later in this dissertation.

The above definition places webometrics as a LIS specific term in line with bibliometrics and informetrics. This domain lineage is stressed by the formulation "drawing on bibliometric and informetric approaches" because "drawing on" denotes a heritage without limiting further methodological developments of web-specific approaches, including the incorporation of approaches of web studies in computer science, social network analysis, hypertext research, media studies, etc. In the present framework, *cybermetrics* is proposed as a generic term for:

> *The study of the quantitative aspects of the construction and use of information resources, structures and technologies on the <u>whole</u> Internet, drawing on bibliometric and informetric approaches*.

Cybermetrics thus encompasses statistical studies of discussion groups, mailing lists, and other computer-mediated communication on the Internet (e.g., Bar-Ilan, 1997; Hernández-Borges *et al.*, 1997; Matzat, 1998; Herring, 2002) *including* the Web. Besides covering all computer-mediated communication using Internet applications, this definition of cybermetrics also covers quantitative measures of the Internet backbone technology, topology and traffic (cf. Molyneux & Williams, 1999). The breadth of coverage of cybermetrics and webometrics implies large overlaps with proliferating computer-science-based approaches in analyses of web contents, link structures, web usage, and web technologies. A range of such approaches has emerged since the mid-1990s with names like *Web Mining* (e.g., Etzioni, 1996; Cooley *et al.*, 1997; Kosala & Blockeel, 2000), *Web Ecology* (e.g., Pitkow, 1997; Chi *et al.*, 1998; Huberman, 2001), *Cyber Geography* / *Cyber Cartography* (e.g., Girardin, 1995, 1996; Dodge, 1999b; Dodge & Kitchin, 2001; 2002)[8], *Web Graph Analysis* (e.g., Clever Project, 1999; Kleinberg *et al.*, 1999; Broder *et al.*, 2000), *Web Dynamics* (e.g., Levene & Poulovassilis, 2001), and *Web Intelligence* (e.g., Yao *et al.*, 2001).

The raison d'être for using the term *webometrics* in this context could be to denote a close lineage to bibliometrics and informetrics and stress a LIS perspective on Web studies as noted above. In this context, the earlier mentioned term (cf. Section 2.1), *web bibliometry*, as used by Chakrabarti *et al.* (2002), is especially interesting because computer scientists thus recognize the heritage in bibliometric research to be drawn upon in web studies. Other computer science approaches to link structure analysis also pay tribute to inspiration from citation studies (e.g., Pitkow & Pirolli, 1997; Kleinberg, 1999a; Chakrabarti *et al.*, 1999; Kosala & Blockeel, 2000; Efe *et al.*, 2000; Vázquez, 2001; Albert & Barabási, 2002).

[8] Cf. http://www.cybergeography.org/

There are different conceptions of informetrics, bibliometrics and scientometrics. The diagram in Fig. 2.1 below shows the field of informetrics embracing the overlapping fields of bibliometrics and scientometrics following widely adopted definitions by, e.g., Brookes (1990), Egghe & Rousseau (1990) and Tague-Sutcliffe (1992). According to Tague-Sutcliffe (1992), *informetrics* is "the study of the quantitative aspects of information in any form, not just records or bibliographies, and in any social group, not just scientists". Furthermore, *bibliometrics* is defined as "the study of the quantitative aspects of the production, dissemination and use of recorded information" and *scientometrics* as "the study of the quantitative aspects of science as a discipline or economic activity" (ibid.). In the figure, politico-economical aspects of scientometrics are covered by the part of the scientometric ellipse lying outside the bibliometric one.

The figure further shows the field of webometrics entirely encompassed by bibliometrics, because web documents, whether text or multimedia, are *recorded* information (cf. Tague-Sutcliffe's abovementioned definition of bibliometrics) stored on web servers. This recording may be temporary only, just as not all paper documents are properly archived. Webometrics is partially covered by scientometrics, as many scholarly activities today are web-based whilst other such activities are even beyond bibliometrics, i.e. non-recorded, like person-to-person conversation. Furthermore, webometrics is totally included within the field of cybermetrics as defined above.

In the diagram, the field of cybermetrics exceeds the boundaries of bibliometrics, because some activities in cyberspace normally are not recorded, but communicated synchronously, like in chat rooms. Cybermetric studies of such activities still fit in the generic field of informetrics as the study of the quantitative aspects of information "in any form" and "in any social group" as stated above by Tague-Sutcliffe (1992).

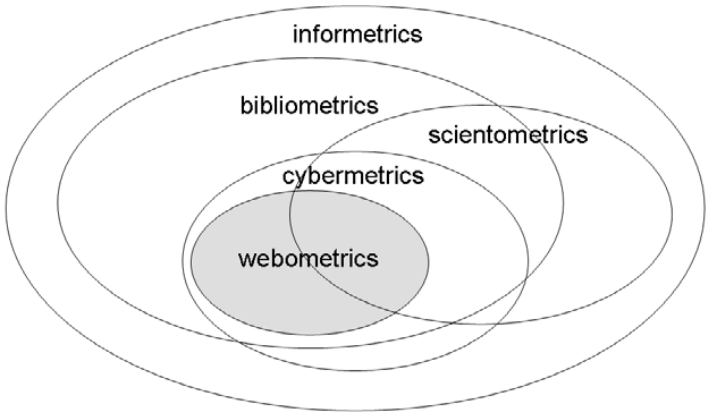


Figure 2.1. Relationships between the LIS fields of infor-/biblio-/sciento-/cyber-/webo-metrics. Sizes of the overlapping ellipses are made for sake of clarity only.

Naturally, the inclusion of webometrics expands the field of bibliometrics, as webometrics inevitably will contribute with further methodological developments of web-specific approaches. As ideas rooted in bibliometrics, scientometrics and informetrics contributed to the emergence of webometrics, ideas in webometrics might now contribute to the development of these embracing fields.

2.3 Conceptual framework

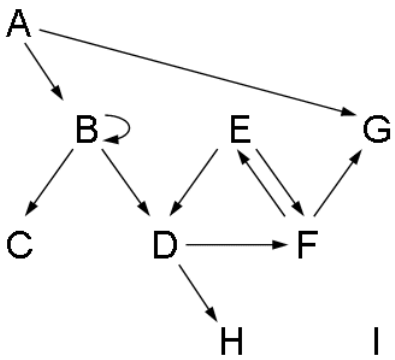
> *"The problem is not that the hypertexts lack structure but rather that we lack words to describe it."* (Bernstein, 1998)

A detailed link typology and terminology was developed in the course of the PhD project in the attempt to build a basic consistent conceptual framework for webometric link structure analysis. As a part of this conceptual framework, a novel web node diagram notation is further proposed in order to fully appreciate and investigate link structures and other webometric characteristics of the Web. The basic link terminology is outlined in Section 2.3.1, whereas the basic web node terminology and diagrams are presented in Section 2.3.2, followed by advanced web node diagrams in Section 2.3.3.

2.3.1 Basic link terminology

The initial exploratory phases of an emerging field like webometrics inevitably lead to a variety in the terminology used. For instance, a link received by a web node (the network term 'node' here denotes a unit of analysis like a web page, directory or web site but could also be an entire top level domain of a country) has been named, e.g., *incoming link*, *inbound link*, *inward link, back link*, and *sitation*; the latter term (McKiernan, 1996; Rousseau, 1997) with clear connotations to bibliometric citation analysis. An example of a more problematic terminology is the two opposite meanings of an *external link*: either as a link pointing out of a web site or a link pointing into a site.

Fig. 2.1 below illustrates the terminology used in the thesis to describe different link topologies, i.e. different degrees of cohesiveness and connectedness of link structures on the Web. In this context, the figure presents an attempt to create a consistent basic webometric terminology for link relations between web nodes. The figure reflects that the Web may be viewed as a *directed graph* (cf. Sections 1.2 and 3.1) using a graph theoretic term. In such a web graph, web nodes are connected by directed links. The proposed terminology in the legend of Fig. 2.1 has linkages to graph theory, social network analysis, and bibliometrics.



- B has an *inlink* from A; B is *inlinked*; A is *inlinking*; A is an *in-neighbor* of B
- B has an *outlink* to C; B is *outlinking*; C is *outlinked*; C is an *out-neighbor* of B
- B has a *selflink*; B is *selflinking*
- A has no inlinks; A is *non-linked*
- C has no outlinks; C is *non-linking*
- I has neither in- nor outlinks; I is *isolated*
- E and F have *reciprocal links*; E and F are *reciprocally* linked
- D, E and F have in- or outlinks connecting each other; they are *triadically interlinked*
- A has a *transversal* outlink to G: functioning as a shortcut
- H is *reachable* from A by a directed *link path*
- C and D are *co-linked* by B; C and D have *co-inlinks*
- B and E are *co-linking* to D; B and E have *co-outlinks*
- Co-inlinks and co-outlinks are both cases of *co-links*

Figure 2.2. Basic webometric link terminology. Letters A-I may represent different web node levels such as web pages, web directories, web sites, or top-level domains of countries or generic sectors.

The terms *outlink* and *inlink* are commonly used in computer-science-based Web studies (e.g., Pirolli *et al.*, 1996; Chen *et al.*, 1998; Broder *et al.*, 2000). The term *outlink* implies that a directed link and its two adjacent nodes are viewed from the source node providing the link, analogous with the use of the term *reference* in bibliometrics. A corresponding analogy exists between the terms *inlink* and *citation*, with the target node as the spectator's perspective, cf. Fig. 2.3 below. Similar considerations of consistent terminology have been put forward in bibliometrics, for instance, by Price (1970) emphasizing a conceptual difference between the terms *reference* and *citation*, matching the difference between *outlink* and *inlink*.

Figure 2.3. Different link terminology for the same link depending on the spectator's perspective as denoted by the eyes.

The terms *out-neighbor* and *in-neighbor* in the proposed terminology are also used in graph theoretic web research (e.g., Chakrabarti *et al.*, 2002).

On the Web, *selflinks* are used for a wider range of purposes than self-citations in scientific literature. This reflects a special case of the general difference between outlinks/inlinks and references/citations. Page selflinks point from one section to another within the same page. Site selflinks (also known as *internal links*) are typically navigational pointers from one page to another within the same web site.

Due to its dynamic and distributed nature the Web often demonstrates web pages reciprocally linking to each other – a case not normally possible in the traditional print-based citation world. *Reciprocal links* as between nodes E and F in Fig. 2.2 is a widespread existing web term for mutual inlinks and outlinks between two web nodes. This reciprocity is not necessarily completely symmetrical as there may be more links in one direction between two web nodes. Sometimes such reciprocal links may be deliberately agreed by two web site creators for attempting to obtain higher ranking in search engines employing inlink counts in ranking algorithms as in Google (Brin & Page, 1998; also cf. Walker, 2002).

In the figure, the *triadically linked* nodes D, E, and F correspond to the social network analytic term *triadic closure* (e.g., Skvoretz & Fararo, 1989), for instance, describing the probability that nodes E and F are connected if there are already links between D and E, and between D and F. In social networks, such simple *triadic structures* or *triads* are the building blocks of larger social structures (Wasserman & Faust, 1994; Scott, 2000)[9]. Triadic closure between sets of web nodes is essential with regard to the size of so-called *clustering coefficients* used in measuring small-world properties in a web graph. This measure is further described in Sections 3.3 and 5.3.2.

The concept of *transversal links* (Björneborn, 2000; 2001a; Björneborn & Ingwersen, 2001) is a key concept in the dissertation for denoting links that span across dissimilar topical domains on the Web. Such transversal links may affect small-world phenomena in the shape of short link paths as will be investigated in this dissertation. Most links on the Web connect web pages containing cognate topics (Davison, 2000). However, some links in a web node neighborhood may break such topical linkage patterns. Such transversal links function as shortcuts between dissimilar topical domains. In mathematics, the term *transversal* denotes a line that intersects a system of other lines.[10] In the PhD project, the concept of transversal links is inspired by Bush's (1945) earlier mentioned vision that researchers should be able to create and interchange associative 'trails' similar to hyperlinks and link paths that interlink text paragraphs *transversely* to classification hierarchies with implications for scientific creativity and innovativeness.[11]

---

[9] Milo *et al.* (2002) use the term '*motif*' for simple triadic building blocks of complex networks, e.g., in biochemistry, neurobiology, ecology, and engineering.

[10] 'Transversal' comes from Latin *transversus* for 'across'. Interestingly, *de transverso* means 'unexpectedly' in Latin – appropriate in the context of possibilities for serendipitous information encountering when following transversal links on the Web.

[11] Another source of inspiration was Sandbothe's (1996) media-philosophical analysis of "interactivity, hypertextuality, and transversality" on the Web, including so-called "*transversal reason*" and *rhizomatic* (i.e. networked entanglements) structures on the Web. The term *rhizome* originates from a botanical term for a kind of stem that burrows underground, sending out shoots and roots thus connecting plants into living networks (*www.rhizome.org*) also cf. e.g., Moulthrop (1994), Aarseth (1997), and Hardy (2001).

The use of the term 'transversal links' in the dissertation – as links between topically different web sites – is different from Tsikrika & Lalmas (2002) who use the term to denote links between different directories *within* a site, that is, site *selflinks* in the present terminology. In graph theoretic literature, a so-called *transversal* covers a sequence of distinct nodes that 'interconnect' all subsets (with overlapping nodes) belonging to a so-called *family* of subsets (cf. Gross & Yellen, 1999). In hypertext research, the term 'transversal links' has been used to describe links connecting dissimilar topic clusters by combining the topics, for example, a hypertext link between the topics 'digestion' and 'mammals' may illustrate a so-called *'transversal argument'*: "digestion in mammals" (Cossentino & Faso, 2001). The dissertation conceptualization of transversal links is more elaborated in Sections 6.5 and 7.1.2.

The concepts of *reachability* and link *paths* as illustrated in Fig. 2.2 are both used in graph theory (e.g., Gross & Yellen, 1999) and will be extensively used in the empirical chapters of the dissertation.

The two *co-linked* web nodes C and D in the figure with *co-inlinks* from the same source node are analogous to the bibliometric concept of *co-citation* (Small, 1973). Correspondingly, two *co-linking* nodes B and E having *co-outlinks* to the same target node are analogous to a *bibliographic coupling* (Kessler, 1963). *Co-links* is proposed as a generic term covering both the novel concepts of co-inlinks and co-outlinks. The underlying assumption for the use of both the bibliometric and webometric concepts is that two documents (or two authors/link creators) are more similar, i.e. more semantically related, the higher frequency of shared 'outlinks' (references) or shared 'inlinks' (citations). This assumption has not been investigated in the dissertation.

2.3.2 Basic web node terminology and diagrams

In webometric studies, it may be useful to visualize relations between different units of analysis, for example, in the so-called *Alternative Document Models* (Thelwall, 2002b; Thelwall & Harries, 2003), cf. Section 2.4.2.

Fig. 2.4 below shows a diagram illustrating some basic building blocks in a consistent web node framework (cf. Björneborn & Ingwersen, forthcoming) that will be used in the dissertation. In the diagram, four basic web node levels are denoted with simple geometrical figures: *quadrangles* (web pages), *diagonal lines* (web directories), *circles* (web sites) and *triangles* (country or generic top level domains, TLDs). Sublevels within each of the four basic node levels are denoted with additional borderlines in the corresponding geometrical figure. For example, a triangle with a double borderline denotes a generic second level domain (SLD), also known as a sub-TLD, assigned by many countries to educational, commercial, governmental and other sectors of society, for instance, *.ac.uk, .co.uk, .ac.jp, .edu.au.*

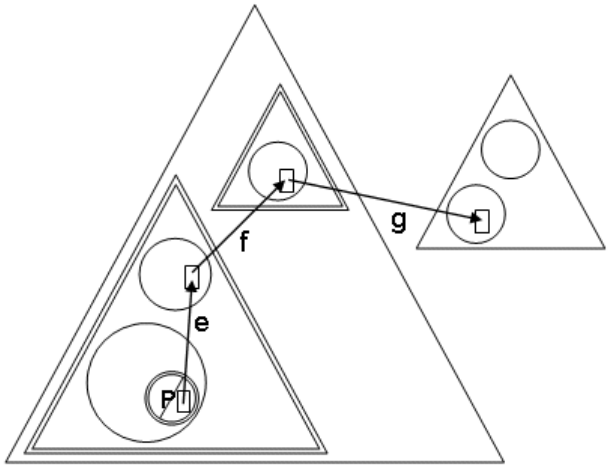


Figure 2.4. Simplified web node diagram illustrating basic web node levels.

The simplistic web node diagram in Fig. 2.4 above shows a page *P* located in a directory of a sub-site in a sub-TLD. The page has a site outlink, *e*, to a page at a site in the same sub-TLD. The outlinked page in turn is outlinking to a page at a site in another sub-TLD in the same country. The link path *e-f-g* ends at a page at a site in another TLD.

Zooming in on a single web site, this may comprise several sub-units in the shape of sub-sites, sub-sub-sites, etc., as indicated by hierarchically derivative domain names. For instance, as shown in Fig. 2.5 below, the sub-sub-site of *The Image, Speech and Intelligent Systems Research Group* (*isis.ecs.soton.ac.uk*) is located within the *Department of Electronics and Computer Science* (*ecs.soton.ac.uk*), one of many sub-sites at the *University of Southampton,* UK (*soton.ac.uk*). Sub-sites and sub-sub-sites are denoted as circles with double and triple borderlines, respectively. Subordinate sublevels would logically be denoted with additional number of borderlines. For sake of simplicity, the diagram does not reflect actual numbers and sizes of elements.

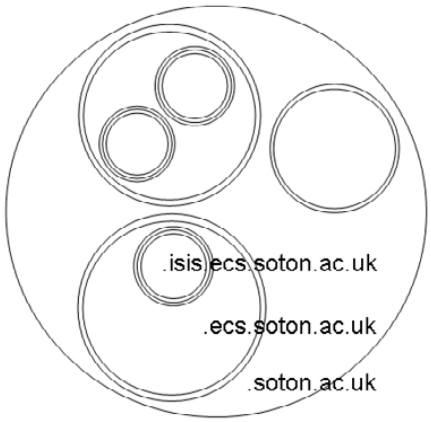


Figure 2.5. Simplified web node diagram of a web site containing sub-sites and sub-sub-sites.

While some web sites subdivide into derivative domain names, as shown above, other web sites locate the same type of subunits into folder directories in their web site file hierarchy. Obviously, such diverse allocation and naming practices complicate comparability in webometric studies. In Fig. 2.6a & 2.6b, one or more diagonal lines (resembling URL slashes and reflecting the number of directory levels below the URL root level) denote directories, sub-directories, etc.

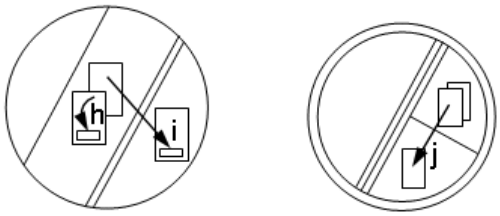


Figure 2.6a&b. Simplified web node diagrams of a web site and a sub-site, respectively, with links between different directory levels including page sub-elements.

Web pages may also consist of sub-elements such as text sections, frames, etc. Additional bands illustrate such page sub-elements as in the targets of the page selflink *h* and the page outlink *i* from the two sibling web pages in the same directory in Fig. 2.6a. More numerous and complex linkages within a site or sub-site, etc., can be illustrated by combinations of elements in Fig. 2.6a & b, showing links between pages located either at different directory levels (Fig. 2.6a) or in sibling directories at the same level (Fig. 2.6b) in the web site file hierarchies.

Naturally, any diagrammatic representation of *large-scale* hypertext structures will get too tangled to be of any practical use – less to be interpreted in any quantitative way. However, the proposed web node diagrams with their simple and intuitive geometrical figures are intended to be used in order to emphasize and illustrate *qualitative* differences between investigated web node levels in a webometric study. The dissertation will use the web node diagrams in this more abstract way. However, the diagrams will also be used in the dissertation to illustrate actual important structural aspects of *limited* subgraphs of the investigated UK academic web space.

2.3.3 Advanced link terminology and diagrams

The Web can be studied at different granularities employing what here will be called *micro*, *meso* and *macro* level perspectives (cf. Björneborn & Ingwersen, forthcoming). In this proposed framework, *micro level webometrics* consists of studies of the construction and use of web pages, web directories and small sub-sites, etc., for example, constituting individual web territories. *Meso level webometrics* is correspondingly concerned with quantitative aspects of larger sub-sites and sites, and *macro level webometrics* comprises studies of clusters of many sites, or focuses on sub-TLDs or TLDs. Several webometric studies, including classic ones by Larson (1996) and Almind & Ingwersen (1997), have used *meso level* approaches concerned with site-to-site interconnectivity as well as *macro level* TLD-to-TLD analysis, primarily applying page level link counts. However, in order to extract useful information, links

may also be aggregated on different node levels as in the earlier mentioned *Alternative Document Models* (Thelwall, 2002b; Thelwall & Harries, 2003), also cf. Section 2.4.2.

An adequate terminology for aggregated link relations should capture both the link level under investigation and the reach of each link. Such a terminology should reflect at least three elements: (1) the investigated link level; (2) the highest-level web node border crossed by the link; and (3) the spectator's perspective (cf. Fig. 2.3 in Section 2.3.1). For sake of simplicity, the perspective from the outlinking nodes is chosen in the following examples showing higher and higher link aggregations.

Fig. 2.7 below shows 14 *page level links* including a *page level sub-site outlink*, k_p (also being a *page level site selflink*). The subscript in k_p denotes *page* level. If a webometric study comprises just one level of links, the terminology can be simplified to cover merely the link reach. In such a case, l_p is a *site outlink*, m_p a *sub-TLD outlink*, and n_p a *TLD outlink*.

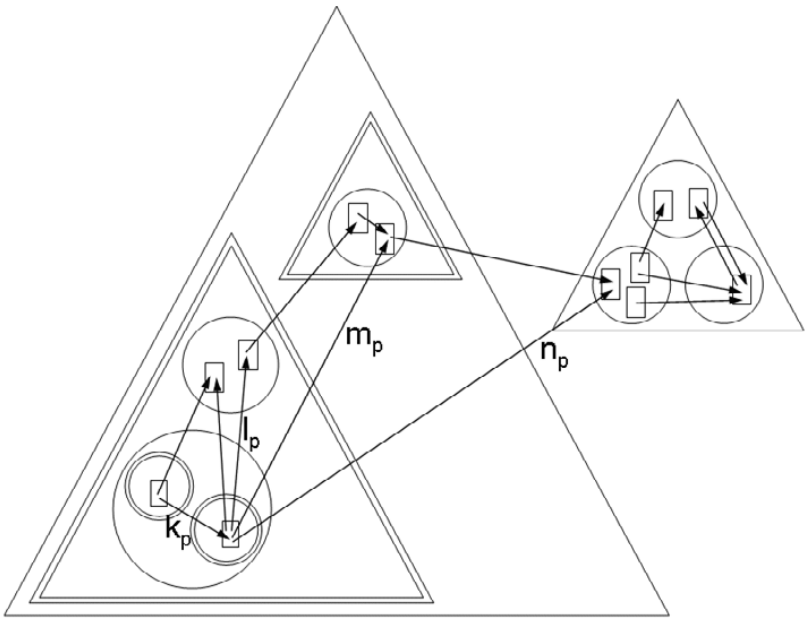


Figure 2.7. Web node diagram with *page* level links.

For sake of simplicity, *directory* and *sub-site* level links will not be treated here. However, the terminology for these levels would parallel the other levels included.

Fig. 2.8 below illustrates 11 *site level links*. For example, o_s is a *site level site outlink* aggregating 3 page level links from Fig. 2.7. Site selflinks are denoted with curved arrows.

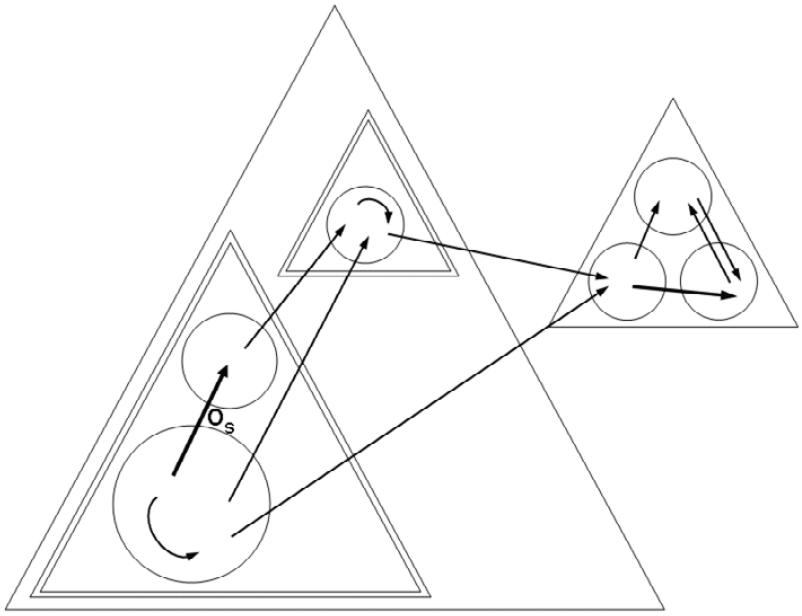


Figure 2.8. Web node diagram with *site* level links.

This line of higher and higher link aggregations ends with *sub-TLD level links* as shown in Fig. 2.9 and *TLD level links* in Fig. 2.10. Terminology for these levels parallel the other levels included.

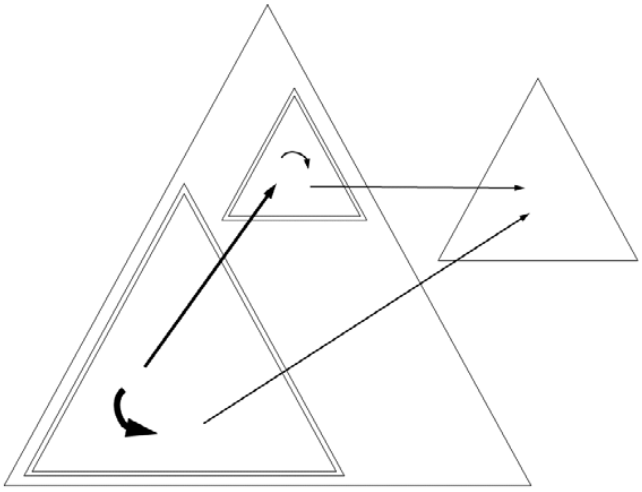


Figure 2.9. Web node diagram with *sub-TLD* level links.

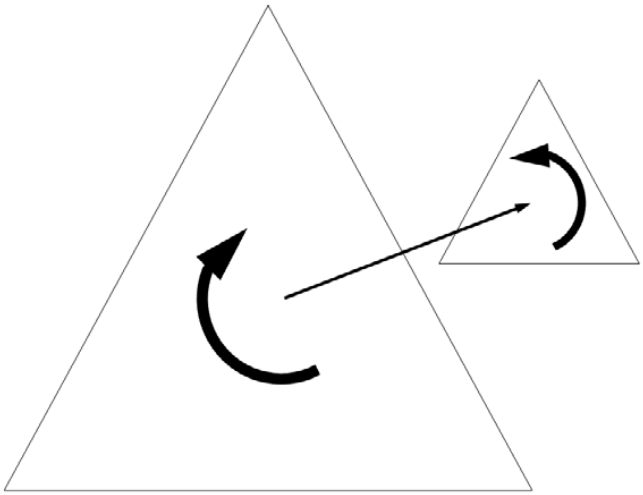


Figure 2.10. Web node diagram with *TLD* level links.

As further outlined in the empirical Chapters 4, 5 and 6, the dissertation investigates *page* level links and *site* level links (Fig. 2.7 and 2.8) in the UK academic web space.

2.4 Literature review

Webometrics is still in its infancy as a scientific domain – *"with its own different theories to be built, tasks to be done, units to be defined, methods to be developed and problems to be solved"* (Aguillo, 2002).

As already alluded to in Sections 2.1 and 2.2, there has been an increasing amount of published research since the mid-1990s concerned with the aforementioned four main webometric research areas: web content, link structures, users' searching and browsing behavior on the Web, and search engine performance. A forthcoming *ARIST* chapter on webometrics (Thelwall, Vaughan & Björneborn, forthcoming) gives an extensive review of this evolving scientific domain with regard to basic webometric concepts; methods for data selection, sampling, and collection; as well as webometric research in general, commercial, and academic web spaces. The chapter further includes topological modeling and mining of the Web, for instance, regarding small-world link structures. Another *ARIST* chapter gives a review on the related topic of "Measuring the Internet" (Molyneux & Williams, 1999). However, due to the rapid changes on the Internet, much of the latter chapter's coverage is subsequently unavoidably out of date. In his paper on 'Who can count the dust of Jacob', Rowlands (1999) reviews cybermetric approaches to quantitative aspects of document production and use on the Internet and locates these within a bibliometric framework.

A special issue in the 50th volume of *Scientometrics* was dedicated to Internet studies containing, amongst others, a paper on different perspectives of webometrics (Björneborn & Ingwersen, 2001) attempting to point to selected areas of webometric research that demonstrate interesting progress and space for development, for instance with regard to graph theoretic approaches to web studies including small-world

phenomena, as well as to some more problematic areas, for instance, the so-called *Web Impact Factor* (Ingwersen, 1998) facing methodological difficulties both with regard to reliability due to the dependence on secondary data from commercial search engines with opaque data coverage and performance, and with regard to validity due to problems with defining comparable units of analysis. These matters including the Web Impact Factor are further elaborated in Section 2.4.2.

A special issue of *JASIST* on webometrics is set to appear in 2004, containing a wide variety of different approaches to the quantitative study of the Web, including a basic conceptual framework of webometrics proposed by Björneborn & Ingwersen (forthcoming). Bar-Ilan & Peritz (2002) give an excellent review of "Informetric theories and methods for exploring the Internet" with focus on general informetric techniques to be applied in both web studies and non-web Internet research. Furthermore, Bar-Ilan (forthcoming) reviews search engine research in a forthcoming *ARIST* chapter. Data collection techniques in general on the Web are covered by Bar-Ilan (2001) and Thelwall (2002f).

Henzinger (2001) reviewed link structures analysis from a computer science perspective, showing how links could be used in search engine ranking algorithms. Barabási (2002) and Huberman (2001) have written popular science books explaining current research into mathematical modeling of the topology and growth of the Web including graph theoretic approaches to small-world phenomena on the Web. Chapter 3 contains a literature review regarding such graph theoretic as well as social network analytic approaches to the Web.

Another webometric review article, by Park & Thelwall (2003), compared information science approaches to studying the Web to those from social network analysis. It was found that information science tended to emphasize data validation and the study of methodological issues, whereas social network analysis suggested how its existing theory could transfer to the Web.

Besides the few selected examples above of more broadly covering webometric reviews, etc., it is beyond the scope of the dissertation to provide a more detailed review of the rich diversity of the increasing amount of webometric research. Instead, Section 2.4.2 further below gives a brief overview of webometric research performed in academic web spaces, since an academic web space – in the UK – is in focus in the dissertation. (Details of the UK academic data set are outlined in Chapter 4). In the preceding Section 2.4.1, a brief background on the sociology of academic web spaces is outlined.

2.4.1 Sociology of academic web spaces

As noted earlier in Section 1.3, a webometric 'tradition' has evolved – in less than 10 years of webometric research – for investigating academic web spaces. This webometric 'tradition' may be traced to similar bibliometric and scientometric focus on scholarly publication activities (e.g., White & McCain, 1989; Borgman & Furner, 2002). Being a research field that has grown out of bibliometrics as described in Section 2.2, it is thus not surprising that webometrics shows similar inclinations. The fact that the Web was initially developed for scholarly use (Berners-Lee & Cailliau, 1990), and today has

become an exceedingly important platform for both formal and informal scholarly communication and collaboration as mentioned earlier (e.g., Cronin *et al.*, 1998; Hurd, 2000; Zhang, 2001; Thelwall & Wilkinson, 2003a; Wilkinson *et al.*, 2003), naturally has contributed to this webometric research interest in academic web spaces.

The Web has thus had a significant – some say revolutionary (e.g., Goodrum *et al.* 2001) – impact on the entire scholarly communication process. According to many researchers in the sociology of science (e.g., Cronin & McKim, 1996; Cronin *et al.*, 1998; Hurd, 2000), the Web is reshaping the ways in which scholars communicate with each other. As Cronin & McKim (1996) put it with regard to the ways in which the Web may support and alter the conduct of scholarship:

> *"The Web is much more than a virtual analogue of existing archival and library institutions. It is a dynamic, interactive and evolving environment that supports new kinds of foraging and communication, in which scholars are anything but passive participants."* (p. 163)

In an earlier work, the early LIS recognizers of hypertext potentials, Davenport & Cronin (1990) emphasize how hypertext may affect the conduct and creativity of science by *"the freedom of movement inside and across texts"* enabled by hypertext that thus allows readers to see referenced sources instantaneously. Early pre-Web hypertext researchers as Yankelovich, Meyrowitz & van Dam (1985) also build on the 'Memex' vision of Bush (1945) for creating what they call *"webs of information"* as hypertextual research literatures enabling scholars *"to both create connections and follow those made by others"* and thus *"link scholars together"* (Yankelovich, Meyrowitz & van Dam, 1985, p. 16).

Today, the Web has enabled such world-wide interlinkage of scholars. On the Web, new kinds of scholarly and proto-scholarly publishing are emerging, implying that work-in-progress, early drafts, preprints and refereed articles are now almost immediately sharable. The Web thus provides fast and efficient means of disseminating and accessing scientific information, with scientists, institutions, and archives making formal research as well as work-in-progress publicly available on their web sites (Goodrum *et al.*, 2001). In other words, the Web offers scholars *"instantaneous and interlinked access" (*Miles-Board *et al.*, 2001) to large research literatures and other scholarly resources available on the Web. In this context, there is a clear trend, especially for younger researchers, to bypass subscription barriers and rely almost exclusively on what they can find free on the Web which often includes working versions posted on the home pages of the authors (Björk & Turk, 2000). This finding is supported by Lawrence (2001) who found a clear correlation between the number of times an article is cited and the probability that the article is online. As also stated by Zhang (2001):

> *"[...] scholars are using e-sources* [Internet-based electronic resources, LB] *as a channel to communicate with colleagues, known or unknown; to elicit research ideas from exchanges on mailing lists or newsgroups; to download preprints or reprints; and to seek research-related information. [...] communicating through the network allows researchers to reach broader audiences in an efficient way; hence, it extends the traditional "invisible colleges" model for scholarly communication in the networked environment. Scholars are also relying on e-*

> *sources as unique, useful, and current sources of information for research. They often consult e-sources when they need to find some factual, background, or contact information for their research. E-sources also provide efficient ways for scholars to track the progress of related research to stay current."* (p. 644)

One may thus say, that the Web – *in spite of* the ongoing massive 'colonization' by commercial and other non-academic web players – *also* still reflects the original idea behind the World-Wide Web developed by Berners-Lee (1989/1990): as a global platform for interlinking academic research by facilitating researchers' information sharing through easy access to online publishing and browsing. In that respect, the Web still reflects and facilitates scholarly activities. In other words – and in continuation of the quotation from Zhang (2001) above – the Web provides a richer and more easily accessible – *as well as* more diversified, muddled and cluttered – picture of scholars' scholarly *and* non-scholarly activities than print media do:

- curriculum vitae; personal research interests and profiles; ongoing, finished, or planned research projects; links to research partners (*'invisible college'* cf. Crane, 1972), publication lists, links to work-in-progress, preprints, conference presentations, course syllabi, tutorials, resource guides, bookmark lists, etc.
- personal hobby interests, family relations, friends, etc.

This blending of scholars' scholarly and non-scholarly activities made visible on the Web is also stressed by Thelwall (2002d) who states that the Web "*often provides a public unrefereed creative space that is used for informal research, teaching and recreational information, for example in personal home pages*" (p. 563).

Academic web sites thus are populated by pages designed for a mixture of purposes and targeted at different audiences (Middleton *et al.*, 1999; Thelwall, 2001a). According to Middleton *et al.* (1999), university web sites function as a tool for communication, providing access, and promotion targeted at a variety of users, both internal and external. The latter audience includes prospective students, prospective staff, other academics, alumni, news media, donors/benefactors, and legislators (ibid.)

Furthermore, web links reflect a diversity of interests, preferences, navigation means and actions of web actors. Thus, motives for making links are more diverse than motives for making references in scientific articles (cf. Kim, 2000; Wilkinson *et al.*, 2003; Thelwall, 2003d). In other words, link structures represent human annotations reflecting cognitive and social structures more extensive than those represented in scientific citation networks because there exists no convention for link creations as for citations in the scientific world.

As stated by Wilkinson *et al.* (2003), the lack of understanding why web links are created is a major obstacle in webometrics and one *"that must be directly addressed in spite of its evident complexity"* (ibid., p. 50). Further, they state that the study *"has really only scratched the surface of the topic of academic linking motivations"* (p. 54). Using a random sample of 414 inter-university links from the UK academic web space, i.e. the ac.uk domain, Wilkinson *et al.* (ibid.) investigated web authors' motivations for creating links between university web sites. The study showed that over 90% of the links were created for broadly scholarly reasons, including teaching activities. In

Section 7.1.1, these findings are compared with findings from the present dissertation study of small-world link structures in the UK academic web space.

In this context, it is also important to note that the sociology of academic web spaces differs between scientific domains. Different scientific domains have developed and use distinctly different communicative forums, both in paper and electronic arenas (including the Web), cf. Kling & McKim (2000) and Jacobs (2001). The sharing of preprints and other unrefereed papers is thus frequent in some fields, for example, in physics or computer science, but not universal for all fields. Webometrics may cast light on such domain differences in use of the Web.

As made clear in the present section – and also noted in Section 2.1 – the Web offers obvious new possibilities of tracking and 'mining' aspects of scientific endeavor traditionally more hidden for bibliometric or scientometric studies, for instance, the use of research results in teaching and by the general public (Björneborn & Ingwersen, 2001; Thelwall & Wilkinson, 2003a; Thelwall, forthcoming). The realization of such possibilities for new approaches on studying the sociology of science strongly spurred the emergence of the new research field of webometrics.

The next section gives a brief overview of webometric research performed in academic web spaces.

2.4.2 Webometrics in academic web spaces

Thelwall, Vaughan & Björneborn (forthcoming) state that the field of webometrics grew out of a realization that quantitative methods originally designed for bibliometric analysis of citation patterns of scientific journal articles could be applied to the Web by using commercial search engines to provide the raw data. Especially, AltaVista's (*www.altavista.com*) search interface that allowed complex Boolean search strings including properties of links and URLs triggered this approach.

As further outlined in the review on webometrics by Thelwall, Vaughan & Björneborn (ibid.), a considerable number of research articles have been published concerning scholarly communication on the Web, mostly originating in the hope that web links could be used to provide similar kinds of information to that extracted from journal citations (e.g., Larson, 1996; Almind & Ingwersen, 1997; Rousseau, 1997; Ingwersen, 1998; Davenport & Cronin, 2000; Cronin, 2001; Borgman & Furner, 2002; Thelwall, 2002b). The major difference between the two is that journal citations occur in refereed documents and therefore their production is subject to quality control and they are part of the mainstream of academic endeavor, whereas hyperlinks are none of these things. This makes web links – also in the multi-purpose sociology of academic web spaces as outlined in the previous section – a more complex phenomenon than journal citations. Thus, several authors like Meyer (1999), Egghe (2000), van Raan (2001), Björneborn & Ingwersen (2001) and Prime *et al.* (2002) warn against taking the analogy between citation analyses and link analyses too far.

Bossy (1995) suggested how *netometrics*, as she called it, could supplement bibliometrics and scientometrics in observing *"science in action"* on the Internet, enabling *"new ways of measuring the impact of scientific contribution that take into*

account the cooperative aspect of science". However, she made no empirical investigation of academic web spaces in the paper.

In his paper 'Bibliometrics of the World Wide Web: an exploratory analysis of the intellectual structure of Cyberspace', Larson (1996) was one of the first information scientists to perform an investigation of link structures in academic web spaces. He used AltaVista in a co-citation analysis of a set of Earth Science related web sites and could produce clusterings of web sites that had topical similarities.

Shortly after, Almind & Ingwersen (1997), in a paper introducing the term *webometrics*, applied a variety of bibliometric-like methods to the Nordic portion of the Web in order to observe the kinds of page connections and define the typology of web pages found at national Nordic level. The methodology involved stratified sampling of web pages and download for local analysis purposes. The contribution also attempted a comparison between the estimated share of scientific web pages and the distribution found in the citation indexes between the Nordic countries. Clearly, the visibility on the Web was quite different from that displayed in the citation databases. Norway, for instance, was much more visible on a Web scale than in the printed scholarly world at the time of analysis.

In his article on *'sitations'* – using a term coined by McKiernan (1996) for site inlinks – Rousseau (1997) analyzed the patterns of distribution of web sites, site inlinks and site selflinks *('self-sitations'*). Rousseau's study operated with 343 web sites retrieved in AltaVista with the search string, *informetrics OR bibliometrics OR scientometrics*. The study showed that the distribution of TLDs (top level domains, such as .edu, .uk, .dk) for the investigated sites followed the ubiquitous power-law-like Lotka distribution. Similarly, Rousseau demonstrated that the distribution of inlinks to the 343 sites also followed a Lotka distribution.

Ingwersen (1998) introduced the concept of the *Web Impact Factor* (WIF) for national domains and individual web sites with parallels to the *Journal Impact Factors* published by the Institute of Scientific Information (ISI) for scientific journals receiving citations from scientific journals indexed in the ISI citation databases (cf., e.g., Hjortgaard Christensen, Ingwersen & Wormell, 1997). In this context, it should be noted that prior to Ingwersen, Rodriguez i Gairin (1997) had introduced the concept of information impact on the Internet in a Spanish documentation journal.

The so-called *external* WIF for a given web site (or TLD, top level domain) was defined by Ingwersen (1998) as the number of external pages (i.e. pages in other sites or TLDs) with links to the given site (or TLD) divided by the number of web pages at the site (or TLD). However, the fluctuating performance of AltaVista at the time of the study yielded problematic variations in the calculated WIF measures.

Subsequently, Thelwall has developed the WIF measure in several papers in order to find possible correlations to traditional research productivity indicators (e.g., Thelwall, 2000; 2001a; 2001d; 2001e; 2002a; 2003a; Smith & Thelwall, 2001; 2002; Thelwall & Tang, 2003).

As stated by Thelwall, Vaughan & Björneborn (forthcoming), the goal underlying almost all of the research reported above was to validate links as a new information source. One of the key tasks is to compare the link data with other related data in order to establish the degree of correlation and overlap between the two. With links between university web sites, for instance, a positive correlation between link counts and a

measure of research would provide some evidence that link creation was not completely random and could be useful for studying scholarly activities.

Thelwall (2001a) showed that the counts of inlinks to a set of 25 UK universities correlated significantly with their average research productivity using the five-yearly UK government Research Assessment Exercise (RAE) cf. HERO (2001). The WIF (Web Impact Factor) delivering the highest correlation with the RAE research rankings was the ratio of web pages with links pointing at research-based pages to the number of full-time academic members of staff. This finding provided the first concrete evidence of a real association between research and links, although no cause-and-effect relationship was claimed. A comparable relationship was later found for Australia (Smith & Thelwall, 2002) and Taiwan (Thelwall & Tang, 2003), using different national measures of research productivity.

Thelwall (2001a; 2001e) introduced an important methodological improvement for webometric investigations of academic web spaces; the employment of specially designed web crawlers for collecting *primary* web data directly from the investigated academic web sites, instead of having to rely on *secondary* data collected in the big commercial search engines with opaque coverage, update frequency, indexing rules, computing performance, and ranking algorithms, etc. (cf. e.g., Lawrence & Giles, 1998; Snyder & Rosenbaum, 1999; Björneborn & Ingwersen, 2001; Bar-Ilan, 2002).

The interest in the Web Impact Factor thus catalyzed an avalanche of webometric research, especially into links in academic web spaces. In parallel with studies of interlinking between universities, there have been studies of departments within a scientific domain. Thomas & Willett (2000) studied UK library and information science departments, finding no significant correlation between inlink counts and research ratings. An earlier small study of 13 Scottish computer science departments (Chen *et al.*, 1998) revealed a number of correlation relationships between structural connectivity measures and the organizational profile based on research assessment exercise ratings, teaching quality assessments, student-staff ratios and funding levels. Furthermore, linkage patterns from the 13 Scottish academic sites to commercial sites in UK and America highlighted the impact of culture and the appropriateness of information technologies on the acceptance of the Web. The study by Chen *et al.* (ibid.) has later been criticized by Thelwall, Vaughan & Björneborn (forthcoming) for not taking departmental size into sufficient account.

In another domain study, significant associations between inlink counts and newspaper rankings (US News) were found for US LIS schools (Chu, He & Thelwall, 2002), giving the first statistical evidence that departmental level studies could give information about scholarly communication (Thelwall, Vaughan & Björneborn, forthcoming). Subsequently, significant research and inlink count correlations have been found for UK computer science departments (Li *et al.*, 2003), in US psychology and US chemistry departments (Tang & Thelwall, forthcoming). The latter study found that interlinking between US history departments was too low for patterns to be extracted and that there were significant disciplinary differences in patterns of interlinking between all three domains. This finding supports the earlier mentioned Kling & McKim (2000) who stress the large differences between different scientific fields in the way electronic media, including the Web, are implemented and utilized.

Geographic factors for interlinking in academic web spaces have also been investigated. For example, the degree of interlinking between pairs of UK universities decreases with geographic distance as found by Thelwall (2002d). In particular, neighboring institutions were much more likely to interlink than average. This shows that despite the existence of *collaboratories* (cf. Finholt, 2002) and other tools for virtual collaboration on the Internet, and its undoubted use for global computer-mediated communication, *"the Web is not divorced from the physical reality"* (Thelwall, Vaughan & Björneborn, forthcoming).

The above examples are mostly from studies employed *within* national university systems. However, the aforementioned study by Chen *et al.* (1998) also included cross-national *and* cross-sectoral link connectivity studies. Another example of a cross-national link structure analysis – while *within* academia – is the co-inlink (called *"co-sitation"*) analysis of 791 university sites from 15 European countries by Polanco *et al.* (2001). Smith & Thelwall (2002) compared linking patterns between UK, Australian and New Zealand universities, and found that New Zealand was relatively isolated on the Web, in line with a previous bibliometric study for journals (Glänzel, 2001). A larger follow-up study mapped the interlinking between universities in the Asia-Pacific region (Thelwall & Smith, 2002) showing that Australia and Japan were central web players in the region, with smaller countries attracting attention disproportionate to their size (cf. Thelwall, Vaughan & Björneborn, forthcoming).

When counting objects on the Web, a decision must be made about what is the most appropriate unit of counting (ibid.). In order to select, sample, filter, quantify and investigate different units of analysis on the Web (cf. the different web node levels described in Section 2.3.3), Thelwall (2002b) introduced the so-called *Alternative Document Models* (ADMs), so far particularly applied in academic web spaces as aggregated units of analysis.[12] Using the terminology by Thelwall (ibid.), there are four main ADMs in use, the *page*, *directory*, *domain* and *site* ADM. The *page* ADM is the default unit consisting of individual files on a web server. In the *directory* ADM, all pages in the same web server directory identified through the URL file name path, are counted as one unit – as an aggregated 'macro document', hence the name of the model. The *domain* ADM – corresponding to a *subsite*[13] as *ecs.soton.ac.uk* in Fig. 2.5 in Section 2.3.2 – aggregates all pages with the same (subsite) domain name in their URL into a common unit of analysis. The *site* ADM (originally called the *university* ADM) aggregates all derivative domain names into a single unit of analysis by specifying only the domain name ending. In the ADM framework, the site of University of Southampton in the abovementioned Fig. 2.5 would thus function as a single unit of analysis embracing all pages from subsites, sub-subsites, etc., with URLs that contained domain names ending in *.soton.ac.uk.*

When counting documents or links, the alternative document models thus give the option to count at different aggregation levels of both source and target web pages, each of which corresponds to a *"URL segmentation strategy"* as noted by Thelwall (forthcoming). For instance, the number of domain (subsite) ADMs in a university site

[12] As stated by Thelwall (2002b), the ADMs are extensions of Björneborn's (2001b) technique for aggregating links at the subsite and site level as a data filtering tool.

[13] The two terms *domain* ADM and *site* ADM may be too easily confused. Perhaps, *subsite* ADM would be a more unambiguous term than *domain* ADM.

could be counted, as could the number of directory ADMs that contain a link to a given web site. The advanced link terminology and web node levels introduced in Section 2.3.3 are especially developed to handle this kind of aggregated units of analysis. For instance, the site ADM corresponds to the aggregated site level links (cf. Fig. 2.8 in Section 2.3.3) in the conceptual framework of the dissertation.

The ADM models have been useful for circumventing anomalies in link data and conceptual problems with counting pages. Anomalies are outliers where individuals or automated processes have created huge numbers of outlinks. For example, in a study by Thelwall (2002b), a biochemistry database at Warwick University contained tens of thousands of links to a similar online database at the University of Cambridge, dwarfing the other link counts in the study. Using a directory or site ADM on such a case would dissolve the anomaly since the source and target pages would respectively be treated as a single 'macro-document'. Thelwall (2002b) applied the ADM models to the UK academic web space with a data set of 108 universities. In particular, the aggregated *directory* and *domain* (subsite) ADMs produced more significant results that correlated better with the aforementioned RAE governmental research productivity measures. These two types of ADMs thus dissolved the earlier outlier counts of *page* level links.

Adamic & Huberman (2000) have used a similar method for aggregating web documents. They studied a crawl of 260,000 web sites each representing a separate domain name. Two sites were considered connected if any of the pages at one site linked to any page in the other. Adamic & Huberman (ibid.) found that the distribution of inlinks between the sites followed a power law.

The majority of studies in academic web spaces have analyzed *inter*-university links as outlined above, either within a discipline, within a national university system, or in international university comparisons. However, some studies have focused instead on links between universities and other sectors of society, such as commerce, industry and government, for example, the aforementioned study by Chen *et al.* (1998). Another example of a cross-sectoral study is Leydesdorff & Curran (2000) who mapped university-industry-government relations on the Internet.

Another important evolving webometric research area in academic web spaces is concerned with scientific journals available on the Web, also known as *e-journals* (cf. Harter & Ford, 2000; Kling & Callahan, forthcoming). Since much of webometrics has been motivated by citation analysis, a natural step has been to see if the kinds of techniques that are applied to journals and authors could also be applied to e-journals (Thelwall, Vaughan & Björneborn, forthcoming). For example, Smith (1999) and Harter & Ford (2000) have investigated e-journals finding no significant correlation between link measures and ISI impact factors for the journals. However, in a large study that incorporated different degrees of web site content, Vaughan & Thelwall (2003) compared inlink counts and ISI's Journal Impact Factors of 88 law and 38 LIS journals indexed by ISI. The findings confirmed that in both law and LIS, counts of links to journal web sites correlated with Journal Impact Factors. Not surprisingly, journals with more extensive online content tended to attract more links than older journal web sites.

Direct web citations between scientific papers were investigated by Vaughan & Shaw (forthcoming). Using Google to collect web citation data to 46 LIS journals, the authors found predominantly significant correlations with traditional citations, suggesting that online and offline citation impact are similar phenomena. A

classification of 854 web citations to papers in the LIS journals indicated that many web citations "represented intellectual impact, coming from other papers posted on the Web (30%) or from class readings lists (12%)" (ibid.).

Other webometric studies of academic web spaces have chosen *scholars* rather than journal web sites or journal papers as the basic unit of analysis, for example, collecting data based upon *invocations*: the mentioning of a scholar's name in any context in a web page. Cronin *et al.* (1998) found academics to be invoked on the Web in a wide variety of contexts, including informal documents such as conference information pages and course reading lists. This was used to support a claim that web invocations could help to highlight *"the diverse ways in which academic influence is exercised and acknowledged"* (ibid.).

In this context, a study by Thelwall & Harries (forthcoming) found that it was the quantity of research produced by scholars that was the main reason for attracting inlinks: universities with better researchers attract more inlinks because the researchers produce more web content, rather than because the content produced is of a higher *'link attractiveness'* (Thelwall, Vaughan & Björneborn, forthcoming). This is in contrast to the case for formal scholarly publications, where better scholars tend to produce papers that attract more citations (Borgman & Furner, 2002).

After this review of webometric research in academic web spaces, the next chapter outlines research on small-world networks – including small-world web spaces.

3 Small-world networks

"The earth to be spann'd, connected by network,
The races, neighbors, to marry and be given in marriage,
The oceans to be cross'd, the distant brought near,
The lands to be welded together."
(Walt Whitman, excerpt from 'Passage to India' in *Leaves of Grass*, 1900)

The dissertation is concerned with small-world phenomena in the shape of short distances along link paths in academic web spaces. This chapter describes small-world research drawing on theories and methodologies developed in graph theory and social network analysis. Section 3.1 gives a brief overview of graph theory and social network analysis. In Section 3.2, early small-world research from the 1960s and onward is described. Section 3.3 outlines the revival of small-world approaches to a wide range of real-world networks from the late 1990s. The chapter is concluded with Section 3.4 on small-world informational networks and Section 3.5 on small-world approaches to Web studies – on how the *"distant"* can be *"brought near"*, as expressed by the American poet Whitman above.

3.1 Graph theory and social network analysis

Theories and empirical work on small-world phenomena stems from research in social network analysis and the related mathematical discipline of graph theory.

A *graph* is a mathematical representation of a network structure consisting of network *nodes* (or *vertices*) connected by *edges* (undirected relations between nodes) or *arcs* (directed), cf. Gross & Yellen (1999). A graph can thus represent, for instance, a social network with the nodes corresponding to people and undirected edges corresponding to acquaintanceships, friendships and kinships. The nodes could also be ecological actors (in a food web), Internet servers, documents (in a citation network), concepts (in a thesaurus or semantic network), and all other types of interrelated network objects. In a *directed* graph the edges represent directional relations between the nodes. As noted in Section 1.2, an example of a directed graph is the World-Wide Web with web pages (or directories, sites or even top level domains) as the nodes and hyperlinks as directed arcs (cf., e.g., Kleinberg *et al.*, 1999; Broder *et al.*, 2000).

The work of the Swiss-born mathematician Leonhard Euler (1707-1783) is regarded as the beginning of graph theory as a mathematical discipline. In 1736 Euler solved the puzzle of the seven bridges in the East Prussian town Königsberg (now Kaliningrad), cf. Fig. 3.1 below. The puzzle dealt with whether it was possible to take a

walk through the city and cross each bridge exactly once before returning to the starting point (cf. Gross & Yellen, 1999; Hayes, 2000a).

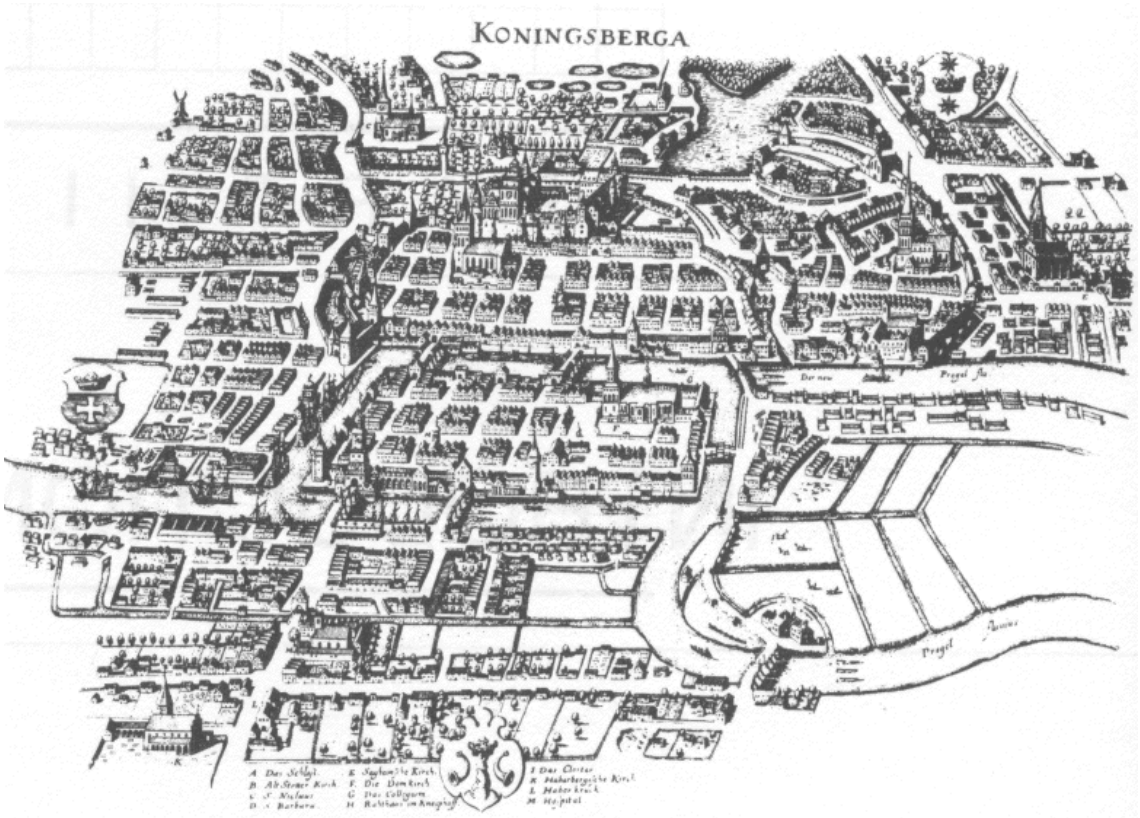


Figure 3.1.* The seven bridges of Königsberg (map from Wilson & Watkins, 1990). (* an asterisk denotes the figure also is shown in the color prints placed before the appendices)

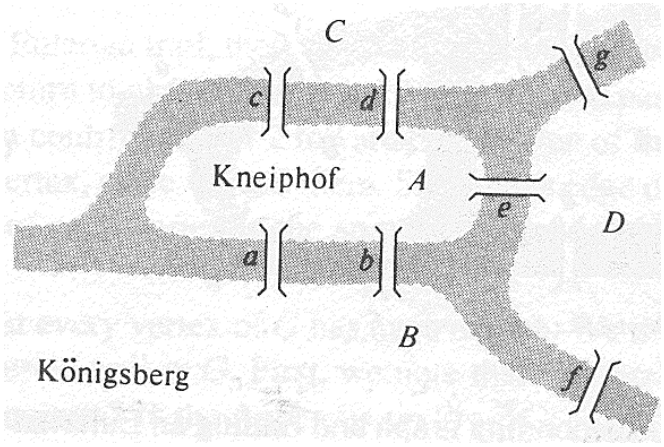


Figure 3.2. Seven bridges (a-f) and four town parts (A-D) in Königsberg (Wilson & Watkins, 1990).

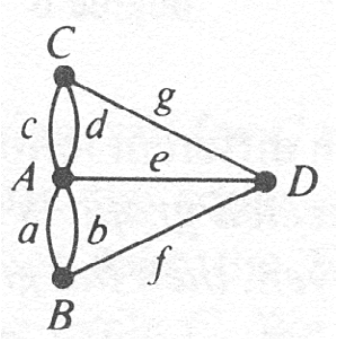


Figure 3.3. Graph with seven edges (bridges a-f) and four vertices (town parts A-D) (Wilson & Watkins, 1990).

Euler constructed a graph by representing the town parts as vertices (nodes) and the bridges as undirected edges, cf. Fig. 3.2 and 3.3 above. He then proved mathematically

that no walk as outlined above was possible because the graph had vertices with odd degree (the degree is the number of edges attached to a vertex), cf. Gross & Yellen (1999). In fact, in the Königsberg graph all four vertices have odd degree (the degree of vertex A in Fig. 3.3 is 5, while the other three vertices have degree 3). Only a graph with vertices all having even degree would allow a so-called *Eulerian walk* as the citizens of Königsberg had tried to make in vain.

Today, graph theory has developed a large mathematical framework (cf. Bollobás, 1998; Gross & Yellen, 1999) useful in a very wide variety of scientific disciplines for mathematical modeling of networks, for instance, in biology, chemistry, physics, sociology, psychology, and technology. Graph theory has also been applied in information sciences for modeling and analyzing, for instance, *citation networks* (e.g., Garner, 1967; Doreian & Fararo, 1985; Hummon & Doreian, 1989; Shepherd, Watters & Cai, 1990; Egghe & Rousseau, 1990; Fang & Rousseau, 2001; Egghe & Rousseau (2002; 2003a; 2003b), *information systems* (e.g., Korfhage, Bhat & Nance, 1972; Nance, Korfhage & Bhat, 1972), *intertextual networks* (Leazer & Furner, 1999) and *hypertextual networks* (e.g., Botafogo & Shneiderman, 1991; Botafogo, Rivlin & Shneiderman, 1992; Smeaton, 1995; Furner, Ellis & Willett, 1996).

One of the scientific disciplines that make extensive use of graph theory is social network analysis (Knoke & Kuklinski, 1982; Scott, 1992; Wasserman & Faust, 1994). As alluded to above, small-world theory was developed in social network analysis. This issue is further described in Section 3.2 below. The pioneers of what today is called social network analysis came from sociology and social psychology. In the 1930s the Austrian-American social psychologist Jacob Levy Moreno (1894-1974) founded *sociometry* as the quantitative measurement of interpersonal relations in small groups (cf. Wasserman & Faust, 1994). Moreno introduced so-called *sociograms* for depicting the interpersonal structure in a group. In a sociogram, people (or any social unit) are represented by nodes[14] in a two-dimensional space, and relations between pairs of nodes are denoted by lines (edges). Recognition that sociograms could be used to study social structure led to a rapid development of techniques and measurements (cf. Wasserman & Faust, 1994) including that mathematical matrices could be used to represent social network data. Early sociometricians also discovered graph theory and applied its mathematical framework to social networks.

In figures 3.4 and 3.5 below, the matrices represent the Königsberg graph in Fig. 3.3 showing in binary digits if the town quarters A-D are *adjacent*, in this case connected by a bridge, or not. However, similar so-called *adjacency matrices* might also represent social networks where the nodes A-D could correspond to, for example, scientists and the 1s and 0s whether a pair of scientists have co-authored or not. In an undirected graph as outlined in the two examples, the corresponding matrix is logically symmetrical and the upper-right or lower-left half of the matrix may thus be omitted. However, in a directed graph like a citation network or the Web, the direction of the citations and links makes a difference. Author A may cite author B, but not necessarily vice versa. Web page A receives an inlink from B, but does not necessarily provide an outlink back to B. An adjacency matrix for 7669 UK university subsites constituted the

[14] In the rest of the dissertation, the term *node* is used instead of other synonyms in graph theory and social network analysis, such as *vertex*, *point*, or *actor*.

basic foundation for the graph computations and empirical investigation in the dissertation as further described in Section 4.2.5.

	A	B	C	D
A	0	1	1	1
B	1	0	0	1
C	1	0	0	1
D	1	1	1	0

Figure 3.4. Matrix representing the Königsberg graph in Fig. 3.3.

$$\begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$

Figure 3.5. The same matrix without row and column headings.

Concepts, techniques and measures developed in social network analysis are today used in a wide array of social and behavioral science disciplines, e.g., social psychology, sociology, social anthropology, cultural geography, economy, and political sciences. Wasserman & Faust (1994) present a long list of research topics where social network analysis is applicable, for example, in occupational mobility, global political and economic systems, social support, problem solving in groups, diffusion of innovations (cf. Rogers, 1995), 'invisible colleges' in the sociology of science, power and exchange, and coalition formation. Otte & Rousseau (2002) give an overview of applications and potentials of social network analysis in the information sciences with regard to studies of, for example, citation and co-citation networks, collaboration structures and other forms of social interaction networks. In Section 3.4 below is shown how several researchers in webometrics and other approaches to Web studies today apply social network analysis, including small-world theories.

Besides the notion of small-world phenomena of special interest in this dissertation, social network analysis has also developed a range of other concepts that are appropriate to include in this context. Especially, the concepts of *weak ties* and *betweenness centrality* are intriguing. In Section 7.1.2, a hypothesis is pursued that transversal links, that is, links that connect topically dissimilar web sites (cf. Section 2.3.1) may function as *weak ties* on the Web. In social network analysis, the concept of weak ties (Granovetter, 1973; 1982) is used to explain macro-level social cohesion and interconnectedness as well as possibilities for rapid diffusion of ideas and epidemics across social boundaries through peripheral and *heterophilous* (socially different) social contacts, so-called *weak ties*. Transversal links may thus give a new significance to the social network analytic notion of *"the strength of weak ties"* (ibid.) for explaining the cohesiveness of a small-world academic web space.

The so-called *betweenness centrality* is an important measure with regard to small-world properties in a graph because it quantifies how many shortest paths pass through each node in the graph (Freeman, 1977). The idea of applying graph theory to analyze the connection between structural centrality and social group processes was introduced in social network analysis by Bavelas (1948), cf. Eppstein & Wang (2001). In Section 6.3.2.4, the measure of betweenness centrality is further outlined as well as calculated for all 7669 UK university subsites in the investigated data set.

In the next sections, the central concept of small-world phenomena is outlined.

3.2 Small-world background

The 'small-world' theory stems from research in social network analysis (Milgram 1967; Pool & Kochen, 1978/1979; Kochen, 1989; Deutsch, 1989) concerned with short distances between two arbitrary persons through intermediate chains of acquaintances. The American social psychologist Stanley Milgram (1933-1984) formulated the "small-world problem" in the following way:

> *"Given any two people in the world, person X and person Z, how many intermediate acquaintance links are needed before X and Z are connected."* (Milgram, 1967, p. 62)

In a famous but disputed experiment (cf. Kleinfeld, 2000), Milgram (1967) asked a sample of persons in Nebraska to forward a letter to a Boston stockbroker not known to them.[15] The letter was to be mailed to a personal acquaintance (on a first-name basis) more likely to know the target person. Of the 160 chains that started in Nebraska, only 44 (27.5%) were completed (the rest dropped out). On the completed chains, a median chain length of 5 intermediate acquaintances was needed to pass on the letter to the stockbroker. This result contributed to the popularized but unverifiable notion of 'six degrees of separation' between two arbitrary persons in the whole world through intermediate chains of acquaintances, for example, as stated in the play 'Six degrees of separation' by Guare (1990), cinematized in 1993 (cf. Collins & Chow, 1998).

Manfred Kochen (1928-1989), a visionary information scientist, was a pioneer in small-world theorizing. The first mathematical analysis of small-world phenomena in social networks was conducted already in 1958, however, first published 20 years later because *"we raised so many questions that we did not know how to answer"* (Pool & Kochen, 1978/1979, p.5). Small-world phenomena in the shape of short distances between people in social networks affect possibilities of diffusion and accessibility of knowledge. Not surprisingly, Kochen also was very interested in bibliometric studies of the growth, diffusion and retrieval of scientific knowledge (e.g., Kochen, 1967). Furthermore, in the current context of webometrics, it is noteworthy that Kochen was highly concerned with early ideas on developing world-encompassing information systems (Kochen, 1972; Garfield, 1989) as well as he was interested in the mathematics of self-organizing systems (Deutsch, 1989).

Another innovative information scientist and bibliometric pioneer, the founder of the citation indexes, Eugene Garfield, envisioned that small-world approaches could be used to identify gatekeepers and 'invisible colleges' in informal scholarly communication networks (Garfield, 1979). Thus, one may say that the ring now closes as small-world approaches again enter the domain of bibliometrics in the shape of webometric studies of small-world link structures in academic web spaces as in this dissertation.

[15] In a parallel study, the starting persons were from Kansas and the target person from Cambridge, Massachusetts (Milgram, 1967). Kleinfeld's (2000) critique points are concerned with the lack of appropriate randomization of starting persons as well as the low percentage of completed chains.

Early small-world studies (cf. Kochen, 1989) of social networks encountered large methodological problems, because of the difficulties to collect and compute large-scale data (Watts, 1999b). In particular, it was difficult to give a rigid definition of an acquaintance. Furthermore, people may have difficulties to list all their acquaintances. Moreover, it is unfeasible in large-scale studies to trace all possible chains of acquaintances of acquaintances. However, at the end of the 1990s, small-world research experienced a breakthrough with the seminal paper by Watts & Strogatz (1998) outlined in the next section. According to Albert & Barabási (2002), this breakthrough was spurred by the circumstance that the computerization of data acquisition in many scientific fields had led to the emergence of large databases on the topology of various real-world networks. Moreover, the increased computing power allowed investigation of complex networks containing millions of nodes, "exploring questions that could not be addressed before" (ibid., p. 48).

3.3 Small-world revival

Traditionally, approaches to complex networks were primarily based on *random* network models introduced by the mathematicians Erdös and Rényi in the 1960s (cf. Albert & Barabási, 2002), with randomly distributed edges (i.e., links, relations). However, the novel small-world network model introduced by Watts & Strogatz (1998) demonstrated how many real-world complex networks in biological, social and man-made systems simultaneously possess properties both of random graphs and regular graphs. Drawing on the mathematical framework of graph theory, the two researchers in applied mathematics quantified the structural properties of networks by the so-called characteristic path length and clustering coefficient. The *characteristic path length* is measured by the number of edges in the shortest path between two nodes, averaged over all pairs of nodes. The *clustering coefficient* gives an average measure of the extent to which the neighbors of each node in a network also are mutually interconnected by edges, for example, if friends of a person are also friends of each other (cf. the *triadic closure* of nodes D, E and F in Fig. 2.2 in Section 2.3.1). In Section 5.3, these measures are further described, as well as calculated for the investigated UK academic web space.

Watts & Strogatz (ibid.) showed, that if a few number of edges is randomly 'rewired' in a regular network, like a ring lattice in Fig. 3.6 below, a so-called *small-world network* emerges, still being highly clustered (i.e. high clustering coefficient), as in the regular lattice, yet having short characteristic path lengths between pairs of nodes, as in a random graph. In this context, the authors found that in a small-world network it is sufficient with a very small percentage of 'long-range' connections (e.g., neural dendrites, social relations, electricity transmission lines) functioning as shortcuts connecting 'distant' nodes of the network.

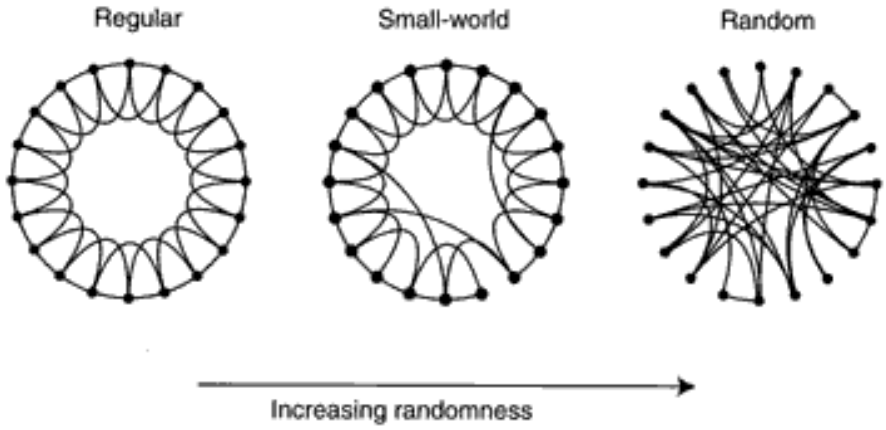


Figure 3.6. Small-world network as a merger between regular and random network graphs (Watts & Strogatz, 1998).

Measuring clustering coefficients and characteristic path lengths, Watts & Strogatz (1998) found small-world topology in the completely mapped neural network of the nematode worm *C. Elegans*, in a collaboration graph of film actors[16], and in the electrical power grid system of the western United States. Their results made them suggest that similar small-world properties probably would turn out to be generic for many large real-world networks. This prediction has subsequently been confirmed in a very wide variety of networks studied in a diversity of scientific disciplines. Furthermore, the revival of small-world theory commenced with Watts & Strogatz' (1998) article catalyzed an avalanche of research not only into small-world networks but also into underlying topological and dynamical principles of complex networks in general.

Small-world phenomena characterized by "the coincidence of high local clustering and short global separation" (Watts, 1999a) have been verified in many complex networks ranging from physics, biochemistry, biology, social networks, communication and technology, for example, in *genotype evolution* (Bagnoli & Bezzi, 2001), *food webs in ecological networks* (Montoya & Solé, 2002), *neural networks* and *associative memory* (Sporns, 2003; Bohland & Minai, 2001), *protein folding* (Jespersen *et al.*, 2000), *metabolic networks* (Bilke & Peterson, 2001), *fire spreading* (Moukarzel, 1999), *epidemics* (Kleczkowski & Grenfell, 1999; Moore & Newman, 2000), *economics: reputation management* (Venkatraman *et al.*, 2000), *labor markets* (Tassier & Menczer, 2001), *wealth distribution* (Souma *et al.*, 2001), *bilateral trade* (Wilhite, 2001), *'old boys' networks* (Davis *et al.*, 2002), *information diffusion* (Ahmed & Abdusalam, 2000; Zanette, 2001), *scientific networks* (Newman, 2001; Barabási *et al.*, 2002), *citation networks* (Bilke & Peterson, 2001), *Internet* (Watts, 1999c; Yook *et al.*, 2002; Jin & Bestavros, 2002), *WWW* (Albert, Jeong & Barabási, 1999; Adamic, 1999; Barabási, 2001; Albert & Barabási, 2002), *e-mail graph* (Ebel *et al.*, 2002), *telephone*

[16] This collaboration graph was based on data at the Internet Movie Database (*www.imdb.com*) also used at the University of Virginia (*www.cs.virginia.edu/oracle/star_links.html*) to compute degrees of separation along intermediate co-actors between any two film actors from the film history of the whole world.

call graph (Albert & Barabási, 2002), *computer circuits* (Ferrer i Cancho *et al.*, 2001), *semantic networks* and *thesauri* (Steyvers & Tenenbaum, 2001; Ferrer i Cancho & Solé, 2001; Sigman & Cecchi, 2002; Kinouchi *et al.*, 2002).

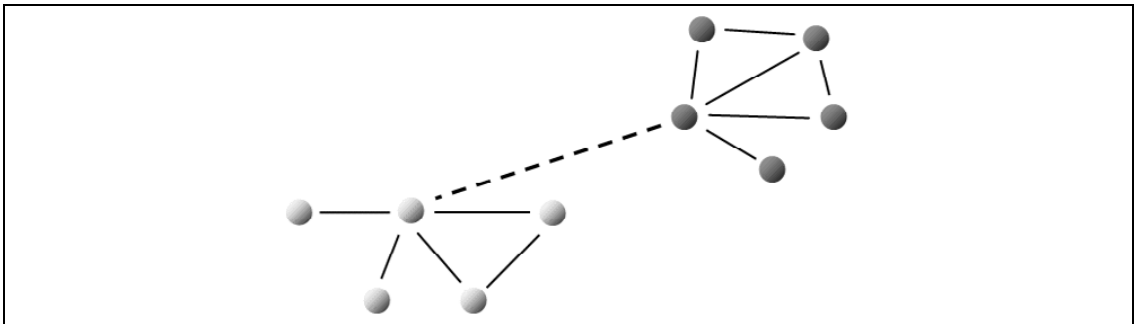
Small-world network features combining high clustering and short link distances thus affect the diffusion speed of properties such as, for example, data, energy, signals, contacts, ideas, influence, economic values or epidemics across the networks covered in the listing above. Watts (1999b), Kleinberg (1999b), Newman (2000), Amaral *et al.* (2000), Barrat & Weigt (2000), Dorogovtsev & Mendes (2002), Mathias & Gopal (2000), Strogatz (2001), and Albert & Barabási (2002) give good overviews of small-world phenomena and other topological properties in a wide range of networks. In Section 3.5, small-world approaches to the Web are further outlined.

The initial small-world model introduced by Watts & Strogatz (1998) was based on simple ring lattices with only small differences in the degree distribution, that is, the number of edges attached to each node in the lattice as illustrated in Fig. 3.6 above. However, many real-world networks are so-called *scale-free*. In a scale-free network, there is no 'typical' node, that is, no characteristic 'scale' to the degree of connectivity. Scale-free degree distribution show long *power-law* tails (cf. Sections 3.5 and 5.4), implying that only a small share of nodes are connected by many edges, whereas the bulk of nodes has quite few edges attached. In Section 3.5 further below is described a range of power-law distributions identified on the Web as well as outlined the connection between scale-free properties and small-world properties. But first, Section 3.4 gives a brief review of small-world approaches to other informational networks than the Web.

3.4 Small-world informational networks

As outlined in the previous section, there has been found small-world properties in many informational networks of interest in library and information science in the wake of Watts & Strogatz' (1998) paper. This section gives a brief review of some of these findings as well as some earlier works in the area in order to provide a contextual background for the dissertation.

In library and information science there is still a lack of research on small-world phenomena and their possible usabilities regarding different types of nodes and relations in informational networks such as the Web, bibliographic and citation networks, semantic networks, thesauri, etc. As suggested in Fig. 3.7 below, there may be possible small-world phenomena when *nodes* in an informational network are defined as corresponding either to documents, document aggregations (e.g., journals, web directories, web sites), terms, authors, scientific domains, institutions, or countries, etc., and distance-reducing, small-world-creating *relations* correspond to either references, citations, outlinks, inlinks, bibliographic couplings, co-cited authors, documents or journals, co-links, co-authors, cross-institutional collaborations, related terms, co-term occurrences, descriptors, etc.



nodes	relations
• authors, documents, document aggregations	– references/citations, outlinks/inlinks
• authors, documents, document aggregations	– bibliographic couplings, co-outlinks
• authors, documents, document aggregations	– co-citations, co-inlinks
• authors	– co-authorships
• terms	– thesaural, semantic relations
• journals	– co-cited journals
• ...	– ...

Figure 3.7. Examples of nodes and relations in informational networks with possible small-world phenomena. The dashed relation in the figure denotes a transversal relation connecting two dissimilar subsets of the concerned network. The relations are illustrated without direction because some are directional (e.g., links) whereas others are bidirectional (e.g., co-authorships and co-links).

An interesting case of early interest in small-world phenomena in a library and information science setting is given by Egghe & Rousseau (1990) who used shortest path algorithms and graphs to measure, for instance, the shortest physical distances between locations within a library (p. 142) as well as to minimize the number of library shelves needed for storing books by height (p. 150).

As already alluded to, other visionary information scientists like Kochen and Garfield early on envisaged possible bibliometric and scientometric applications of small-world properties, especially in scientific collaboration networks. Such networks were investigated by Newman (2001) who used bibliographic data from a 5-year window in databases like NCSTRL, MEDLINE, and the Los Alamos e-print archives, and found that co-authorship patterns in scientific collaboration networks in, for instance, computer science, biomedicine, astrophysics, and high-energy physics, all had small-world properties. In other words, there were short distances along chains of intermediate co-authors between arbitrary researchers within each field. Barabási *et al.* (2002) show similar small-world findings for co-authorship networks in mathematics and neuro-science.

A famous example related to small-world phenomena in scientific collaboration networks is concerned with the so-called *Erdös-numbers*. Before his death in 1996, the Hungarian mathematician Paul Erdös (e.g., random network theory, cf. Section 3.3) wrote almost 1500 papers with 472 different co-authors (Bar-Ilan, 1998). Erdös-

numbers are measured as the number of intermediate co-authorships between a given scientist and Erdös. For example, a scientist has Erdös-number 2 if he/she has been co-author to someone being co-author with Erdös. (Erdös had Erdös-number 0). Grossman & Ion (1995), de Castro & Grossman (1999), Batagelj & Mrvar (2000), and Grossman (2003) give extensive coverage of Erdös-numbers including lists of small Erdös-numbers for famous researchers both within and outside mathematics, as well as for Nobel Prize laureates in physics, chemistry, economics and medicine.

Pool & Kochen (1978/1979) anticipated how bibliometric data could be used in graph theoretical models of small-world phenomena, for example, using Erdös-numbers or co-citations (p. 32). With regard to co-citations, Small (1999; 2000) investigates pathways of strong co-citations crossing disciplinary boundaries in science and the cross-fertilizing creativity that can emerge at such boundary crossings which can be exploited in computer-supported knowledge discovery (cf. Björneborn & Ingwersen, 2001). Small is concerned with 'strong ties' rather than with 'weak ties' (cf. Section 3.1) in the shape of strong co-citations for creating pathways – called *co-citation chains* in the dissertation – through the scientific literature. According to Small, in scientific literature it is possible to travel from any topic or field to another because of the interconnected fabric of scientific disciplines. This is illustrated by using a specific co-citation chain starting in economics and ending in astrophysics, cf. Fig. 3.8. However, as also noted by Small (1999), such 'strong-tie' co-citation chains do not necessarily follow the shortest paths between the domains as in the small-world studies conducted by Milgram or Kochen, where 'weak ties' may create shortcuts between heterogeneous groups.

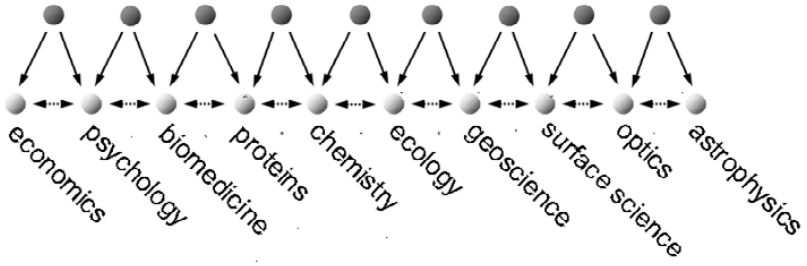


Figure 3.8. Co-citation chain (bidirectional arrows) illustrating Small's (1999) example of pathways of strong co-citations between nodes representing scientific literatures starting in economics and ending in astrophysics.

Qin & Norton (1999), commenting on Small, envisaged that

> *"in future retrieval systems, a user could pick two topics or documents and generate a path of documents or topics that connect them, which could be used for information discovery and hypothesis generation".*

As shown in Section 6.3, the so-called *path nets* used for identifying transversal links in the present study are constructed in this way by juxtaposing two topically dissimilar start and end nodes.

Another knowledge discovery method used in bibliographic databases, with possible applicability to small-world approaches in library and information science, is Swanson's research on so-called *'undiscovered public knowledge'* (1986), developed over the years (e.g., Swanson & Smalheiser, 1997; 1999). Swanson (1986) stated,

"[k]nowledge can be public, yet undiscovered, if independently created fragments are logically related but never retrieved, brought together, and interpreted". Swanson used a systematic trial-and-error strategy to reveal intermediate relations between two literatures. Using Swanson's (1986) own example, cf. Fig. 3.9, if literature C_1 is concerned with fish oil and literature C_3 is about Raynaud's disease (a peripheral circulatory disorder), then literature C_2 on blood platelets can be the missing, transitive relation. If $C_1 \rightarrow C_2$ and $C_2 \rightarrow C_3$, then $C_1 \rightarrow C_3$. Later medical experiments verified that fish oil actually had a beneficial effect on Raynaud's disease (ibid.).

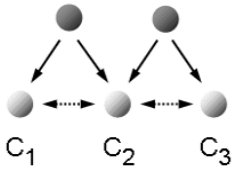


Figure 3.9. Short co-citation chain (bidirectional arrows) illustrating Swanson's (1986) example on 'undiscovered public knowledge' including literature on fish oil (C_1), blood platelets (C_2), and Raynaud's disease (C_3).

This literature-based knowledge discovery method is thus used to find *"interesting but previously unknown implicit information"* within the scientific literature (Swanson & Smalheiser, 1999), revealing connections between ideas or concepts that were not considered before (Garfield, 1994). Applying this method on small-world academic web spaces, a hypothesis that functioned as a strong motive power in the PhD project was that transversal links as well as *transversal co-links* may give useful hints for finding unexpected relations between scientific disciplines in order to identify fertile areas for cross-disciplinary exploration and thus stimulate scientific creativity. Inspired by the Swanson approach, the original idea in PhD project thus included investigations of *small-world co-linkage* (Björneborn, 2001a) as short distances along co-linkage chains between different scientific domains on the Web.

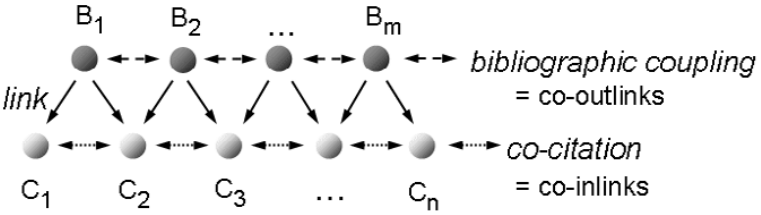


Figure 3.10. Long *co-linkage chain* (bi-directional arrows) of co-linking and co-linked web nodes.

In a co-linkage chain as in Fig. 3.10 above, both co-outlinks (analogous to bibliographic couplings, cf. Section 2.3.1) and co-inlinks (co-citations) on the Web are considered. There are obvious complementarities between the two paths of co-linking nodes (B_i) and co-linked nodes (C_j) marked with bi-directional arrows in the figure. The two paths generate each other. A path of co-linked nodes (co-citations) generates a path of co-linking nodes (bibliographic couplings) - and vice versa.

In a pilot study of possible small-world phenomena in the shape of short distances along co-linkages between researchers with topically dissimilar research interests, a co-linkage chain of co-linked (co-cited) researchers' homepages and co-linking (bibliographically coupled) researchers' bookmark lists was constructed, cf. Fig. 3.11. The co-linkage chain was manually constructed by using AltaVista to find researchers' personal homepages receiving inlinks from bookmark lists published on the Web by researchers. Topical dissimilarity between nodes in the chain was assessed using subjective criteria on research interests as reflected on the investigated homepages and bookmark lists. The co-linkage chain in Fig. 3.11 consists of five co-linked homepages and four co-linking bookmark lists with research interests spanning from small-world link structures to distributed knowledge systems, interdisciplinary studies, philosophy of mind, education research, and linguistics. As mentioned before, the idea behind this approach of constructing co-linkage chains was that direct transversal co-linkages, e.g., between cybernetics and philosophy of mind – or chains of transversal co-linkages, e.g., between small-world phenomena and linguistics – could be used in computer-supported knowledge discovery ('web mining') in order to identify fertile scientific areas for cross-disciplinary exploration and hypothesis generation, cf. the quote by Qin & Norton (1999) earlier in this section. However, as alluded to in Section 1.3, this ambitious initial approach in the PhD project was deserted due to difficulties in developing tractable methodologies for computer-supported data extraction and objective selection criteria.

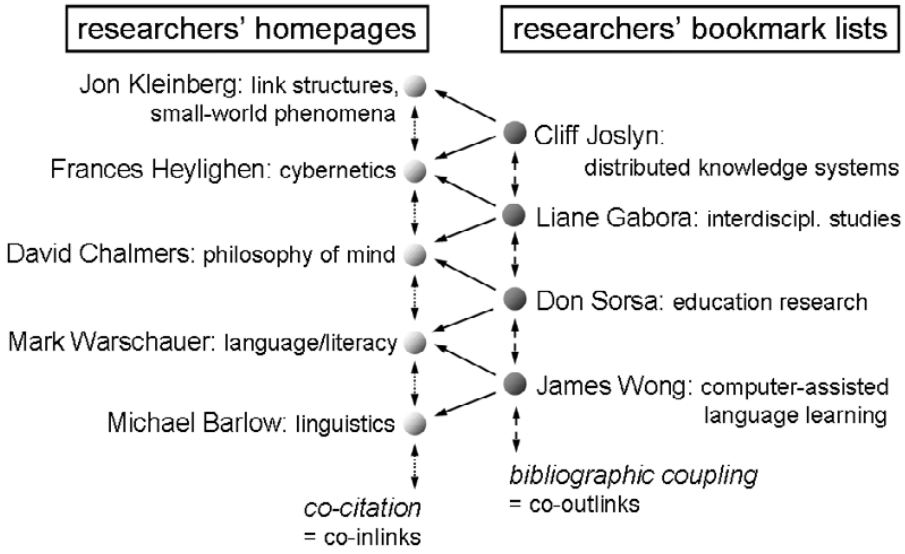


Figure 3.11. Example of *co-linkage chain* (bi-directional arrows) spanning dissimilar research interests reflected on co-linked researchers' homepages and co-linking bookmark lists on the Web.

The creator of the Web in 1989-1990, Tim Berners-Lee, envisaged how link structure analysis as mirrors of human interactions could be used for computer-supported knowledge discovery or 'data mining' on the Web. One important incentive for him to develop the Web was thus the possibility to keep track of *"the complex web of relationships between people, programs, machines and ideas"* (Berners-Lee, 1997).

An early study related to the combination of small-world approaches and computer-supported knowledge discovery in an informational network was conducted by Schwartz & Wood (1993) who showed it was possible to discover shared interest relationships by analyzing the graph structure of connections deriving from posted and received emails extracted from email logs of about 3,700 different Internet sites. The investigated email graph comprised 50,834 nodes (people) and 183,833 edges (emails).

In the revival of small-world research following Watts & Strogatz' (1998) study, citation networks has been considered in a number of papers, for instance, Dorogovtsev & Mendes (2000) and Bilke & Peterson (2001). Building on an earlier study by Redner (1998) who found the probability that a paper is cited k times follows a power law, Bilke & Peterson (2001) used the same SPIRES database as Redner (ibid.) for investigating topological properties of citation networks among high energy physics publications. Treating the citations as undirected, Bilke & Peterson (ibid.) found small-world properties in the investigated citation networks – as well as in the parallel study in the paper of so-called *metabolic* networks of chemical reactions in living cells.

Leazer & Furner (1999) suggest small-world approaches to study topologies of so-called *textual identity networks*, i.e., sets of documents that share common semantic or linguistic content, for use in information retrieval systems. As listed earlier, small-world properties of informational networks as semantic networks and thesauri have been investigated by several researchers (e.g., Steyvers & Tenenbaum, 2001; Ferrer i Cancho & Solé, 2001; Sigman & Cecchi, 2002; Kinouchi *et al.*, 2002). For instance, Steyvers & Tenenbaum (2001) analyzed the large-scale structure of three types of semantic networks based on free word associations, WordNet, and Roget's thesaurus. The analysis showed how the semantic networks have a small-world structure characterized by short average path-lengths between words, and strong local clustering.

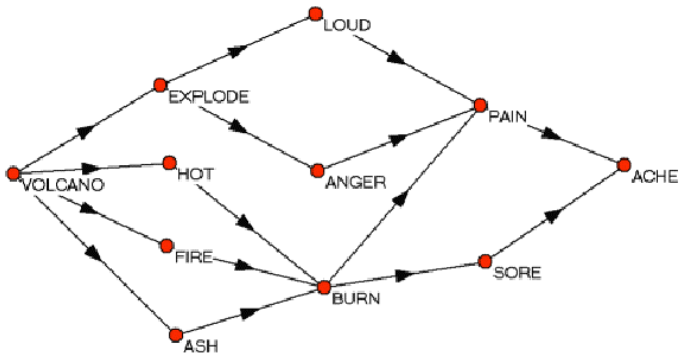


Figure 3.12. All shortest paths between the terms *volcano* and *ache* in a semantic network formed by free word associations (Steyvers & Tenenbaum, 2001).

According to the authors, traditional tree-structured hierarchical models impose severe limitations as a general model of semantic structure. Instead, they suggest that small-world structures may provide a more appropriate model of semantic structures, cf. Fig. 3.12 above. The figure shows a subgraph of all the shortest paths between the terms *volcano* and *ache* in a semantic network formed by free word associations. Section 6.3.2

presents topologically congruent subgraphs, called *path nets* in the dissertation, of all shortest link paths between topically dissimilar subsites in the UK academic web space.

Finally, in the context of this review of small-world approaches to informational networks, it should also be noted that the concept of 'small world' has been used in a completely different meaning in studies of information seeking behavior. In several famous studies, especially of under-privileged American social subgroups like female lifetime prisoners or janitorial workers, Elfreda Chatman (e.g., Chatman, 1991; Pendleton & Chatman, 1998; Burnett, Besant & Chatman, 2001; Huotari & Chatman, 2001) has used the concept to describe how most of the investigated persons were not active seekers of information outside their most familiar social environment. In the Chatman framework, the 'small world' concept is used to designate such narrow-bounded and limited information environments in everyday life information seeking. This use of the concept is thus diametrically opposed to the customary conceptualization of small-world phenomena in social network analysis for denoting *'boundary-breaking'* – and not *'boundary-delimiting'* (as Chatman has it) – phenomena in social networks.

The next section gives a review on small-world approaches to the Web – the informational network investigated in this dissertation.

3.5 Small-world web graphs

In recent years, there has been a strongly increasing research interest in investigating the dynamics and intricate structures of web link topologies. As mentioned earlier, researchers from especially computer science, physics and mathematics as well as information scientists have applied methods from graph theory and social network analysis to treat the Web as a directed graph consisting of nodes in the shape of web pages or web sites connected by directed edges in the shape of hyperlinks. As noted in Section 1.2, such approaches have been used, for example, for identifying web graph components (Broder *et al.*, 2000), inferring web communities (Gibson, Kleinberg & Raghavan, 1998; Clever Project, 1999), identifying hub-like and authoritative web pages (Kleinberg, 1999; Cui, 1999), topic distillation (Bharat & Henzinger, 1998), or improving search engine ranking algorithms as in Google (Brin & Page, 1998).

Furthermore, this research has shown that the Web contains structural properties in the shape of small-world phenomena and so-called *scale-free* features typically including skewed power-law distributions of connections resembling those found in other dynamic and complex networks such as social networks, ecological food webs, neural networks, etc. Such scale-free features and power laws are further outlined below.

In current research on complex networks, the Web plays an important role, because the Web is the largest real-world network for which topological information presently is available. As mentioned earlier, the Web has thus become a *testing ground* for many current research efforts to build models of the emergence and dynamics of complex networks. This explains the large interest in Web link topologies also from physicists, mathematicians, and other researchers in complex networks. There has thus

been an ever-increasing amount of literature on link topology research since the late 1990s. Excellent reviews are given by, for example, Kleinberg *et al.* (1999), Deo & Gupta (2001), Albert & Barabási (2002), Barabási (2002), and Scharnhorst (2003). This section presents a selection of important works on the topic, relating both to more overall graph theoretic approaches to hypertextual network structures and to more specific research on small-world web spaces.

Instead of counting the number of intermediate acquaintances between two persons in the social networks as described in Section 3.2, small-world approaches to the Web are concerned with the number of links along link paths between two web nodes, for example, between two web sites or web pages.

In one of the first papers on small-world properties of the Web, Adamic (1999) investigated a so-called *Strongly Connected Component* (explained below) of 3,400 web sites in the *.edu* top level domain and found a characteristic (i.e. average) path length of 4.1 links along link paths between these sites. In Section 5.3.1, this small-world finding from the American educational web space is compared to the UK academic web space investigated in the dissertation.

Another early and widespread small-world web study by Albert, Jeong & Barabási (1999) was based on a complete mapping of the web site of the University of Notre Dame (*nd.edu*) comprising 325.729 web pages and 1.469.680 links. Extrapolating results from the connectivity patterns of this single web site, the paper presented a conjecture that any pair of web *pages* (and not *sites* as in the Adamic study above) on the whole Web could be connected by a short path of links, with an average path length of only "19 clicks" (ibid.). However, this finding was later discovered not to be correct (Broder *et al.*, 2000) because many pairs of web pages are not connected by link paths at all. In their so-called *'bow-tie'* model of the Web, Broder *et al.* (2000) showed that small-world properties in the shape of short link paths are only present in specially well-connected areas of the Web graph, in the so-called *Strongly Connected Component* (SCC) – the 'bow-tie knot' or core in Fig. 3.13 below.

The 'bow-tie' model of the graph structure of the Web was based on two large AltaVista crawls in 1999 and specially constructed software to process the huge data set of about 200 million web pages and 1.5 billion links each. Broder *et al.* (ibid.) found that over 90% of the links formed one huge *weakly connected component* (*WCC*) if the direction of the links was ignored. This component splits into four roughly equally large parts. The 'bow-tie' core in the model is the abovementioned *Strongly Connected Component* (*SCC*) in which any pair of web pages can be connected by directed link paths, cf. Fig. 3.13 above. The SCC covers about 28% of the web pages in the Broder study. The *IN* component (21%) comprises pages that can reach the SCC by directed link paths but cannot in turn be reached from the SCC. Correspondingly, pages in the *OUT* component (21%) can be reached by directed link paths from the SCC but cannot reach back. Pages in the so-called *Tendrils* and *Tube* (together 22%) are connected with the IN and OUT components but cannot reach to the SCC or be reached from the SCC. The remaining *Disconnected* component (8%) are not connected in any way with the main 'bow-tie'.

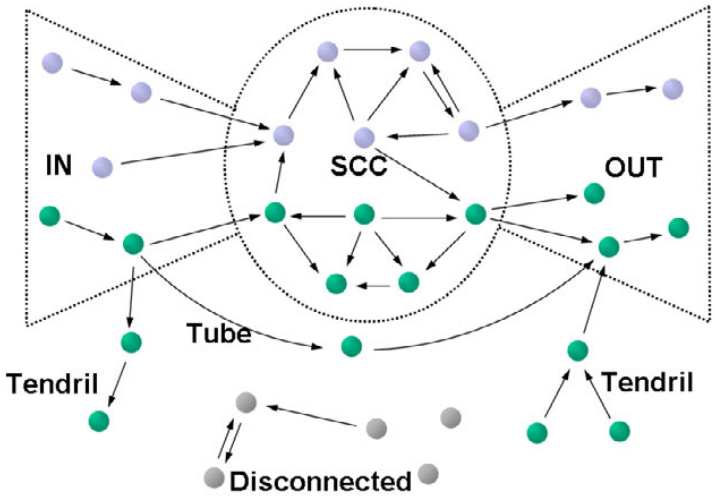


Figure 3.13.* The 'bow-tie' model (modified after Broder *et al.*, 2000) of different graph components in the Web. (*cf. color prints placed before appendices).

According to the 'bow-tie' model by Broder *et al.* (ibid.), small-world phenomena in the shape of short link paths between web nodes primarily occur within the SCC because any pairs of nodes in this component can reach each other with directed link paths, whereas a web node, for example, in the OUT component cannot reach a node, for example, in the IN component by link paths. Indeed, according to the Broder study, there are *no* directed link paths between over 75% of all pairs of web pages. However, when there *is* a link path, the average path length is about 16.

According to Thelwall, Vaughan & Björneborn (forthcoming), a problem with the Broder study is that the model is difficult to extrapolate to the whole Web because of the inherent bias in the underlying data set from AltaVista. AltaVista finds pages partly from user submissions of URLs, but mainly by following links from previously visited and indexed web pages. As a result, pages that are not well linked to are more likely to be missed by AltaVista's crawler. Thus, the *IN* and *Disconnected* components are likely to be far larger for the whole Web. However, it is not possible to estimate how large they are since there is no practical way to automatically find web pages that receive no inlinks.

Similar 'bow-tie'-looking link structures have been identified in subgraphs of the Web, including national university systems (Thelwall & Wilkinson, 2003b) and individual countries (Baeza-Yates & Castillo, 2001). Baeza-Yates and Castillo extended the 'bow-tie' model by finding that some of the components could be subdivided. In this context, it should be noted that the graph components are not made up of homogeneous and coherent sets of web pages, as also pointed out by Thelwall, Vaughan & Björneborn (forthcoming). For example, as shown by Thelwall & Wilkinson (2003b), the OUT component will contain many pages with no outlinks but with one inlink from an SCC page; being pages on the same web site as an SCC page.

In Section 5.1, a modified 'bow-tie' model – the *'corona'* model' – developed in the PhD project is presented based on the graph components identified in the investigated subgraph of the Web comprising the UK national university system.

In a graph theoretic and so-called *cyber-geographic* approach on the connectivity patterns of the UK university web graph, Dodge (1999a) investigated 450,000 links between 122 UK universities. The inlink data was gathered using AltaVista. The constructed connectivity graph was analyzed in order to find the most central web site using a specially constructed metric measured by the number of outlinks and inlinks to each site. The node with greatest connectivity to all the other 121 universities was the University of Oxford. Oxford also had a central connectivity position in the dissertation study, however, using a different centrality measure than in the Dodge study. Thus, the subsite of *users.ox.ac.uk* containing personal web pages at Oxford had the highest so-called *betweenness centrality* (cf. Section 3.1) of the 7669 UK university subsites investigated in the dissertation, as further outlined in Section 6.3.2.4.

In this context, it is interesting that hypertext research early on applied graph theory and centrality measures to investigate topological properties in large hypertextual systems. This research is thus relevant also in a Web context. Botafogo & Shneiderman (1991) and Botafogo, Rivlin & Shneiderman (1992) developed a compactness measure for the degree of interconnectedness or cohesion based on a so-called *distance matrix* which gives the length of the shortest link path between each pair of nodes in a hypertext graph.[17] Smeaton (1995) applied this compactness measure to determine how the addition of extra links affect the cohesiveness of the overall hypertext topology. As stated by Smeaton (ibid.):

> "*To make a hypertext easy for a reader to navigate it should require few hypertext jumps in order to move from any node to any other. A compact hypertext is convenient in the sense of requiring few hypertext jumps to travel within the hypertext*."

This focus by Smeaton on short link distances thus resembles small-world approaches to the cohesiveness of link structures on the Web. Interestingly, in the context of the methodology developed in the dissertation of juxtaposing topically dissimilar web nodes, Smeaton (ibid.) also investigated how several so-called 'mini-hypertexts' containing different topics were connected with *"cross-topic links"* to make a *"coherent overall structure"* including a global overview node that provided *"a path from any area to any other area."*

In a paper in *Scientometrics*, De Bra (2000) suggests how the compactness measure and other hypertext metrics from Botafogo, Rivlin & Shneiderman (1992) may be applied in graph theoretic analysis of differences in publication and citation habits in different scientific communities. Subsequently, Egghe & Rousseau (2003a; 2003b) present a generalization of the compactness measure by Botafogo, Rivlin & Shneiderman (ibid.) with regard to the degree of cohesion or interconnectedness of citation graphs, co-citation graphs, scientific collaboration networks and the Internet. Egghe & Rousseau (2003a) suggest that it would be interesting to test whether the compactness measure can distinguish between real-world random and small-world networks. As stated by Egghe & Rousseau (2003b) with a reference to Khan & Locatis (1998): *"the density and cohesion of links in a hypermedia environment influences the retrieval efficiency of users"*. In this context, Egghe & Rousseau (ibid.) also include

[17] Cf. the description of matrices in Section 3.1.

Pritchard's (1984) approach on how a high level of accessibility between nodes improves the transfer of information.

This statement and approach is in accordance with the earlier suggestion (cf. Section 1.3) of how small-world properties of link structures may affect navigability and accessibility of information across vast document networks on the Web. For instance, how short connectivity distances along link paths may influence the speed and exhaustivity with which web crawlers can reach and retrieve web pages when following links from web page to web page.

The emergence of small-world topologies on the Web and in other evolving complex networks can be attributed to so-called *scale-free* network features (Barabási & Albert, 1999; Barabási Albert & Jeong, 2000). As noted earlier, there is no 'typical' node, that is, no characteristic 'scale' to the degree of connectivity in a scale-free network. Scale-free degree distributions thus show long *power-law* tails implying that only a small share of nodes are connected by many edges, whereas the bulk of nodes has quite few edges attached.

A range of power-law distributions have been identified on the Web in, for example, the scattering of TLDs (top level domains) on a given topic (Rousseau, 1997); inlinks per site (Albert, Jeong, & Barabási, 1999; Adamic & Huberman, 2000; 2001); outlinks per site (Adamic & Huberman, 2001); pages per site (Huberman & Adamic, 1999; Adamic & Huberman, 2001); visits per site (Huberman *et al.*, 1998; Pirolli & Pitkow, 1999; Pitkow, 1999; Adamic & Huberman, 2001); and visited pages within a site (Huberman *et al.*, 1998; Pitkow, 1999). Faloutsos *et al.* (1999) and Medina *et al.* (2000) have identified corresponding power-law distributions in the network of interconnected routers that transmit traffic on the Internet.

As noted by Albert & Barabási (2002), a heterogeneous scale-free topology is very efficient in bringing network nodes close to each other. Small-world structures may thus arise from a scale-free organization in which a relatively small number of well-connected nodes serve as hubs (cf. Kleinberg, 1999a; Steyvers & Tenenbaum, 2001). Well-connected web nodes often have 'multiple memberships' providing strong and weak ties (cf. Granovetter, 1973; 1982; Section 3.1) across many topical clusters. Such 'far-reaching' bridging ties may thus function as *transversal links* that contribute to contracting link distances in a network to form a small world.

In this context, it should be noted that the scale-free small-world topologies in a network also affect *network robustness* (Albert, Jeong & Barabási, 2000) in the shape of high *tolerance* to random node errors due to the high degree of redundant local clustering and thus multiple independent pathways, but with high *vulnerability* to malicious attacks on the well-connected hub-like nodes described above.

According to Barabási & Albert (1999), scale-free link distributions are rooted in two generic mechanisms of many real-world networks: *continuous growth* and *preferential attachment* ("rich-get-richer"). In this framework, the Web is an open self-organizing system that grows by the continuous addition of new nodes and links where the probability of connecting to a node depends on the number of links already attached to the node. This significance of preferential attachment for power-law distributions is well known in bibliometrics as *'the Matthew effect'*: "unto every that hath shall be given" (Merton, 1968) and *'cumulative advantage'* (Price, 1976).

Barabási & Albert's (1999) original model on explaining the emergence of scale-free power-law properties in networks has later been augmented with additional factors, including *initial attractiveness* (Dorogovtsev, Mendes, & Samukhin, 2000), *competition* and *fitness* (Bianconi & Barabási, 2001), *optimization* (Valverde, Cancho & Solé, 2002), *uniform attachment* (Pennock *et al.*, 2002), *transitive linking* (Ebel, Mielsch & Bornholdt, 2002) and *lexical distance* (Menczer, 2002). Moreover, Amaral *et al.* (2000) explain deviations in scale-free distributions by factors like *ageing*, *cost* and *capacity constraints* of network nodes.

The scale-free properties of the Web imply a *fractal* structure where cohesive sub-regions display the same characteristics as the Web at large (Dill *et al.*, 2001) – for example, with regard to the aforementioned 'bow-tie' graph components also present in Web subregions like national academic web spaces (Thelwall & Wilkinson, 2003b) and countries (Baeza-Yates & Castillo, 2001).

Such scale-free properties are often regarded as a fingerprint of *self-organization* (van Raan, 2000). According to Barabási & Albert (1999), large complex networks self-organize into a scale-free state, a feature unpredicted by the traditionally applied random network models. As noted in Section 1.1, analyses of the Web show a remarkable degree of self-organization in the shape of clustered hyperlink structures that reflect topic-focused interest communities (Gibson, Kleinberg & Raghavan, 1998; Kumar *et al.*, 1999; Kleinberg & Lawrence, 2001; Flake *et al.*, 2002). The coincidence of high local clustering and short global separation means that small-world networks simultaneously consist of small local and global distances, leading to high efficiency in propagating information both on a local and global scale (Marchiori & Latora, 2000; also cf. Sun & Ouyang, 2001[18]). However, as mentioned in Section 1.2, web links do not *directly* channel information flows like social networks, neural networks or computer networks. On the other hand, web links *indirectly* reflect information diffusion among link creators, because added or removed links may reflect changes in topical interests and social preferences of link creators. On an *aggregated macro level*, such dynamic link adaptations thus could reflect cognitive, cultural and social currents and formations, including the emergence of scholarly networks and the diffusion of scientific ideas across topical domains in such networks.

As computer-supported networks such as the Web connect people and organizations, they can host social networks (cf. Erickson, 1996; Garton, Haythornthwaite & Wellman, 1999; Wellman, 2001; Kumar *et al.*, 2002; Thelwall, Vaughan & Björneborn, forthcoming). According to Kumar *et al.* (2002), the earlier mentioned fractal self-similarity with subsets of the Web that display the same power-law-like connectivity distributions as the Web at large are also pervasive in social networks.

Garton, Haythornthwaite & Wellman (1999) argue for the usefulness of a social network analysis approach to the study of computer networks in general and the Internet in particular. Jackson (1997) focused particularly on web links as a communication tool and discussed how to interpret web link structures through social network analysis. Otte

[18] Sun & Ouyang (2001) found that in a small-world network, the longest distance (the so-called *diameter* of the network, cf. Section 5.3) between two nodes is just slightly longer than the average distance, indicating that the efficiency of network propagation is absolute rather than in average.

& Rousseau (2002) also pointed out the applicability of social network analysis to information science, especially to studies of the Internet and the Web. In this context, they declare that *"the World Wide Web represents a social network of a scale unprecedented in history"* (ibid., p. 441).

Adamic & Adar (2003) find that the link structures between personal homepages of students, faculty and staff at the MIT and Stanford University form small-world social networks at least within their local university web spaces, thus showing that *"personal homepages provide a glimpse into the social structure of university communities*" (ibid.).

Garrido & Halavais (2003) also apply social network analysis to social phenomena as reflected by link structures on the Web. The authors studied the networks of support for the Zapatista peasantry movement in the southern state of Chiapas in Mexico, a contemporary social movement in which the Internet plays a central role for gaining global attention. The authors collected data on inlinks to the Zapatista web site and mapped these links into a Zapatista connectivity network on the Web. An examination of this hyperlinked network of web sites provided a unique insight into the character of the Zapatista's phenomenal success and particularly the degree to which they have become a catalyst for a transnational network of activists.

Park & Thelwall (2003) note that social network analysis approaches to link structures on the Web are based upon the assumption that hyperlinks may serve as social symbols or signs of communication:

> *"In other words, hyperlinks are considered not simply as a technological tool but as a newly emerging social and communicational channel. There is a tie through hyperlinks that connects individuals, organizations, or countries on the Web."*

Hyperlink networks on the Web may thus in some circumstances reflect *off-line* connections among social actors in social networks. The authors suggest how webometrics can benefit from adopting the extensive social networks analysis tool set – including small-world approaches.

In the following three empirical chapters, such webometric approaches drawing on social network analysis tools are applied to the UK academic web space. There has not been found any studies – neither in library and information science nor in other disciplines – investigating small-world properties across topical domains in an academic web space, or identifying and characterizing what types of links, pages and sites contribute to such small-world properties.

4 UK link data

The empirical investigation in the dissertation spans three chapters. The present chapter gives a starting point by outlining the original collection of link data from a harvest in 2001 of 109 UK academic web sites as well as methodological considerations regarding the extraction of a data subset comprising 7669 university subsites to be analyzed in the dissertation.

The subsequent Chapter 5 implements a broad range of graph measures on how cohesively interconnected are the link connectivity structures in the UK academic web space in order to address the first two research questions in the dissertation. The empirical investigation is finished in Chapter 6 with a developed five-step methodology for sampling, identification and characterization of small-world properties in an academic web space in order to answer the last two research questions in Section 1.4.

4.1 Original data set

Early contacts in the PhD project to proprietary search engines with requests of access to large-scale link data did not yield any positive response. For example, the Connectivity Server (cf. Bharat *et al.*, 1998; Broder *et al.*, 2000) was contacted because they use link data from AltaVista for modeling link structures containing hundreds of millions of links. The Internet Archive (*www.archive.org*) that has harvested 10 billion web pages since 1996 was also contacted because they give research access to subsets of their databases. The insoluble obstacle here was that the requested UNIX programming skills in order to process the data were not available at the Royal School. The lack of available computing skills at the Royal School also optioned out an alternative approach for the PhD project to launch a web crawler to harvest web pages and extract links.

Fortunately, the problematic lack of accessible large-scale databases with web link data needed in the PhD project was solved when the author was contacted in July 2001 by Dr Mike Thelwall at the Statistical Cybermetrics Research Group, University of Wolverhampton, UK, with a generous offer to get access to a database with large-scale link data that had just been made publicly available.

The database included extensive link data sets from the UK academic web space. As outlined in Section 1.3, academic web spaces were of special interest in the PhD project because of the author's interest in how the Web may be used to stimulate scientific creativity across topical domains as well as the webometric 'tradition' for investigating academic web spaces with lineages to bibliometric and scientometric traditions (cf. Sections 2.4.1 and 2.4.2).

The access to the UK link data turned out to be very beneficial as the UK academic web space is large enough to be interesting whilst not being too large in order to compute link topologies with the available network analysis tool (cf. Section 4.2).

Furthermore, the UK academia was early adopters of the Internet (cf. Kirstein, 1999) as well of the Web (cf. Kelly, 1995; Day, 2003) resulting in a relatively well-developed and mature web presence. This was of special interest in the present study of how micro-level properties of interconnectedness may emerge in an academic web space. Another beneficial aspect was that the UK academic web space makes use of a consistent naming practice, using the sub-TLD (sub top level domain), *ac.uk*, that enables convenient data filtering of UK academic web sites.

The link data used in the investigation of small-world link structures in the UK academic web space was thus downloaded from a publicly accessible database containing large-scale outlink data harvested from 109 UK university web sites as of June and July, 2001 (Thelwall (2001b; 2001c).[19]

The 2001 UK database is one of several similar databases based on extensive web crawls, the first of which started in 2000, harvesting link data from British, Australian, New Zealand, Chinese, Taiwanese and Spanish university web sites. The original purpose of the databases was to provide link data for investigations of so-called Web Impact Factors (Ingwersen, 1998), cf. Section 2.4.2, in this case concerned with link counts between university sites and research assessments of the universities (e.g., Thelwall, 2001a; 2001d; 2001e; Smith & Thelwall, 2001; 2002).

For this purpose, Dr Thelwall designed a web crawler in order to harvest primary link data directly from the Web in order to avoid secondary data from commercial search engines with opaque and inconsistent results (cf. Ingwersen, 1998; Snyder & Rosenbaum, 1999; Rousseau, 1998/1999; Mettrop & Nieuwenhuysen, 2001). This circumstance reflects a general methodological problem in webometrics, that is, the difficult selection and sampling of data from the Web as a basis for empirical investigations.

4.1.1 Universities included

The UK universities and other Higher Education (HE) institutions included in the 2001 web crawl, were based on 108 HE institutions listed in the *Times Higher Education Supplement* (THES, 2001) special section on "League Tables 2001". The THES list includes almost all official UK universities as well as most of the largest non-university HE institutions. The list does not include all universities. For example, the Open University had been omitted, because the THES list is aimed at students studying full-time and the Open University specializes in part-time distance learning education. In order to include all the largest UK universities in the web crawl, Dr Thelwall added the Open University to the web crawl. Thus, a total of 109 UK universities and other HE institutions (hereafter all called universities) were included in the 2001 web crawl. Appendix 3 gives an overview of the 109 UK university web sites included in the data set.

It should be noted that the universities of London and Wales are officially single institutions, but their constituent colleges, etc., were included as separate institutions in

[19] The database is accessible at *http://cybermetrics.wlv.ac.uk/database*

the web crawl, "following the THES list, public perception, financial reality and web site naming conventions" (Thelwall, personal e-mail 27.6.2003).

The 109 universities included comprise a limited proportion of the UK HE institutions. One of the most inlinked web pages in the UK academic web space (Thelwall, 2002c) contains an extensive clickable map of "all recognised Universities, University Colleges and Higher Education Colleges in the United Kingdom except for residential colleges within universities" (Burden, 2003) – see Appendix 1. A list of all the universities and colleges on the map as of 2001 (Burden, 2001)[20] contains 132 unique domain names for UK universities including the constituent colleges, etc., at the universities of London and Wales.[21] The 2001 web crawl covered 101 (76.5%) of these universities and colleges. Furthermore, the list (Burden, 2001) comprises 74 unique domain names of university sector colleges, international colleges and professional institutions. Of these, eight (10.8%) were covered by the web crawl. In Appendix 2, the included 109 universities and colleges are highlighted on the list.

The universities and colleges not included on the THES list (2001) and thus excluded from the 2001 web crawl are typically smaller universities and colleges covering a limited variety of disciplines, for example, schools in arts, business, medicine or music. For the purpose of the present study, it could have been interesting if such more topical focused institutions had been included in the web crawl in order to provide as varied topics as possible to be included on the shortest link paths investigated in the study. From a LIS perspective, it would also have been interesting if the Cranfield University had been included in the web crawl, because of the famous Cranfield experiments conducted in the 1950s and 60s on the precision and recall in information retrieval.

However, the 109 included universities turned out to comprise a rich diversity of topics covered by their departments, centers, research groups, etc. This circumstance was exploited in the present study in order to investigate possible small-world phenomena in the shape of short link paths between such university subunits reflecting dissimilar topics.

4.1.2 Original web crawl

The original web crawler software was programmed by Dr Thelwall to start at the top homepage of each of the 109 included universities, and then follow all links from page to page on each university web site. Some university sites had no embedded links on the home page, instead using a pull-down menu for selecting pages, a feature ignored by the web crawler. In order to allow a web crawl to take place on such sites, an alternative web page was selected as the starting point for the crawl, usually a page with a list of links to the department home pages at the university. A page was judged to be on the

[20] A page indexed in July 2001 at the Internet Archive (*www.archive.org*) containing the list was downloaded in order to get an overview over the universities and colleges at the time of the 2001 web crawl: *http://web.archive.org/web/20010707114102/http://www.scit.wlv.ac.uk/ukinfo/alpha.html* [visited 27.9.2003]

[21] The two generic web sites of these universities *www.lon.ac.uk* and *www.wales.ac.uk* were not included in the 2001 web crawl.

same university site if its URL contained the same three rightmost suffix-segments in their domain name as the home page, for instance, *.man.ac.uk*. This procedure allowed subsites with derivative domain names like *lib.man.ac.uk* and *www.maths.man.ac.uk* to be included in the data set. As will be more elaborated in Section 4.2.1 (cf. Table 4-2), the larger universities contained hundreds of derivative domain names of this form, often one for each school, department or research group. Some universities also had non-derivative domain names for separate sites, for example, *www.mcc.ac.uk* for Manchester University's Computing Centre. These sites were also crawled when identified. However, some secondary web sites will probably have been missed if there were no outlinks to them or because they were registered in a non-academic domain, for example, *co.uk* or *org.uk*. For instance, some universities have registered commercial domain names for industry-related projects, and some individual academics even have personal domains, such as the theoretical physicist and cosmologist Stephen Hawkings, *www.hawking.org.uk*. These web sites were all ignored by the original web crawler because it was not practical to check the identity of the owner of each web site in a non-academic domain (Thelwall, 2001a).

The web crawler also excluded, when identified, so-called *mirror*[22] sites (such as entire collections of web pages located on multiple web sites, for example, containing copies of computer documentation), online e-journals, and hosted web sites of external organizations because they do not represent content created at the host institution (Thelwall, 2003a; forthcoming). Furthermore, anomalies like web pages containing thousands of automatically generated links were excluded because they would bias inter-university link counts. Human intervention was needed in order to identify such undesirable web pages and subsites (Thelwall, 2001b).

A web site may contain many web pages with different URLs but identical contents. Such duplicate pages were also identified and excluded, where possible in the original link data set (Thelwall, 2001a).

The web crawler obeyed the convention of ignoring pages specified by site owners in the so-called *robots.txt* file (World Wide Web Consortium, 1999)[23] . However, most web sites had a limited number of such banned areas (Thelwall, 2001a).

Wrongly protocolled outlinks were excluded as non-valid by the web crawler. Only outlinks using the *http*-protocol (*HyperText Transfer Protocal*) were harvested. Furthermore, wrongly HTML-tagged outlinks, for instance, without an end tag ** were excluded. However, it was not feasible to validate all site outlinks in order to check if they were accessible. It is thus not possible in the present study to verify if such outlinks were outdated already when harvested in 2001. Because link targets thus were unvetted in the original data set, some of the *target* web pages in the present study

[22] A mirror site is "a replica of an already existing site, used to reduce network traffic (hits on a server) or improve the availability of the original site" (*www.webopedia.com/TERM/M/mirror_site.html* [visited 27.6.2003]).

[23] The robot exclusion protocol (REP) is a method implemented on web servers to control access to server resources for robots that crawl the Web (Kelly & Peacock, 1999). Web crawlers will look for a robots.txt file at the root of the server file hierarchy. This robots.txt file can be customized to apply only to specific robots, and to disallow access to specific directories or files (cf. World Wide Web Consortium, 1999): www.w3.org/TR/html4/appendix/notes.html#h-B.4.1.1). Furthermore, web crawlers do not index web pages containing the HTML tag: *<META name="robots" content="noindex, nofollow">* (ibid.).

belonged to mirror sites, e-journals, and hosted web sites of external organizations, that had been excluded from the *source* pages as mentioned above.

The web crawler harvested 3.40 million outlinking web pages containing 39.34 million outlinks. Only the count of source pages with outlinks is used here as a pragmatic approximation of the number of web pages, because data on the so-called MIME type that unambiguously can identify file formats, including true HTML extensions, had not been extracted by the original web crawler. The lack of unambiguous file format identification also meant that non-HTML file formats were included as targets, for example, PostScript files. Section 5.4.2 gives more statistical details on the counts of pages and links from the data set, including counts of site selflinks and site outlinks.

The downloaded 2001 UK database consisted of 109 separate plain text files (total 2.3 GB), one for each university web site, containing long lists (cf. Fig. 4.8, Section 4.2.3.1) of the URLs of all crawled outlinking *source* pages including the URLs of all identified outlinked *target* URLs with duplicate URLs removed and all URLs truncated the first '#' segment when present (for bookmarked text paragraphs, pictures, etc. within a web page). This last point means that a source page cannot contain harvested links to two or more parts of the same target page (Thelwall, 2002b). In order to save space, the database did not contain other page data such as page title, body text, anchor text, file format (MIME type), etc.

The database providing the basic data set in this dissertation thus does not cover the entire web sites of the included universities, and thus is not a definitive crawl because of web pages deliberately excluded and because other pages failed to be included, for example, due to lack of inlinks.

This extensive description of included and excluded pages and subsites in the web crawl also illustrates some of the difficulties when extracting data from the Web.

When processing the link data files in order to count inter-university links as the original purpose was – or in order to construct the so-called adjacency matrix (cf. Section 4.2.5) as in the present study – all source and target URLs were converted to canonical domain name forms (Thelwall, 2002b) in order to obtain data comparability. For example, *.doc.edinburgh.ac.uk* was converted to *.doc.ed.ac.uk*. This precaution was necessary because many universities have restructured their domain names, leaving the old versions active for existing links. Also other types of domain names of subunits at the same university were converted to a canonical name form, for instance, *mcc.ac.uk* to *man.ac.uk,* because the Manchester Computing Centre was a subunit of the Manchester University as noted earlier. Domain names starting with a *www.*-prefix were truncated to avoid duplicate pages through the common practice of having multiple equivalent versions of web domain names, for example, *www.wlv.ac.uk* and *wlv.ac.uk*, both giving access to the University of Wolverhampton. Appendix 4 contains a list of the canonical and variant domain names of the 109 universities in the 2001 data set.

4.1.3 Web terminology

The terms *web site*, *web server, domain name*, and *URL* are frequently used in the dissertation. Some brief terminological definitions are thus appropriate.

A *web site* is a location given a domain name on the Web containing one or more web pages created and maintained by an individual, company or organization (cf. *www.webopedia.com/TERM/w/web_site.html*)[24]. A web site may be conceived as a *web* term for an allocation of web *documents*, whereas the term *web server* is an *internet* term representing one or more computer *machines.* This conceptual distinction is essential as the Web and the Internet are two distinct entities, because the Web is a network of *documents* connected by hyperlinks, whereas the Internet is a network of *computers* connected along backbones of routers.

A *web server* is "a computer that delivers (serves up) web pages" (*www.webopedia.com/TERM/W/Web_server.html*). Every web server has an IP (Internet Protocol) address and possibly a domain name (see definition below). Any computer can be turned into a web server by installing server software and connecting the machine to the Internet (ibid.)

A *domain name* functions as an addressing system and identifier with an alphanumeric name used to identify one or more so-called IP addresses. Because the Internet is based on IP (Internet Protocol) numbers, not domain names, every web server requires a Domain Name System (DNS) server to translate domain names into IP addresses (cf. *http://www.webopedia.com/TERM/D/domain_name.html*). For example, 130.226.186.6 is an IP number with the alias domain name *www.db.dk* at the Royal School.

A basic domain name consists of three segments, *www.xxx.yy*. The rightmost domain name segment *yy* is either a generic top level domain (gTLD) like *aero, biz, com, coop, edu, gov, info, int, mil, museum, name, net, org*, or one of the over 240 country code top level domains (ccTLD)[25]. In the above example, *xxx* is the main site name (aka host name) and *www* is the server name that sometimes has an alias server name used by local computer staff (for example, *ex.db.dk* is an alias server name for *www.db.dk* at the Royal School).

At the second rightmost part of the domain name next to the TLD, many countries like the United Kingdom add a generic second level domain (SLD, also called a sub-TLD) for selected societal sectors; for instance, some UK sub-TLDs are *ac* (academic web sites), *co* (commercial), and *org* (non-profit organizations).[26] Other countries like Denmark do not use SLDs, thus impeding data filtering, because SLDs as mentioned earlier enable useful webometric data filtering options as in the present study of UK academic web sites.

---

[24] The Webopedia web site on *www.webopedia.com* used in this section was visited 4.7.2003.
[25] Cf. 'Domain name registries around the world': *www.norid.no/domenenavnbaser/domreg-alpha.html* [visited 4.7.2003]
[26] Cf. 'Nominet UK - Second Level Domains': *www.nic.uk/news/guides/reg4.html* [visited 4.7.2003]

More extensively segmented derivative domain names are frequent in the investigated data set of *ac.uk* web sites, for example, *teachernetuk.ultralab.anglia.ac.uk* This domain name consists of a TLD (*uk*), SLD (*ac*), the main site name (*anglia*), the subsite name (*ultralab*), the sub-subsite name as the leftmost segment of the domain name (*teachernetuk*).

URLs, Uniform Resource Locators, "identify resources in the Web: documents, images, downloadable files, services, electronic mailboxes, and other resources. They make resources available under a variety of naming schemes and access methods such as HTTP, FTP, and Internet mail addressable in the same simple way." (World Wide Web Consortium, 2002).

In the URL *http://teachernetuk.ultralab.anglia.ac.uk/v7/news/news03.htm* the first part of the URL indicates what Internet protocol to use (http, HyperText Transfer Protocol). The URL string after the domain name specifies the exact location of the web page *news03.htm* within the subdirectory *news* within the directory *v7* in the web site file hierarchy.

4.2 Methodological considerations and delimitations

"Theory suggests what should be measured. Data limits what can be measured."
(WISER, 2001: part B, p.18)

In order to study the Web or a subset of the Web with quantitative methods, given the distributed and dynamic nature of page and link creation, it will not always be possible to find or analyze every page or link and so consideration must be given to the selection of the sample of pages and links to be processed (cf. Thelwall, Vaughan & Björneborn, forthcoming). The inherent distributed and dynamic properties of the Web thus crucially limit which webometric methodologies are feasible – also in the current study, in accordance with the opening quote.

A range of necessary initial methodological considerations and delimitations were needed regarding the selection of units of analysis and what inclusion criteria to employ in the empirical investigation conducted in this dissertation. In order to answer the research question concerned with identifying and characterizing web links, web pages and web sites providing small-world shortcuts across dissimilar topics in an academic web space, an appropriate and feasible methodological approach had to be developed with regard to the given data set. The methodology developed in the dissertation comprises five steps of 'zooming in' deeper and deeper into the data set. This stepwise methodology is described in more detail in Chapter 6 based on a graph model of the UK academic Web presented in Chapter 5. In the course of describing and substantiating the stepwise methodology, the consequences of methodological decisions at each step are addressed.

The methodology was developed to comply with limits imposed by the given data set; by the computing capability of available tools of network analysis; and by the limited programming expertise available at the Royal School.

4.2.1 Focus on university subsites

The limits imposed by the given data set were concerned with problems including the lack of consistent naming practices of university domain names, the multi-disciplinarity of web sites, the unavailability of outdated URLs and with problems of non-valid domain names. These problems will be addressed in the following sections.

Different network analysis tools were considered for investigating link structures extracted from the data set. The tool *NetMiner* (*www.netminer.com*) was excluded because of the low number of network nodes (200-300) that could be analyzed. Instead, *Pajek*, a software program for large network analysis was chosen because it could handle large-scale data, was freeware, had a good reputation, and contained an intuitive user interface and extensive help manuals and tutorials.[27] Pajek can handle matrices of network data containing several thousands of nodes depending on the memory capacity of the computer used. It was thus out of question to investigate link structures among the 3.4 million outlinking web pages in the raw data set because of the computer memory implications. Instead, a feasible approach was to use Pajek to compute link structure properties including shortest link paths between the identified 7669 subsites, sub-subsites and sub-sub-subsites (hereafter called subsites for sake of simplicity) at the 109 universities. The 109 university main web sites were not included in the study, as will be further discussed in next section.

All three levels of subordinate subsite units listed in Table 4-1 below were included in the study in order to provide a rich population of nodes representing as many different topics as possible. The intention with this inclusion was to obtain sufficient topical heterogeneity among the nodes to facilitate the identification of small-world shortcuts that was in focus in this study of the UK academic subsite web space (called *subweb* in the dissertation).

	# stemmed domain name segments	prototype example	#
main univ. sites	3	ddd.ac.uk	109
subsites	4	ccc.ddd.ac.uk	4751
sub-subsites	5	bbb.ccc.ddd.ac.uk	2852
sub-sub-subsites	6	aaa.bbb.ccc.ddd.ac.uk	64
			[28]**7776**

Table 4-1. Distribution of stemmed domain name segments in the UK data set.

Pilot tests showed that naming practices were quite inconsistent between the universities. It would thus have made no sense to exclude any of the lower levels if the objective had been to obtain a more homogeneous population. For example, research groups could consist of four, five or six segments in their stemmed domain names, as in

[27] The network analysis tool *Pajek* (Slovenian for 'spider') has been developed since 1996 in more and more sophisticated versions by Professor Batagelj and one of his students, Andrej Mrvar, at the University of Ljubljana in Slovenia (*http://vlado.fmf.uni-lj.si/pub/networks/pajek/*).

[28] 7776-109 = 7667. This count deviates from the earlier mentioned number of 7669 subsites because two domain names with 3 segments turned out to be typos. This problem is elaborated in Section 4.2.3.2.

speech.essex.ac.uk (Speech Group, Department of Language and Linguistics, University of Essex), *ep.cs.nott.ac.uk* (Electronic Publishing Research Group, School of Computer Science and Information Technology, University of Nottingham), *james.lsr.ph.ic.ac.uk* (Laser Optics & Spectroscopy Group, Department of Physics, Imperial College).[29]

Universities employ different policies to locate subunit information. Some universities use an extensive range of subsites, sub-subsites and further subordinate derivative domain names – like Chinese boxes within boxes – as web territories for schools, departments, centers, laboratories, research groups, individual researchers, etc. For example, the Palaeontology Research Group at the University of Bristol has a sub-subsite with the stemmed domain name *palaeo.gly.bris.ac.uk*, the Department of Earth Sciences at the same university has the domain name *gly.bris.ac.uk*.

Table 4-2 shows an excerpt of the universities with the highest and lowest number of identified subsites. The five universities of Cambridge, Oxford, Edinburgh, Glasgow and Manchester comprised 25% of all harvested subsites. Furthermore, 16 (14.7%) of the 109 universities cover over 50% of the 7669 subsites. The Surrey Institute of Art and Design had no subsites harvested by the web crawler. See full list in Appendix 5.

rank	domain name	university	# sub-sites	%	cum. %
1	cam.ac.uk	Cambridge	582	7.59	7.59
2	ox.ac.uk	Oxford	515	6.72	14.31
3	ed.ac.uk	Edinburgh	321	4.19	18.49
4	gla.ac.uk	Glasgow	297	3.87	22.36
5	man.ac.uk	Manchester	266	3.47	25.83
6	ic.ac.uk	Imperial College	251	3.27	29.11
7	soton.ac.uk	Southampton	249	3.25	32.35
8	ucl.ac.uk	University College London	227	2.96	35.31
9	open.ac.uk	Open University	174	2.27	37.58
10	strath.ac.uk	Strathclyde	171	2.23	39.81
11	bris.ac.uk	Bristol	167	2.18	41.99
12	bham.ac.uk	Birmingham	164	2.14	44.13
13	leeds.ac.uk	Leeds	162	2.11	46.24
14	ncl.ac.uk	Newcastle	149	1.94	48.18
15	umist.ac.uk	Manchester Inst. of Science and Technology	125	1.63	49.81
16	nott.ac.uk	Nottingham	119	1.55	51.36
	…	...	…	…	
99	lmu.ac.uk	Leeds Metropolitan	7	0.09	99.53
100	uclan.ac.uk	Central Lancashire	7	0.09	99.62
101	lamp.ac.uk	Lampeter	6	0.08	99.70
102	lgu.ac.uk	London Guildhall	6	0.08	99.78
103	worc.ac.uk	Worcester	6	0.08	99.86
104	northampton.ac.uk	Northampton	4	0.05	99.91
105	bathspa.ac.uk	Bath Spa	3	0.04	99.95
106	buckingham.ac.uk	Buckingham	2	0.03	99.97
107	chichester.ac.uk	Chichester	1	0.01	99.99
108	harper-adams.ac.uk	Harper Adams	1	0.01	100.00
109	surrart.ac.uk	Surrey Institute of Art and Design	0	0.00	100.00
			7669	**100.00**	

Table 4-2. UK universities with most and least number of subsites. Full list in Appendix 5.

Though highly skewed as illustrated in Table 4-3 below, the distribution of subsites per university does not 'hug' the x and y axes sufficiently tight to follow a power law (cf.

[29] Affiliations are from web pages indexed in the Internet Archive (*www.archive.org*) as close to the original data harvest in June-July 2001 as possible. Present institutional names may thus have changed.

Section 5.4). This was confirmed by the special *LOTKA* software program developed by Rousseau & Rousseau (2000) for fitting power-law distributions. The lack of power-law distribution may perhaps be due to a combination of different university sizes and different research productivities, giving two important factors rather than just one.

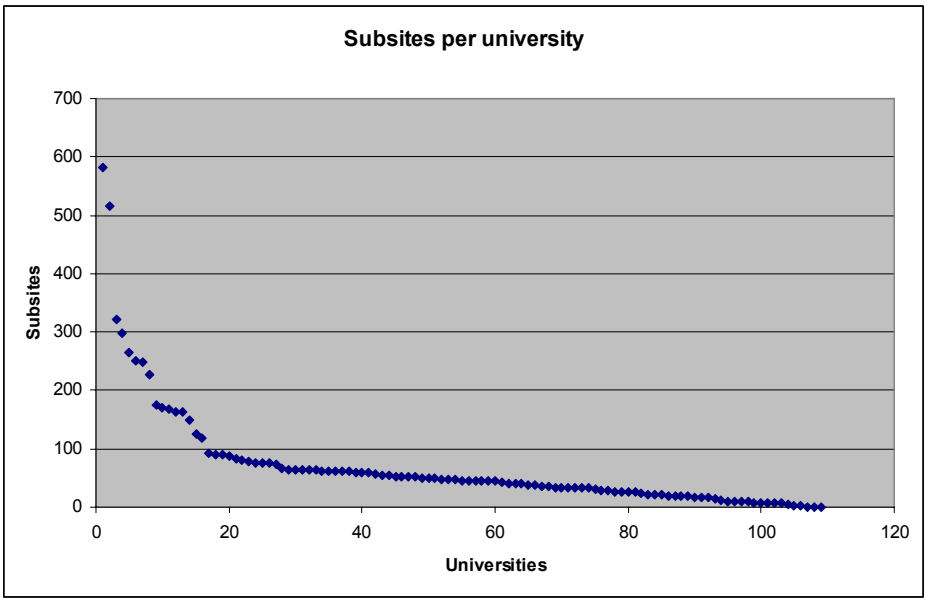


Table 4-3. Distribution of 7669 subsites on 109 UK universities.

Other universities locate the above type of subunits into different sublevels of folder directories in the server file hierarchy. Many universities use a combination of derivative domain names and directories. For example, *School of Computing, Information Systems & Mathematics* at South Bank University has a homepage at *scism.sbu.ac.uk* as well as at *www.sbu.ac.uk/scism/*. Some universities like South Bank locate departments and other subunits of the university in directories just below the root URL of the university. Other universities locate the same type of units further down in the hierarchical URL structure. For instance, the Department of Visual Arts at Keele University is located in a sub-directory at *www.keele.ac.uk/depts/va/*. Obviously, such diverse allocation and naming practices complicate comparability in webometric studies including the present one. If stemmed web directories, for instance, down to two sublevels in order to incorporate departments at the university main sites as the ones of Keele, had been used as units of analysis in the dissertation, the number of nodes to be analyzed in the network analysis tool Pajek would probably have exceeded perhaps 20,000-30,000, which was well beyond the capacity of Pajek. The number of web directories in the raw data set has not been counted in the present study. An indication of the count can be found in a survey of link data and impact factors among 111 UK universities as of 2002 conducted by Thelwall (2003c) who investigated link counts between 94,983 web directories stemmed at all sublevels under the URL roots. However, it is not link counts but identification of all shortest link paths between two nodes, which are much more demanding to compute, that provide the key figures in the dissertation.

The selection of stemmed domain names as the unit of analysis in the dissertation imply that universities prone of locating academic subunits in web directories will not be as visible nor appear as topically diversified as universities placing the same kind of subunits in subsites with derivative domain names.

4.2.2 Exclusion of university main sites and site selflinks

Logically, degrees of specificity are lost when merely the domain names and not the web directories nor the full page URLs are used as units of analysis. If page URLs had been used, a page *p* as in Fig. 4.1 would function as an intermediate node with both a received inlink and a provided outlink on a shortest link path between two other web pages. The figure illustrates a node diagram as introduced in the conceptual framework in Section 2.3.2 where main sites are denoted as circles with a single borderline, subsites with double borders and sub-subsites with triple.

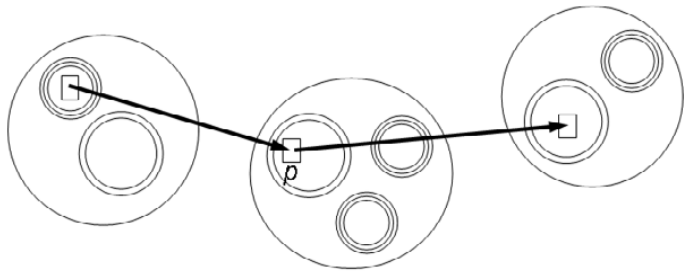


Figure 4.1. A shortest link path between web pages.

However, in the employed methodology it is domain names – and not pages – that function as interlinked nodes on shortest paths. Contrary to the case in Fig. 4.1, this means that it may be different pages *q* and *r* on an intermediate subsite node on a shortest path as in Fig. 4.2 that receives the inlink and provides the outlink, respectively.

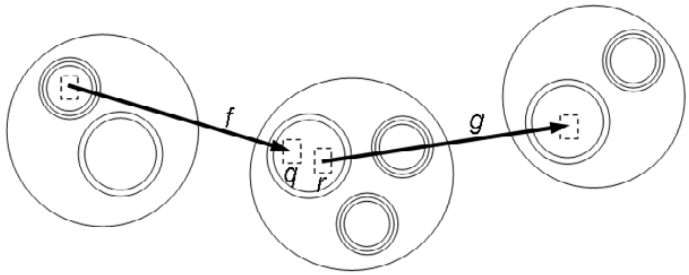


Figure 4.2. A shortest link path between nodes representing domain names.

This difference between inlinked and outlinking pages implies that if the page contents in, for instance, a subsite do not predominantly deal with a shared overall topic but instead comprise a diversity of topics, then an inlinked page *q* can be topically distant from the outlinking page *r*. The non-linked '*gap*' between *q* and *r* means that

transversal links across dissimilar topics as searched for in this dissertation can occur *within* site nodes – hidden and unidentifiable in the 'gap' – and not *between* nodes. Naturally, the pages *q* and *r* may be connected by a link path, for example, including navigational links within the site. Furthermore, link paths traversed by human web surfers or digital web crawlers between *pages* on the real-world Web as in Fig. 4.1 may cross topical boundaries within sites. However, because of the necessary delimitation of investigated nodes to be web sites and not pages, only links like *f* and *g* in Fig. 4.2 between nodes representing domain names were available and investigable in this dissertation. It would thus be an advantage if each included node in the dissertation represented as topically focused contents as possible so potentially transversal links would occur between nodes where they can be identified and not within nodes where they cannot be identified.

This meant that university *main sites* posed a problem with regard to being included or not in the dissertation. University main web sites typically function as gateways to a diversity of scientific disciplines and other resources at each university, including a variety of schools, departments, centers, research groups, administrative units, etc. It was decided to exclude links to and from the 109 university main sites because the multi-disciplinary contents of resources located in different folder directories at the main sites would blur where transversal links occur between the included nodes. This would have made it difficult to answer the central research question concerned with identifying what kind of links, pages and web sites provide transversal shortcuts in small-world link structures. If the university main sites had been allowed to be included, many shortest paths between pairs of nodes in the data set would probably pass such multi-disciplinary main sites. For example, a shortest link path connecting a subsite in linguistics and a subsite in physics could pass a university main site containing both topics, a situation illustrated in Fig. 4.3.

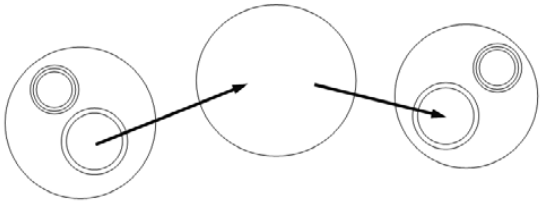


Figure 4.3. A shortest link path between two topically dissimilar subsites passing a multi-disciplinary university main site.

The exclusion of links to and from the 109 university main web sites means that shortest link paths instead will have to pass subsites, sub-subsites, etc., as in Fig. 4.4 below. The exclusion thus enables more clear-cut sampling and identification of transversal link structures because subsites were more likely not to be multi-disciplinary but of a more mono-disciplinary kind, such as those of university departments, research centers and research groups.

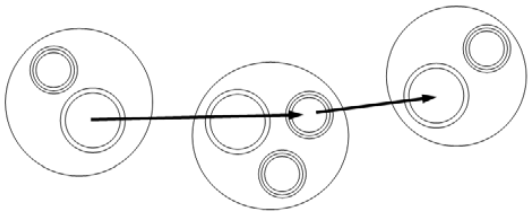


Figure 4.4. A shortest link path between two subsites passing a sub-subsite.

Admittedly, subunit nodes could also be multi-disciplinary as subsites representing homepages of colleges or faculties. However, it was not practical to identify such nodes among the over 7000 included nodes. These kinds of nodes will nevertheless be detected during the investigations of specific shortest paths undertaken later in the study.

Another effort to obtain a more clear-cut picture of cross-disciplinary connectivity was the exclusion of *site selflinks* such as link *a* in Fig. 4.5 below. Site selflinks were excluded from the data set in order to avoid subsites belonging to the same university functioning as cross-topic shortcuts due to embedded university navigation links within subsite pages. The above example of a shortest link path connecting a subsite in linguistics and a subsite in physics at another university could this time pass a physics department at the first university through such a navigational site selflink.

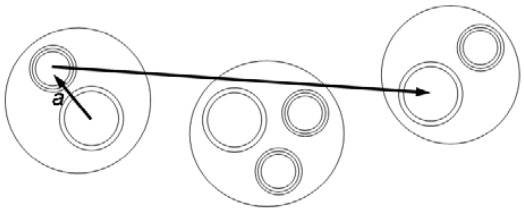


Figure 4.5. A site selflink, *a*, on a shortest path between two subsites.

The exclusion of site selflinks follows a standard procedure in webometric studies to exclude navigational links in order to obtain a more clear-cut picture (Thelwall, 2002b; 2002e; Wilkinson *et al.*, 2003).

The node diagram in Fig. 4.6 below gives an overview of the types of included and excluded links in the study. Thick lines denote included links between subsites, sub-subsites, etc., located at different universities. Thin lines denote excluded site selflinks and excluded links to and from university main sites. Lines represent links in both directions, that is, both outlinks and inlinks.

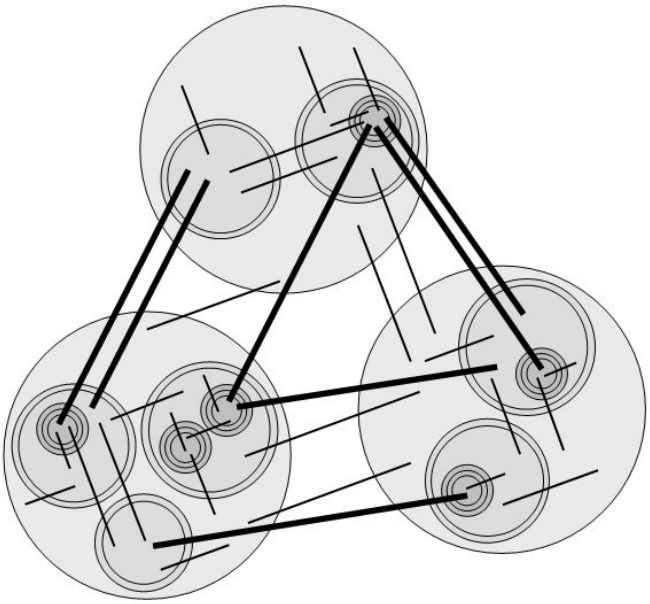


Figure 4.6. Links included and excluded in the study. *Thick* lines denote *included* links between subsites, sub-subsites, etc. (circles with two or more borders) located at different universities. Thin lines denote excluded site selflinks and excluded links to or from university main sites (circles with single border). Lines represent links in both directions.

So far, this discussion about necessary initial methodological considerations and delimitations in the dissertation has dealt with what to include or exclude from the given raw link data set from the 109 UK universities. However, inclusion criteria for the original raw data set also imposed some delimitations that should be mentioned. The data set constitutes a 'frozen' snapshot picture of the publicly available link structure at the 109 universities as of June-July 2001. This *temporal delimitation* does not capture the dynamics of the investigated link structures. However, this issue is not in focus in the present dissertation, but would clearly be interesting to follow up in future longitudinal studies of snapshots of the same population of web sites, including how transversal links change over the years.

Another delimitation is constituted by all the web sites not harvested by the web crawler in the original raw data set. Fig. 4.7 below contains the same included links (thick lines) as Fig. 4.6 representing the delimited data set. Compared with Fig. 4.6, the figure below shows other links not included in the data set. These non-included links comprise links to and from (thin lines) non-harvested nodes such as UK academic web sites outside the 109 included universities and colleges (cf. Section 4.1.1) – as well as other UK sub-TLDs as commercial or governmental web sites and foreign TLDs. This *population delimitation* causes disconnected link structures due to non-included sites. Of importance for this dissertation, there could be shorter link paths via non-included nodes than via included nodes. The results of the dissertation should thus be interpreted in the context of this information.

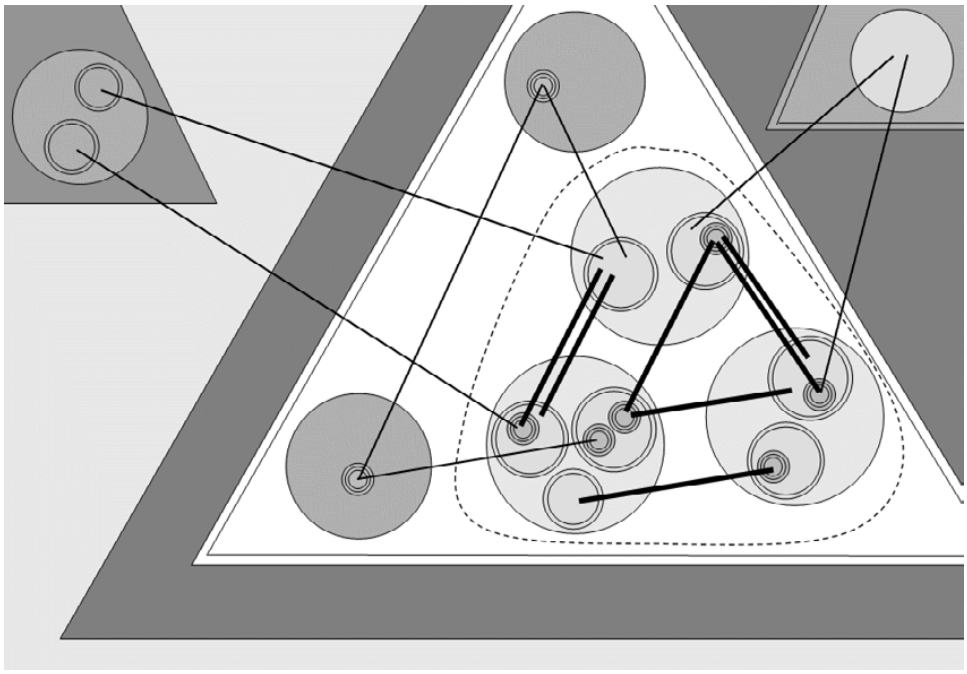


Figure 4.7. Links included (thick lines) and excluded (thin lines) in the data set (within dashed border). Thin lines denote links to and from non-harvested nodes in the UK academic sub-TLD (*.ac.uk*) outside dashed border, other UK sub-TLDs (such as *.co.uk*), or in foreign TLDs (as *.edu*).

Naturally, the employed exclusions create delimited link structures. However, the included link structures still reflect real-world link connectivity patterns in the investigated academic web space. Furthermore, as stated earlier in Section 1.3, the objective of the dissertation is not to produce representative and generalizable findings, but to identify and characterize phenomena especially in relation to links that cross disciplinary borders in an academic web space. The overall objective in the present study is thus identification and characterization of phenomena in connection to small-world link structures in an academic web space.

4.2.3 Data problems in included subsites

There were some data problems in the original data set with regard to the included subsites. Some of the data problems were due to the circumstance that the purpose of the original link data set was to investigate correlations between inlink counts and research assessments of universities. As stated by Thelwall & Wilkinson (2003a), it was *"not assumed that the target page actually exists at the time of testing. The reason for the lack of checking is that the intention to link is viewed as more important than whether there was a typo in the URL or if the target had disappeared."*

In an analysis of *link counts* one can thus argue that 'non-clickable' links reflect intentions by page authors and hence should not be filtered from the data set. However, in a *link connectivity* analysis as the present study concerned with the accessibility and navigability of such link structures, links should be valid and traversable, that is, they should be 'clickable'. If links are not valid and 'clickable' they do not allow web surfers

or web crawlers to access target web pages. The rigidity of data requirements in a study thus depends on the purpose with the data analysis, in this case that *link traversability* – and not *link impact* – is in focus. Indeed, the opening quote in Section 4.2 is very appropriate: *"Theory suggests what should be measured. Data limits what can be measured."* (WISER, 2001: part B, p.18)

In the following section the origins of the data problems are outlined in order to possibly find tractable ways to circumvent them – as well as to realize what problems are insoluble given the circumstances.

4.2.3.1 Origins of subsite link data

As stated earlier, the original web crawler did not cover the entire web sites of the included universities. Pages were excluded by the crawler for different reasons. Some source pages were *deliberately not included* because they belonged to mirror sites or non-academic domains as mentioned earlier. Other pages *failed to be included* due to robot exclusion, lack of inlinks, etc. However, some outlinks on source pages also *failed to be excluded* because of typographical errors, *typos*, in the URLs to target pages. This illustrates the methodological 'slippery' problems when attempting to extract valid link data from the Web.

The problems with *failed-to-be-included* and *failed-to-be-excluded* nodes affected the data set comprising the subsites. In order to understand the problems of validating the included subsites, the origins of the included URLs will be briefly outlined below.

In the raw data set, source pages harvested by the web crawler were listed in long plain text files for each of the 109 universities. Harvested source pages were flagged with a '1' as *palaeo.gly.bris.ac.uk/links1.html* below in the data excerpt in Fig. 4.8. Preceding (and not succeeding as would perhaps have been more logical) the source URL is an indented list of all outlinks provided by that source page. For example, on the source page *palaeo.gly.bris.ac.uk/links1.html* is an outlink targeted to *geolsoc.org.uk*.

```
        . . .
        ibs.uel.ac.uk/ibs/palaeo/indexst.htm
        geo.arizona.edu/palynology/ifps.html
        quercus.ge.man.ac.uk/PalSoc.html
        .geology.gla.ac.uk/palaeo/syst99/
        .nhm.ac.uk/hosted_sites/hennig/
        .spb.wau.nl/jeb/
        .geolsoc.org.uk
        .geosociety.org/index.html
        . . .
        .nhm.ac.uk/hosted_sites/paleonet/
        .pitt.edu/~mattf/PaleoRing.html
        www-odp.tamu.edu/
        .york.biosis.org/
        .uhmc.sunysb.edu/anatomicalsci/paleo/
        geosci.uchicago.edu/paleo/csource/
        life.bio.sunysb.edu/morph/
        .nhm.ac.uk/hosted_sites/paleonet/ftp/ftp.html
        evolution.genetics.washington.edu/phylip/software.html
        .museum.state.il.us/svp/methods/
        palaeo.gly.bris.ac.uk/default.html
palaeo.gly.bris.ac.uk/links1.html       1
```

Figure 4.8. Excerpt (denoted by three dots) from raw data set file. A source page URL is flagged with a '1' and is placed below the preceding indented list of URLs of outlinks extracted from the source page. URLs with an initial dot denote that a prefix *www.* was omitted by the harvester program.

There were 7669 subsites identified as belonging to the 109 universities in the raw data set (cf. Appendix 5). Some of these subsites were derived from only source pages URLs, other subsites were derived from only target URLs of outlinks on the source pages, and yet others existed as both source pages and target pages in the raw data set. In Fig. 4.9 and 4.10 and Table 4-4 the different origins of the 7669 subsites are showed.

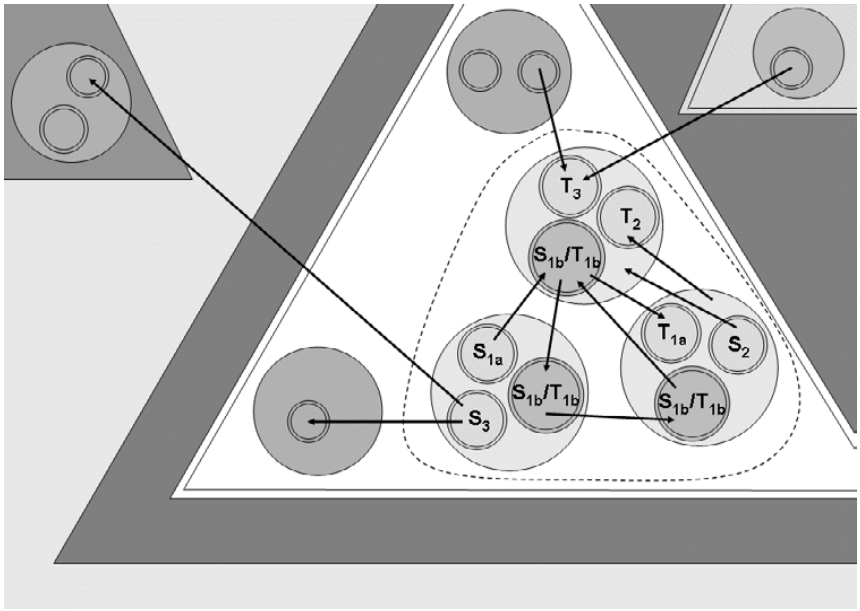


Figure 4.9. Origins of included subsites (cf. legend in Table 4-4).

code	definition	# sub-sites
S_{1a}	subsites with source pages with outlinks to - but no inlinks from - subsites at other univ.	*653
T_{1a}	subsites with target pages with inlinks from - but no outlinks to - subsites at other univ.	2686
S_{1b}/T_{1b}	subsites with source pages *and* target pages connected with subsites at other univ.	*2018
	subsites connected with subsites at other univ.	*5357*
S_2	subsites with source pages with site outlinks to detached univ. main sites	*155
T_2	subsites with target pages with site inlinks from detached univ. main sites	507
	subsites connected with detached univ. main site but not connected with subsites at other univ.	*662*
	subsites with links within 109 univ.	**6019**
S_3	subsites with source pages with site outlinks only to targets outside 109 univ.	*1650
	total number of identified subsites	**7669**
T_3	unknown number of subsites with target pages with inlinks only from outside 109 univ.	?

Table 4-4. Legend of origins of 7669 included subsites. Asterisks denote subsites with source pages all of which inevitable have valid domain names.

In Fig. 4.9, the raw data set of the 109 universities is represented by the subsites within the dashed borderline. As shown in Table 4-4, 2018 subsites were identified because they existed as subsites with both source pages (S_{1b}) and target pages (T_{1b}) connected by links with subsites at the other 108 universities. Clearly, the link structure of the 5357 subsites in S_{1a}, S_{1b}/T_{1b}, and T_{1a} with links to and/or from the other universities was of most interest in this dissertation, as they possibly contained small-world properties that could be identified.

There were 155 subsites with source pages (S_2) with no outlinks to subsites at the other universities but only linking to the detached university main sites. Although they contained no outlinks of interest in this dissertation as described earlier, they were anyhow included in order to obtain an estimate size of the UK academic subweb. This is also the case with the 1650 subsites having outlinks to neither other subsites nor main sites at the other 108 universities but only containing site selflinks or outlinks to web sites outside the harvested universities. An unknown number of subsites (T_3) were isolated or received inlinks only from sources outside the universities.

It was considered whether to use AltaVista to get an estimate of the number of T_3 subsites. The advanced search facility of AltaVista was used (on Feb 26, 2003) in a pilot test to identify such failed-to-be-included subsites from the University of Aberdeen (*abdn.ac.uk*), using the following search string with all 40 identified *abdn.ac.uk* subsites:

> host:abdn.ac.uk AND NOT (host:www.abdn.ac.uk OR host:admin.abdn.ac.uk OR host:azumaya.eng.abdn.ac.uk OR host:azumaya.maths.abdn.ac.uk OR host:biochem.abdn.ac.uk OR host:biomed.abdn.ac.uk OR host:bms.abdn.ac.uk OR host:car.abdn.ac.uk OR host:cgi.csd.abdn.ac.uk OR host:clues.abdn.ac.uk OR host:csd.abdn.ac.uk OR host:dcs.abdn.ac.uk OR host:docs.csd.abdn.ac.uk OR host:eng.abdn.ac.uk OR host:erg.abdn.ac.uk OR host:f15-1.langcen.abdn.ac.uk OR host:gauss.maths.abdn.ac.uk OR host:george.qmlib.abdn.ac.uk OR host:hist.abdn.ac.uk OR host:hutton.geol.abdn.ac.uk OR host:info.abdn.ac.uk OR host:info.biomed.abdn.ac.uk OR host:jbm.lang.abdn.ac.uk OR host:kelvin.eng.abdn.ac.uk OR host:map.eng.abdn.ac.uk OR host:maths.abdn.ac.uk OR host:mcb1.ims.abdn.ac.uk OR host:nightingale.eng.abdn.ac.uk OR host:oceanlab.abdn.ac.uk OR host:pc3.cpd.abdn.ac.uk OR host:psyc.abdn.ac.uk OR host:schools.csd.abdn.ac.uk OR host:scieng.abdn.ac.uk OR host:scratchy.csd.abdn.ac.uk OR host:sula.mill.abdn.ac.uk OR host:sysb.abdn.ac.uk OR host:thistle2.eng.abdn.ac.uk OR host:vcs.abdn.ac.uk OR host:w3.abdn.ac.uk OR host:webpac.qmlib.abdn.ac.uk OR host:wwwcad.eng.abdn.ac.uk)

The result was three unique non-included domain names. None could be verified in the Internet Archive (*www.archive.org*), which could indicate that these were three new domains not existing when Thelwall's web crawler visited the University of Aberdeen in 2001. This idea of estimating the number of T_3 subsites was not implemented on all 109 universities due to the time such a manual task would take and because the results from AltaVista would be indicative only as the search engine covers only a minority of the Web (cf. Lawrence & Giles, 1999).

As to the 7669 included subsites, all URLs and thus also their stemmed domain names of the subset of 4476 subsites derived from source pages (S_{1a}, S_{1b}, S_2, S_3) were all valid due to the fact that they had been identified and harvested by the web crawler. The remaining 3193 subsites were derived from target URLs only (T_{1a}, T_2) not visited by the web crawler and could potentially be non-valid, for example, being outdated 'dead links' or containing typographical errors, typos.

4.2.3.2 Typos in domain names

Typographical errors, typos, typically originate from misspellings made by source page creators when they manually insert URLs pointing to target pages. Some typos may also have been created by script errors in web editor programs. Typos may be copied into many web pages. Several obvious types of typos were manually identified as exemplified in Table 4-5.

identified typo characters	example of domain name typo	correct domain name
%20 (=space)	%20www.bath.ac.uk	www.bath.ac.uk
%7E (= tilde)	%7Eassem.shef.ac.uk	assem.shef.ac.uk
.	.aber.ac.uk	www.aber.ac.uk
..	wombat..doc.ic.ac.uk	wombat.doc.ic.ac.uk
,	cs,.rdg.ac.uk	cs.rdg.ac.uk
:	website:www.lse.ac.uk	www.lse.ac.uk
"	"www.bangor.ac.uk	www.bangor.ac.uk
_	www_control.eng.cam.ac.uk	www-control.eng.cam.ac.uk
@	pmg1001@econ.cam.ac.uk	**however not typos:** www@bits.bris.ac.uk ggg@ggg.qub.ac.uk
\	www-mcdonald.ar\ch.cam.ac.uk	www-mcdonald.arch.cam.ac.uk
ww	ww.aber.ac.uk	www.aber.ac.uk
wwww	wwww.ch.ic.ac.uk	www.ch.ic.ac.uk
wwwwww	wwwwww.maths.nott.ac.uk	www.maths.nott.ac.uk
redundant or omitted characters	www.grapeviine.bris.ac.uk	www.grapevine.bris.ac.uk
reversed characters	sol.utlralab.anglia.ac.uk	sol.ultralab.anglia.ac.uk

Table 4-5. Examples of obvious and possible typos in target domain names.

Table 4-5 shows some characters occurring in identified typos. For example, if a web page creator by mistake adds a keystroke space in front of a domain name in a manually inserted URL, this space is typically converted by the web editing software into its hexadecimal character code *%20* as in the example *%20www.bath.ac.uk*. Other characters are not necessarily typos. For instance, the character @ used in a URL does not have to be an email address wrongly protocolled with *http://* instead of *mailto:* as in the typo example *pmg1001@econ.cam.ac.uk*.[30] For example, *ggg@ggg.qub.ac.uk* (the electronic journal *Glacial Geology & Geomorphology* hosted at The Queen's University of Belfast) and *www@bits.bris.ac.uk* (*Bristol Information Technology Society*, University of Bristol) are both valid domain names.

As will appear from Table 4-5, some typos lead to a duplication of subsites as shown by the added correct domain names. 53 domain names with typos were manually identified by searching the 7669 subsites for the different types of possible typo characters in the table. However, many domain names consist of special proper names,

[30] Furthermore, the @ symbol can be used for user names and passwords, to gain access to protected areas of a site. For instance, *lennart:password@www.db.dk* would send the user name 'lennart' and the password 'password' to the domain name *www.db.dk*.

acronyms or abbreviations making it impossible automatically to identify typos with omitted, redundant or reversed characters.

In order to validate the domain names, special heuristics were employed as described below.

4.2.4 Data validation

A special software script was used to validate as many of the domain names as possible using the Internet Archive (*www.archive.org*). Since its start in 1996, the Internet Archive has crawled the Web and made continuous snapshots building a publicly available database containing over 10 billion web pages saved since 1996 as mentioned earlier. The Archive thus provides means for so-called *"web archaeology"* (Björneborn & Ingwersen, 2001) for the retrieval and verification of links, web pages and web sites that meanwhile may have disappeared from the dynamic Web. In the case of the present study web pages and links may have disappeared between the time of the original web crawl in June/July, 2001, and the time of the data runs and data analysis from September 2002 and onward.

Validating the domain names in the Internet Archive turned out to be quite complex with many non-trivial factors to take into account, for example, the different variant domain names converted into canonical domain names in the data set. The excluded prefix *www.* in the converted URLs in the raw data set posed an extra factor to be included in the programming script. It was possible to verify 6868 (89.5%) of the 7669 domain names this way in the Archive.

Another problem was that the Internet Archive stripped of parts of domain names containing unusual characters and hence displayed results of indexed pages not searched. For instance, a search for the subsite *http://www@bits.bris.ac.uk* resulted in the Internet Archive displaying indexed pages from *http://bits.bris.ac.uk*. All domain names with unusual characters were thus manually checked in the Internet Archive.

Like other web crawler-based databases such as search engines like Google and AltaVista, the Internet Archive cannot index all web pages because their harvester does not detect the sites or because the sites have made restrictions against web crawlers (the so-called 'robots.txt query exclusion' cf. Section 4.1). Coping with this particular problem, a subsite from the raw data set that could not be validated in the Internet Archive thus also was searched for directly on the present Web with a special script. Only another 133 subsites were verified this way. Again, variant domain names had to be searched for additionally, including the www-prefix.

Some subsites do not have a default home page. For example, such subsites represent servers containing several different research groups or projects each having subordinate directories and home pages but with no top entry page to the whole site. For example, a search for the URL *cgi.csd.abdn.ac.uk* rendered no matches in the Internet Archive, because this subsite had no default top entry page – at least not indexed in the Archive. However, a truncation wildcard search for the same URL: *cgi.csd.abdn.ac.uk** gave three matches in the Archive because three subordinate pages of this subsite were indexed in the Archive. Adding a truncation wild card to the subsite URL thus allowed Internet Archive to retrieve all indexed pages from the subsite and not just the top

homepage. Extra 143 subsites were identified this way in the Archive. Among the still non-verified subsites, additional two subsites were identified in the Archive as banned by robot exclusion.[31]

After the above heuristics to validate the domain names among the 7669 subsites in the data set, and minus the already identified 53 obvious typos (Section 4.2.3.2), a residual of 470[32] (6.1%) domain names in the data set thus could be found neither in the Internet Archive nor directly on the Web. Among these 470 domain names there will most likely be several typos, for example, with omitted, redundant or reversed characters, which only a time-consuming manual validation procedure could possibly identify. However, many of the 470 non-verified domain names could also have existed when visited by the original web crawler in June/July 2001 and just have been removed from the dynamic Web later on.[33]

The attempt to validate *all* the 7669 subsites thus did not succeed. A total of 523 subsites (53 obvious typos plus 470 non-verifiable domain names), that is 6.8% of the data set could not be verified due to a combination of human typing errors, access barriers on the Web, as well as the general elusive nature of the Web.

This demonstrates a fundamental methodological problem in webometrics. It must handle data of a much more messy, non-standardized, diverse and dynamic nature than traditional bibliographic data used in bibliometrics and scientometrics, even though data validation also is required in these fields in order to obtain adequate comparable units of data as a basis for empirical investigations. For example, in traditional co-citation analysis, it is necessary to check and edit references (e.g., Persson, 1994) in order to investigate which references are co-cited in reference lists. Without sufficient data editing, diverse name variants of author names, journal names, etc., used in references would obstruct exact match tools used for data analysis.

One positive methodological finding in the present webometric setting was that even if the Internet Archive does not cover the entire Web, a remarkably high percentage – over 90% – of the investigated UK academic subsites (minus the 53 obvious typos) had top home pages indexed in the Archive. Searching with wildcard in the Internet Archive yielded over 93% of the subsites.[34] The Internet Archive is thus an excellent *web archaeological tool* – at least for investigating the UK academic web space.

[31] The Internet Archive result page displays the message 'Robots.txt Query Exclusion' if a site owner has banned access for web crawlers – including the Archive. This message was extracted by the special validation script in the present study and was interpreted as an indirect verification that the queried domain name actually existed even though the Archive was not allowed access.

[32] 470 = 7669 – 53 typos – 6868 homepage in Internet Archive – 133 found on Web – 143 wildcard in Internet Archive – 2 robot exclusion in Internet Archive.

[33] Searches in search engines might have verified some of the 470 domain names, for example, as part of outlinks on indexed web pages in the search engines. However, this heuristic was dropped due to time problems.

[34] 7047 subsites were found with the wildcard search in the Internet Archive. (Of these, 34 had precedingly been validated among the 133 subsites verified directly on the Web).

4.2.5 Adjacency matrix

Due to the different programming problems described above, the data validation took a long time. Thus, it was not feasible to eliminate the non-valid domain names from the data set *prior* to the construction of the so-called adjacency matrix that formed the foundation for all later necessary data runs and data analyses in the dissertation. The basic adjacency matrix thus necessarily had to include non-validated subsites and links. Of course, this circumstance means that caution has to be exercised when dealing with the resulting statistics. However, as will be demonstrated in the subsequent sections, the consequences of this circumstance were minimized, because the subsites selected for further investigation in the dissertation were all chosen from a special part of the data set that was inevitably valid (cf. Section 6.1).

The adjacency matrix is a mathematical representation showing which of the subsites are *adjacent* to each other, that is, link to each other. Fig. 4.10 illustrates an example excerpt from the adjacency matrix where source subsite *f* has three outlinks to target subsite *b*. In the adjacency matrix constructed in the dissertation, the order of rows and columns reflected the unique id number that had been assigned to each subsite domain name.

		b			
	0	0	0	0	0
	4	0	0	0	0
	0	0	0	0	0
	0	1	0	0	0
	0	0	0	0	0
f	0	3	0	0	0
	0	0	0	0	0
	0	0	0	1	0

Figure 4.10. Excerpt from the adjacency matrix. The row and column headings *f* and *b*, respectively, have been added to exemplify a source subsite *f* having 3 outlinks to target subsite *b*.

The constructed adjacency matrix contained 6128 × 6128 rows and columns. In order to preserve the originally assigned id numbers equivalent with row and column numbers in the matrix, the rows and columns of the 109 detached university main sites were not deleted but all counts of links to and from these 109 web sites were zeroed in the matrix. The adjacency matrix thus represented the link structures connecting the remaining 6019 (6128 – 109) subsites. All 6019 were subsites at the 109 universities and they were identified in the raw data set as source or target pages as outlined earlier in Table 4-4. As described in Table 4-4, another 1650 subsites (1650 + 6019 = 7669) were identified as source pages in the raw data set. However, none of these extra subsites contained outlinks to subsites at other universities in the data set. As totally disconnected subsites, they were of no interest in the present study and thus not included in the matrix. However, as mentioned earlier they were included in the total

number of university subsites hence increased to 7669 in order to obtain a more complete overall picture.

In order to obtain comparability in the adjacency matrix, all URLs of source and target pages in the raw data set were stemmed to the domain name format with the earlier mentioned conversion of any variant name into the canonical domain name of each university and deletion of any *www.*-prefix.

The network analysis software Pajek could treat the adjacency matrix as unweighted when computing shortest paths. This was a useful functionality in Pajek, as the link counts in the matrix otherwise would affect the shortest paths computations, giving priority to link paths containing low link counts.[35]

In the next chapter, the adjacency matrix is used to model the connectivity patterns of the investigated UK academic web space, including the computation of all shortest link paths between pairs of subsites.

[35] Before this 'no-weight' functionality was detected in Pajek, an alternative *binary* matrix had been constructed with no link counts and just 1s and 0s for denoting adjacency or not between all pairs of subsite nodes.

5 Basic graph measures of the UK academic subweb

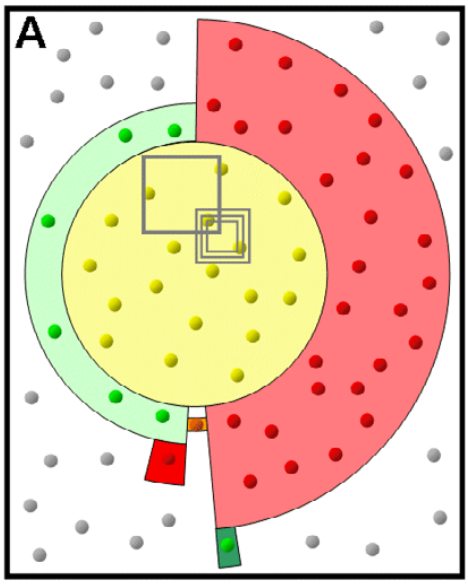


Figure 5.1. Step A in a five-step methodology: identification of graph components in the UK academic sub-web.

The first step in a five-step methodology for sampling, identifying and characterizing possible small-world properties in the UK academic web space links was to establish a graph model as a 'mapping' of the link structures among the 7669 subsites of UK universities in order to enable the selection of a suitable sample for further investigation. The whole five-step methodology is outlined in chapter 6.

As noted in Section 4.2.5, it was not feasible to identify non-valid subsites from the data set prior to the construction of the graph model. The data set of 7669 subsites will thus be used throughout chapters 5 and 6 implying that caution has to be exercised when dealing with the resulting statistics. The present chapter primarily addresses the first two research questions in the dissertation (cf. Section 1.4):

1. *How cohesively interconnected are link structures in an academic web space?*
2. *In particular, to what extent can so-called small-world properties be identified in this web space?*

These research questions are addressed by implementing a range of graph measures. Section 5.1 presents a *'corona'* graph model of link connectivity structures in the UK data set. Indicative 'ages' of the identified graph components are shown in Section 5.2. Small-world properties of the UK data set are investigated in Section 5.3 by measuring characteristic path lengths and clustering coefficients. Distributions of in-neighbors/out-neighbors and inlinks/outlinks are examined in Section 5.4 in order to identify possible power laws in the data set.

5.1 'Corona' graph model

The adjacency matrix of the delimited data set was used to compute a graph model showing components of interconnected subsites based on the so-called *'bow-tie'* model of Broder *et al.* (2000). As outlined earlier in Section 3.5, Broder *et al.* (2000) give a 'bow-tie'-looking model of the graph structure in the Web shown in Fig. 5.2.

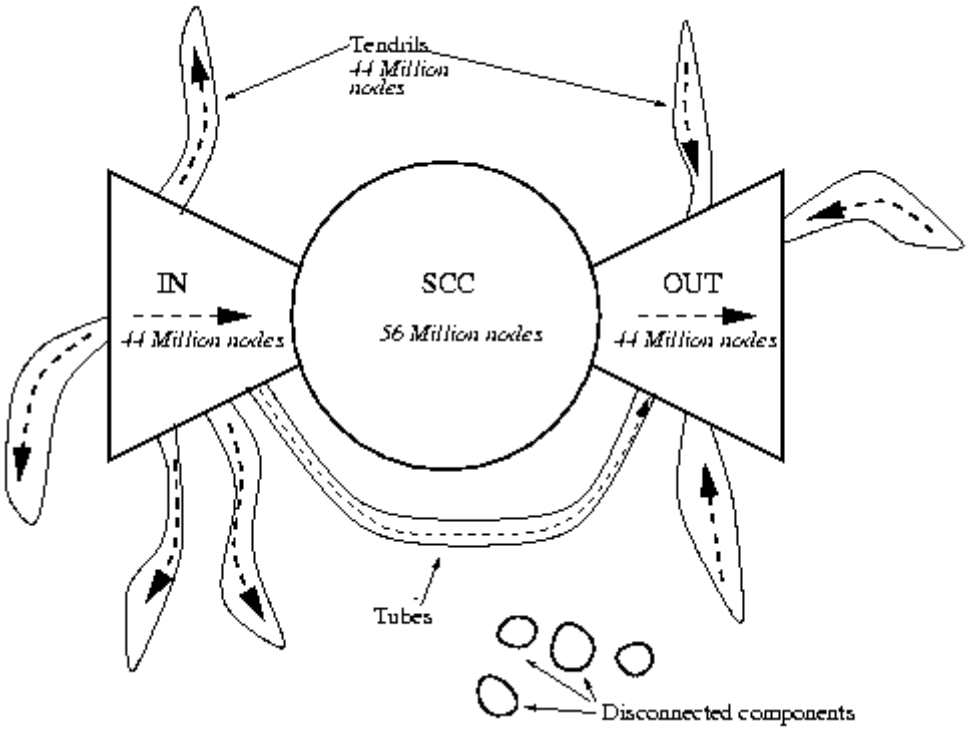


Figure 5.2. The 'bow-tie' model in Broder *et al.* (2000) of the graph structure in the Web.

As described earlier in Section 3.5, the 'bow-tie' core in the model is the so-called *Strongly Connected Component* (SCC) in which any pair of web pages can be connected by directed link paths, cf. Fig. 5.3 below. The *IN* component consists of pages that can reach the SCC through directed link paths but cannot in turn be reached from the SCC. Correspondingly, pages in the *OUT* component can be reached through directed link paths from the SCC but cannot reach back. Pages in the so-called *Tendrils* and *Tube* are connected with the IN and OUT components but cannot reach to the SCC or be reached from the SCC. The remaining *Disconnected* component are not connected in any way with the main 'bow-tie'. According to the 'bow-tie' model by Broder *et al.* (2000), small-world phenomena in the shape of short link paths between web nodes primarily occur within the SCC because all pairs of nodes in this component can reach each other with directed link paths, whereas a web node, for example, in the OUT component cannot reach a node, for example, in the IN component by link paths.

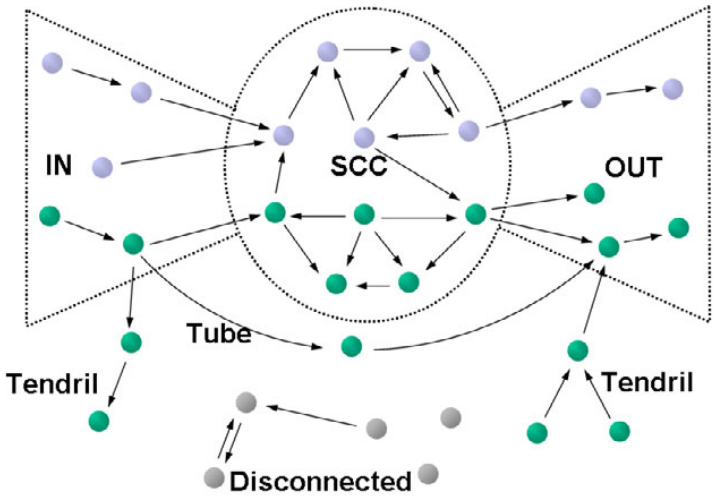


Figure 5.3. The 'bow-tie' model (modified from Broder *et al.* 2000) showing simplified link structures between web nodes in the different graph components in the Web.

The graph components based on the delimited data set of 7669 subsites are presented in Table 5-1 and Fig. 5.4 below. These graph components reflect link structures between web *sites* and not between web *pages* as in the original 'bow-tie' model.

	# subsites	%
IN	626	8.2
SCC	1893	24.7
OUT	2660	34.7
IN-Tendrils	96	1.3
Tube	7	0.1
OUT-Tendrils	55	0.7
Disconnected	2332	30.4
	7669	**100.0**

Table 5-1. Distribution of graph components among 7669 UK university subsites.

It was considered whether to use samples from each graph component to estimate the number of non-valid subsites described in Section 4.2.3. Alas, this validation was not feasible due to different reasons, including the time-consuming task to manually identify the subsites in the Internet Archive. However, as will be more elaborated below, only subsites from the SCC component were selected for further investigation in the dissertation. Subsites in the SCC are all valid because they by definition must have outlinks (*and* inlinks) and thus their domain names derive from URLs of outlinking source pages visited and thus validated by the web crawler. Recapitulating Section 4.2.3; the potentially non-valid subsites stem from typos in the domain names of the URLs of outlinked *target* pages not visited by the web crawler. Besides the 1893 subsites from SCC, all 626 subsites in the IN component (cf. Table 5-1) also were 100% valid by necessity because these subsites per definition must have outlinks and thus had outlinking source pages visited and validated by the web crawler. The potentially non-valid subsites thus belong to the remaining graph components, with the OUT and the

Disconnected components as the two largest ones. The statistics from these remaining components must thus be treated with caution.

The sizes of the IN, SCC, OUT and Disconnected components shown in Table 5-1 were derived from a special computation of the adjacency matrix of the data set, which also provided an aggregated count for the Tendrils and the Tube. This aggregated count was dissolved by a manual process of elimination of what nodes were interlinked with the IN and OUT component.

The same overall component terms are used as in the original 'bow-tie' model. However, in the present study the tendrils are assigned more specific terms depending on whether they are connected with the IN or OUT component. A closer analysis of the components in the data set revealed connectivity patterns within and between the components not evident in the 'bow-tie' model. Especially, the frequent *direct* links from the IN to the OUT component are not clearly illustrated in the 'bow-tie' model where the 'bow-tie wings' do not touch each other.[36] Instead, a so-called '*corona*' model as in Fig. 5.4 below seemed more appropriate to depict actual inter-component adjacencies in the graph than does the 'bow-tie' model – at least in the investigated UK web space.

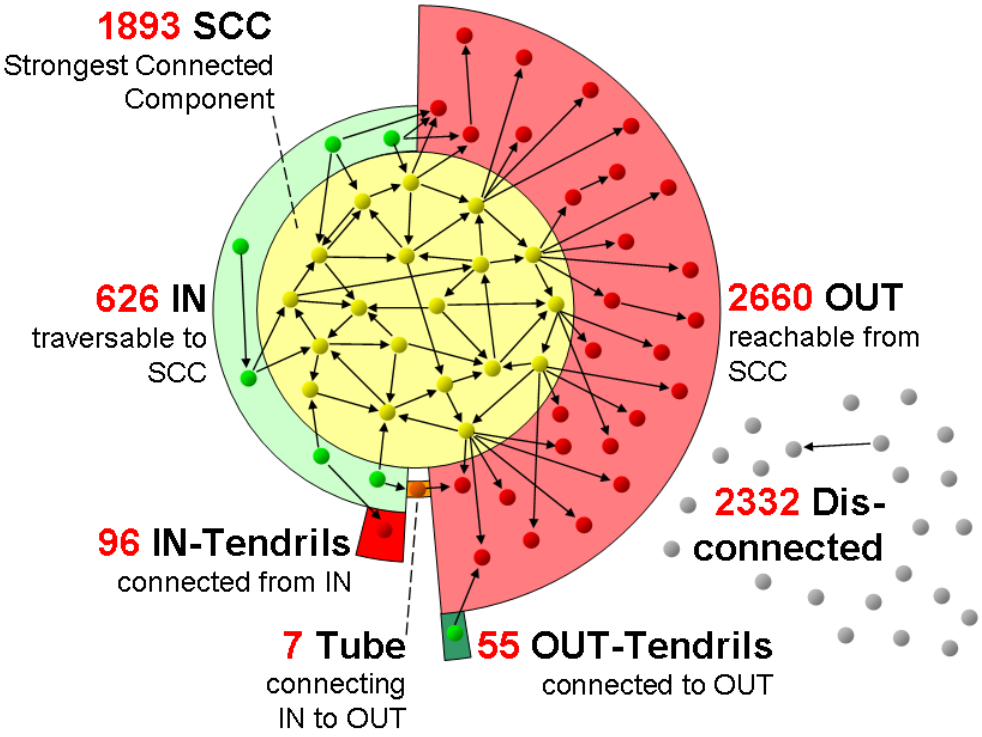


Figure 5.4.* 'Corona' model of graph components among 7669 UK university subsites. The number of nodes and sizes of components in the figure roughly reflect actual numbers and sizes. Green and red graph colors symbolize where link paths may start and stop, respectively. (* cf. color prints placed before the appendices.)

The 'corona' term denotes the figure's resemblance with a solar corona with protuberances. The number of colored nodes and sizes of components in the figure

[36] 457 different nodes in the OUT component received inlinks directly from the IN component.

roughly reflect the actual numbers and sizes. Green and red colors of nodes and components symbolize where link paths may start and stop, respectively (cf. color prints placed before the appendices).

No other web studies have been found with a similar close investigation of micro-level link structures within and between web graph components. However, it is beyond the scope of the dissertation to give detailed descriptions of the different components revealed in the close investigation of the 'corona' model'. A few examples though may give an impression of the intricate link structures.

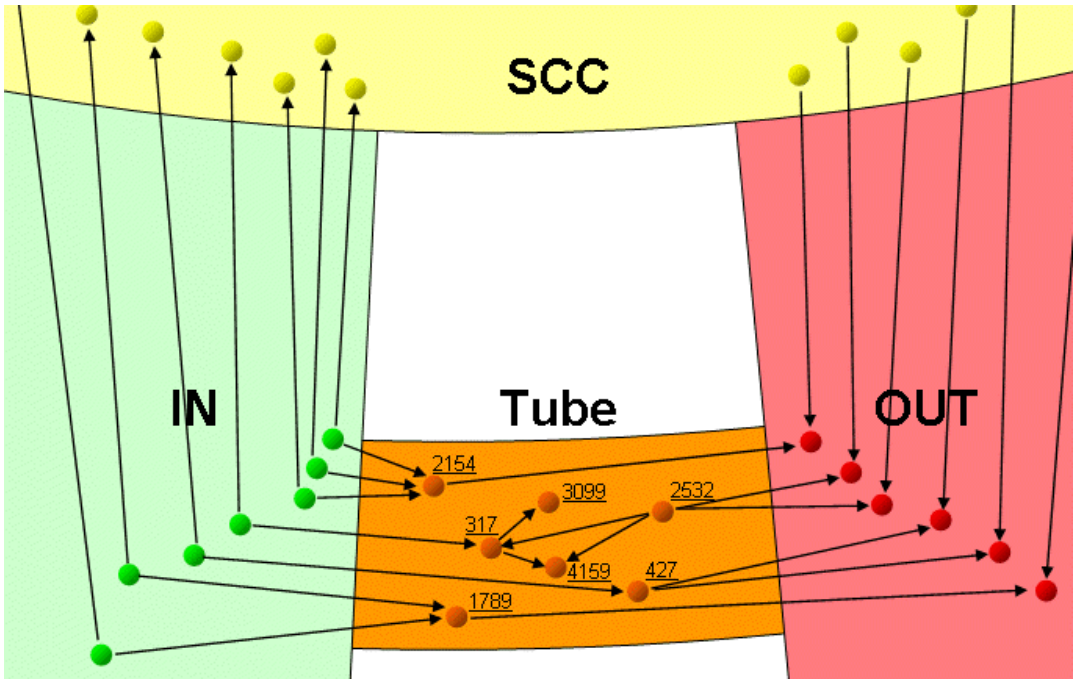


Figure 5.5.* Link structures within the Tube component of the 'corona' graph model of the UK academic subweb 2001. The seven Tube nodes have id numbers assigned the 7669 subsite nodes.

An interesting observation is that a Tube-node not necessarily functions as an intermediary node on a directed link path between the IN and OUT components. For instance, node 317 (id number among the 7669 subsites) in Fig. 5.5 above is connected to node 2532. However, there is no directed link path leading from 317 to 2532, nor via node 4159. A correct definition of the Tube would thus be that it contains nodes directly or indirectly *connected* with both the IN and OUT components but not necessarily *traversable* from IN to OUT.

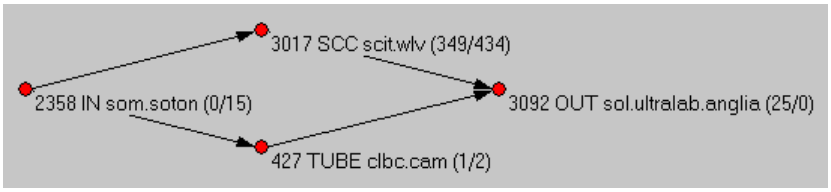


Figure 5.6. All shortest link paths between IN-node 2358 and OUT-node 3092.

Fig. 5.6 above shows a close-up of a subgraph with all shortest link paths between an IN-node and an OUT-node showing another interesting detail: how an IN-node may be

connected to an OUT-node through both a Tube-node and a SCC-node. The brackets in the figure show the number of in-neighbors/out-neighbors of the subsites. For example, IN-node 2358 has no in-neighbors but 15 different out-neighbors in the data set including nodes 3017 and 427.[37]

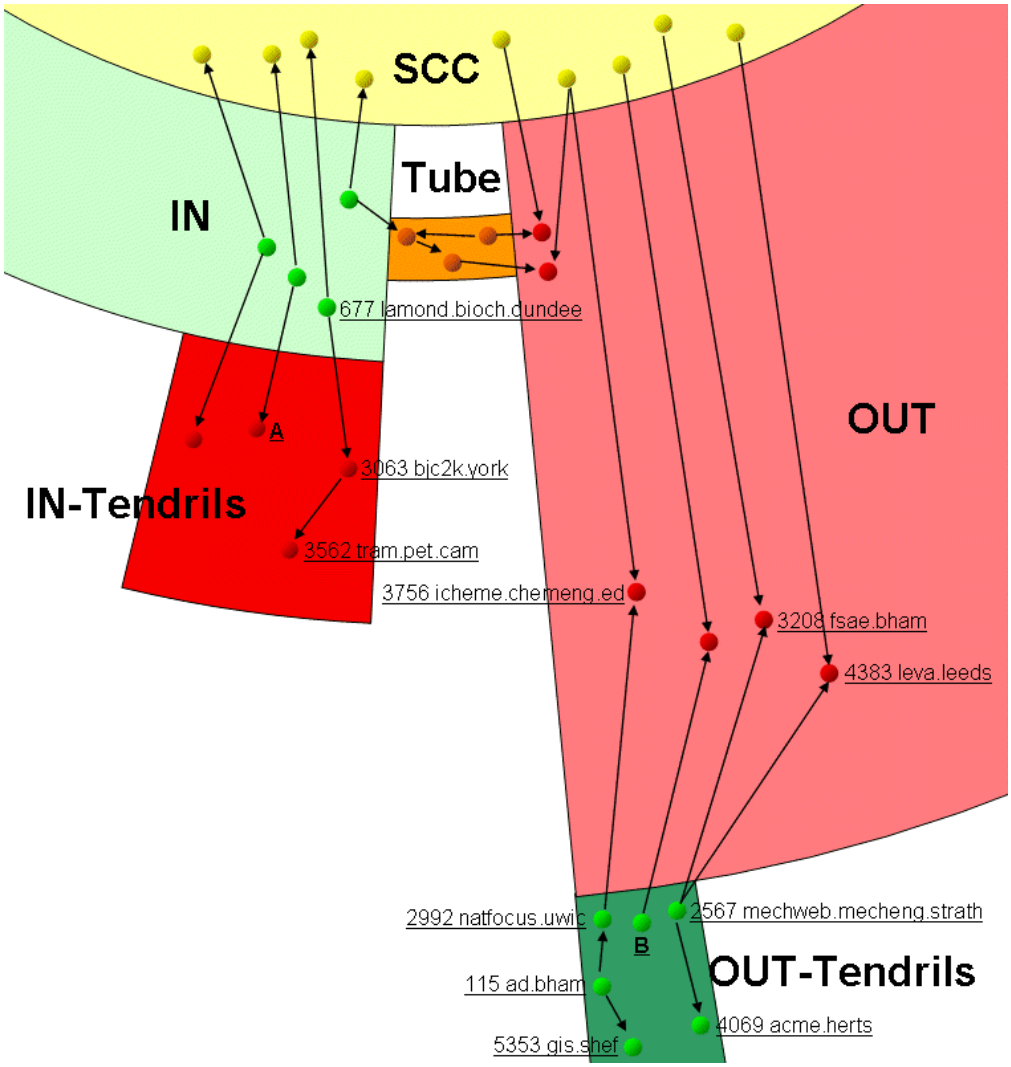


Figure 5.7.* The actual link structures of nodes in the IN-Tendrils and OUT-Tendrils with intra-component links. Nodes A and B represent the majority of Tendril nodes with no intra-component links. (* cf. color prints placed before the appendices).

Fig. 5.7 above illustrates a final detail, that is, an OUT-Tendril node does not necessarily – as perhaps usually conceived – have to reach the OUT component. For example, nodes 4069 and 5353 are connected with other OUT-Tendril nodes but cannot themselves reach the OUT component.

[37] Node 2358 is the School of Medicine, Univ. of Southampton; node 3017, School of Computing and Information Technology, Univ. of Wolverhampton; node 427, Department of Clinical Biochemistry, Cambridge; and node 3092, Ultralab, a learning technology research centre at the Anglia Polytechnic University.

The 'corona' model of the UK academic subweb supports the notion of a fractal 'self-similar' Web (Dill *et al.*, 2001; Kumar *et al.*, 2002) with subsets of the Web displaying the same graph properties as the Web at large, including 'bow-tie'-like structures and power-law like distributions.

As stated in Section 4.2, the delimitations of the data set in the present study were deliberately imposed in order to enable more clear-cut sampling and identification of transversal link structures between university subsites. However, the exclusion of site selflinks between subsites at the same university, as well as links to and from the 109 university main sites from the adjacency matrix as described in Section 4.2, naturally influenced the resulting graph model of link structures in the UK academic web space.

First, the amount of disconnected nodes increased, because many subsites only had site selflinks or links to or from university main sites. The 662 subsites in Table 4-4 (the S_2 and T_2 categories) in Section 4.2.3.1, comprising subsites with source or target pages connected with the detached universities main sites but not connected with subsites at other universities, all belong to the Disconnected component in the present delimited graph model. More subsites would thus have been connected by directed link paths in an undelimited graph model that included all the otherwise excluded links. This assumption is supported by a specially computed adjacency matrix that allowed links to and from the 109 university main sites (it was not feasible to include site selflinks as well). In this special case, the strongly connected component (SCC) consisted of 2354 subsites instead of 1893 subsites in the graph model based on the delimited data set.

Second, distances on link paths between the subsites would have been shorter in an undelimited graph model because site selflinks and links to and from university main sites would provide shortcuts. Once more, this assumption is supported by an analysis of the special adjacency matrix above including links to and from the university main sites (again, it was not feasible to include site selflinks), yielding an average link path length of 2.93 between connectable nodes in the resulting graph instead of 3.46 in the present model of the UK *subsite* web graph and a so-called *diameter* (the length of the longest of the shortest link paths) of 7 instead of 10 in the present graph.

The measures of shortest path lengths and other small-world properties of the present graph model are further outlined in Section 5.3 below.

5.2 Indicative ages of graph components

The specially developed programming script that validated 6868 of the subsites in the Internet Archive (cf. Section 4.2.4) also was used to identify when they were indexed the first time in the Archive. Fig. 5.8 below shows an example screenshot from the Archive's *Wayback Machine* (*www.archive.org*) displaying that the SCC subsite *bssv01.lancs.ac.uk* (Institute of Environmental and Biological Sciences, Lancaster University) in the UK data set was indexed in the Archive the first time on Feb 22, 1996 (the earliest indexing date of all the subsites as noted in Table 5-2 below). Naturally, a subsite may very well have existed years prior to when it is harvested by Internet Archive. For example, as shown in Table 5-2, some of the subsites were not indexed by the Archive before 2002, even though they clearly existed in 2001 when harvested by

Thelwall's web crawler. The date of first indexing in the Internet Archive is thus *no proof of the real age* of a subsite and should be interpreted in an indicative and cautionary way.

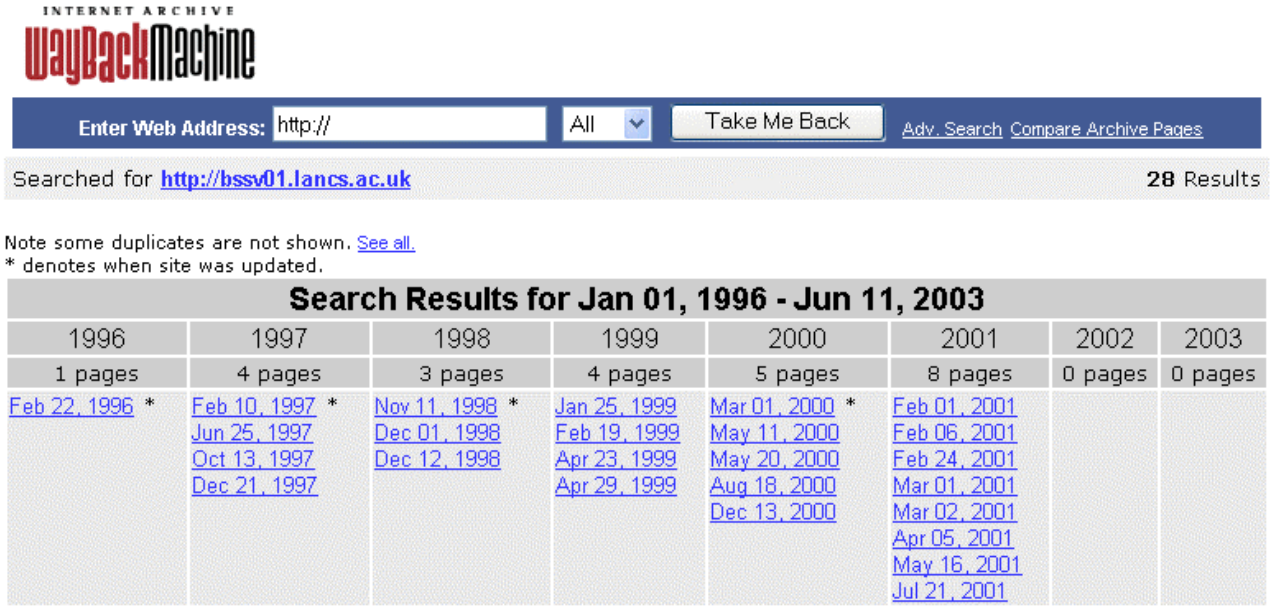


Figure 5.8. Example screenshot from the Internet Archive (*www.archive.org*). Homepage of *bssv01.lancs.ac.uk* first indexed in the Internet Archive on Feb 22, 1996.

Using the special date facility of Excel, all the first indexing dates extracted from Internet Archive were converted to a special number, for example, 'May 08, 1997' to '35558' (the number of days since Jan 01, 1900) in order to compute indicative average ages of subsites in the different graph components. In Table 5-2, the average first time indexing dates of the 6868 subsites are distributed on the graph components. No other Web studies have been found using this special approach of exploiting data from the Internet Archive.

	# subsites	# subsites identified in IA	% identified subsites	average first time indexed in IA	earliest first time indexed in IA	latest first time indexed in IA
IN	626	606	96.8	11.06.2000	30.10.1996	28.03.2002
SCC	1893	1874	99.0	20.08.1998	22.02.1996	02.06.2002
OUT	2660	2222	83.5	06.07.1998	17.10.1996	06.06.2002
IN-Tendrils	96	76	79.2	19.07.1999	30.10.1996	25.06.2001
Tube	7	7	100.0	28.10.1999	16.04.1997	19.02.2001
OUT-Tendrils	55	53	96.4	01.04.2000	25.12.1996	04.06.2002
Disconnected	2332	2030	87.0	16.05.2000	11.05.1996	09.06.2002
	7669	**6868**	**89.6**	**17.04.1999**		

Table 5-2. Average first time indexing in the Internet Archive (IA) of 6868 subsites.

As shown in Table 5-2, only a very small percentage of subsites in the IN and SCC components were not found in the Archive. Therefore, even if the Internet Archive does not cover the entire Web, a remarkably high percentage of UK academic subsites were actually indexed in the Archive. If at least 53 typos (cf. Section 4.2.3.2) are excluded from the 7669 investigated subsites, over 90% of the remaining subsites were verified in the Archive – and 99% of the SCC subsites. Some of the subsites not found in the Archive were due to the robots exclusion (cf. Section 4.2.4) where site owners ban access of web crawlers. However, the large percentage of non-identified subsites in the OUT, In-Tendrils and Disconnected components may partly be caused by *non-existing* subsites with erroneous domain names due to typos.

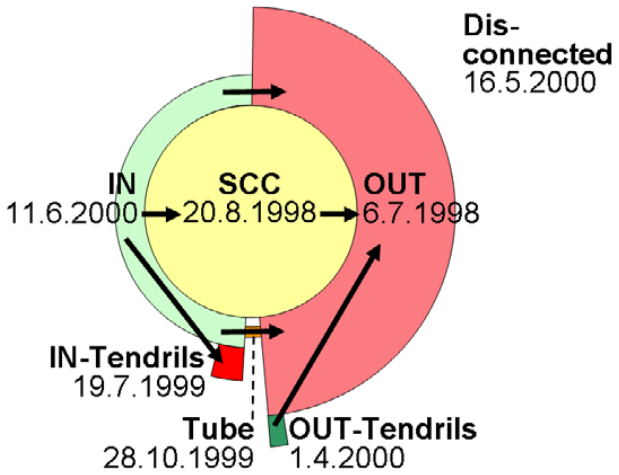


Figure 5.9.* Indicative ages of graph components based on average first time indexing in the Internet Archive of 6868 subsites (cf. Table 5-2).

Fig. 5.9 above presents indicative 'ages' of the investigated graph components in the UK academic subweb as indicated by the average first time indexing in the Internet Archive of the verified 6868 subsites in the data set (cf. Table 5-2). Looking at the direction of inter-component links (following the component definitions) in the figure, it

is apparent that 'younger' components – without exception – only link to 'older' ones. No older components link to younger ones. This makes sense, as new subsites will make links to already existing and 'popular' ones – the so-called *preferential attachment* (Barabási & Albert, 1999) or *Matthew effect: "unto every that hath shall be given"* (Merton, 1968) as described in Section 3.5. Moreover, it takes time before new subsites are visible enough to receive inlinks from older ones. The OUT component contained the oldest subsites, the IN component the youngest, and the SCC subsites were on average slightly younger than the OUT subsites.

These indicative component 'ages' apply to the investigated UK subsites. It remains to be tested whether other web spaces show similar 'ageing' patterns.

5.3 Small-world properties of the UK academic subweb

The basic structure of a graph, in this case, the UK academic subweb space, may be characterized by four graph theoretic measures (Steyvers & Tenenbaum, 2001):
(1) the characteristic path length L, (2) the diameter D (the longest of the shortest paths), (3) the clustering coefficient C, and (4) the distribution of so-called in-degrees $P(k_i)$ and out-degrees $P(k_o)$, that is, the distribution of in-neighbors/out-neighbors and inlinks/outlinks attached to the nodes. In the following sections, these basic graph measures for the UK academic subweb are calculated. In particular, it is investigated whether this subweb has small-world properties. A small-world graph contains the following properties (Watts & Strogatz, 1998), cf. Section 3.3:

- The *clustering coefficient C* is much larger than that of a random graph with the same number of nodes and average number of edges (node level links to adjacent neighbors) per node.
- The *characteristic path length L* is almost as small as L for the corresponding random graph.

5.3.1 Characteristic path length

The two graph-theoretic measures of the characteristic path length distance L and the diameter D are closely associated: L denotes the average of the shortest link path lengths between all pairs of nodes, while D refers to the longest of these shortest paths. In other words, at most D steps are required to move between any two nodes, but on average only L steps are necessary.

Fig. 5.10 below shows two variants of a simplified network of web sites covering three different topical web clusters S, T and U. In Fig. 5.10a, the shortest link path along directed links between the two network nodes S_1 and U_5 has path length 8. In Fig. 5.10b, the shortest link path along directed links between the same two network nodes S_1 and U_5 only has path length 4, because a transversal link S_5-U_4 has added a direct shortcut between topics S and U.

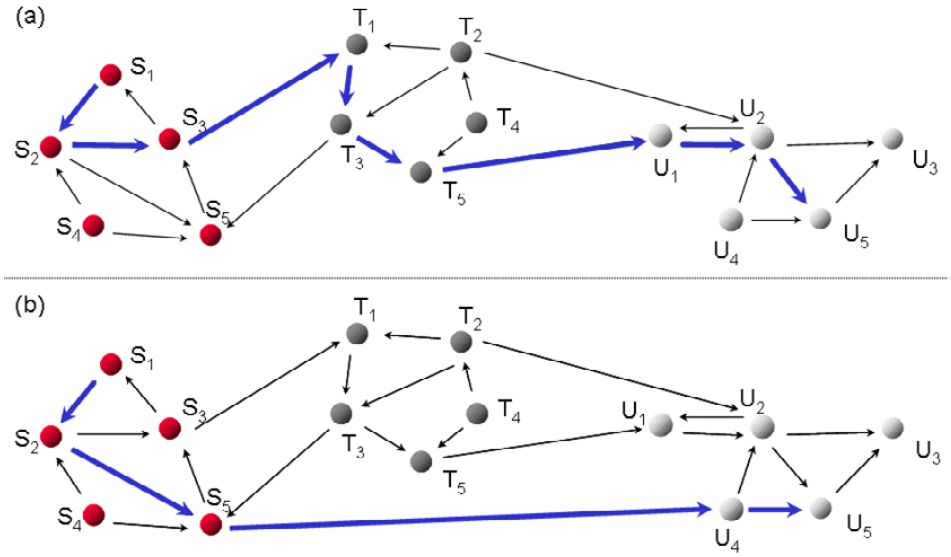


Figure 5.10a&b. The shortest link paths (bold links) between two network nodes S_1 and U_5 before (a) and after (b) the addition of a transversal link S_5-U_4.

Table 5-3 and Fig. 5.11 shows the total distribution of the length of shortest link paths between pairs of subsites in the delimited data set. A special software program extracted 11,469,432 interconnected pairs of subsites in the data set. This means that only 19.5% of all 58,805,892 (7669×7668) pairs of subsites could be connected by a directed link path. The low percentage is affected by the high share (65.1%) of subsites belonging to the OUT and Disconnected component – cf. Table 5-1. In Tables 5-4 and 5-5 further below, the distribution of shortest paths between and within the different graph components is listed.

path length	# subsite pairs	%	length × count	average path length
1	48,902	0.43	48,902	
2	1,205,077	10.51	2,410,154	
3	4,995,679	43.56	14,987,037	
4	4,041,513	35.24	16,166,052	
5	1,016,426	8.86	5,082,130	
6	143,914	1.25	863,484	
7	16,446	0.14	115,122	
8	1,389	0.01	11,112	
9	83	0.00	747	
10	3	0.00	30	
	11,469,432	**100.00**	**39,684,770**	**3.46**

Table 5-3. Distribution of lengths of existing shortest paths between pairs of subsites.

According to Table 5-3 above, 48,902 pairs of subsites were directly linked to each other, that is, they were neighbors in the graph (subsites with reciprocal links count as 2 pairs), whereas 1,205,077 subsite pairs were connected with a link path of length 2 (cf. the link path between S_1 and S_3 in Fig. 5.10 above). As shown in Table 5-3, the average

path length, that is, the *characteristic path length* in the UK academic subweb was 3.46 among the subsites that could reach each other with directed link paths. The *diameter* of the UK academic subweb was 10, the longest of the existing shortest paths between subsites.

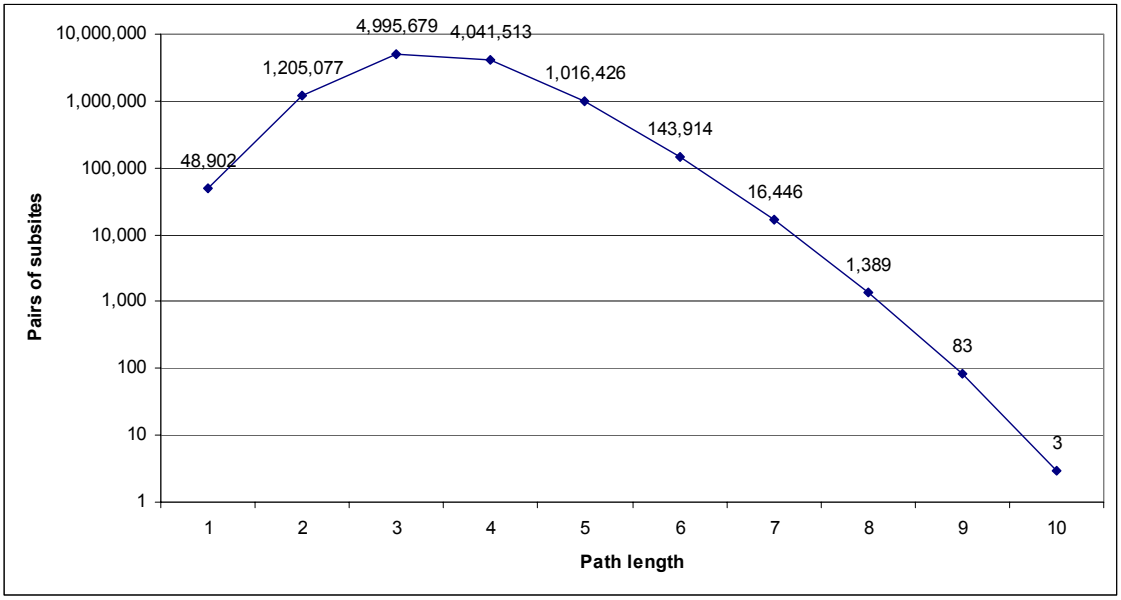


Figure 5.11. Distribution of lengths of existing shortest paths between pairs of subsites. Semi-log scale.

The semi-log scale plot of the distribution curve in Fig. 5.11 is similar to other distributions of shortest link paths, for example, the distribution of lengths of shortest paths via co-actors between pairs of film actors in a small-world graph of over 350.000 film actors (Hayes, 2000a; 2000b).

An example of a pair of subsites connected by shortest link paths of length 10 – one of the three subsite pairs with this longest path length in the data set noted in Table 5-3 and Fig. 5.11 above – is illustrated in Fig. 5.12. The figure shows all the shortest link paths of length 10 starting at the Hitachi Cambridge Laboratory at the Department of Physics in Cambridge (*www-hcl.phy.cam.ac.uk*) and ending at the Aston Centre for Asian Business and Management, University of Aston (*asian-mgt.abs.aston.ac.uk*). The network analysis software Pajek was used to extract all the shortest paths. The figure was manually created using drawing functions in Pajek.

The end node at Aston belonged to the OUT component, whereas all the remaining subsite nodes in the figure were located in the SCC component. (See Appendix 6 for the affiliations of the nodes in the figure). However, case studies of 10 similar so-called 'path nets' are investigated in later sections with regard to topics and genres both on the subsite levels and the page levels of shortest link paths.[38]

[38] As explained in section 6.3, the term 'path net' is used in the dissertation for a subgraph containing all shortest link paths between a single pair of nodes.

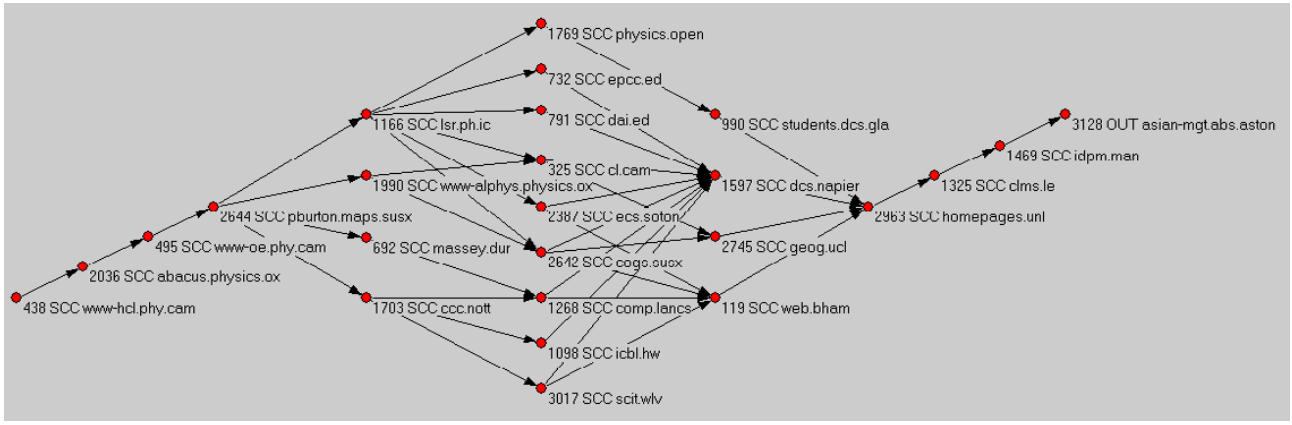


Figure 5.12. All shortest link paths (path length 10) between node 438 (*www-hcl.phy.cam.ac.uk*) and node 3128 (*asian-mgt.abs.aston.ac.uk*). (See Appendix 6 for affiliations of the nodes on the link paths).

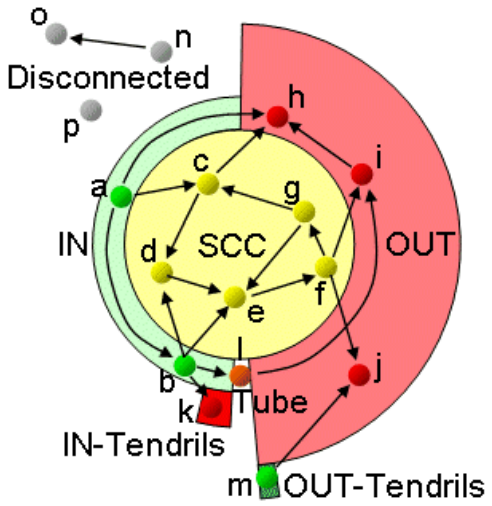


Figure 5.13. Simplified example of some possible link paths within and between graph components.

Fig. 5.13 shows a simplified example of link structures illustrating some of the link paths possible between the subsites in the different graph components in the 'corona' model. For example, IN-subsite *a* is connected to OUT-subsite *j* by a shortest link path of length 4 (*a-b-e-f-j*), and the SCC-subsite *g* has a shortest path of length 2 to another SCC-subsite *f* (path *g-e-f*).

In Tables 5-4 and 5-5, the distribution of link paths between and within the different graph components is listed. For example, there is a permutation of 1,185,018 subsite pairs connected by link paths starting from the 626 IN-subsites leading to the 1893 SCC-subsites. Almost a third (31.2%), 3,581,556, of the link paths are between SCC subsites. Furthermore, at least 85.5% of the existing link paths pass SCC subsites (IN→SCC; SCC→SCC; SCC→OUT). This percentage is probably higher than 95% when including link paths IN → OUT. However, it has not been possible to compute a specified count of how many link paths go directly from IN to OUT or pass through the SCC or the Tube. Probably, the very large majority pass the SCC because of the high connectivity IN→SCC and SCC→OUT. The low number of link paths within the IN

and OUT components as indicated in Table 5-5 further below implies that a link path starting in IN and ending in OUT typically will consist of only one subsite in IN and one in OUT and all the remaining subsites on the link path will belong to the SCC.

It should be emphasized that the counts in Table 5-4 denote the number of subsite pairs being connected by link paths. As shown in Fig. 5.13 above and further demonstrated in Section 6.3, there may exist several *different* shortest link paths of the *same* path length between a pair of subsites. For example, IN-subsite a and SCC-subsite d in Fig. 5.13 can be connected by two different shortest link paths of length 2: *a-c-d* and *a-b-d*. In Tables 5-4 and 5-5 some other example subsite pairs from Fig. 5.13 have been included.

link paths	**# subsite pairs**		**%**	**examples (Fig. 5.13)**
IN → SCC	626 × 1893	1,185,018	10.33	a-b-e
IN → (direct or via SCC or Tube) → OUT	626 × 2660	1,665,160	14.52	a-h; a-b-l-i
SCC → SCC	1893 × 1892	3,581,556	31.23	c-d-e-f-g
SCC → OUT	1893 × 2660	5,035,380	43.90	g-e-f-i
residue		*2,318*	*0.02*	
		11,469,432	**100.00**	

Table 5-4. Distribution of all subsite pairs connected by directed link paths within and between different graph components.

The count of 2,318 residual subsite pairs in Table 5-4 above was calculated by subtracting the cumulated known counts of the other pairs from the known total count of pairs (11,469,432: cf. Table 5-3 further above). Table 5-5 below lists the combinations of components providing the residual subsite pairs. It has not been feasible to compute the number of paths in these components. Many subsites are not connected *within* these components, for example within the IN and OUT components but only connected to subsites in other components. This explains the small total count of the residual subsite pairs.

link paths	**# subsite pairs**	**examples in Fig. 5.13**
IN → IN	n/a	a-b
IN → IN-Tendril	n/a	b-k
IN → Tube	n/a	b-l
OUT → OUT	n/a	i-h
OUT-Tendril → OUT	n/a	m-j
Tube → OUT	n/a	l-i
Tube → Tube	n/a	-
IN-Tendril → IN-Tendril	n/a	-
OUT-Tendril → OUT-Tendril	n/a	-
Disconnected → Disconnected	10	n-o
	2,318	

Table 5-5. Distribution of 2,328 subsite pairs connected by directed link paths within and between different graph components.

As noted earlier, one of the criteria for possible small-world properties of the UK academic subsite web was that the characteristic path length should be almost as small

as for the corresponding *random* graph. In order to investigate this matter, the network analysis program Pajek was used to construct a random graph containing 7669 nodes and 48,902 edges (node level links) just as the UK academic subsite web. A special program then extracted all shortest link paths between all pairs of nodes as in the UK academic subsite graph. The characteristic path length of the random graph was 5.04 compared with 3.46 for the UK graph. In other words, the characteristic path length of the UK graph is not only "almost as small", but actually smaller than the corresponding random graph. This circumstance may be due to the lower percentage (19.5% or 11,469,432) of UK subsites connected by link paths than the percentage (99.7% or 58,606,678) of the random graph nodes (Table 5-6 below) among all possible 58,805,892 (7669×7668) pairs of subsites.

The diameter of the random graph was 9 (the longest of the shortest path lengths) close to the diameter of 10 for the UK academic subsite graph.

path length	# subsite pairs	length × count	average path length
1	48,902	48,902	
2	310,886	621,772	
3	1,930,713	5,792,139	
4	10,420,901	41,683,604	
5	29,106,432	145,532,160	
6	15,639,350	93,836,100	
7	1,122,828	7,859,796	
8	26,440	211,520	
9	226	2,034	
	58,606,678	**295,588,027**	**5.04**

Table 5-6. Distribution of lengths of shortest paths between pairs of nodes in a *random* graph with 7669 nodes and 48,902 edges.

Broder *et al.* (2000) concluded that the *whole* bow-tie graph model of web link structures did not have small-world properties, only the strongly connected component (SCC) was a small-world web. In the Broder study, only about 24% of all nodes could be connected by directed link paths. Correspondingly, the UK academic subsite web of 2001 only had true small-world properties within the SCC, because just 19.5% of all pairs of nodes in the investigated UK graph could be interlinked by link paths.

In a similar study of a national academic web space (cf. Section 3.5), Adamic (1999) investigated the strongly connected component (SCC) of 3,400 web sites in the *.edu* top level domain and found a characteristic path length of about 4.1 among these sites and a diameter of 13. It has not been feasible in the present study to compute the characteristic path length of the SCC comprising 1,893 sites in the investigated UK web space.

5.3.2 Clustering coefficient

A measure introduced by Watts & Strogatz (1998) for understanding the structural properties of small-world graphs is the so-called *clustering coefficient*, *C*, that depict the average interconnectedness or 'cliquishness' of all local so-called *neighborhoods* in the graph.

In graph theory, a neighborhood of a network node *v* is the subgraph consisting of *v* and all nodes directly connected with *v*, and all edges (i.e., undirected links) connecting these nodes. Two network nodes that are directly connected by an edge are said to be *neighbors*. In order to compute the clustering coefficient, links in the graph are treated as undirected edges. If a network node *v* has k_v neighbors (not including node *v*), then the maximum number of possible undirected edges in the neighborhood is $k_v(k_v - 1)/2$. If T_v denotes the number of connections between the neighbors of node *v*, the *local clustering coefficient*, C_v, then may be calculated by the equation (Steyvers & Tenenbaum, 2001):

$$C_v = T_v \Big/ \binom{k_v}{2} = 2T_v / k_v(k_v - 1)$$

The measure may vary between 0 (disconnected node with no neighbors) and 1 (all neighbors are interlinked with each other). The clustering coefficient reflects the probability that nodes connected with a node *v* also are connected with each other.

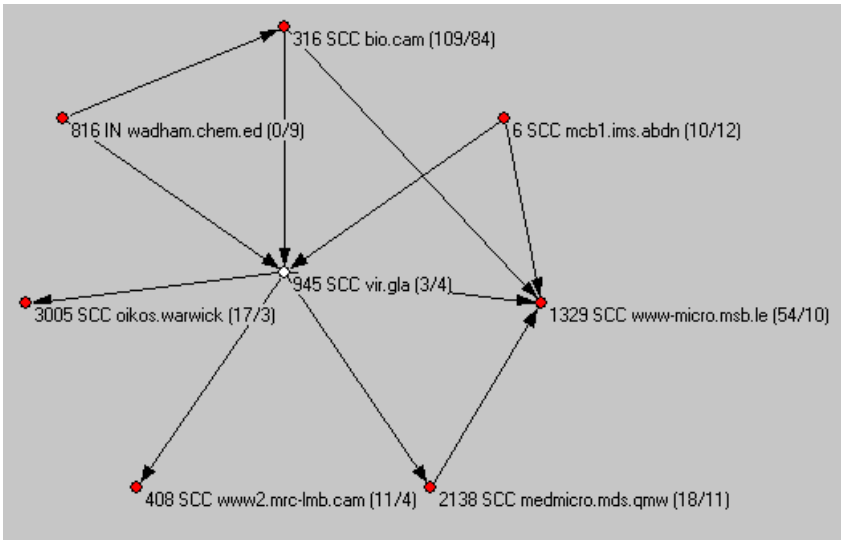


Figure 5.14. Neighborhood of SCC node 945 comprising all in-neighbors (e.g., node 816) and out-neighbors (e.g. node 2138). Brackets show total number of in-neighbors/out-neighbors in data set. Cf. footnote next page for affiliations.

Fig. 5.14 shows an example of a neighborhood in the present data set. The neighborhood of SCC node 945 (*vir.gla.ac.uk* : Division of Virology, University of Glasgow) from the data set consists of 7 subsites nodes (3 in-neighbors and 4 out-neighbors as denoted in the brackets in the figure). One in-neighbor, node 816, is from

the IN component. The remaining neighbors all belong to the SCC. All neighbor nodes were related to microbiology and biological sciences[39].

At most 21 (=7×6/2) edges are possible in the neighborhood of node 945. Since there are only 4 edges among the neighbors (not counting edges connected with node 945), for instance, between nodes 2138 and 1329, the local clustering coefficient of node 945 is 0.1904761 (=4/21).

The neighborhood of node 945 in the figure may be divided into *triadic* structures, using a social network analytic term (Wasserman & Faust, 1994; Scott, 2000 – cf. Section 2.3.1). For instance, node 945 has links to both node 1329 and 2138, that in turn also are interlinked with each other. Such *triadic closure* (Skvoretz & Fararo, 1989) between sets of three nodes that all are interlinked with each other is thus essential with regard to the size of the clustering coefficient of a neighborhood subgraph.

The clustering coefficient, C, of a whole graph is the average of all local clustering coefficients C_v over all nodes v in the graph. The network analysis software Pajek was used to compute all the local clustering coefficients for the UK academic subsite graph with 7669 nodes giving an average $C_{ac.uk}$ of 0.09038. In other words, if a subsite node v_1 was connected with the two subsites nodes v_2 and v_3, there was 9.0% probability that v_2 and v_3 were also connected. The corresponding clustering coefficient C_{SCC} for the SCC alone was 0.1299 reflecting the more interlinked link patterns in that graph component.

The clustering coefficient for the corresponding random graph with 7669 nodes and 48,902 edges, C_{random} was 0.00084. The clustering coefficient $C_{ac.uk}$ for the UK academic subsite graph thus is much larger (over 100 times larger) than that of a random graph with the same number of nodes and edges.

The short characteristic path length of the UK web graph identified in Section 5.3.1 and the large clustering coefficient identified above thus meet the requirements for a small-world network introduced by Watts & Strogatz (1998) and listed at the start of Section 5.3. This finding hence answers the second research question concerned with whether small-world properties could be identified in the UK academic web space.

[39] Affiliations clockwise: node 316 (*bio.cam.ac.uk*) School of the Biological Sciences, Cambridge;
node 6 (*mcb1.ims.abdn.ac.uk*) Department of Molecular and Cell Biology, Aberdeen;
node 1329 (*www-micro.msb.le.ac.uk*) Department of Microbiology & Immunology, Leicester;
node 2138 (*medmicro.mds.qmw.ac.uk*) Department of Microbiology, Queen Mary University of London;
node 408 (www2.mrc-lmb.cam.ac.uk) MRC Laboratory of Molecular Biology, Cambridge;
node 3005 (*oikos.warwick.ac.uk*) Department of Biological Sciences, Warwick;
and node 816 (*wadham.chem.ed.ac.uk*) Edinburgh Biomolecular NMR Unit.

5.4 Distribution of links and neighbor nodes

As the link structures of the UK academic subweb are in focus in the dissertation, some more detailed statistics on this matter are outlined below.

As indicated earlier, not all 7669 subsites had inlinks or outlinks. Table 5-7 below shows that 2965 (38.7%) subsites had no inlinks and 4998 (65.2%) no outlinks. These counts reflect the links to and from university main sites as well as site self links excluded in the delimited data set as described in Section 4.2.2. Furthermore, outlinks from mirror sites, e-journals, hosted web sites of external organizations were also excluded by the web crawler (cf. Section 4.1.2). In the present data set it is especially Disconnected subsites that neither have inlinks nor outlinks, as well as the large share of OUT subsites with no outlinks as listed in Table 5-7. Furthermore, it can be derived from Table 5-8 that 2018 subsites (including the 1893 SCC subsites) have both in- and outlinks; 2686 have only inlinks; 653 only outlinks; and 2312 no links at all (only in Disconnected).

	sub-sites	%	with inlinks	%	no inlinks	%	with outlinks	%	no outlinks	%
IN	626	8.2	36	0.8	590	19.9	626	23.4	imposs.	-
SCC	1893	24.7	1893	40.2	imposs.	-	1893	70.9	imposs.	-
OUT	2660	34.7	2660	56.5	imposs.	-	83	3.1	2577	51.6
IN-Tendrils	96	1.3	96	2.0	0	0.0	1	0.04	95	1.9
Tube	7	0.09	6	0.1	1	0.03	5	0.2	2	0.04
OUT-Tendrils	55	0.7	3	0.06	52	1.8	53	2.0	2	0.04
Disconnected	2332	30.4	10	0.2	2322	78.3	10	0.4	2322	46.5
	7669	100.0	**4704**	100.0	**2965**	100.0	**2671**	100.0	**4998**	100.0

Table 5-7. Number of subsites in the different graph components with outlinks and inlinks.

subsites with	outlinks	%	no outlinks	%		%
inlinks	2018	26.3	2686	35.0	4704	61.3
no inlinks	653	8.5	2312	30.1	2965	38.7
	2671	34.8	4998	65.2	**7669**	100.0

Table 5-8. Number of subsites with outlinks and inlinks.

The 7669 subsites were interlinked by 48,902 subsite level links and 207,865 page level links using the link terminology presented in Section 2.3.3. The 48,902 subsite level links represent the total number of interlinked pairs of subsites in the adjacency matrix. If subsite A is adjacent with subsites B and C as in Fig. 5.15, it means that A has a subsite level link to each of the two out-neighbors. In Fig. 5.16, the two subsite level links are dissolved into six page level links.

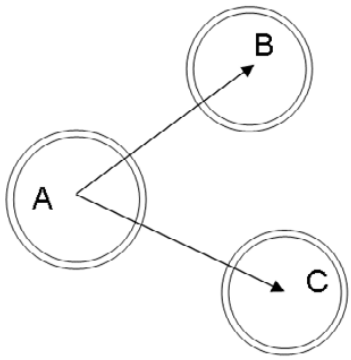


Figure 5.15. Subsite level links.

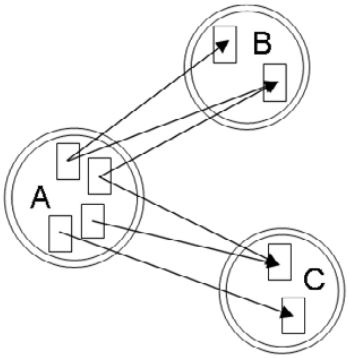


Figure 5.16. Page level links.

Table 5-9 below shows the distribution of in-neighbors and out-neighbors on the graph components in the 'corona' model. For example, the 626 subsites in the IN component only received inlinks from 43 in-neighbors. This is not necessarily *unique* in-neighbors, since there may be redundancy, as illustrated in Fig. 5.15 where subsite A would be counted twice as an in-neighbor from the viewpoint of subsites B and C. The few in-neighbors to subsites in the IN component were all located in this component. However, the subsites in the IN component had 3953 out-neighbors as shown in Table 5-9 below. This count can be specified in Table 5-10 as 2910 out-neighbors in SCC, 895 in OUT, 98 in IN-Tendrils, 43 in IN (i.e. equals the count of in-neighbors), and 7 in the Tube.

	# sub-sites	%	# in-neigh-bors	%	average in-neighbors / subsite	# out-neigh-bors	%	average out-neighbors / subsite
IN	626	8.2	43	0.1	0.07	3,953	8.1	6.3
SCC	1893	24.7	34,315	70.2	18.1	44,754	91.5	23.6
OUT	2660	34.7	14,421	29.5	5.4	111	0.2	0.04
IN-Tendrils	96	1.3	99	0.2	1.0	1	0.002	0.01
Tube	7	0.1	11	0.02	1.6	10	0.02	1.4
OUT-Tendrils	55	0.7	3	0.01	0.05	63	0.1	1.1
Disconnected	2332	30.4	10	0.02	0.004	10	0.02	0.004
	7669	**100.0**	**48,902**	**100.0**	**6.4**	**48,902**	**100.0**	**6.4**

Table 5-9. Distribution of in-neighbors and out-neighbors among the 7669 subsites.

The SCC component with 1893 subsites comprised 24.7% of all the subsites, yet had 70.2% of all in-neighbors and 91.5% of all out-neighbors in the graph as shown in Table 5-9 above. According to Table 5-10 below, 31,405 of the 34,315 in-neighbors were located in SCC and 2910 in the IN component. Of the 44,754 out-neighbors, 31,405 were in SCC and 13,349 in the OUT component. According to Table 5-9 above, a SCC subsite received inlinks from on average 18.1 other subsites and provided outlinks to on average 23.6 other subsites. These high degrees of connectivity in the SCC by far exceed the in-degrees and out-degrees of subsites in the other components. This kind of detailed investigation of connectivity patterns in a web graph structure as presented here has not been found in other Web studies.

intra- & inter-component connectivity		# subsite level links	%
SCC →	SCC	31,405	64.22
SCC →	OUT	13,349	27.30
IN →	SCC	2,910	5.95
IN →	OUT	895	1.83
OUT →	OUT	111	0.23
IN →	IN-Tendrils	98	0.20
OUT-Tendril →	OUT	60	0.12
IN →	IN	43	0.09
Disconnected →	Disconnected	10	0.02
IN →	Tube	7	0.01
Tube →	OUT	6	0.01
Tube →	Tube	4	0.01
OUT-Tendrils →	OUT-Tendrils	3	0.01
IN-Tendrils →	IN-Tendrils	1	0.00
		48,902	**100.00**

Table 5-10. Intra- & inter-component connectivity of 48,902 subsite level links.

Zooming in on the page level links (cf. Fig. 5.16), Table 5-11 below shows that the SCC subsites had 80.0 % of all inlinks and 95.7 % of all outlinks. This means, a SCC subsite received on average 87.8 inlinks from other subsites and provided on average 105.1 outlinks to other subsites. Once more, these high SCC counts surpassed the other components. Due to non-trivial computations, there is no connectivity study of how the 207,865 page level links are distributed within and between the components as the one made for subsite level links in Table 5-10 above. Thus, there is no count of how large share of the inlinks and outlinks that was *within* the SCC.

	# sub-sites	%	# inlinks	%	average inlinks / subsite	# outlinks	%	average outlinks / subsite
IN	626	8.2	58	0.03	0.1	8,588	4.1	13.7
SCC	1893	24.7	166,285	80.0	87.8	198,986	95.7	105.1
OUT	2660	34.7	41,342	19.9	15.5	182	0.1	0.1
IN-Tendrils	96	1.3	151	0.07	1.6	4	0.002	0.04
Tube	7	0.1	12	0.01	1.7	11	0.01	1.6
OUT-Tendrils	55	0.7	6	0.00	0.1	83	0.04	1.5
Disconnected	2332	30.4	11	0.01	0.005	11	0.01	0.005
	7669	**100.0**	**207,865**	**100.0**	**27.1**	**207,865**	**100.0**	**27.1**

Table 5-11. Distribution of inlinks and outlinks among the 7669 subsites.

5.4.1 Distribution of in-neighbors and out-neighbors

The distributions of in-neighbors and out-neighbors among the 7669 subsites were highly skewed as illustrated in Figures 5.17 and 5.18 further below reflecting that few subsites have many in-neighbors or out-neighbors, whereas the bulk of subsites only have very few in-neighbors or out-neighbors.

The distributions for each different graph component will not be brought here, but the statistics of the means, medians, ranges and standard deviations in Tables 5-12 and 5-13 illustrate that the SCC and OUT components have the most skewed distributions of in-neighbors, and the SCC and IN component correspondingly for the out-neighbors.

	# sub-sites	# with inlinks	# in-neigh-bors	mean # in-neighbors / subsite	median # in-neighbors / subsite	range # in-neighbors / subsite	std. dev.
IN	626	36	43	0.07	0	0 – 3	0.3
SCC	1893	1893	34315	18.1	7	1 – 387	32.8
OUT	2660	2660	14421	5.4	2	1 – 411	16.2
IN-Tendrils	96	96	99	1.0	1	1 – 2	0.2
Tube	7	6	11	1.6	2	0 – 3	1.0
OUT-Tendrils	55	3	3	0.05	0	0 – 1	0.2
Disconnected	2332	10	10	0.004	0	0 – 1	0.06
	7669	**4704**	**48902**	**6.4**	**1**	**0 – 411**	**20.2**

Table 5-12. Statistics of *in-neighbors* per subsite.

	# sub-sites	# with outlinks	# out-neigh-bors	mean # out-neighbors / subsite	median # out-neighbors / subsite	range # out-neighbors / subsite	std. dev.
IN	626	626	3953	6	3	1 – 163	13.4
SCC	1893	1893	44754	23.6	9	1 – 518	43.9
OUT	2660	83	111	0.04	0	0 – 7	0.3
IN-Tendrils	96	1	1	0.01	0	0 – 1	0.1
Tube	7	5	10	1.4	1	0 – 4	1.4
OUT-Tendrils	55	53	63	1.1	1	0 – 3	0.5
Disconnected	2332	10	10	0.004	0	0 – 1	0.06
	7669	**2671**	**48902**	**6.4**	**0**	**0 – 518**	**24.3**

Table 5-13. Statistics of *out-neighbors* per subsite.

There were power-law-like distributions of in-neighbors and out-neighbors as reflected by the approximately linear-shaped distributions in the log-log graphs of Fig. 5.17 and 5.18. However, the deliberate exclusion of site selflinks and links to and from university main web sites from the data set, thus excluding neighbor subsites at the same university and neighbor main sites at all universities naturally affect the distributions. For example, in the undelimited data set, main university sites with very high numbers of in-neighbors or out-neighbors would perhaps make the upper-left part of the distributions steeper thus allowing the curve to be more linear-shaped. It has not been feasible to compute distributions and possible power laws for an undelimited data set.

Another point that must be taken into account, is the aggregation of links from the page level to the subsite level, as illustrated earlier in Fig. 5.15 and 5.16, because the number of in-neighbors and out-neighbors on the subsite level will reduce the extreme number of connections which can be found on the page level when counting inlinks and outlinks. On the *page level*, power-law distributions have been verified in large-scale Web studies for inlinks to web sites (e.g., Albert, Jeong & Barabási, 1999; Adamic & Huberman, 2001), and outlinks from web sites (e.g., Adamic & Huberman, 2001).

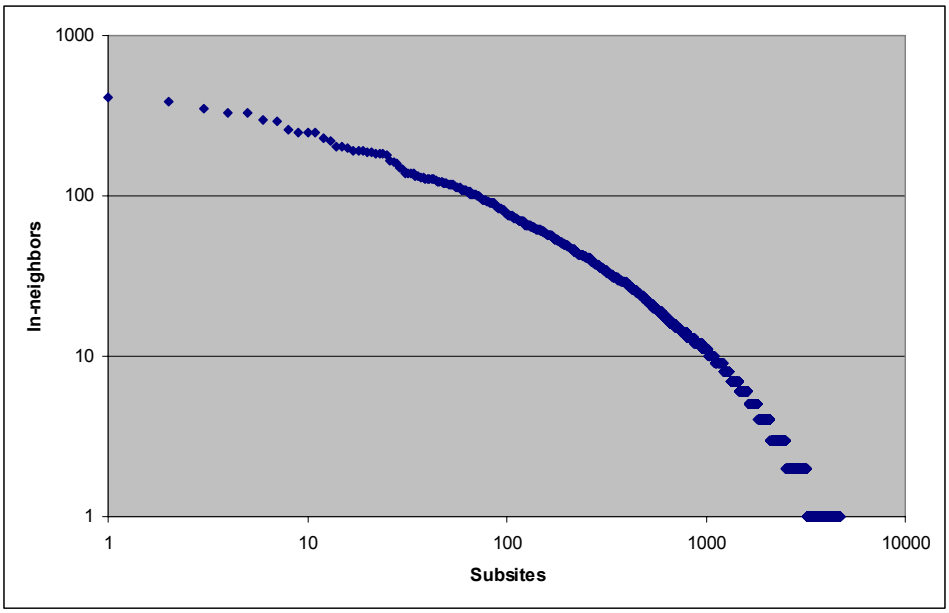


Figure 5.17. Distribution of *in-neighbors* for 7669 subsites. Log-log scale.

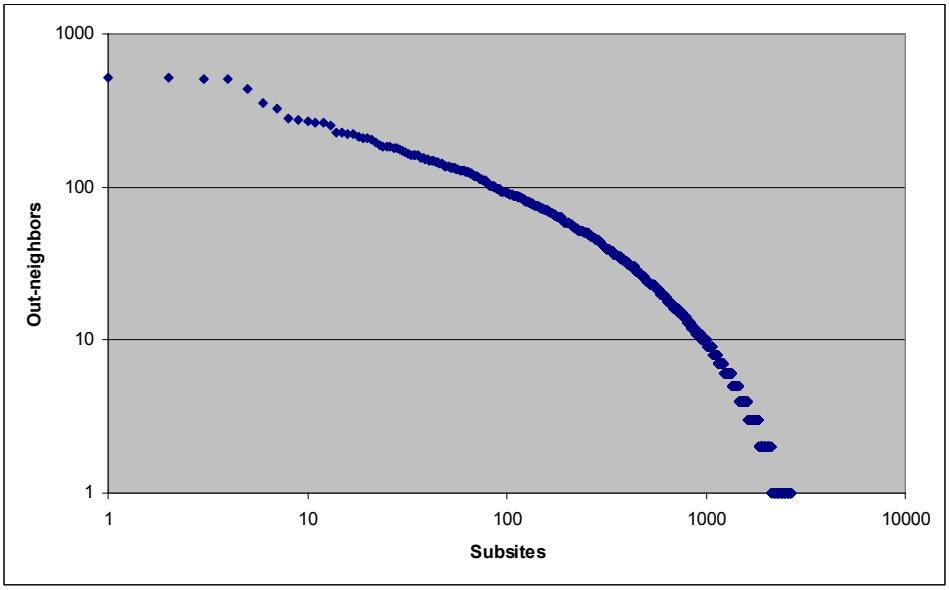


Figure 5.18. Distribution of *out-neighbors* for 7669 subsites. Log-log scale.

Tables 5-14 and 5-15 below give an impression of the type of subsites with most in-neighbors and out-neighbors. Among the 15 subsites with most in-neighbors are 9 computer science departments and related computer science institutions, 2 mirror sites with archives for software and Usenet discussion groups, 2 generic type university web sites, and 1 publisher. Almost all the subsites in the two tables belong to the SCC. The subsites belonging to the OUT component in Table 5-14 have no out-neighbors because source pages of mirror sites and publishers (including scientific journals) were excluded by the original web crawler (cf. Section 4.1.2). However, the OUT subsite of

comlab.ox.ac.uk was not crawled, probably because the official domain name had been changed to *web.comlab.ox.ac.uk* (Thelwall, personal e-mail 27.8.2003). Thus, only outlinks to *target* pages of *comlab.ox.ac.uk* were extracted in the crawl, leaving the subsite in the OUT component of the data set. The high number of in-neighbors to the two subsites belonging to the School of Computing and Information Technology, University of Wolverhampton (the affiliation of Dr Thelwall) is due to the many UK universities making links to the clickable image map (cf. Appendix 1) of all UK universities and higher education institutions (Thelwall, 2002b).

rank	id	subsite	comp.	# in-neigh-bors	# out-neigh-bors	affiliation
1	4124	src.doc.ic.ac.uk	OUT	411	0	'Sun Site' mirror site (Usenet archive) at Imperial College, London
2	1357	cbl.leeds.ac.uk	SCC	387	91	Computer Based Learning Unit, Univ. of Leeds
3	3017	scit.wlv.ac.uk	SCC	349	434	School of Computing and Info.Technology, Univ. of Wolverhampton
4	4129	sunsite.doc.ic.ac.uk	OUT	331	0	'Sun Site' mirror site (Usenet archive) at Imperial College, London
5	1821	users.ox.ac.uk	SCC	330	507	Personal web pages at Univ. of Oxford
6	2760	cs.ucl.ac.uk	SCC	300	265	Dept. of Computer Science, Univ. College London
7	4928	comlab.ox.ac.uk	OUT	289	0	Computing Laboratory (CS dept.), Univ. of Oxford
8	1866	info.ox.ac.uk	SCC	259	120	Former server with official web pages of Univ. of Oxford
9	3020	scitsc.wlv.ac.uk	SCC	249	8	School of Computing and Info. echnology, Univ. of Wolverhampton
10	3339	cup.cam.ac.uk	OUT	249	0	Cambridge University Press
11	325	cl.cam.ac.uk	SCC	246	141	Computer Laboratory (CS dept.), Univ. of Cambridge
12	2642	cogs.susx.ac.uk	SCC	231	268	School of Cognitive and Computing Sciences, Univ. of Sussex
13	1466	cs.man.ac.uk	SCC	218	224	Dept. of Computer Science, Univ. of Manchester
14	925	dcs.gla.ac.uk	SCC	203	511	Dept. of Computing Science, Univ. of Glasgow
15	3010	csv.warwick.ac.uk	SCC	202	354	Univ. of Warwick Information Service

Table 5-14. 15 subsites with most *in-neighbors* in the UK data set.

rank	id	subsite	comp.	# in-neigh-bors	# out-neigh-bors	affiliation
1	1088	cee.hw.ac.uk	SCC	148	518	Dept. of Computing and Electrical Engineering, Heriot-Watt Univ.
2	1572	doc.mmu.ac.uk	SCC	127	514	Dept. of Computing and Mathematics, Manchester Metropolitan Univ
3	925	dcs.gla.ac.uk	SCC	203	511	Dept. of Computing Science, Univ. of Glasgow
4	1821	users.ox.ac.uk	SCC	330	507	Personal web pages at Univ. of Oxford
5	3017	scit.wlv.ac.uk	SCC	349	434	School of Computing and Info. Technology, Univ. of Wolverhampton
6	3010	csv.warwick.ac.uk	SCC	202	354	Univ. of Warwick Information Service
7	2387	ecs.soton.ac.uk	SCC	117	327	Dept. of Electronics and Computer Science, Univ. of Southampton
8	791	dai.ed.ac.uk	SCC	137	280	Department of Artificial Intelligence, Univ. of Edinburgh
9	2291	afm.sbu.ac.uk	SCC	2	277	'Virtual library', Centre for Applied Formal Methods, South Bank Univ
10	2642	cogs.susx.ac.uk	SCC	231	268	School of Cognitive and Computing Sciences, Univ. of Sussex
11	1268	comp.lancs.ac.uk	SCC	183	265	Computing Dept., Univ. of Lancaster
12	2760	cs.ucl.ac.uk	SCC	300	265	Dept. of Computer Science, Univ. College London
13	19	users.aber.ac.uk	SCC	34	250	Personal web pages at Univ. of Wales, Aberystwyth
14	3042	www-users.york.ac.uk	SCC	94	226	Personal web pages at Univ. of York
15	1597	dcs.napier.ac.uk	SCC	127	226	School of Computing, Napier Univ.

Table 5-15. 15 subsites with most *out-neighbors* in the UK data set.

There are six overlaps between Tables 5-14 and 5-15: *cogs.susx.ac.uk, cs.ucl.ac.uk, csv.warwick.ac.uk, dcs.gla.ac.uk, scit.wlv.ac.uk, users.ox.ac.uk*; five of which related to computer science. Not surprisingly, many of these high-ranked subsites also belong to the *hubs* and *authorities* (Kleinberg, 1999a) in the UK academic web space as identified further below in Section 6.3.2.4.

It should be noted that figures 5.17 and 5.18 above show the overall link distributions for *all* scientific domains in the UK academic subweb. It would be interesting to investigate possible differences in distributions in different domains, reflecting different Web use between disciplines (Kling & McKim, 2000; Jacobs, 2001).

5.4.2 Links in the UK data set

Before zooming into a closer look at the distribution of page level links in the present study, some background statistics may yield an appropriate context. As noted in Section 4.1.2, the original web crawler harvested 3.40 million outlinking web pages containing 39.34 million outlinks at the 109 included universities. In other words, a total of 3.40 million web pages in the *undelimited* dataset of the 109 universities (including pages at the university main sites) contained outlinks either pointing at other pages within the same subsite or same university or pointing out to the whole Web.

Fig. 5.19 may illustrate some of the different link targets in the original undelimited UK data set.

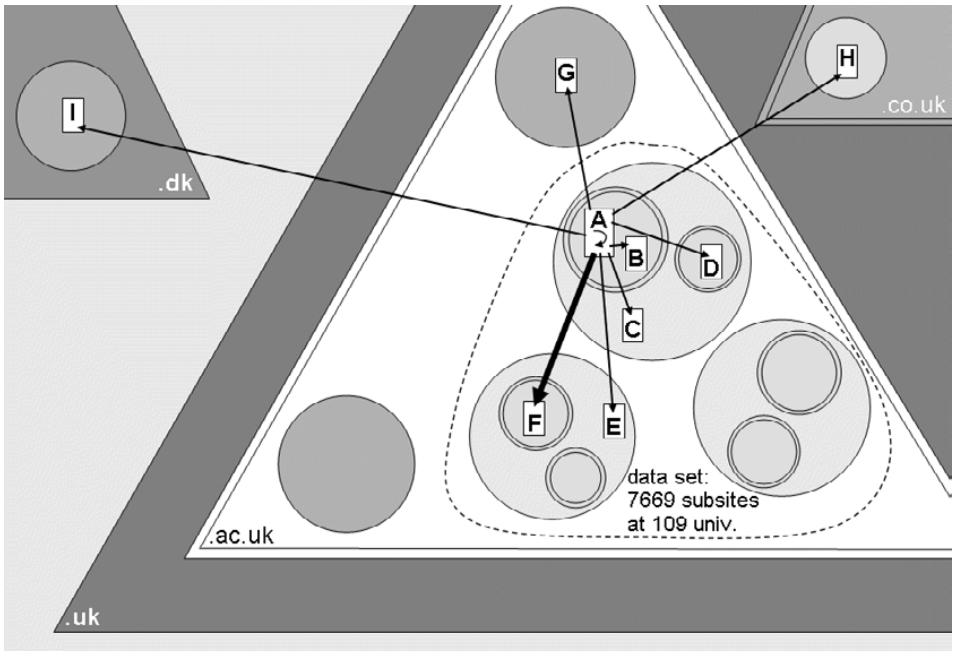


Figure 5.19. Pages and links in original undelimited UK data set. Bold link AF between pages A and F represent 207,865 page level links *between subsites* at different universities in the data set (within dashed borderline).

There were 34.39 million *site selflinks* (links AA[40], AB, AC and AD in Fig. 5.19) and 4.94 million *site outlinks* (links AE, AF, AG, AH, and AI), that is, a total of 39.34 million links in the undelimited data set (including links to and from the university main sites). The 3.40 million outlinking UK university web pages in the UK data set thus had on average 11.6 outlinks comprising 10.1 *site selflinks* and only 1.5 site outlink. Not surprisingly, most university links thus point to pages within the same university.

Only 3.1% (105.817) of the 3.40 million outlinking web pages at the 109 universities had links (links AE and AF) to the other 108 universities and their subsites.

[40] Link AA is a page selflink on page A.

It has not been feasible to compute number of pages belonging to the 7669 subsites with site outlinks to these subsites.

Of the 4.94 million site outlinks to all the Web, only 380,898 (7.7%) were targeted to the other 108 universities and their subsites in the *undelimited* data set. Looking only at site outlinks targeted between subsites at different universities; there were 207,865 such links (link AF in Fig. 5.19) in the *delimited* data set (cf. Table 5-11, Section 5.4). In other words, the data set investigated in the present study comprised 4.2% of all site outlinks at the 109 universities. The vast majority of site outlinks in the study thus were targeted to academic, commercial, and other targets *outside* the data set.

5.4.3 Distribution of inlinks and outlinks

As noted above, there were 207,865 *page level* links connecting source and target pages among the 7669 subsites. Compared with the 48,902 *site level* links, this gives average 4.25 page level links per subsite level link in the UK academic subweb. In other words, if two subsites are interlinked, there are on average 4.25 *page level links* between them. However, the distribution of page level links is highly skewed. Again, power-law-like distributions were found for both the distributions of inlinks (Fig. 5.20) and outlinks (Fig. 5.21).

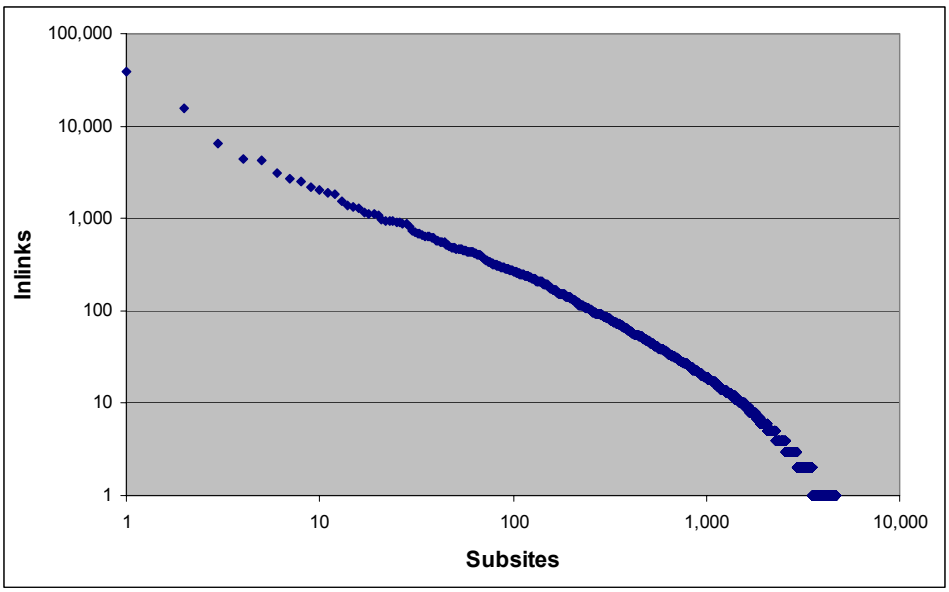


Figure 5.20. Power-law-like distribution of *inlinks* among 7669 subsites.

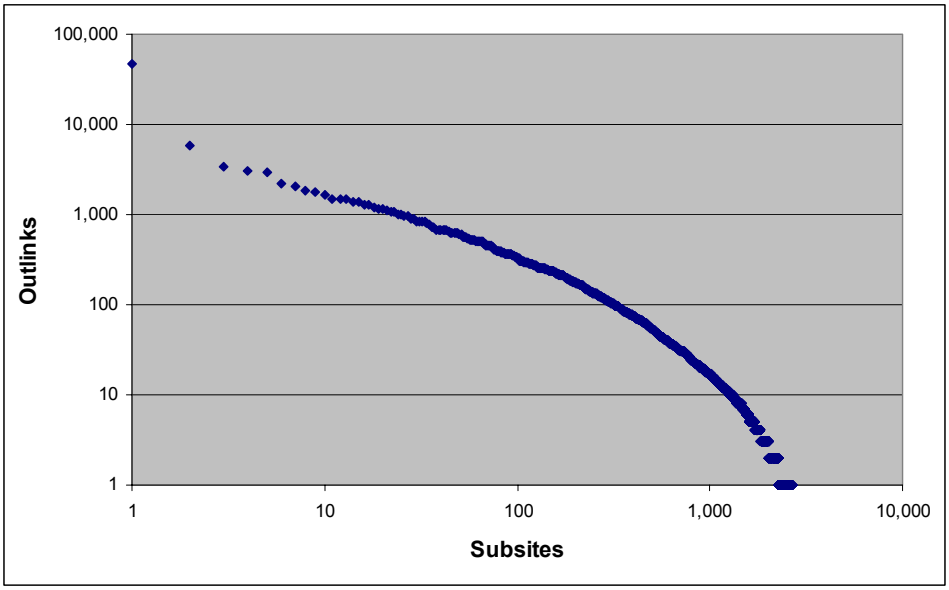


Figure 5.21. Power-law-like distribution of *outlinks* among 7669 subsites.

Tables 5-16 and 5-17 below show the 15 subsites with most inlinks and the 15 subsites with most outlinks, respectively.

rank	id	subsite	comp.	# inlinks	# outlinks	affiliation
1	2756	biochem.ucl.ac.uk	SCC	38619	497	Dept. of Biochemistry and Molecular Biology, University College London
2	345	scop.mrc-lmb.cam.ac.uk	SCC	15533	8	SCOP (Structural Classification of Proteins), Univ. of Cambridge
3	5401	xxx.soton.ac.uk	OUT	6437	0	E-print archive mirror of arXiv.org, hosted at Univ. of Southampton
4	*1357	cbl.leeds.ac.uk	SCC	4475	245	Computer Based Learning Unit, Univ. of Leeds
5	2184	cs.rdg.ac.uk	SCC	4246	1063	Dept. of Computer Science, Univ. of Reading
6	*4928	comlab.ox.ac.uk	OUT	3118	0	Computing Laboratory (CS dept.), Univ. of Oxford
7	*925	dcs.gla.ac.uk	SCC	2696	2197	Dept. of Computing Science, Univ. of Glasgow
8	1714	ukads.nott.ac.uk	OUT	2539	1	NASA Astrophysics Data System, database mirrored at Univ. of Nottingham
9	318	statslab.cam.ac.uk	SCC	2209	375	Statistical Laboratory, Univ. of Cambridge
10	*325	cl.cam.ac.uk	SCC	2027	888	Computer Laboratory (CS dept.), Univ. of Cambridge
11	*4124	src.doc.ic.ac.uk	OUT	1868	0	'Sun Site' mirror site (archive for software and Usenet groups) hosted at Imperial College, London
12	*2760	cs.ucl.ac.uk	SCC	1802	1155	Dept. of Computer Science, Univ. College London
13	2484	www-groups.dcs.st-and.ac.uk	SCC	1513	57	MacTutor History of Mathematics Archive, Univ. of St Andrews
14	*1466	cs.man.ac.uk	SCC	1377	828	Dept. of Computer Science, Univ. of Manchester
15	*2642	cogs.susx.ac.uk	SCC	1319	1264	School of Cognitive and Computing Sciences, Univ. of Sussex

Table 5-16. 15 subsites with most *inlinks*. Overlapping subsites with Table 5-14 (most *in-neighbors*) are marked with an asterisk.

rank	id	subsite	comp.	# inlinks	# outlinks	affiliation
1	3008	globin.bio.warwick.ac.uk	SCC	42	47798	Protein Bioinformatics Group, Univ. of Warwick
2	*2291	afm.sbu.ac.uk	SCC	20	5733	'Virtual library' about Formal Methods, Centre for Applied Formal Methods, South Bank Univ., London
3	*1088	cee.hw.ac.uk	SCC	608	3380	Dept. of Computing and Electrical Engineering, Heriot-Watt Univ.
4	2112	dcs.qmw.ac.uk	SCC	470	3097	Dept. of Computer Science, Queen Mary University of London
5	588	soi.city.ac.uk	SCC	248	2947	School of Informatics, City Univ.
6	*925	dcs.gla.ac.uk	SCC	2696	2197	Dept. of Computing Science, Univ. of Glasgow
7	1968	archive.comlab.ox.ac.uk	SCC	554	2059	Archive Service, Oxford University Computing Laboratory[41]
8	339	ast.cam.ac.uk	SCC	931	1848	Inst.of Astronomy, School of Physical Sciences, Cambridge
9	1384	bioinf.leeds.ac.uk	SCC	14	1763	Bioinformatics Research Group, Univ. of Leeds
10	418	www-cryst.bioc.cam.ac.uk	SCC	29	1676	Crystallography and Biocomputing Group, Univ.of Cambridge
11	*791	dai.ed.ac.uk	SCC	1062	1506	Dept. of Artificial Intelligence, Univ. of Edinburgh
12	*1268	comp.lancs.ac.uk	SCC	709	1497	Computing Dept., Univ. of Lancaster
13	*3017	scit.wlv.ac.uk	SCC	895	1482	School of Computing and Information Technology, Univ. of Wolverhampton
14	*1821	users.ox.ac.uk	SCC	1106	1405	Personal web pages at Univ. of Oxford
15	*1597	dcs.napier.ac.uk	SCC	576	1404	School of Computing, Napier Univ.

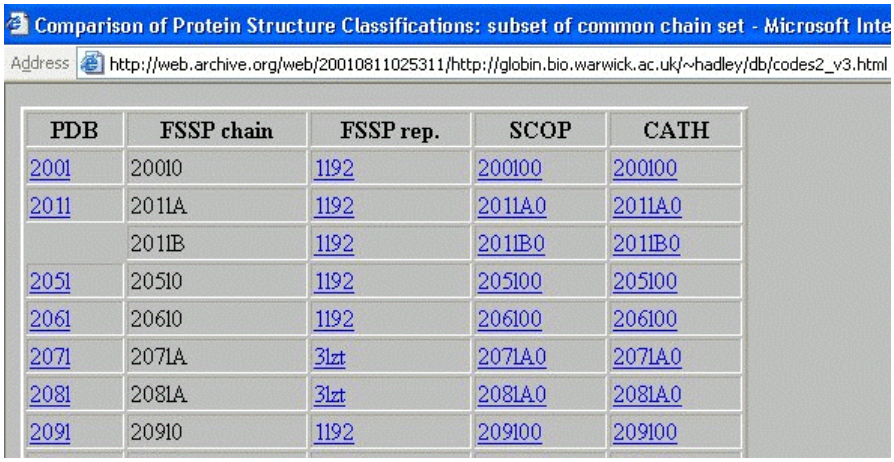
Table 5-17. 15 subsites with most *outlinks*. Overlapping subsites with Table 5-15 (most *out-neighbors*) are marked with an asterisk.

There were eight overlaps between the 15 subsites with most *inlinks* in Table 5-16 and the 15 subsites with most *in-neighbors* (Table 5-14), overlaps marked with an asterisk at the id number in Table 5-16 above. Eight of the 15 subsites with most inlinks were related to computer science. As was the case for the distribution of in-neighbors and out-neighbors, also almost all the subsites with most inlinks and outlinks in Tables 5-16 and 5-17 belong to the SCC graph component. In Table 5-16, the four subsites belonging to the OUT component with many inlinks were all excluded by the original web crawler for the same reasons as in Table 5-14 (mirrored contents of databases, e-print archives, etc.), thus the lack of outlinks from these subsites.

Two overall topical groups are prevailing among the 15 subsites with most *outlinks*: computer science and bioinformatics. Only one subsite overlapped with the subsites with most inlinks: node 925 (*dcs.gla.ac.uk*), Department of Computing Science, Glasgow.

A closer investigation of the data set revealed an interesting relation between the subsite that had most outlinks, Protein Bioinformatics Group at the University of Warwick (*globin.bio.warwick.ac.uk*) and the subsite with most inlinks, the Department of Biochemistry and Molecular Biology at the University College London (*biochem.ucl.ac.uk*) also noted by Thelwall (2002d). Of the 47,798 outlinks from the subsite at Warwick, 33,174 were pointing as automatically generated outlinks to a database of protein structure classification results located at the subsite at UCL. Fig. 5.22 below shows a very small excerpt from a web page at Warwick full of such automatically generated outlinks targeted to the protein structure database at UCL.

[41] Including mirrored material from the Centre for Applied Formal Methods, South Bank University (cf. rank 2).



PDB	FSSP chain	FSSP rep.	SCOP	CATH
2001	20010	1192	200100	200100
2011	2011A	1192	2011A0	2011A0
	2011B	1192	2011B0	2011B0
2051	20510	1192	205100	205100
2061	20610	1192	206100	206100
2071	2071A	31zt	2071A0	2071A0
2081	2081A	31zt	2081A0	2081A0
2091	20910	1192	209100	209100

Figure 5.22. Small excerpt of web page (retrieved at the Internet Archive) on protein structure classification at *globin.bio.warwick.ac.uk* with automatically generated outlinks to a database on the same topic at *biochem.ucl.ac.uk*.

This example with extreme outliers and automatic link generation illustrates that raw link counts are highly unreliable as indicators of the degree of web interconnectivity between universities (Thelwall, ibid.). This realization spurred Thelwall's development of the ADMs, the *Alternative Document Models*, outlined in Section 2.4.2 employing aggregated units of analysis in order to circumvent anomalies such as the abovementioned when conducting link connectivity analysis.

6 Five-step methodology

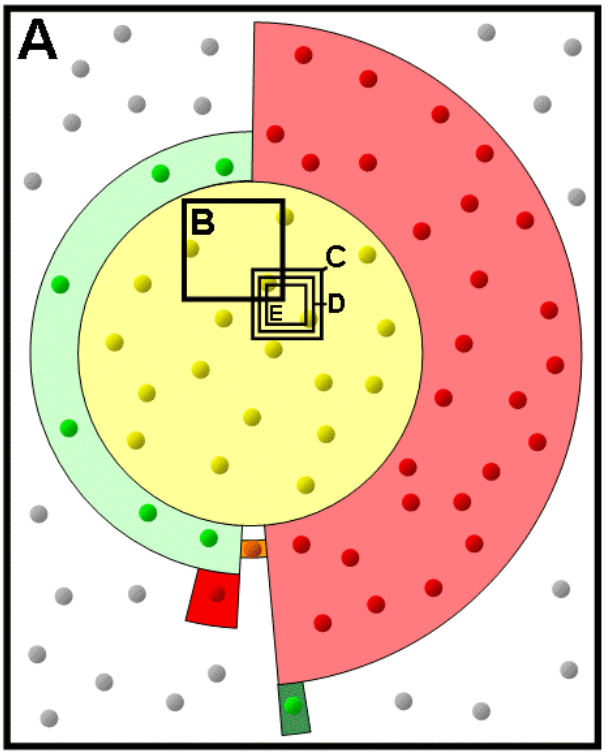


Figure 6.1.* Five-step methodology (A-E) for sampling, identifying and characterizing transversal links. (*cf. color prints placed before appendices)

The extraction of 7669 subsite nodes and the subsequent investigation of the distribution of graph components in the identified 'corona'-model of the UK academic web sub-sites in the previous chapter constitutes the first step A in a five-step methodology. The objective of the methodological steps is to lead up to the final step E (Section 6.5) concerned with the main research question put forward in this dissertation, that is, identifying what types of web links, web pages and web sites function as transversal (cross-topic) connectors in small-world academic web spaces.

The five-step methodology can be briefly outlined as follows:

A. Chapter 5: Selection and extraction of subsite nodes from the raw data set. Distribution of subsites in graph components ('corona' model). Validation of domain names in the Internet Archive. Section 6.1: Substantiated focus on the strongest connected component (SCC) for further investigation.
B. Section 6.2: Random sampling of 189 (10%) subsites belonging to the SCC. Retrieval of the subsites in the Internet Archive for classification of subsite

topics and genres. Subdivision of the subsites in 'nat/tech' and 'hum/soc' meta-topics.

C. Section 6.3: Selection of five random node pairs from a stratified sampling of 'nat/tech' subsites and 'hum/soc' subsites, respectively, from the 189 SCC subsites in step B. Extraction of 10 subgraphs, called *path nets*, comprising all shortest link paths in both directions between each of the five pairs of subsites.
D. Section 6.4: Extraction from the raw data set of URLs of source and target pages belonging to subsites along shortest paths in the 10 path nets. Retrieval of the pages in the Internet Archive for classification of page topics and genres.
E. Section 6.5: Identification, investigation and characterization of links, pages and subsites providing *transversal* (cross-topic) connections across dissimilar topics along shortest link paths in the 10 path nets (i.e. research question 4)

The five-step methodology is described and substantiated in more detail successively throughout Sections 6.1 - 6.5 since each step depends on the results from the preceding step. In Fig. 6.1 above, the five steps are figuratively illustrated. The sizes of the five inserted areas indicate the number of subsite nodes investigated in the different steps (cf. Table 6-1 below). The overlap between areas B and C includes the five pairs of subsites used as seed nodes in the 10 path nets, i.e., 'all shortest paths'-subgraphs.

step	number of analyzed subsites
A	7669 subsites in UK academic web space
B	sampled 189 SCC subsites
C	141 subsites in 10 path nets, comprising 104 unique subsites, 17 of which overlap with B, incl. 10 start and end seed nodes in path nets
D	78 visited unique path net nodes used for retrieving 352 page level links between 281 unique source pages and 249 unique target pages
E	48 unique source subsites with identified 112 transversal links

Table 6-1. Number of analyzed subsite nodes in the five-step methodology (A-E)

6.1 Focus on SCC subsites

The subsites in the strongest connected component (SCC) in the UK academic subweb graph were selected as a basis for further investigation in the dissertation, cf. Fig. 6.2 below. This decision was based on the following conditions:

(1) 100% validity of SCC domain names (cf. Section 5.1);
(2) the circumstance that only within the SCC there can be link paths in both directions between *all* subsites – irrespective of topical dissimilarity of the subsites – thus allowing easier identification of small-world properties across topical boundaries;
(3) the high percentage, at least 85.5% of link paths passing the SCC (Table 5-4**,** Section 5.3.1);
(4) the large share, 64.2% (Table 5-10, Section 5.4) of all subsite-to-subsite connections located within the SCC;
(5) the SCC contained only 24.7% (1893) of the subsites in the delimited data set, thus making it feasible to investigate a relatively large sample (10%, i.e. 189) of the SCC subsites.

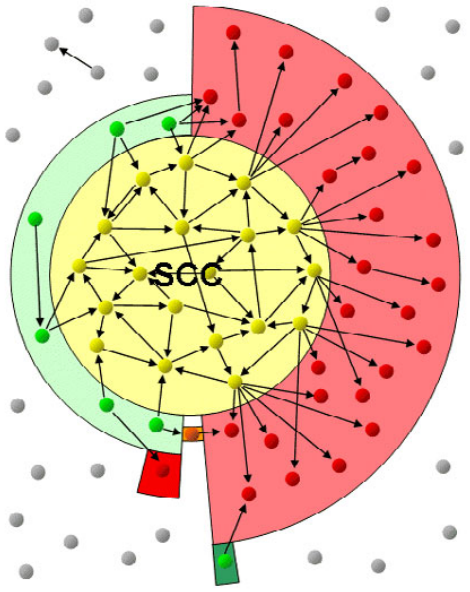


Figure 6.2. Focus on SCC subsites.

The distributions of in-neighbors and out-neighbors in Figures 6.3 and 6.4 among the 1893 SCC subsites show the same power-law-like shapes as the distributions for all 7669 subsites outlined in Section 5.4.1. This similarity is logical, since many of the outliers with many in-neighbors or out-neighbors belonged to the SCC. The SCC subsites with most in-neighbors and out-neighbors were listed in Tables 5-14 and 5-15 in Section 5.4.1 where they make up most of the 30 included subsites.

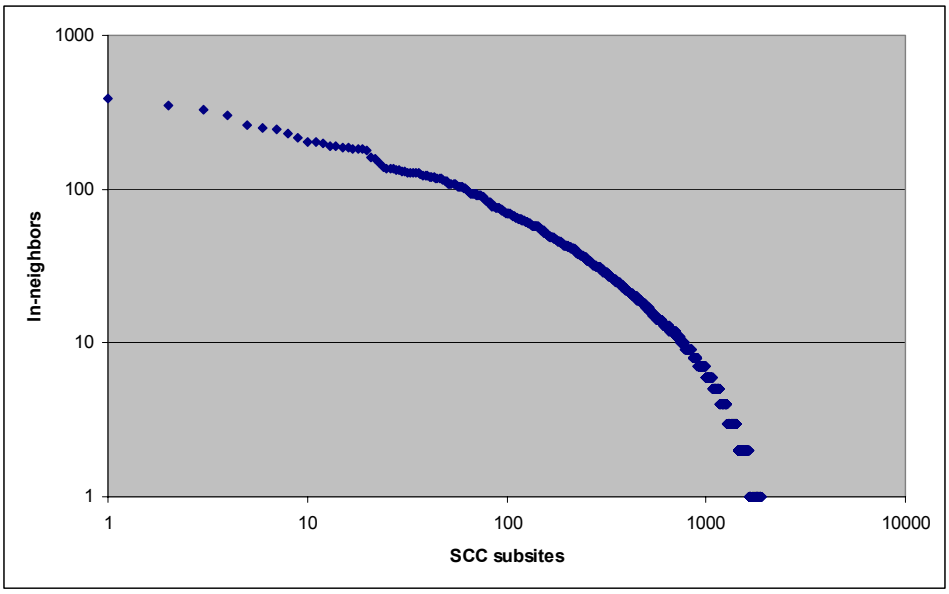


Figure 6.3. Distribution of *in*-neighbors for 1893 SCC subsites. Log-log scale.

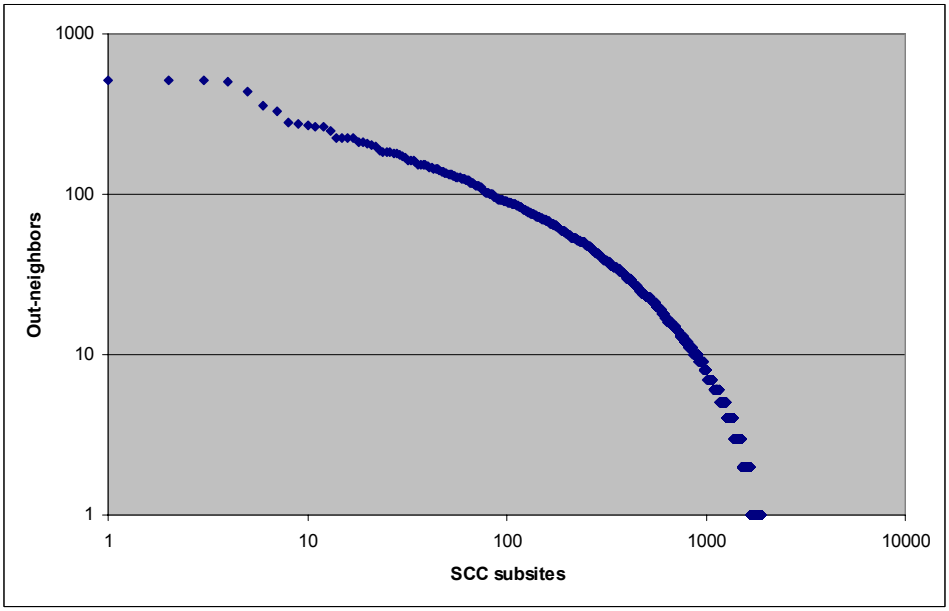


Figure 6.4. Distribution of *out*-neighbors for 1893 SCC subsites. Log-log scale.

A SCC subsite received inlinks from on average 18.1 other subsites and provided outlinks to on average 23.6 other subsites as shown in Tables 5-12 and 5-13, respectively, in Section 5.4.1. As noted earlier, these high degrees of connectivity in the SCC by far exceeded the in-degrees and out-degrees of subsites in the other graph components in the UK data set. However, the highly skewed distributions also imply a high number of SCC subsites with just a few in-neighbors or out-neighbors. For example, 255 SCC subsites had just one in-neighbor and 220 had one out-neighbor. Of these subsites, 80 SCC subsites had only one in-neighbor *and* one out-neighbor. Such

low connectivity degrees indicate how vulnerably close to isolation a web node can be even if it belongs to the strongest connected component.

The distributions of *inlinks* and *outlinks* in the SCC will not be shown here, as they are very similar to the distributions shown in Section 5.4.3. Most of the subsites in Tables 5-16 and 5-17 in Section 5.4.3 with many inlinks and outlinks belonged to the SCC.

6.2 Sample of 189 SCC subsites

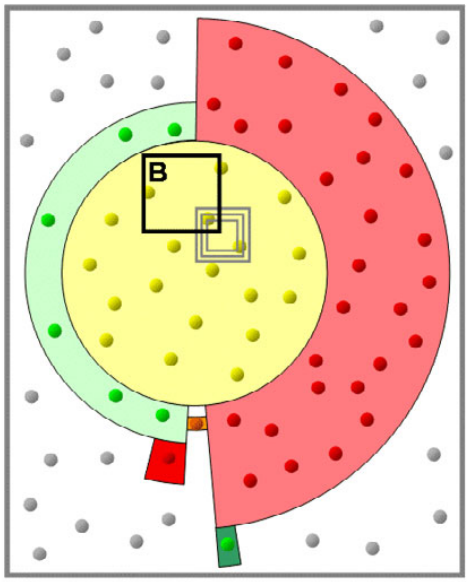


Figure 6.5. Step B in the five-step methodology: topics and genres of sampled 189 SCC subsites

This section is concerned with the second step in the five-step methodology. This step comprises a random sample extracted by the statistical program SPSS of 10% of the 1893 SCC subsites. Sections 6.2.1 and 6.2.2 below outline how the overall topics and 'genres' (the overall function and type of the subsite), respectively, of the 189 sampled SCC subsites were classified by visiting each sample URLs indexed in the Internet Archive as close to July 2001 as possible. Even if the original web crawler had harvested the link data set between June and July in 2001, it was not practical to identify the individual harvest dates for each subsite and use these dates for retrieval in the Internet Archive. Instead, July 2001 was selected as a common point of time. If several dates in July were indexed in the Internet Archive, an early date was chosen. If nothing was indexed in July, the closest preceding month was chosen unless a subsequent indexed month was closer. The same heuristics were followed if indexed web pages were not accessible due to errors (e.g., 'Failed Connection', 'Path Index Error') in the Internet Archive. The reason for using the Internet Archive, instead of finding the sample subsites directly on the present Web, was that subsites including their topics might have changed since the web crawler visited them in 2001. This way of using the Internet Archive is an example of *'web archaeology'* necessary on the ever-

changing Web as suggested by Björneborn & Ingwersen (2001; forthcoming). Only one (0.5%) of the 189 sampled SCC subsites was not available in the Internet Archive because the subsite owner had banned access for web crawlers by using robot exclusion (cf. Section 4.2.4).

During the classification process in the Internet Archive, it was often necessary to retrieve subordinate web pages at the subsites in order to acquire more detailed information about the overall topic and genre. Because of this time-consuming task of manual identification of topics and genres of subsites, it was not feasible to make similar samples from the other graph components as well. However, in future studies it would be interesting to compare subsite topics and genres of different graph components.

The classifications of topics and genres were conducted by the author alone. Using more indexers would probably have resulted in other classifications, however not necessarily more noncontradictory, confer the so-called *inter-indexer inconsistency* concerned with how indexers often disagree on what descriptors to assign to the same documents (Cooper, 1969; Wilkinson *et al.*, 2003).

6.2.1 Topics of 189 SCC subsites

The overall objective with the classification of the sample subsite topics was to use them for selecting pairs of subsites with dissimilar topics in the next third step in the five-step methodology. As shown in Table 6-2 below (cf. extensive table in Appendix 7), the sample of 189 SCC subsites contained 36 subsites (19.0%) in the humanities and social sciences ('hum/soc') and 119 subsites (63.0%) in the natural sciences and technology ('nat/tech'). The remaining 34 subsites (18.0%) were generic nation-wide or campus-wide service style (cf. Section 6.2.2).

In Table 6-2, the two rough blocks of meta-topics 'hum/soc' and 'nat/tech' are further subdivided; the 'hum/soc'-related subsites in 5 broad topical groups (A-E) and the 'nat/tech'-related subsites in 7 groups (F-L):

subsite topics	# sub-sites	% n=189
HUM/SOC	**36**	**19.0**
A	**6**	**3.2**
Architecture + Landscape architecture	2	
Art & Media	3	
Humanities & Social sciences	1	
B	**12**	**6.3**
Business	4	
Economics	3	
Law	2	
Law + Economics	1	
Political science	2	
C	**5**	**2.6**
Education: Learning technology	1	
Library & Information science	2	
Linguistics	2	
D	**6**	**3.2**
Archaeology	2	
Ethnography	1	
Geography	3	

E	**7**	**3.7**
Psychology	3	
Sociology	3	
Social medicine	1	
NAT/TECH	**119**	**63.0**
F	**14**	**7.4**
Agriculture: rural studies	1	
Earth sciences	3	
Environmental studies	8	
Zoology	2	
G	**25**	**13.2**
Biochemistry	3	
Bioscience	5	
Medicine	15	
Pharmacology	1	
Psychology	1	
H	**10**	**5.3**
Chemistry	5	
Chemical engineering	2	
Materials science	3	
I	**19**	**10.1**
Astronomy	3	
Physics & Astronomy	2	
Physics	14	
J	**11**	**5.8**
Mathematics	7	
Mathematics & Statistics	3	
Statistics	1	
K	**16**	**8.5**
Engineering	5	
Electronical engineering + Telematics	11	
L	**24**	**12.7**
Computer science	11	
Computer science + Electronics	1	
Computer science + Engineering	1	
Computer science + Management science	1	
Computer science + Mathematics	1	
Informatics	7	
Informatics + Psychology	1	
Library & Information science	1	
	155	**82.0**

Table 6-2. Topics of 155 SCC subsites divided in 5 'hum/soc' groups (A-E) and 7 'nat/tech' (F-L).

The term *topic* is used in a pragmatic common-sense way in the dissertation when identifying the overall topic of a subsite. The topical classifications, as well as the division of groups and allocation of subsites in the groups may be questioned. However, the main purpose of the groups was to function as rough 'bags' in order to select diversified topical pairs for the subsequent case studies of shortest paths.

The groups were constructed in a 'bottom-up' way induced by the actual topics of the included subsites. Subsites were grouped together with related topics in a pragmatic attempt to form 10-12 groups, neither too big nor small, that could form a basis for stratified sampling. Even if the purpose thus not was to obtain a perfect noncontradictory classification, some schemes and guidelines were consulted. The classification scheme of science fields and subfields designed by Glänzel & Schubert (2003) (see Appendix 8) was helpful for classifying some of the 'nat/tech' topics, for example, biochemistry and materials science. There were some deviations in classifications. For example, astronomy was not placed together with earth sciences as

Glänzel and Schubert recommend, but together with physics because of the common affiliate conjunction, for instance, in the Department of Physics and Astronomy, Open University (*yan.open.ac.uk*).

The classification scheme by Glänzel and Schubert is not as expanded in the 'hum/soc' categories as in the 'nat/tech'. However, the extensive topical categories in all scientific domains included in the Research Assessment Exercise (RAE) of UK higher education institutions (HERO, 2001; cf. Appendix 9) functioned as supplementary guidelines when grouping the 189 SCC subsites.

Many of the topics of the included 189 subsites were interdisciplinary. In case of doubt, the departmental affiliation of the subsite could be decisive. For example, the subsite of the International Boundaries Research Unit, Department of Geography, University of Durham (*www-ibru.dur.ac.uk*) overlaps with topics in International law (group B in Table 6-2), but was placed under geography (group D) because of the departmental affiliation. The interdisciplinary domain of geography encompasses both physical and cultural geography. This disciplinary span is reflected in institutional names encountered in the subsequent case studies, for example, the CTI Centre for Geography, Geology and Meteorology, University of Leicester (*www.geog.le.ac.uk/cti*) and the School of Geography, Politics and Sociology, University of Newcastle upon Tyne (*www.ncl.ac.uk/geps/*). In the present study, geography was grouped as a domain belonging to the social sciences following the RAE units of assessment above.

Some interdisciplinary domains were placed in both 'hum/soc' and 'nat/tech' groups. The subsite of the research project 'Business Information and the Internet' at the Department of Information Science, University of Strathclyde (*business.dis.strath.ac.uk*) was placed in group C with another 'hum/soc'-related library and information science project on thesaurus construction, together with linguistic and educational subsites. However, the subsite of the Information Retrieval Group, Department of Information Studies, University of Sheffield (*ir.shef.ac.uk*), was judged to belong to a more computer-science-related field of LIS and hence placed in group L together with subsites in informatics and computer science.

Three psychology subsites were placed together with social medicine and sociology in 'hum/soc' group E. However, the subsite of the research group of Computational Neuroscience, Department of Psychology, University of Stirling (*cn.stir.ac.uk*), was grouped together with medicine subsites in group G.

6.2.2 Genres of 189 SCC subsites

As described above, the main objective at this methodological step was to classify the overall topics of the subsites. However, it may give a useful impression of the subsites in the investigated SCC component also to identify the subsite genres, that is, the overall function and categorical type of the subsites. The term *genre* is here used in a broad sense in accordance with contemporary web terminology for describing types of web sites as well as web pages (e.g., Koehler, 1999a; Nilan, Pomerantz & Paling, 2001; Agatucci, 2001; Jackson-Sanborn *et al.*, 2002). According to Agatucci (2001), "the genre or 'form' of an academic website should form its 'function' or communication purpose". A web site may be conceived as a meta-document constituted by a collection

or aggregation of web pages as argued by Thelwall (2002b) in his *Alternative Document Models* (cf. Section 2.4.2). Genre classification on the Web is an inherently complex task due to non-established, muddled and overlapping genres (cf. Crowston & Williams, 2000; Dillon & Gushrowski, 2000; Rehm, 2001; Thelwall, forthcoming). These points are further elaborated below in Section 6.4.5.

In the current study, the classification of the overall genres of 189 SCC subsites was primarily based on the web site creators' own terminology, for example, *'students' union', 'department homepage', 'conference homepage'*, etc. This induced 'bottom-up' classification of the given body of subsites is shown in Table 6-3. The subsites were divided into two main categories: generic and research/teaching, as discussed below.

subsite genres	# sub-sites	% n=189
GENERIC	**34**	**18.0**
Library service	11	5.8
University campus-wide service	8	4.2
College homepage	5	2.6
Students' union	5	2.6
University main web site	3	1.6
National university service	2	1.1
RESEARCH & TEACHING	**155**	**82.0**
Department homepage	36	19.0
Research group homepage	30	15.9
Centre homepage	19	10.1
School homepage	18	9.5
Research project homepage	8	4.2
Institute homepage	7	3.7
Teaching resource pages	5	2.6
Faculty homepage	4	2.1
Division homepage	3	1.6
Personal resource pages	3	1.6
Research group resource pages	3	1.6
Startpage without content	3	1.6
Conference homepage	2	1.1
Personal homepages	2	1.1
Lab resource pages	2	1.1
School resource pages	2	1.1
Collaborative project homepage	1	0.5
Intranet	1	0.5
Journal homepage	1	0.5
Lab homepage	1	0.5
Online archive homepage	1	0.5
Postgraduate prospectus homepage	1	0.5
Students' union society homepage	1	0.5
N/A in the Internet Archive or on the Web, June 2003	1	0.5
	189	**100.0**

Table 6-3. Genres of 189 SCC subsites.

Of the 189 SCC subsites, 34 (18.0%) were generic subsite genres being nation-wide or campus-wide service style, for example, the national university services EDINA (Edinburgh Data and Information Access: *edina.ed.ac.uk*) and EDEC (Electronic Design Education Consortium: *edec.brookes.ac.uk*). Other generic subsites contained library services or students' unions. There were three university main sites 'disguised' with alias domain names (*ccc.nott.ac.uk, www2.rhul.ac.uk, www3.open.ac.uk*) among the 189 subsites. Five college homepages were also identified, for example, Trinity

College at the University of Oxford (*trinity.ox.ac.uk*). These university and college sites were placed in the generic category because they contained multidisciplinary contents not possible to assign overall topics. It is thus apparent, that even if the 109 official university main sites were excluded in order to reduce multidisciplinarity *within* the investigated subsites to enable identification of cross-disciplinary links *between* subsites as described in Section 4.2.1, such multidisciplinary subsites were nevertheless present in the delimited data set. This circumstance made it necessary to take precautions in the following case studies, identifying such generic subsites on the shortest link paths.

Not only university main sites had alias domain names, also other genres used aliases, for example, the sample showed that the centre homepage of the Environment Office at the Imperial College in London had two domain names with identical contents: *gse.ic.ac.uk* (id 1187) and *iceo.ic.ac.uk* (id 1195). This redundancy of alias domain names is impossible to avoid in a large-scale data set as the present because it is not feasible to conduct close inspection of the contents behind all domain names.

There were 155 subsites containing research-related or teaching-related contents. The teaching-related subsites were concerned with *particular* topics, for example, statistics for psychology students, or learning technology in earth sciences. Teaching-related subsites not concerned with particular topics but functioning as a nation-wide or campus-wide service like the Information and Learning Resource Services at Middlesex University (*ilrs.mdx.ac.uk*) were not included but placed in the generic category.

The research-related or teaching-related subsites are grouped together in Table 6-3, because it was not always possible to make unambiguous distinction as some contents can be related to both research and teaching, and some subsites may contain both types of contents. The subsites of departments, research groups, centers and schools were by far the largest subsite genres, together comprising 54.5% of the sampled SCC subsites.

In Table 6-4 below, a distinction is made between what here is called *'home'-type* and *'support'-type* subsites. *'Home'-type* subsites function as official 'show windows' aimed both at external and internal users. For example, the official homepage of a faculty, conference, or research group – or the homepage of an online archive containing extensive material on the life and research of Charles Booth (1840-1916) mapping poverty in London (*booth.lse.ac.uk*), see Fig. 6.6 further below. It may also be a single person having an entire subsite as a personal web territory, like *starform.infj.ulst.ac.uk* created by Bill McMillan, senior lecturer at the Faculty of Informatics, University of Ulster – or four researchers at the Department of Geological Sciences, University College London, having their personal homepages on their own web server (*slamdunk.geol.ucl.ac.uk*) as illustrated in Fig. 6.7.

The original affiliation terms have been preserved in Table 6-4, even if universities do not use consistent terminology for similar units. What is called 'department' at one university may be called 'school' at another. The term 'research *project'* is used here when a subsite treats *one* specific research project, whereas a subsite of a research *group* may describe several ongoing projects as well as provide other contents.

The term *'support'-type subsite* designates subsites functioning as an assembly of sub-territories and resources, sometimes quite heterogeneous, aimed primarily as support for internal users. An example of such a subsite is the Applied Psychology and

Computing Support Server, at the University of Bournemouth (*xanadu.bournemouth.ac.uk*) shown in Fig. 6.8 below. This subsite contains resources supporting internal staff and students, for instance, 'the Agora discussion space' or 'Stuff to help educate and civilize'. The teaching resource pages include subsites containing learning technology facilities for supporting teaching at specific schools, departments, etc. Some of the 'support'-type subsites are not intended to give access to external visitors browsing through a top entry homepage to reach pages within the whole subsite. Such subsites may have a server default page as top page as shown in Fig. 6.9 giving no access to the rest of the subsite that may contain extensive contents in separate web territories for research groups, individuals, etc.

Of the 155 research & teaching subsites, 135 (87.1%) were 'home'-type, and 19 (12.3%) were 'support'-type. One subsite could not be classified because it was neither available in the Internet Archive nor directly on the Web.

research & teaching-related subsite genres	**# sub-sites**	**% of 155**	**% of 189**
	155	**100.0**	**82.0**
'Home'-type subsites	*135*	*87.1*	*71.4*
Department homepage	36	23.2	19.0
Research group homepage	30	19.4	15.9
Centre homepage	19	12.3	10.1
School homepage	18	11.6	9.5
Research project homepage	8	5.2	4.2
Institute homepage	7	4.5	3.7
Faculty homepage	4	2.6	2.1
Division homepage	3	1.9	1.6
Conference homepage	2	1.3	1.1
Personal homepages	2	1.3	1.1
Collaborative project homepage	1	0.6	0.5
Journal homepage	1	0.6	0.5
Lab homepage	1	0.6	0.5
Online archive homepage	1	0.6	0.5
Postgraduate prospectus homepage	1	0.6	0.5
Students' union society homepage	1	0.6	0.5
'Support'-type subsites	*19*	*12.3*	*10.1*
Teaching resource pages	5	0.0	2.6
Personal resource pages	3	1.9	1.6
Research group resource pages	3	1.9	1.6
Startpage without content	3	1.9	1.6
Lab resource pages	2	1.3	1.1
School resource pages	2	1.3	1.1
Intranet	1	0.6	0.5
N/A in the Internet Archive or the Web	1	0.6	0.5

Table 6-4. 'Home'- and 'support'-type subsite genres among 155 research & teaching subsites.

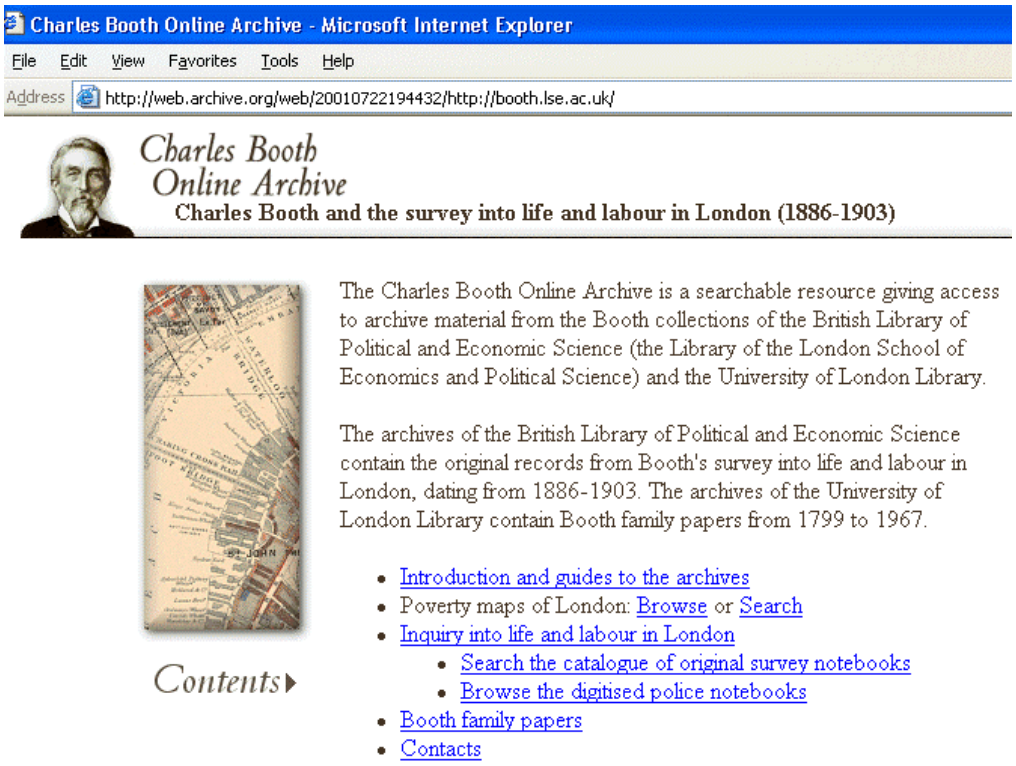


Figure 6.6. Example of 'home'-type subsite, the Charles Booth Online Archive, London School of Economics and Political Science (*booth.lse.ac.uk*). Excerpt of homepage retrieved in the Internet Archive.

Figure 6.7. Example of 'home'-type subsite with four researchers' personal homepages (*slamdunk.geol.ucl.ac.uk*) at the Department of Geological Sciences, University College London. Excerpt of homepage retrieved in the Internet Archive.

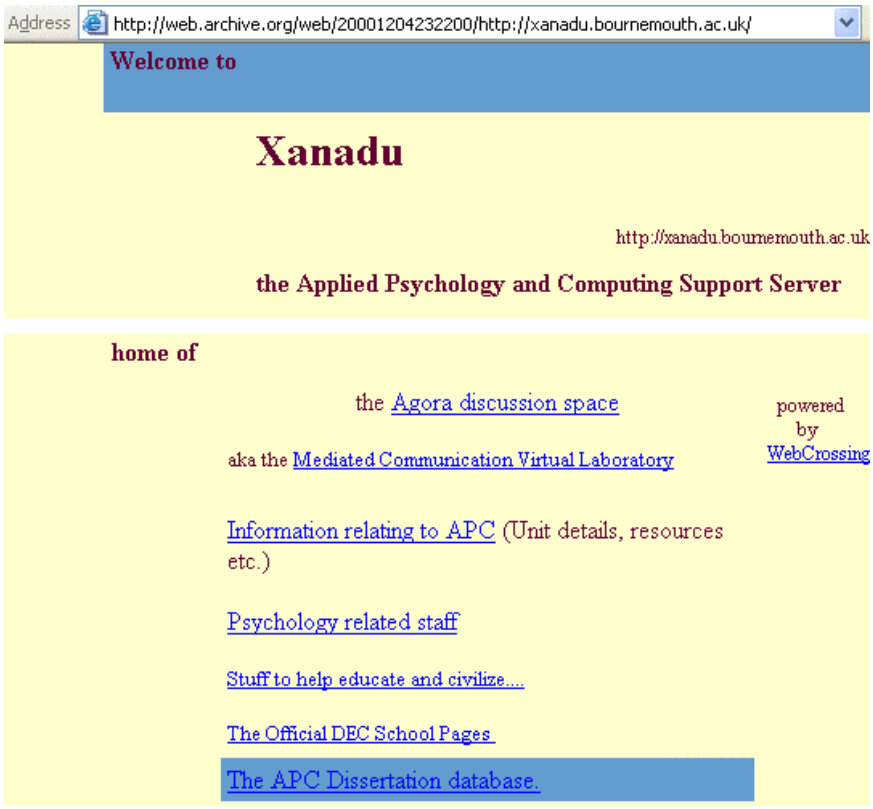


Figure 6.8. Example of 'support'-type subsite (*xanadu.bournemouth.ac.uk*). Excerpt of screenshot from entry page retrieved in the Internet Archive.

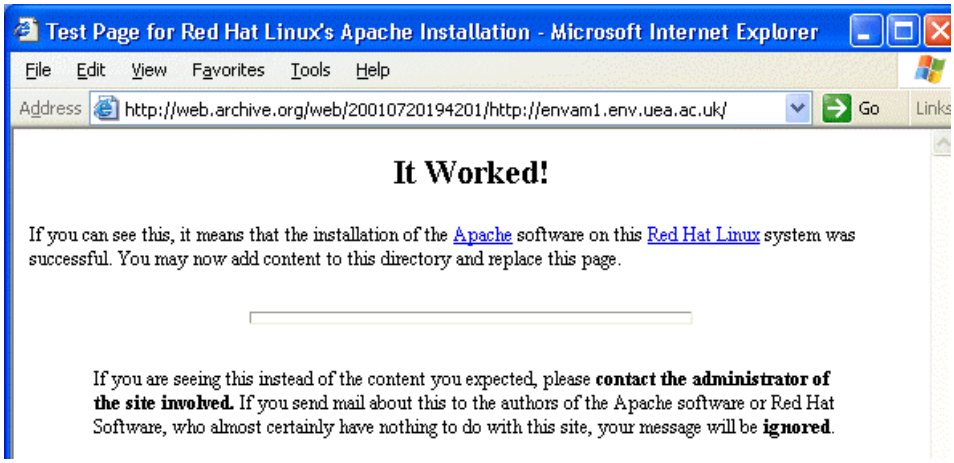


Figure 6.9. Example of 'support'-type subsite (*envam1.env.uea.ac.uk*) with no top entry page, instead using a server default page. Excerpt of screenshot of page retrieved in the Internet Archive.

6.3 Sample of 10 path nets among SCC subsites

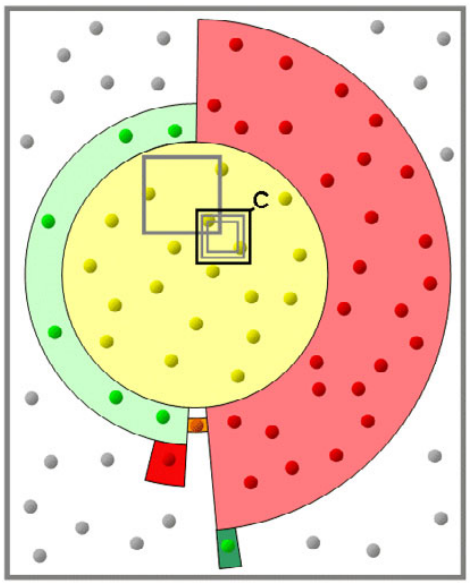


Figure 6.10. Step C in the five-step methodology: sample of 10 path nets among SCC subsites.

The third step, C, in the five-step methodology deals with extracting a sample of SCC subsites belonging to dissimilar topics among the 189 visited SCC subsites in the previous step B. The objective is to investigate all shortest link paths in both directions between each of the subsite pairs in order to enable the subsequent steps D-E concerned with the identification and characterization of links, pages and subsites providing small-world shortcuts across dissimilar topical domains in an academic web space.

6.3.1 Methodology

Different approaches were considered in the PhD project with regard to how to select random subsites for shortest link paths. A pilot test using the network analysis software *Pajek* (cf. Section 4.2.1) to extract all shortest link paths between 10 pairs of randomly selected start nodes from the IN component and end nodes from the OUT component revealed that the resulting link paths contained only one IN-subsite and one OUT-subsite identical with the selected start and end nodes. All the intermediary subsites on the link paths were located in the SCC. Furthermore, no so-called *topic drift* (cf. Section 6.5.1), that is, no cross-topic (transversal) links were identified on the first or the last link in the sample link paths. The link paths *within the SCC* thus contained all the topic drift. This observation is of special interest, since the dissertation is concerned with small-world phenomena affected by topic drift in the shape of transversal links. The pilot test findings thus supported the decision to focus on the SCC component.

However, it should be noted that a larger sample might have yielded different results with regard to possible topic drift within the IN and OUT components.

As stated earlier, the SCC subsites are also interesting because only within the SCC component there can be link paths in both directions between *all* subsites (cf. Section 5.1). Shortest link paths between any pair of subsites containing dissimilar topics may thus be identified in the SCC component. The special feature of reversible link paths within the SCC component is exploited in the present methodological step when identifying transversal links among university subsites.

The pilot test described above also revealed that some subsite pairs could be connected by quite many shortest paths with same path length. A sample size of 10 SCC subsite pairs was thus considered to be tractable with regard to analyzing all shortest paths between the subsites, including the 'zooming-in' manual close inspections of source and target web pages and page level links to be conducted in subsequent steps D and E. The small sample size compared to the very large number[42] of corresponding permutations of subsite pairs in the whole SCC implied that the sample could not be used to generalize any findings. Instead, the sample will be used as case studies for identification of phenomena and generation of concepts and hypotheses (cf. Yin, 1994; Andersen, 1997; Dubé & Paré, 2001). According to Yin (1994) a "case study is an empirical inquiry that investigates a contemporary phenomenon within its real-life context, especially when the boundaries between phenomena and context are not clearly evident" (cited by Dubé & Paré, 2001). Case studies are thus useful when a phenomenon is broad and complex, where the existing body of knowledge is insufficient to permit the posing of causal questions, and when a holistic, in-depth investigation is needed (Dubé & Paré, 2001). All these conditions are to a high degree present in the new research area of small-world link structures in academic web spaces explored in the present dissertation. As elaborated earlier (cf. Sections 1.2 and 3.5), link structures also reflect social activities, the traditional research object in case studies. Thus, a case study approach is suitable in the present study, where the small sample of 10 SCC subsite pairs will be used as case studies for identification of phenomena and generation of concepts and hypotheses, as stated above.

The overall objective with the sample of deliberately juxtaposed pairs of topically dissimilar subsites was to construct confined and thus investigable small-world *subgraphs* or *'mini small worlds'* in the shape of so-called *path nets* (see definition in Section 6.3.2) where all shortest link paths could be analyzed between the juxtaposed nodes.

The sample of 10 topically dissimilar SCC subsite pairs was extracted in the following way. From the sample of 189 SCC subsites in the previous methodological step (Section 6.2.1), a stratified sample was extracted consisting of five subsites

[42] Using the estimates from the sample of 189 SCC subsites (cf. section 6.2.1), there are approx. 19% 'hum/soc' and 63% 'nat/tech' subsites in the whole SCC (not including confidence intervals). These estimates would mean that there are approx. 360 'hum/soc' subsites and 1193 'nat/tech' subsites yielding over 400,000 possible pairs of subsites with a 'hum/soc' start node and a 'nat/tech' end node. (Using a 90% confidence interval, the percentage of 'hum/soc' subsites in the whole SCC component was estimated to be 19.0±4.7; the percentage of 'nat/tech' 63.0±5.8; and 'generic' subsites 18.0±4.6.)

belonging to the formed categories in humanities and social sciences ('hum/soc') and five subsites in natural sciences and technology ('nat/tech'). More specifically, from each of the five 'hum/soc' categories A-E in Table 6-2 in Section 6.2.1 was randomly selected one subsite. From the seven 'nat/tech' categories F-L were randomly selected five categories, and from each of these five categories was randomly selected one subsite. The selected ten *seed* subsites were then randomly grouped into five pairs with a start node in 'hum/soc' and an end node in 'nat/tech' categories as listed in Table 6-5 below.

hum/ soc	id	random start node	affiliation	nat/ tech	id	random end node	affiliation
A	2099	hum.port	Faculty of Humanities and Social Sciences, Portsmouth	I	1904	atm.ox	Atmospheric, Oceanic and Planetary Physics, Physics Dept, Oxford
B	2394	economics.soton	Economics Dept, Southampton	H	917	chem.gla	Chemistry Dept, Glasgow
E	1494	psy.man	Psychology Dept, Manchester	J	893	maths.gcal	Mathematics Dept, Glasgow Caledonian
C	871	speech.essex	Speech Research Group, Language and Linguistics Dept, Essex	F	245	palaeo.gly.bris	Palaeontology Research Group, Earth Sciences Dept, Bristol
D	2068	geog.plym	Geography Dept, Plymouth	G	1885	eye.ox	Ophthalmology Dept, Oxford

Table 6-5. Selected five pairs of SCC subsites with start node in 'hum/soc' and end node in 'nat/tech'.

The order of start and end node was reversed for the five node pairs. In Table 6-6 the five pairs of subsites starting in 'hum/soc' and ending in 'nat/tech' were named HN01, HN02, etc. Correspondingly, the reversed pairs of subsites starting in 'nat/tech' and ending in 'hum/soc' were named NH01, NH02, etc.

path net	id	start node	id	end node
HN01	2099	hum.port	1904	atm.ox
HN02	2394	economics.soton	917	chem.gla
HN03	1494	psy.man	893	maths.gcal
HN04	871	speech.essex	245	palaeo.gly.bris
HN05	2068	geog.plym	1885	eye.ox
NH01	1904	atm.ox	2099	hum.port
NH02	917	chem.gla	2394	economics.soton
NH03	893	maths.gcal	1494	psy.man
NH04	245	palaeo.gly.bris	871	speech.essex
NH05	1885	eye.ox	2068	geog.plym

Table 6-6. Five pairs of SCC subsites from 'hum/soc' to 'nat/tech' (HN) and the same subsites in reversed order from 'nat/tech' to 'hum/soc' (NH).

The whole procedure thus resulted in a list of ten diversified topical pairs of subsites. The idea is now to extract all shortest link paths between these dissimilar pairs of seed subsites. For example, what is the link distance, i.e. the number of links on the shortest link path, between a department of geography and a department of ophthalmology (eye research) or vice versa, as in HN05 and NH05 – and what kind of subsites, pages and links appear on the shortest link paths functioning as connectors between the topically dissimilar start and end subsite nodes? Answering such questions could yield a better

understanding of how small-world phenomena emerge in an academic web space – and answer the overall research question in this dissertation.

The network analysis program *Pajek* was used to extract all shortest link paths between the five plus five pairs of subsite nodes. In order to achieve this, the original adjacency matrix was transformed to a matrix fit to Pajek. As mentioned earlier, the adjacency matrix could be treated as unweighted in Pajek when computing shortest paths even if the matrix contained page level link counts (cf. Section 4.2.5). This was a useful functionality in Pajek, as the link counts in the matrix otherwise would affect the shortest paths computations, giving priority to link paths containing low link counts.

The Pajek manual[43] does not explicitly mention what algorithm is used for extracting all shortest paths between two nodes. Presumably, it is a so-called *breadth-first search* based on Dijkstra's algorithm (Gross & Yellen, 1998)[44]. A breadth-first-search (BFS) in a directed graph like the present UK web graph starts at a given node *s* in the graph, then visits all out-neighbors of *s*, then their out-neighbors, etc., until all reachable nodes have been visited in the graph. Such a BFS performed by Dijkstra's algorithm solves the so-called *single source shortest path problem* by finding all shortest paths between a given node and *all* other nodes in the graph. The present study is concerned with a limited aspect of this problem, the so-called *single pair shortest path problem,* finding all shortest paths between a *single pair* of nodes, that is, between a given node *s* and another given node *t*. Counter-intuitively, it turns out that the single pair shortest path problem cannot be solved significantly more efficiently than the single source shortest paths problem (Kalorkoti, 2003). This means that a breadth-first-search may take as long time to find all shortest paths between a given node and *all* other nodes in the graph, as it takes to find all shortest paths between the given node and just one other given node in the same graph. This circumstance is due to the fact that the BFS sometimes has to visit all reachable nodes in the graph before it has identified all shortest paths between two given nodes.

Using the abovementioned adjacency matrix of the UK web graph, Pajek found all shortest link paths between the 10 given SCC subsite pairs in split seconds.

6.3.2 Resulting 10 path nets

Figures 6.11 and 6.12 show two of the resulting 10 subgraphs consisting of all shortest link paths between a pair of subsites with dissimilar topics. Such an 'all shortest paths' subgraph is here designated the brief term *path net*.[45] Path net HN05 in Fig. 6.11 consists of all shortest link paths between node 2068 (*geog.plym.ac.uk*), the geography department at the University of Plymouth and node 1885 (*eye.ox.ac.uk*), the ophthalmology (eye research) department in Oxford. The path net consists of four nodes

[43] The manual is included in the Pajek software, but can also be consulted on the Web: http://vlado.fmf.uni-lj.si/pub/networks/pajek/doc/pajekman.htm

[44] One of the two creators of Pajek, Andrej Mrvar, confirms that Dijkstra's breadth-first search algorithm is used in Pajek for finding all shortest paths between two given nodes (personal e-mail 23.1.2004).

[45] Searches in graph theoretic literature has not disclosed detailed terminology for a subgraph constituted by all shortest paths connecting a single pair of nodes. Fig. 3.12 in Section 3.4 shows a subgraph of all shortest paths in a semantic network (Steyvers & Tenenbaum, 2001).

connected by seven *subsite level* links (subsite-to-subsite connections)[46]. The path length, that is, the link distance, between the two seed subsites is 3. In other words, all link paths in the path net between the two departments have the *same* path length 3. The path length was longer (4) in path net NH05 with the reversed pair of start and end nodes, see Fig. 6.12. This path net contained 13 nodes and 18 subsite level links.

Figures of all 10 path nets are shown in Appendix 10, including the affiliations for all 141 subsites in the 10 path nets. Summary node data of the 141 subsites are listed in Appendix 11.

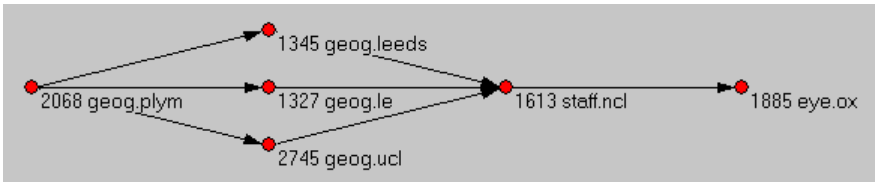


Figure 6.11. Path net HN05: all shortest link paths between *geog.plym.ac.uk* and *eye.ox.ac.uk*.

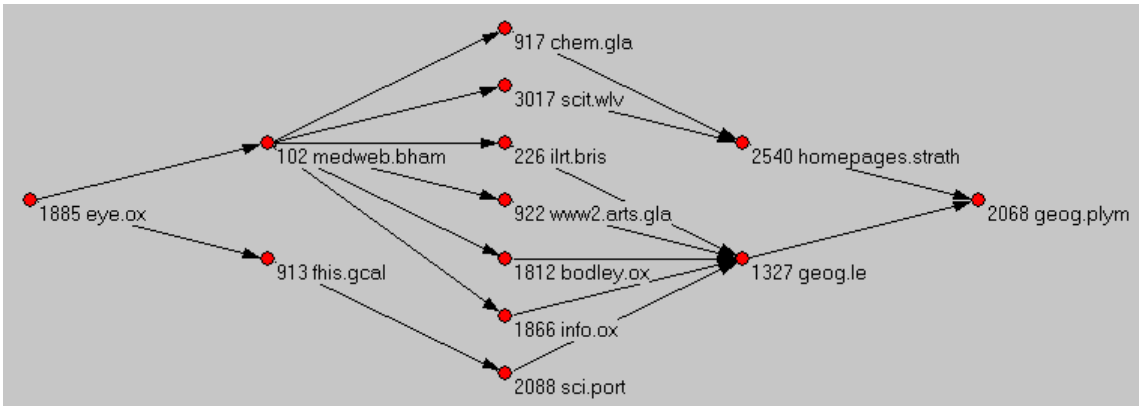


Figure 6.12. Path net NH05: all shortest link paths between *eye.ox.ac.uk* and *geog.plym.ac.uk*.

There were different intermediate nodes and topics in a set of reversed path nets. For example, only node 1327 (*geog.le.ac.uk*)[47], the Department of Geography, University of Leicester, occurred in both path nets HN05 and NH05 in the above figures. This issue is elaborated further below.

It should be noted, that when both a start and end nodes on a shortest link path belong to the SCC, all intermediary nodes on the link path inevitably also belong to the SCC. Even if SCC nodes may have links to the OUT component, no links can point the opposite way from OUT to SCC (cf. Section 5.1). Thus, all intermediary subsite nodes in the 10 path nets belonged to the SCC because so did the start and end nodes.

The term *path net level* is assigned to denote all nodes with the same link distance – that is, the same 'degrees of separation' (cf. Section 3.2) – from the start node. In Fig. 6.13, the four levels of path net HN01 are illustrated. For instance, the seven subsite nodes at level 2 in the figure all have link distance 2 from the start node 2099.

[46] One *subsite level* link may comprise several *page level* links, cf. Section 2.3.3 and Fig. 6.13.
[47] Primarily, stemmed canonical domain names are used in the dissertation, thus excluding any www-prefix or variant domain name of the investigated subsites (cf. Section 4.1.2 and Appendix 4).

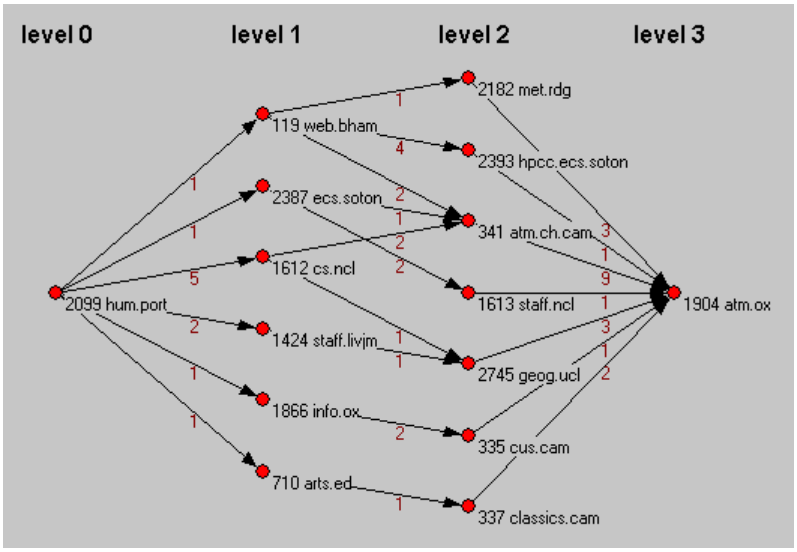


Figure 6.13. Levels in path net HN01. Counts of *page* level links denoted at the *subsite* level links.

The path nets were manually drawn using functions in the network analysis tool Pajek. The placement of nodes within each path net level was affected by an attempt to create easy-to-see graphs by reducing the number of crossing links in the path net.

path net	start node	end node	path length	# subsites	# subsite level links	subsite level links / subsite[48]
HN01	hum	atm	3	15	23	1.6
HN02	econ	chem	3	6	7	1.4
HN03	psy	math	4	8	12	1.7
HN04	speech	palaeo	3	4	3	1.0
HN05	geogr	eye	3	6	7	1.4
			3.2	*39*	*52*	*1.5*
NH01	atm	hum	3	15	25	1.8
NH02	chem	econ	4	28	53	2.0
NH03	math	psy	3	5	5	1.2
NH04	palaeo	speech	4	41	99	2.5
NH05	eye	geogr	4	13	18	1.5
			3.6	*102*	*200*	*2.1*
			3.4	**141**	**252**	**1.9**

Table 6-7. Summary of 10 path nets.

[48] The number of subsite level links per subsite node was calculated with n-1 nodes because there are n-1 outlinking nodes as well as n-1 inlinked nodes in a path net with n nodes. In the subtotals, n-5 nodes were used as denominator, and n-10 in the average count for all 10 path nets.

Table 6-7 above presents the path lengths and counts of nodes and subsite level links in all 10 path nets. Appendix 11 contains a summary table with more statistics (including measures from subsequent sections) on the subsites in the path nets.

The average path length in the 10 path nets was 3.4, close to the average path length of the whole UK academic subweb, 3.46 (cf. Section 5.3.1). It has not been feasible to compute the average path length within the SCC.

The variation of node and link counts is apparent in Table 6-7. The smallest path net, HN04, shown in Fig. 6.14 below, contained only one shortest link path, four subsite nodes and three subsite level links between the Palaeontology Research Group in Bristol (*palaeo.gly.bris.ac.uk*) and the Speech Research Group at the University of Essex (*speech.essex.ac.uk*). The largest path net, NH04, in Fig. 6.15 further below**,** contained all shortest link paths in the reverse direction between the Speech Research Group and the Palaeontology Research Group including 41 nodes and 99 subsite level links.

On average, there were 1.9 subsite level links per subsite node in the 10 path nets. In other words, a subsite had on average almost two in-neighbors and two out-neighbors in the path nets.

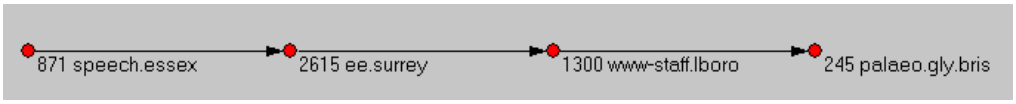


Figure 6.14. Path net HN04: all shortest link paths (path length 3) between *speech.essex.ac.uk* and *palaeo.gly.bris.ac.uk*.

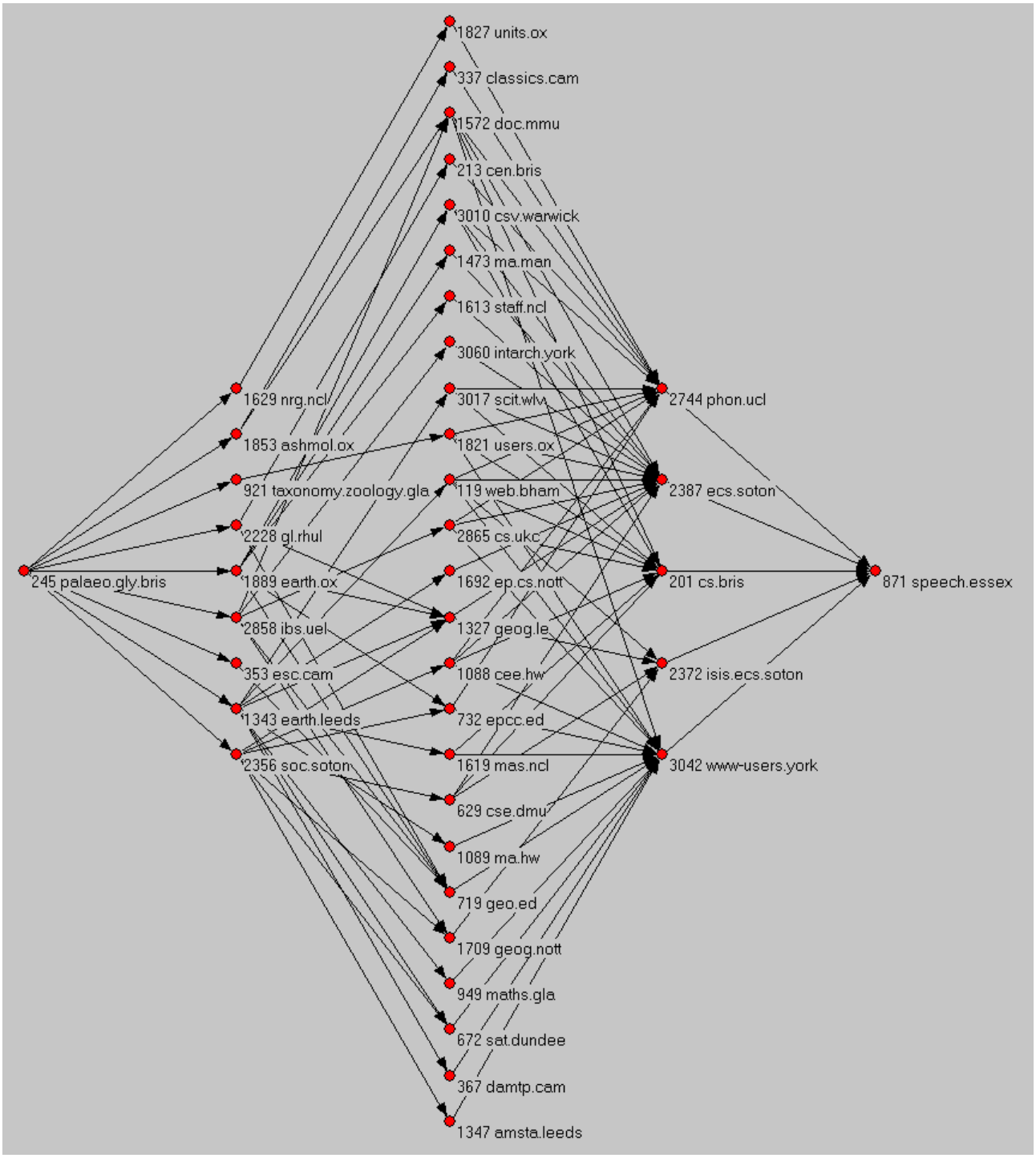


Figure 6.15. Path net NH04: all shortest link paths (path length 4) between *palaeo.gly.bris.ac.uk* and *speech.essex.ac.uk*.

There were 104 unique subsites among the 141 subsites in the 10 path nets. The 32 subsites present in more than one path net logically included the 10 seed start and end subsites that occurred in two path nets. One of the 10 seed subsites, node 917 (*chem.gla.ac.uk*) occurred in path net NH05 (Fig. 6.12) in addition to its 'own' two path

nets HN02 and NH02. The 22 multi-occurring 'non-seed' subsites are listed in Appendix 12. Four of these subsites occurred in three path nets, the rest in two. For example, node 1088 (*cee.hw.ac.uk*), Department of Computing and Electrical Engineering, Heriot-Watt University, occurred in path nets NH04 (Fig. 6.15) as well as in HN02, NH02. This phenomenon of multi-occurrence is related to the so-called *betweenness centrality* measure that reflects the probability that a node occurs on a shortest link path between two arbitrary nodes. This will be more elaborated further below in Section 6.3.2.4. Node 1088 mentioned above thus had the eighth highest betweenness centrality in the data set. Some of the multi-occurring subsites had a somewhat lower betweenness centrality like node 341 (*atm.ch.cam.ac.uk*), Centre for Atmospheric Science, University of Cambridge (cf. path net HN01, Fig. 6.13) with betweenness centrality rank 144. However, such subsites with lower betweenness centrality could play a significant role in local topic-specific web spaces. For instance, node 341 above could have such a position in web clusters concerned with the topic of atmospheric sciences, as indicated by its links to one of the seed subsites, node 1904 (*atm.ox.ac.uk*) in both path nets HN01 and NH01.

Furthermore, it may be added that 17 of the 104 unique path net subsites overlapped with the 189 sample SCC subsites in Section 6.2. Of these 17 overlapping subsites, logically 10 were the seed start and end subsites in the path nets. This overlap is illustrated in Fig. 6.10 at the start of Section 6.3 where the area of C partially overlaps area B.

It has not been possible to find any literature describing similarly close investigations of path nets in the shape of subgraphs of all shortest link paths between node pairs either on the Web or in other real-world networks.

In the next subsections a broad range of graph measures are applied on the 10 path nets in order to reveal as many facets and characteristics as possible regarding the cohesiveness and 'small-world-ness' of their connectivity structures – primarily dealing with research questions 1-3. The next subsections thus supplement the investigations undertaken in Chapter 5.

Section 6.3.2.1 treats so-called *in-distance* and *out-distance*. Distributions of path net *in-neighbors* and *out-neighbors* are measured in Section 6.3.2.2. So-called *assortative mixing* (Newman, 2002) related to the preceding measures of node neighbors is investigated in Section 6.3.2.3. An important measure with regard to small-world properties is the so-called *betweenness centrality* treated in Section 6.3.2.4. This centrality measure is examined in a novel way in relation to Kleinberg's (1999a) concepts of *hubs* and *authorities*. The section also includes identification of cluster-like neighborhoods, so-called *cores*. Finally, section 6.3.2.5 deals with *small-world co-linkage* in the shape of shortest paths along co-linkage chains (cf. Section 3.4) between start and end nodes in the 10 path nets.

6.3.2.1 In-distance and out-distance

Table 6-8 below brings a more detailed account of some parameters of the start and end nodes in the 10 path nets. As will appear from the table, the average path length of the HN path nets with link paths leading from 'hum/soc' to 'nat/tech' was somewhat lower, 3.2, than the average path length, 3.6, of the NH path nets with link paths in the

reversed topical direction. Considering the small sample size, this difference may be arbitrary. However, statistics from the sample of 189 SCC subsites may support the observation of longer path lengths when end nodes are 'hum/soc'. In Table 6-9, the so-called *average in-distance*[49] for the 36 'hum/soc' subsites in the sample are 3.38 as computed by Pajek, meaning that existing shortest link paths with such subsites as end nodes are on average 3.38 long in the UK data set. The average in-distance for the 119 'nat/tech' subsites was somewhat lower, 3.28. Confidence intervals have not been estimated due to non-trivial calculations for highly skewed distributions. Hence, the above differences in means should be treated with caution.

The picture is more opaque with regard to *average out-distance* of 'hum/soc' and 'nat/tech' start nodes in the 10 path nets. The path nets with the lowest average path length, i.e., the HN path nets passing from 'hum/soc' to 'nat/tech', even if their start nodes have the highest average out-distance, 3.43, meaning the shortest link paths between these five 'hum/soc' subsites and all other reachable subsites in the whole UK data set have an average path length of 3.43.

The small sample of 10 path nets does not provide sufficient evidence in order to decide whether *end* node in-distance is more important than *start* node out-distance in determining the shortest path length between start and end node. It has not been possible to find any related work discussing this matter.

path net	path length	id	start node	average out-distance	all out-neighbors of start node	path net out-neighbors	id	end node	average in-distance	all in-neighbors of end node	path net in-neighbors
HN01	3	2099	hum.port	3.09	16	6	1904	atm.ox	3.17	24	7
HN02	3	2394	economics.soton	3.57	2	1	917	chem.gla	2.83	39	3
HN03	4	1494	psy.man	4.28	1	1	893	maths.gcal	3.14	7	3
HN04	3	871	speech.essex	3.19	13	1	245	palaeo.gly.bris	3.24	17	1
HN05	3	2068	geog.plym	3.02	35	3	1885	eye.ox	3.75	2	1
	3.2			*3.43*	*13.4*	*2.4*			*3.23*	*17.8*	*3.0*
NH01	3	1904	atm.ox	2.93	48	8	2099	hum.port	3.27	17	5
NH02	4	917	chem.gla	2.82	46	15	2394	economics.soton	4.03	1	1
NH03	3	893	maths.gcal	3.75	1	1	1494	psy.man	3.04	9	2
NH04	4	245	palaeo.gly.bris	3.52	15	9	871	speech.essex	3.24	5	5
NH05	4	1885	eye.ox	3.80	2	2	2068	geog.plym	3.49	4	2
	3.6			*3.36*	*22.4*	*7.0*			*3.41*	*7.2*	*3.0*
	3.4			**3.40**	17.9	4.7			**3.32**	12.5	3.0

Table 6-8. Average out-distance, all out-neighbors, and path net out-neighbors of start nodes. Average in-distance, all in-neighbors, and path net in-neighbors of end nodes.

[49] The graph theoretic terms *in-distance* and *out-distance* are from Botafogo *et al.* (1992).

data	meta topic	# sub-sites	average in-distance	average out-distance	in-neigh-bors	out-neigh-bors
all 1893 SCC subsites	-	1893	3.30	3.38	18.1	23.6
sample of 189 SCC subsites	generic	34	3.33	3.36	18.9	21.5
	hum/soc	36	3.38	3.52	8.8	10.6
	nat/tech	119	3.28	3.38	21.2	23.0
		189	**3.31**	**3.40**	**18.4**	**20.4**
seed subsites in 10 path nets	hum/soc	5	3.41	3.43	7.2	13.4
	nat/tech	5	3.23	3.36	17.8	22.4

Table 6-9. Average in-distance and out-distance in 189 sampled SCC nodes and in 10 path nets depending on subsite meta topic.

The average in-distance and out-distance for all 1893 SCC subsites is shown in Table 6-9 above. Note the average in-distance for SCC subsites includes shortest link paths originating in the IN component. Correspondingly, the average out-distance for SCC subsites includes shortest link paths ending in the OUT component.

As shown in Table 6-9, an average 'hum/soc' subsite in the SCC sample had 8.8 in-neighbors and 10.6 out-neighbors, whereas an average 'nat/tech' subsite had over twice as high count of 21.2 in-neighbors and 23.0 out-neighbors. This divergence between 'hum/soc' and 'nat/tech' is logically present in the average counts of neighbors to the five plus five seed nodes in the 10 path nets as listed in the table. Recapitulating, the seed nodes were selected by stratified sampling from the SCC sample.

The counts of out-neighbors of start nodes and in-neighbors of end nodes in the 10 path nets are listed in Table 6-8 further above because of the intuition that these properties of start and end nodes may influence possible routes of shortest paths in a graph. Naturally, the same properties of intermediate nodes also influence possible routes. For example, a start node could have just one out-neighbor, but this out-neighbor could in turn have hundreds of out-neighbors influencing the nodes that can be reached from the start node. However, the sample is too small to deduce any significant findings. Future studies on larger samples of path nets could yield more interesting results. Appendix 11 contains data on in-distance and out-distance for all 141 subsites in the 10 path nets.

6.3.2.2 Path net in-neighbors and out-neighbors

As described above, the subsite nodes had different sets of in-neighbors and out-neighbors in the UK data set affecting the pattern of different link paths that could pass the subsites in a path net. The brackets in Fig. 6.16 display the total number of in-neighbors and out-neighbors of each subsite in the data set. For instance, node 1327 had 91 in-neighbors and 175 out-neighbors in the UK data set. Furthermore, node 1327 has five in-neighbors *within* path net NH05 marked with *white* nodes in the figure. As illustrated in Fig. 6.17, node 102 has six out-neighbors within the same path net NH05, some of which overlapping with the in-neighbors of node 1327, thus enabling four link paths between the two nodes.

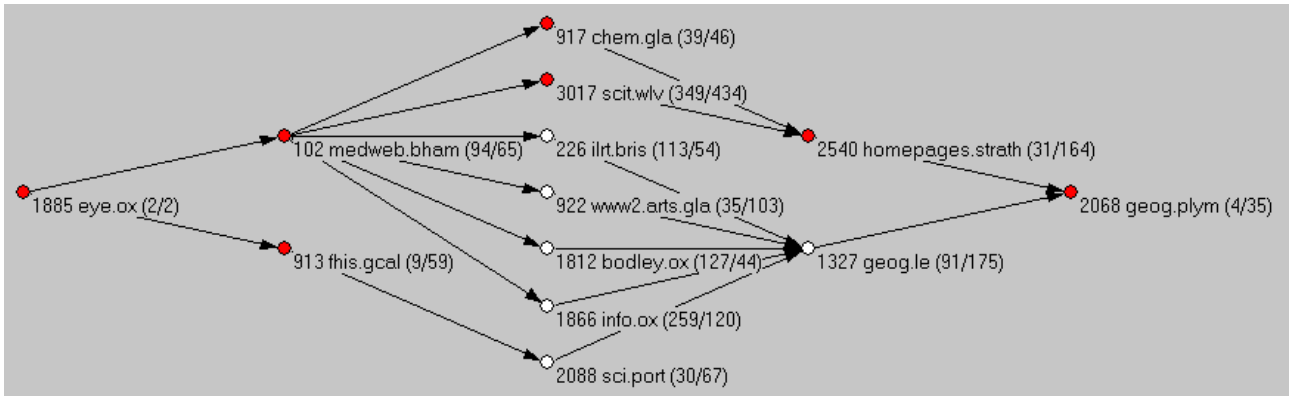


Figure 6.16. Node 1327 has five in-neighbors *within* path net NH05 (white nodes). Brackets display the number of in-neighbors / out-neighbors of each subsite in the whole UK data set.

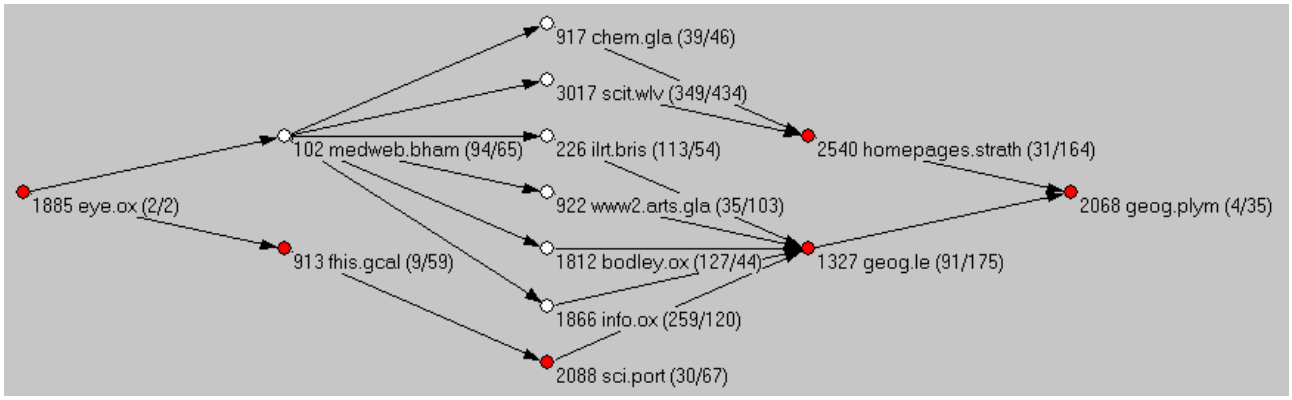


Figure 6.17. Node 102 has six out-neighbors *within* path net NH05 (white nodes).

# path net *in-neighbors*	freq.	%
0	10	7.1
1	95	67.4
2	16	11.3
3	5	3.5
4	4	2.8
5	4	2.8
7	1	0.7
9	1	0.7
10	1	0.7
11	2	1.4
13	2	1.4
	141	**100.0**

Table 6-10. Frequency distribution of in-neighbors *within* the 10 path nets for 141 subsite nodes.

# path net *out-neighbors*	freq.	%
0	10	7.1
1	89	63.1
2	14	9.9
3	13	9.2
4	5	3.5
5	2	1.4
6	3	2.1
7	1	0.7
8	1	0.7
9	2	1.4
15	1	0.7
	141	**100.0**

Table 6-11. Frequency distribution of out-neighbors *within* the 10 path nets for 141 subsite nodes.

Tables 6-10 and 6-11 above show the counts of in-neighbors and out-neighbors *within* the 10 path nets for all 141 subsite nodes. For example, 95 (67.4%) of the 141 subsites had only one in-neighbor within the path nets, and two subsites had 13 in-neighbors.

The tables show that 25.3% of the nodes have more than one in-neighbor, and 29.6% of the nodes more than one out-neighbor. Logically, start nodes have no in-neighbors and end nodes no out-neighbors in the path nets, hence the zeroes in the two tables.

Interestingly, both frequency distributions follow power laws as tested by the *LOTKA* software program (Rousseau & Rousseau, 2000) mentioned earlier in Section 4.2.1.[50]

Appendix 11 contains data on in-/out-neighbors and in-/outlinks both within the path nets and within the UK subweb for all 141 subsites in the 10 path nets.

6.3.2.3 Assortative mixing

The so-called *assortative mixing* of the nodes in the 10 path nets was investigated because this measure relates to the distribution of in-neighbors and out-neighbors. According to Newman (*2002*), a network shows assortative mixing if nodes with many connections tend to be connected to other nodes with many connections. In other words, in an assortatively mixed network, nodes with many neighbors tend to connect to nodes also with many neighbors. Newman shows that social networks are often assortatively mixed, but technological and biological networks tend to be disassortative. The measure used by Newman is the Pearson correlation coefficient of the connectivity degrees of pairs of nodes connected by an edge.[51]

This measure was calculated for the 252 different pairs of subsites (cf. 252 subsite level links, Table 6-7, Section 6.3.2) interconnected in the 10 path nets. The connectivity degree of each node comprises the sum of in-neighbors and out-neighbors of the node. For example, in Fig. 6.18 below, the connectivity degree of node 1885 is 4, and 783 for node 3017.

The Pearson correlation measure for the 252 subsite pairs in the 10 path nets was 0.063, close to the measure -0.065 calculated by Newman (ibid.) based on a large sample of 269,504 web pages[52]. According to Newman, the low correlation measure reflects the *lack* of assortative mixing in networks on the Web. In other words, web nodes with high connectivity degrees do *not* tend to connect to other nodes with many connections. The indicative finding in the small sample of 10 path nets in the present study thus supports Newman.

---

[50] If $f_{in}(x)$ denotes the relative number of subsites in the path nets with x in-neighbors within the path nets, then $f_{in}(x) = 0.7138x^{-2.3619}$ for the 10 path nets according to the *LOTKA* software (Rousseau & Rousseau 2000). The corresponding power law f_{out} for out-neighbors in the 10 path nets was $f_{out}(x) = 0{,}6903x^{-2{,}2704}$.

[51] Usually, the Spearman rank correlation is regarded a better correlation test than the Pearson correlation when frequency distributions are not assumed normal but are skewed as in the present link data. Nevertheless, Newman used the Pearson test – also on web link data. So, for sake of comparability, this correlation test will also be used in the present study.

[52] Based on web data from Barabási & Albert (1999).

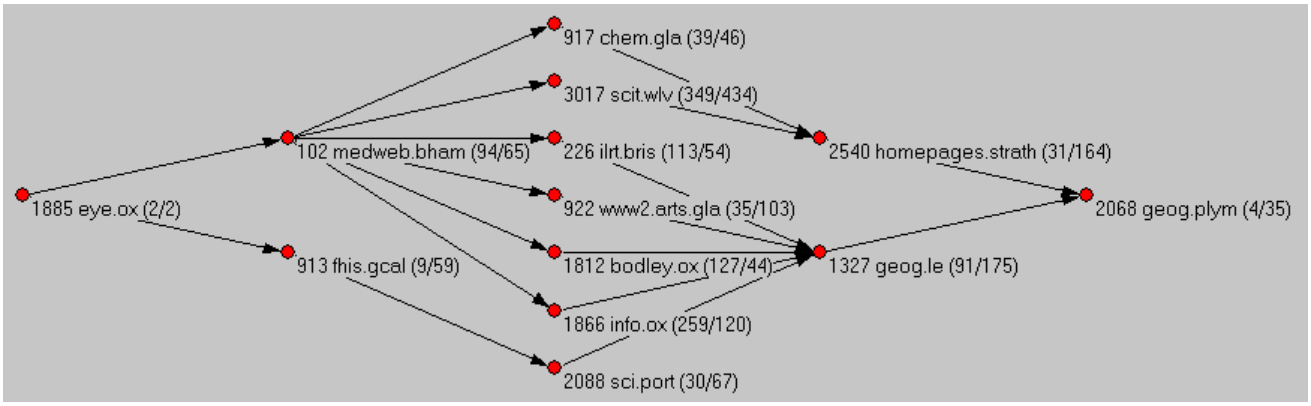


Figure 6.18. Path net NH05. In brackets is the total number of in-neighbors / out-neighbors of each subsite in the investigated UK subweb.

It should be noted that the present calculation was based on node pairs in the 10 confined path net *subgraphs*, whereas Newman's calculations were based on connectivity patterns in *whole* graphs. However, it was not feasible to compute this measure on the whole UK subsite graph.

6.3.2.4 Betweenness centrality, cores and hubs/authorities

This section contains terminology about shortest link paths and node neighborhoods that makes this placing more logical than together with other graph measures in Chapter 5. The section contains statistics on both the whole UK data set of 7669 subsites and the 10 path nets (cf. Appendix 11 for summary node data for the 10 path nets).

An important graph measure with regard to small-world properties is the so-called *betweenness centrality* because it quantifies how many shortest paths pass through each node in a graph. More precisely, the betweenness centrality of a node is a measure giving the probability that the node will occur on a shortest path between two arbitrary nodes in the graph (Freeman, 1977).

The measure of betweenness centrality can be traced back to Bavelas (1948) and was developed in social network analysis and graph theory (cf. Section 3.1). According to Freeman (1977), "a point in a communication network is central to the extent that it falls on the shortest path between pairs of other points" because such a point can function as a broker or gatekeeper in control of information passing between other nodes in the network. In a web graph, betweenness centrality is not concerned with control of information transfers but deals with how web nodes can be accessed and traversed by human web surfers and digital web crawlers.

Pajek was used to calculate the betweenness centrality of all 7669 subsite nodes in the UK data set. In Fig. 6.19, the power-law-like distribution of betweenness centrality is shown. In future studies, it would be interesting to identify possible factors behind the break in the curve below betweenness centrality 0.0001 – and whether a similar break occurs in the distribution of betweenness centrality in other web spaces. One possible factor behind the break may be the exclusion of links to and from main university sites in the delimited data set possibly making some betweenness centrality measures lower

than otherwise. It has not been tractable to calculate the betweenness centrality for the *undelimited* UK data set.

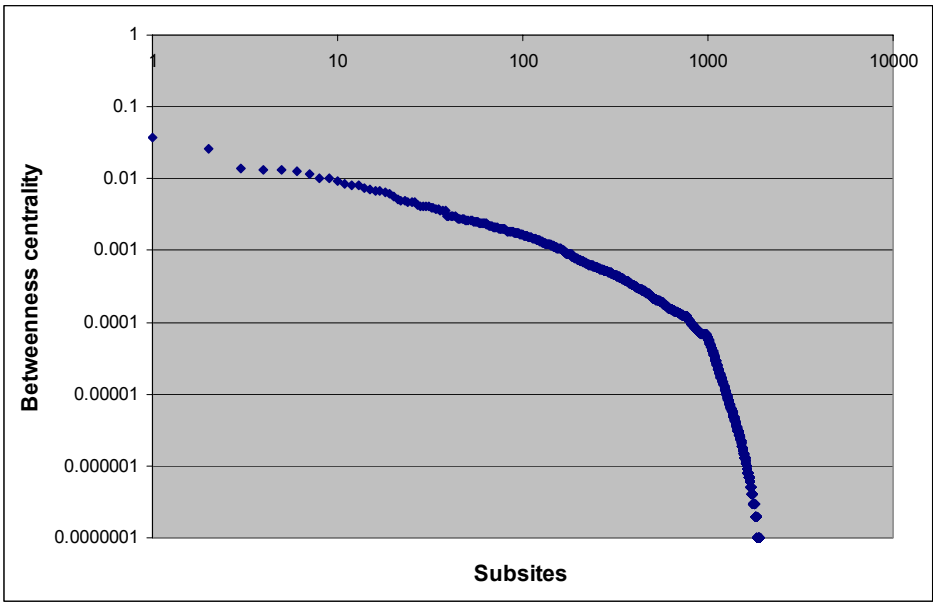


Figure 6.19. Power-law-like distribution of betweenness centrality for 1914 subsites with betweenness centrality > 0 in the UK data set. Log-log scale.

Of the 7669 subsites, 5755 (75.0%) had betweenness centrality 0, that is, they did not occur on any shortest link paths between other pairs of subsites in the UK data set. This result was not unexpected due to the high percentage of subsites not having both inlinks and outlinks (cf. Tables 5-7 and 5-8, Section 5.4). Logically, both inlinks and outlinks are necessary if a node shall function as a *connector* node being both a receiver and provider of links on a shortest link path. However, having both inlinks and outlinks is no guarantee for occurring on any shortest paths. For instance, 75 of the 1893 SCC subsites had betweenness centrality 0. These 'bypassed' SCC subsites all had low counts (≤ 5) of in-neighbors and out-neighbors.

The highest measure of betweenness centrality, 0.0369, in the delimited data set of subsites belongs to node 1821 (*users.ox.ac.uk*) containing staff and student personal web pages at the University of Oxford (cf. node 1821 at level 2 in path net NH04 in Fig. 6.15, Section 6.3.2). In other words, on an arbitrary shortest link path between two arbitrary subsites in the data set, there was 3.7% probability that the shortest path would pass node 1821. Node 1821 had outlinks to 330 different subsites and received inlinks from 507 different subsites in the UK data set.

In Table 6-12 below, the 25 subsites with the highest betweenness centrality in the UK data set is listed, all belonging to the SCC component as could be expected. Of the 25 subsites, 15 were computer-science-related as will appear from the affiliations in the table, reflecting the influence such subsites have on the connectivity patterns of an academic web space. This matter is further elaborated in Section 6.5.4 in relation to subsites providing transversal links in small-world link structures. Table 6-12 also shows the path nets that contained these betweenness-central subsites. The table

indicates a relation between betweenness centrality and Kleinberg's (1999a) concepts of hubs and authorities. This point is elaborated further below in this section.

domain name	id	com-pon-ent	path nets	between-ness centrality	bc rank	core	in-neigh -bors	out-neigh -bors	hub /auth.	affiliation
users.ox	1821	SCC	NH02+04	0.0369	1	53	330	507	H/A	Personal web pages at Oxford Univ.
scit.wlv	3017	SCC	NH04+05	0.0265	2	53	349	434	H/A	School of Computing and Information Technology, Univ. of Wolverhampton
cs.ucl	2760	SCC	NH02+03	0.0140	3	53	300	265	H/A	Dept. of Computer Science, Univ. College London
dcs.gla	925	SCC		0.0130	4	53	203	511	H	Dept. of Computing Science, Univ. of Glasgow
doc.mmu	1572	SCC	NH04	0.0130	5	53	127	514	H	Dept. of Computing and Mathematics, Manchester Metropolitan Univ.
cbl.leeds	1357	SCC	NH01+02	0.0130	6	53	387	91	A	Computer Based Learning Unit, Univ. of Leeds
csv.warwick	3010	SCC	NH04	0.0115	7	53	202	354	H	Alias server for Univ. of Warwick main site
cee.hw	1088	SCC	HN02 NH02+04	0.0104	8	53	148	518	H	Dept. of Computing and Electrical Engineering, Heriot-Watt Univ.
cogs.susx	2642	SCC	NH02	0.0103	9	53	231	268	H/A	School of Cognitive and Computing Sciences, Univ. of Sussex
comp.lancs	1268	SCC	NH02+03	0.0091	10	53	183	265	H	Computing Dept., Univ. of Lancaster
info.ox	1866	SCC	HN01 NH05	0.0083	11	53	259	120		Former server with official web pages of Univ. of Oxford
ee.surrey	2615	SCC	HN04 NH01	0.0081	12	53	187	213		Electronic Engineering Dept., Univ. of Surrey
ecs.soton	2387	SCC	NH01 NH04	0.0079	13	53	117	327	H	Dept. of Electronics and Computer Science, Univ. of Southampton
cl.cam	325	SCC		0.0074	14	53	246	141	A	Computer Laboratory (=Computer Science Dept.), Univ. of Cambridge
cs.man	1466	SCC	HN03 NH02	0.0071	15	53	218	224	A	Dept. of Computer Science, Univ. of Manchester
dcs.ed	772	SCC		0.0068	16	53	191	185		Dept. of Computer Science, Univ. of Edinburgh
dai.ed	791	SCC	HN03 NH04	0.0066	17	53	137	280		Dept. of Artificial Intelligence, Division of Informatics, Univ. of Edinburgh
ma.hw	1089	SCC	NH02+04	0.0065	18	53	191	186		Dept. of Mathematics, Heriot-Watt Univ.
dcs.napier	1597	SCC	NH02	0.0061	19	53	127	226		School of Computing, Napier Univ.
www-users.york	3042	SCC	NH04	0.0056	20	53	94	226		Personal web pages at Univ. of York
biochem.ucl	2756	SCC		0.0052	21	46	98	136		Dept. of Biochemistry and Molecular Biology, Univ. College London
ch.ic.ac.uk	1148	SCC		0.0050	22	46	131	100		Dept. of Chemistry, Imperial College, London
icbl.hw.ac.uk	1098	SCC		0.0050	23	52	200	60		Institute for Computer Based Learning, Heriot-Watt Univ.
damtp.cam.ac.uk	367	SCC	NH04	0.0048	24	53	137	147		Dept. of Applied Mathematics & Theoretical Physics, Univ. of Cambridge
ukoln.bath.ac.uk	91	SCC		0.0048	25	45	123	90		UK Office for Library and Information Networking, hosted at Univ. of Bath
src.doc.ic	4124	OUT		0	4124	53	411	0	A	Sun Site Northern Europe mirror site
sunsite.doc.ic	4129	OUT		0	4129	53	331	0	A	Sun Site Northern Europe mirror site
comlab.ox	4928	OUT		0	4928	53	289	0	A	Computing Laboratory (=Computer Science Dept.), Univ. of Oxford

Table 6-12. 25 subsites with highest betweenness centrality also include the 10 strongest hubs (H) and 7 strongest authorities (A) including 4 combined (H/A) among 7669 subsites. However, 3 of the strongest authorities had betweenness centrality 0.

Pajek was used to compute so-called *k-cores* for all the 7669 subsites in order to identify what kind of cluster-like neighborhoods the subsites belonged to. A *k-core* is a subgraph of a given graph where each node in the core has at least *k* neighbors in the

same core (Scott, 2000). A k-core is thus an area of relatively high cohesion within a graph.

As shown in Table 6-12 above, the top 20 subsites with highest betweenness centrality all belonged to the same 53-core. The 53-core contained 113 subsite nodes with at least 53 out- or inlinks to or from other nodes in the core. This was the most interconnected core in the whole data set. Using a term from the previous section, one can thus say that the 53-core indicates that subsites with high betweenness centrality show high *assortative mixing* by tending to link to other subsites with high betweenness centrality. It has not been feasible to verify this hypothesis due to lack of tractable data and programming facilities.

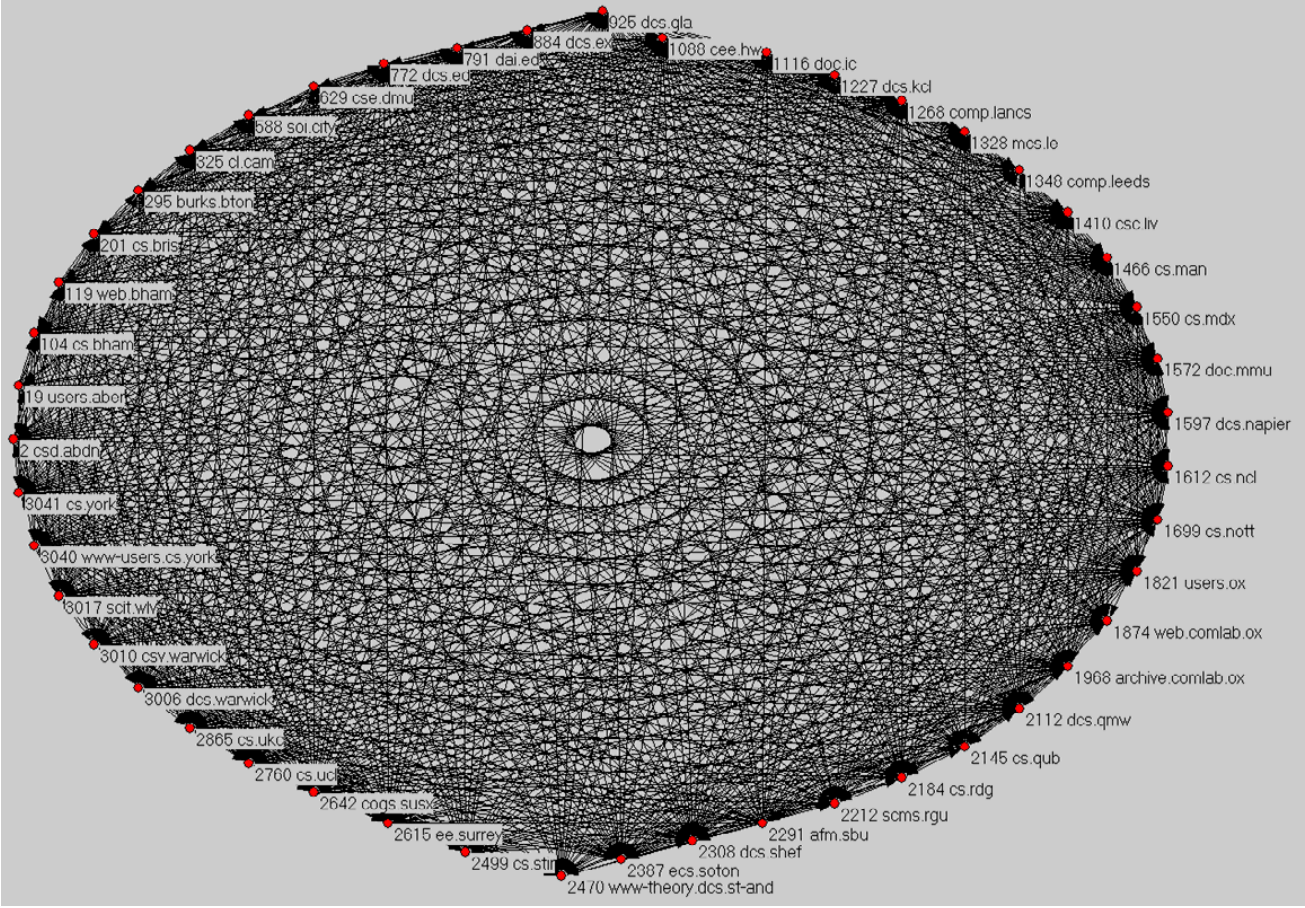


Figure 6.20.* ***Out-23-core*** **containing 47 subsite nodes with at least** ***23 outlinks*** **to other core nodes.**

The 53-core will not be shown here because of the many nodes in the core. However, Fig. 6.20 above gives an impression of a core containing 47 nodes with at least 23 outlinks to the other core nodes (inlinks were thus not counted in this special core variant). Many of the subsites in this so-called *out-23-core* were related to computer science and overlapped the 53-core.

As expected, many of the subsites with high betweenness centrality rank (bc rank) occurred in the 10 path nets as shown in Table 6-12 above. For example, node 1088 (*cee.hw.ac.uk*) – with bc rank 8 – occurred in three path nets, HN02, NH02, and NH04.

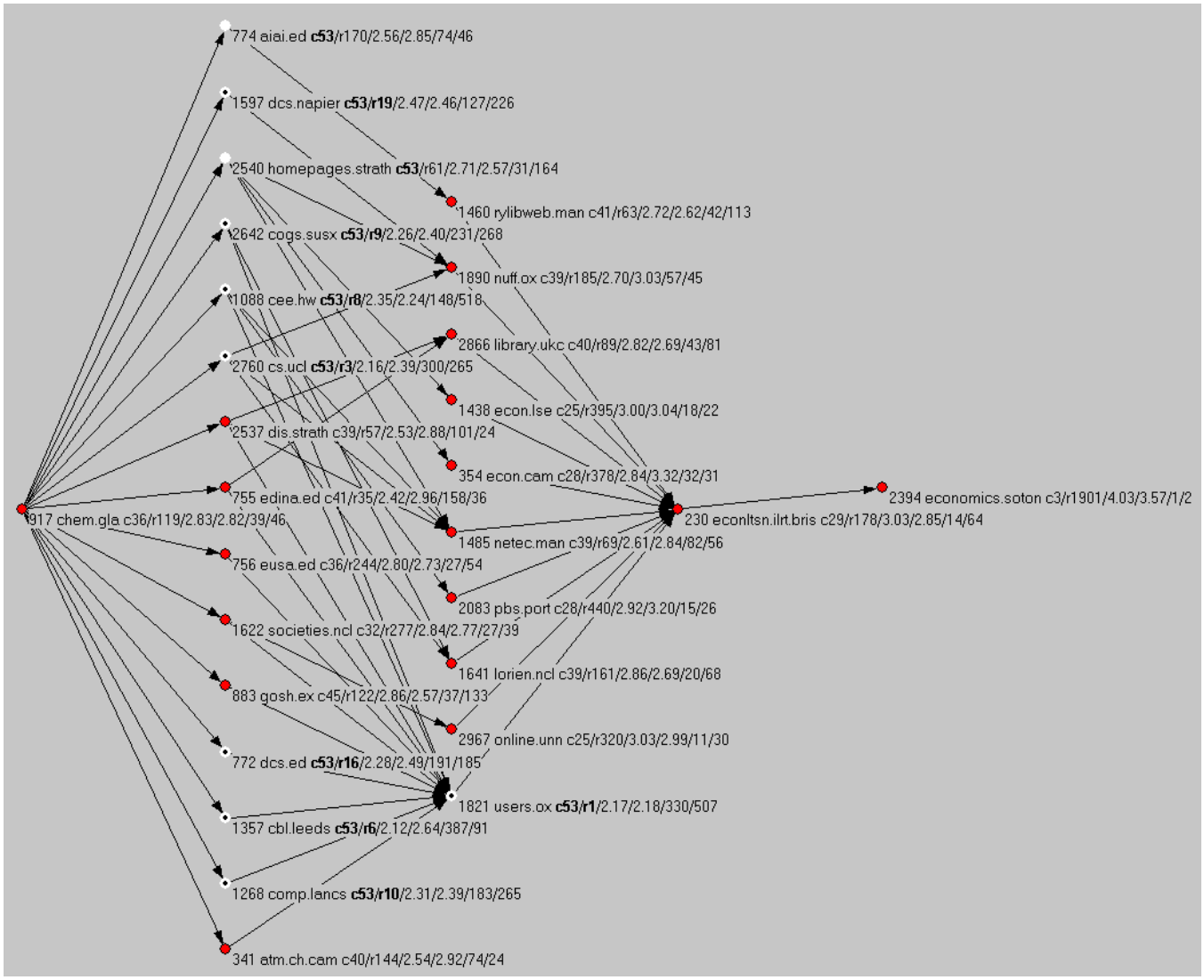


Figure 6.21.* Path net NH02 with graph measures in a string at each subsite node showing core (c), betweenness centrality rank (r), average in-distance/out-distance, and number of in-neighbors/out-neighbors. Subsites belonging to the 53-core are marked in white, and subsites with bc rank < 25 are marked in white with a black spot. See Appendix 10 for affiliations.

Fig. 6.21 above shows path net NH02 with graph measures grouped in a string at each subsite node showing *core* (c), *betweenness centrality rank* (r), average *in-distance/out-distance*, and number of *in-neighbors/out-neighbors*. For example, the string c53/r170/2.56/2.85/74/46 at node 774 at the top of level 1 in the figure gives data on core (c53), bc rank (r170), average in-distance/out-distance (2.56/2.85)[53], and number

[53] Recapitulating Section 6.3.2.1; an average in-distance of 2.56 means that the shortest link paths from other subsites in the whole UK data set that could reach this node have an average path length of 2.56. An average out-distance of 2.85 means that the shortest link paths between this node and all other reachable subsites have an average path length of 2.85. The reason for making path nets including these measures was an idea to examine how they relate to the formation of path nets as mentioned in Section 6.3.2.1. However, this line of inquiry has not been completed.

of in-neighbors/out-neighbors (74/46). In the figure, subsites belonging to the 53-core are marked in white, and subsites with bc rank < 25 are marked in white with a black spot. Of the 141 path net nodes, 52 (36.9%) belonged to the 53-core indicating the importance of this core for providing short distances in the UK academic subweb. In future studies it would interesting to test other clustering measures on an academic web space, including co-linkage, that is, co-citation and bibliographic coupling (cf. next Section 6.3.2.5). Pajek had no facilities for this kind of clustering.

Appendix 11 contains data on betweenness centrality, cores, etc., for all 141 subsites in the 10 path nets.

The notion of betweenness centrality seems to be related to the notion of *hubs* and *authorities* introduced by Kleinberg (1999a). Hubs and authorities on the Web are defined recursively: a web node, such as a web page or a web site, is an *authority* if it receives inlinks from many hubs, and is a *hub* if it provides outlinks to many authorities (ibid.). As noted by Otte & Rousseau (2002), the Kleinberg approach of hubs and authorities is thus related to the influence weight citation measure proposed by Pinski & Narin (1976) and mimics the idea of 'highly cited documents' (authorities) and reviews (hubs) in scholarly literatures.

Pajek had a function for identifying hubs and authorities in a network by using Kleinberg's algorithm. As shown in Table 6-12 above, the 10 strongest hubs and 10 strongest authorities (including 4 combined hubs/authorities) in the data set tend to have high betweenness centrality, that is, they have high probability of occurring on shortest paths connecting nodes in the graph. However, as shown in the table, three authority subsites belonging to the OUT component had betweenness centrality 0. On the other hand, they also had high counts of in-neighbors presumably including inlinks from hubs, as suggested by the circumstance that the strongest hubs and authorities all belonged to the 53-core as shown in the table.

The relation between hubs/authorities and betweenness centrality has not been found discussed in any literature. However, intuitively it makes sense that nodes with high betweenness centrality that function as connectors on many shortest paths between nodes in a network also should tend to be hubs and authorities in the network. This hypothesis remains to be verified.

6.3.2.5 Co-linkage chains

Pajek was used to find all shortest paths consisting of *co-linkage chains* (cf. Section 3.4) in the shape of co-linked nodes (analogous to co-citations) or co-linking nodes (bibliographic couplings) between start and end nodes in the 10 path nets. The computation was based on a specially constructed adjacency matrix representing which nodes were adjacent through co-linkage. All 7669 subsites in the delimited data set were included.

The bi-directional block arrows in Fig. 6.22 shows how the start and end nodes 2394 and 917 in path nets HN02 and NH02 are connected by a shortest path comprising a chain of co-linked nodes, 2394 - C_j - 917. The chain length is 2. Subsites represented by B_i and B_{i+1} function as source nodes generating the co-linked chain. As shown in Table 6-13 further below, there were 34 different subsites functioning as intermediate

nodes C_j on the co-linked chain (the listed 36 nodes in the table includes the two uttermost nodes 2394 and 917).

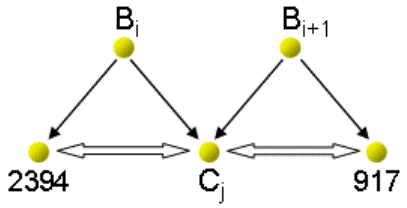


Figure 6.22. Shortest path (bi-directional block arrows) along chain of *co-linked* nodes between start and end nodes 2394 and 917 in path net HN02 and NH02. Chain length 2.

Finding the shortest *co-linked chain* between two nodes C_{j-2} and C_{j+1} (Fig. 6.23a) is analogous of finding the shortest link path consisting of alternate *inlinks* and *outlinks* in precisely that order between the two nodes as illustrated by Fig. 6.23c. Correspondingly, finding the shortest *co-linking chain* between two nodes B_{i-1} and B_{i+1} (Fig. 6.23a) is analogous of finding the shortest link path consisting of alternate *outlinks* and *inlinks* in that order between the two nodes as illustrated by Fig. 6.23b.

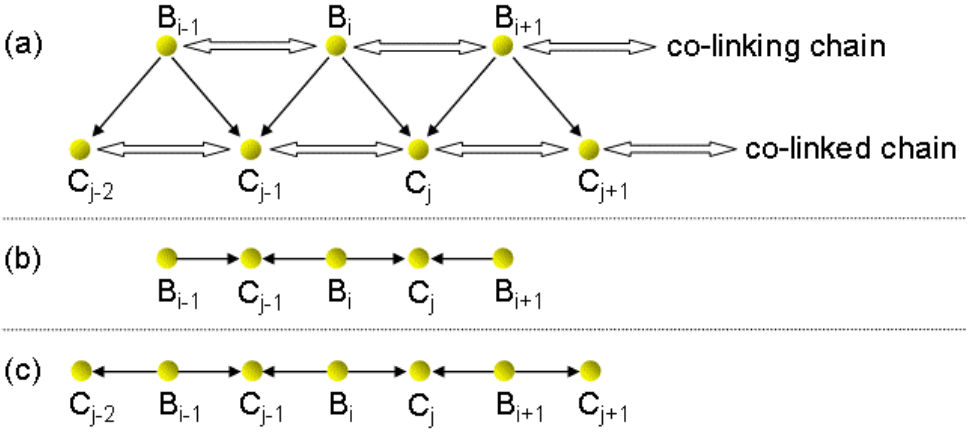


Figure 6.23a-c. (a) Co-linkage chain comprising interwoven co-linking chain and co-linked chain; (b) path of alternate outlinks and inlinks in that order generating co-linking chain in (a); (c) path of alternate inlinks and outlinks in that order generating co-linked chain in (a).

As noted earlier in relation to Fig. 3.10 in Section 3.4, there is a close connection between a co-linked chain and the corresponding co-linking chain, as also illustrated in Fig. 6.23a above: The two chains generate each other. For example, the source nodes in the co-linking chain between B_{i-1}, B_i, and B_{i+1} generate the co-linked chain between target nodes C_{j-1} and C_j. The reverse line of argument is also valid, with the co-linked chain as generator of the co-linking chain.

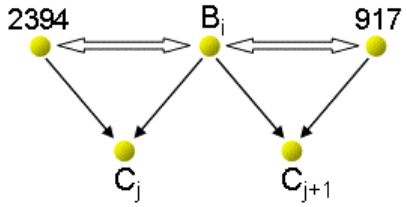


Figure 6.24. Shortest path (bi-directional block arrows) along chain of *co-linking* nodes between start and end nodes 2394 and 917 in path net HN02 and NH02. Chain length 2.

The *co-linking* chain between nodes 2394 and 917 in Fig. 6.24 above has the same chain length 2 as the corresponding *co-linked* chain in Fig. 6.22. There were 225 different subsites functioning as the intermediate node B_i. The counts or id numbers of the source nodes B_i and B_j of the co-linked chain in Fig. 6.22 or the corresponding data of the target nodes C_j and C_{j+1} of the co-linking chain in Fig. 6.24 were not directly available in Pajek. Only time-consuming extractions and comparisons of overlapping in-neighbors of the co-linked chain nodes and overlapping out-neighbors of the co-linking chain nodes yielded such data. However, it is beyond the scope of the present study to present the resulting data here.

					shortest co-linkage chain			
					co-linked		co-linking	
path nets	id			id	chain length	chain nodes	chain length	chain nodes
HN/NH01	2099	hum	atm	1904	1	2	2	444
HN/NH02	2394	econ	chem	917	2	36	2	227
HN/NH03	1494	psy	math	893	2	155	2	7
HN/NH04	871	speech	palaeo	245	2	156	2	118
HN/NH05	2068	geogr	eye	1885	2	55	2	58

Table 6-13. Shortest co-linked chains and co-linking chains in the path nets.

The co-linkage chains are mentioned here because they provide an interesting indication that the 5 pairs of seed subsites were from dissimilar topics as supposed. This is based on a non-verified assumption that two subsites on similar topics would probably be either co-linked or co-linking in the strongly connected SCC component of a relatively limited web space, as the present one. Lack of direct co-linkage may thus be utilized as a *measure of dissimilarity*. However, large-scale studies are needed to test this hypothesis.

Naturally, the reversed argument is not valid; two subsites do not have to be on similar topics just because they are co-linked or co-linking (cf. Thelwall & Wilkinson, forthcoming)[54]. For example, a single topically diversified link list or two different pages at node B_i in Fig. 6.25 below may point to a node in humanities (node 2099) and a node in atmospheric physics (node 1904), respectively.

[54] Also cf. Section 7.1.2, where Thelwall & Wilkinson (forthcoming) paper is discussed in connection to how they investigated topical similarities between 500 random sites and subsites in the UK academic web space with regard to three linkage types: direct links, co-inlinks (co-citation) and co-outlinks (bibliographical coupling).

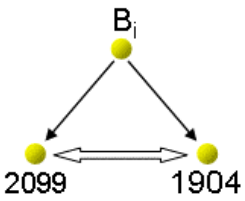


Figure 6.25. Shortest path (bi-directional block arrows) along chain of co-linked nodes between start and end nodes 2099 and 1904 in path net HN01 and NH01. Chain length 1.

As will appear from Table 6-13 above, this was the only pair of seed subsites directly co-linked. None of the other seed subsites were directly co-linked or co-linking but all were separated by co-linkage chains of chain length 2, thus suggesting dissimilar topics.

6.3.2.6 Summary

The previous subsections have yielded interesting findings briefly summarized in the list below (brackets show the subsections concerned). The findings are primarily concerned with answering the *first* research question in the dissertation:

How cohesively interconnected are link structures in an academic web space?

- *Power-law-like* distributions of in-neighbors/out-neighbors within the 10 path nets. (Section 6.3.2.2);
- *Power-law-like* distribution of *betweenness centrality* in the investigated UK web space. (Section 6.3.2.4);
- Indication of close relation between Kleinberg's (1999a) concepts of *hubs and authorities* on the Web and the *betweenness centrality* measure. No literature has been found discussing such a relation. (Section 6.3.2.4);
- Web nodes with high *betweenness centrality* may have high *assortative mixing* by tending to link to other nodes with high betweenness centrality. This hypothesis remains to be verified. (Section 6.3.2.4);
- Low correlation measure indicates a *lack of 'assortative mixing'*: web nodes with high connectivity degrees (many in-neighbors and out-neighbors) do *not* tend to connect to other nodes with many connections. This finding yields indicative support to Newman's (2002) finding regarding no assortative mixing in networks on the Web (Section 6.3.2.3).

Section 6.6 gives a summary of findings from all the empirical chapters and sections. Table 6-14 below gives a summary of some key figures of the four data levels of web nodes investigated so far in this study. Logically, there is a large similarity between the random SCC sample and the whole SCC. Besides this fact, the more 'zoomed-in' the data level is in the data set, with the 10 path nets as the most 'zoomed-in', there is a clear tendency that there will be shorter domain names, older subsites, higher betweenness centrality, shorter in-distance and out-distance, more in-neighbors and out-neighbors, more in-links and out-links as will appear from the table. This is logical,

since the employed 'zoom-in' methodology entails a condensation of the mentioned parameters. By definition, path net nodes (because they occur on shortest link paths) have larger betweenness centrality and shorter average in-distances and out-distances than average nodes in the overall graph.

	# sub-sites	domain name segments	average first time indexed in IA	between ness centrality	in-dist	out-dist	in-neigh-bors	out-neigh-bors	in-links	out-links
10 path nets	141	4.08	17.11.1997[55]	0.003612	2.74	2.82	83.9	113.9	351.9	419.3
SCC sample	189	4.26	02.10.1998[56]	0.000358	3.31	3.40	18.4	20.4	137.6	315.7
SCC subsites	1893	4.27	20.08.1998[57]	0.000396	3.30	3.38	18.1	23.6	87.8	105.1
all subsites	7669	4.39	17.04.1999[58]	0.000098	-	-	6.4	6.4	27.1	27.1

Table 6-14. Key figures from four investigated data levels of subsites.

An interesting curiosity was the length of the domain name segments starting with 4.39 among all 7669 subsites (cf. Table 4-1, Section 4.2.1). Among the 141 subsites in the 10 path nets, 130 had 4 domain name segments and 11 had 5 segments, yielding an average domain name length of 4.08. This circumstance reflects the older (as indicated by the Internet Archive indexing), and thus more central, established and 'short-named' nature of subsite affiliations occurring as connectors in the path nets.

After the focus in the preceding subsections on graph measures in the sample of 10 path nets, the next section 'zooms in' on the fourth step in the five-step methodology, concerned with extracting pages and links in the 10 path nets.

[55] All 141 subsites in the 10 path nets had top homepage identified in the Internet Archive, cf. Appendix 11.
[56] Based on 186 sampled SCC subsites identified with top homepage in the Internet Archive.
[57] 1874 SCC subsites were identified with top homepage in the Internet Archive.
[58] 6868 subsites were identified with top homepage in the Internet Archive.

6.4 Path net pages and links

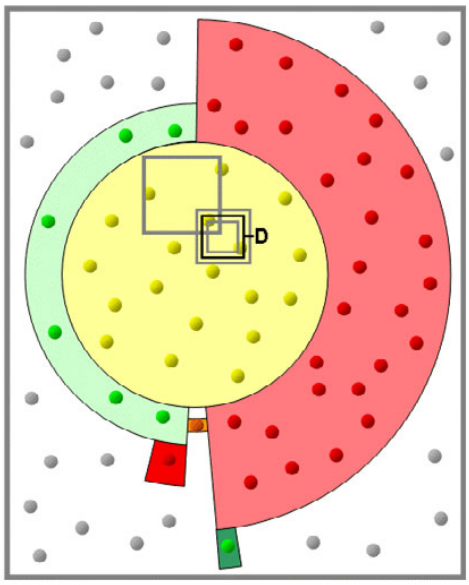


Figure 6.26. Step D in the five-step methodology: path net pages and links.

After the sampling and graph measurements of the 10 path nets containing all shortest paths between start and end subsite nodes belonging to topically dissimilar scientific domains, step D in the five-step methodology is concerned with 'zooming in' on the source pages and target pages with outlinks and inlinks in the 10 path nets. The objective of this step is to identify page genres and topics and thus enable the identification of what kind of pages and links provide transversal connections between subsites in the final step E.

Section 6.4.1 treats the extraction of URLs from the raw data set for source and target pages from the 10 path nets. Section 6.4.2 is concerned with selecting a feasible subset of pages to investigate more closely. Section 6.4.3 outlines how the Internet Archive was used to retrieve the subset of source and target pages and their interconnecting links from the 10 path nets. The retrieved pages and links are further described in Section 6.4.4. Finally, Section 6.4.5 gives a thorough examination of the genres identified on the retrieved pages – suggesting how the Web may be conceived as a *web of genres* with rich *genre connectivity*.

6.4.1 Data extraction of source and target pages

The 109 raw data files – one long plain text file for each university – with data from the original harvest of the UK academic web space in 2001 were used to extract all source pages providing outlinks and all target pages receiving inlinks at the subsites in the 10 path nets. In other words, for each subsite-to-subsite connection in the path net, the

URLs of the matching pairs of interlinked source pages and target pages were extracted from the data set files. For example, finding the two page level outlinks from node 1885 (*eye.ox.ac.uk*) targeted to node 102 (*medweb.bham.ac.uk*) in Fig. 6.27 below, a special script extracted the URLs of the source pages and target pages marked in bold in Fig. 6.28. The figure shows two excerpts from the original raw data files with the source page URLs flagged with a tabulated '1' and the indented lists of target page URLs preceding the source URLs.

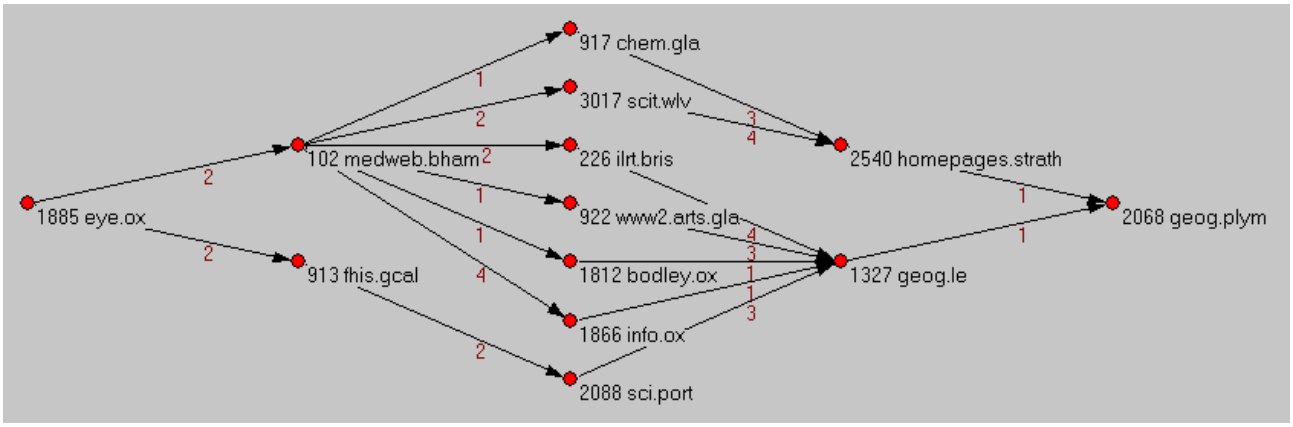


Figure 6.27. Counts of page level links in path net NH05.

```
      .eye.ox.ac.uk/symposium.htm
      medweb.bham.ac.uk/ophthalmology/courses/
      .eye.ox.ac.uk/oxfeye.htm
      .eye.ox.ac.uk/general3.htm
      .eye.ox.ac.uk/clinstaf.htm
      .eye.ox.ac.uk/teach3.htm
      users.ox.ac.uk/~opht0011
      .eye.ox.ac.uk/jobs3.htm
      .eye.ox.ac.uk/links3.htm
      .eye.ox.ac.uk/map3.htm
      .eye.ox.ac.uk/welcome.htm
      .eye.ox.ac.uk/whatsnew.htm
.eye.ox.ac.uk/teach3.htm    1
_____________________________________________

      .eye.ox.ac.uk/symposium.htm
      medweb.bham.ac.uk/ophthalmology/courses/
      .eye.ox.ac.uk/welcome.htm
      .eye.ox.ac.uk/general.htm
      .eye.ox.ac.uk/staff.htm
      .eye.ox.ac.uk/research.htm
      .eye.ox.ac.uk/publicat.htm
      .eye.ox.ac.uk/seminars.htm
      .eye.ox.ac.uk/teach.htm
      .eye.ox.ac.uk/jobs.htm
      .eye.ox.ac.uk/links.htm
      .eye.ox.ac.uk/map.htm
      .eye.ox.ac.uk/oxfeye.htm
      .eye.ox.ac.uk/whatsnew.htm
.eye.ox.ac.uk/teach.htm     1
```

Figure 6.28. Two excerpts from raw text data file with identified source page URLs (bold with tabulated '1') and target page URLs (indented bold).

As described earlier in Section 4.1.2, variant university domain names had all been converted into a canonical form for sake of comparability when used in the adjacency matrix. For example, all domain names ending with *.brighton.ac.uk* had been converted into *.bton.ac.uk*. However, in the original data set files, all URLs were stored with their original domain name. This meant that the special script had to search for all the possible combinations of variant domain names when extracting the URLs of source pages and target pages from the data set files. Another data extraction problem was variant URL-strings representing the same source or target page in the raw data files. For example, URLs with or without default file names:

web.bham.ac.uk/p.jackson/index.html and *web.bham.ac.uk/p.jackson/*.

Other page duplications were URLs with or without transcribed tildes, e.g.,

users.ox.ac.uk/~quarrell/ and *users.ox.ac.uk/%7Equarrell/;*

included/excluded slash endings, e.g.,

cs.bris.ac.uk/~spilio/ and *.cs.bris.ac.uk/~spilio*;

and changed cases, e.g.,

mml.cam.ac.uk/ling/staff.htm and *mml.cam.ac.uk/ling/staff.HTM*.

A special case of page duplication was the two different slightly different URLs (tilde or not) that were treated as duplicate pages and thus just counted once:

.cbl.leeds.ac.uk/~nikos/tex2html/doc/latex2html/latex2html.html

.cbl.leeds.ac.uk/nikos/tex2html/doc/latex2html/latex2html.html

In order to extract *unique* source and target pages only, close time-consuming manual inspection was needed to filter out such page duplications as exemplified above.

As noted earlier, the original web crawler had been programmed to exclude as many duplicate pages as possible (Thelwall, 2002f). However, the non-excluded page duplications exemplified above yet again illustrate the inherent almost ineradicable data filtering problems in webometrics due to innumerous data error sources. As also stated by Heydon & Najork (1999) and Thelwall (2003b), identifying and eliminating duplicate web pages is indeed a technically challenging task.

path net	start node	end node	# sub-sites	# source pages	# target pages
HN01	hum	atm	15	41	28
HN02	econ	chem	6	7	7
HN03	psy	math	8	29	28
HN04	speech	palaeo	4	4	3
HN05	geogr	eye	6	7	12
			39	*88*	*78*
NH01	atm	hum	15	68	47
NH02	chem	econ	28	134	114
NH03	math	psy	5	7	7
NH04	palaeo	speech	41	167	152
NH05	eye	geogr	13	33	27
			102	*409*	*348*
			141	**497**	**425**

Table 6-15. Number of unique source pages and target pages in the 10 path nets.

The data extraction and duplicate removals yielded 497 unique source page URLs and 425 unique target page URLs in the 10 path nets as listed in Table 6-15 above. The discrepancy in the number of source pages and target pages is due to the circumstance that some source pages share targets and vice versa as illustrated in Fig. 6.29 and 6.30.

These phenomena of co-linking *source* pages (analogous to bibliographic coupling) and co-linked *target* pages (co-citation) in the path nets are examined further below.

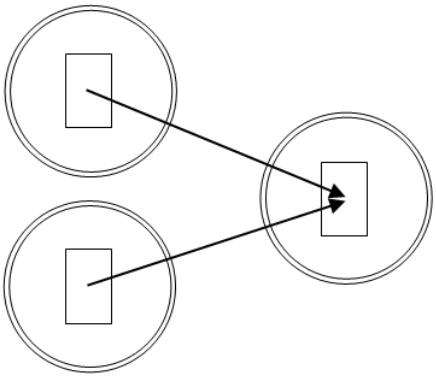


Figure 6.29.Co-linking source pages (analogous to bibliographic coupling).

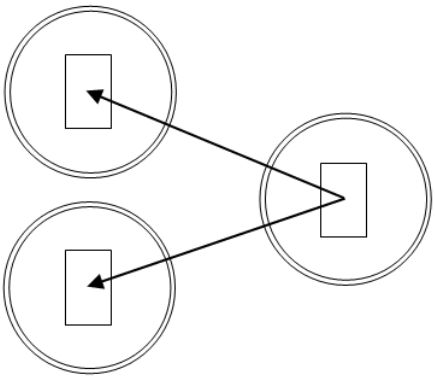


Figure 6.30. Co-linked target pages (analogous to co-citation).

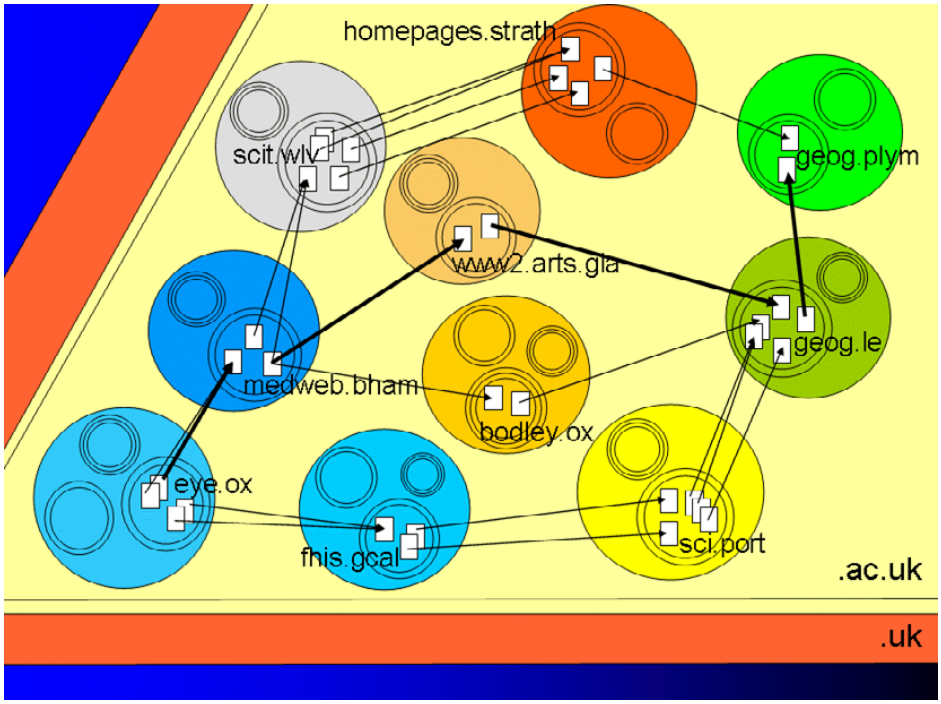


Figure 6.31.* Node diagram with link path visualization. Excerpt from path net NH05 with actual source pages and target pages. All links belong to shortest link paths (path length 4) between start node *eye.ox.ac.uk* and end node *geog.plym.ac.uk*. See Appendix 10 for affiliations. Bold links show one example of such a link path. (*cf. color prints placed before appendices).

Fig. 6.31 above shows a page level node diagram of an excerpt of path net NH05 (cf. Fig. 6.27). The figure gives a realistic illustration of the actual number of links between source pages and target pages along the shortest link paths between the two subsites *eye.ox.ac.uk* and *geog.plym.ac.uk*. One such shortest link path (path length 4) is highlighted with bold links. Stacked pages illustrate so-called sibling web pages located

in the same file directory at a subsite. The data file excerpts in Fig. 6.28 above gives the URLs of the sibling source pages at *eye.ox.ac.uk* with outlinks to the same target page at *medweb.bham.ac.uk*.

In continuation of Fig. 6.29 and 6.30, there are several co-linking and co-linked pages in Fig. 6.31. For example, four co-linking source pages at *bodley.ox.ac.uk* and *sci.port.ac.uk* linking to the same subsite *geog.le.ac.uk* – or three co-linked target pages at *scit.wlv.ac.uk, bodley.ox.ac.uk* and *www2.arts.gla.ac.uk* that are outlinked from a source page at *medweb.bham.ac.uk* (cf. Appendix 10 for affiliations of subsites in path net NH05). The latter source page thus provides connecting links for three different shortest link paths in the path net.

As discussed in Section 4.2.2, the necessary focus on shortest paths between *subsites* and not between web *pages* – in order to create a feasible adjacency matrix for shortest path computation in Pajek – inevitably resulted in non-linked 'gaps' between target pages and source pages *within* the subsites. As argued earlier, sites and subsites with generic or multi-disciplinary contents were filtered out from the study, in an attempt to reduce topic drift across these unavoidable link-'gaps'.

6.4.2 Followed and non-followed link paths

Due to the time-consuming in-depth analysis needed in order to classify page topics and genres, it was not feasible to visit all 922 (497 source + 425 target) pages identified in the 10 path nets. Instead, all pages in the smaller path nets were visited. In the large path nets, NH01, NH02, and NH04 (cf. Table 6-15 above, and Appendix 10), only pages belonging to link paths not passing nation-wide or campus-wide generic-type subsites were visited. The same typology was used as in the classification of subsite genres in the sample of 189 SCC subsites in Section 6.2.2. The delimitation of followed link paths and visited subsites in the large path nets facilitated the later identification of transversal links between different topically more focused subsites.

In Fig. 6.32 below, nine generic-type subsites in the large path net NH02 are denoted with filled white nodes. Some examples: node 756 (Edinburgh University Students' Association), 883 (Guild of Students' Home Page, University of Exeter), 2540 (campus-wide staff and student homepages at the University of Strathclyde), 1460 (university library, University of Manchester), 1821 (campus-wide staff and student homepages at the University of Oxford), 2967 (alias university homepage of the University of Northumbria). The close investigation of all nodes in the smaller path nets had revealed that such generic-type subsites yielded quite trivial links. Transferring this observation to path net NH02 in Fig. 6.32, investigating a link path from 917 (chem.gla.ac.uk) passing 2540 (*homepages.strath.ac.uk*) and then 354 (*econ.cam.ac.uk*) would yield results like a chemist at node 2540 receiving an inlink from another chemist at node 917, whereas an economist with an homepage at the same campus-wide and thus multi-disciplinary node 2540 has made an outlink to an economy web page at node 354. As noted earlier, such intra-subsite topical gaps are inevitably a problem when studying shortest link paths between *subsites* as was the feasible option in this dissertation – instead of examining shortest paths between web *pages* as was beyond the computational possibilities. However, as stated above, some of the limits with shortest

paths between subsites were circumvented by investigating link paths passing non-generic topic-focused subsites.

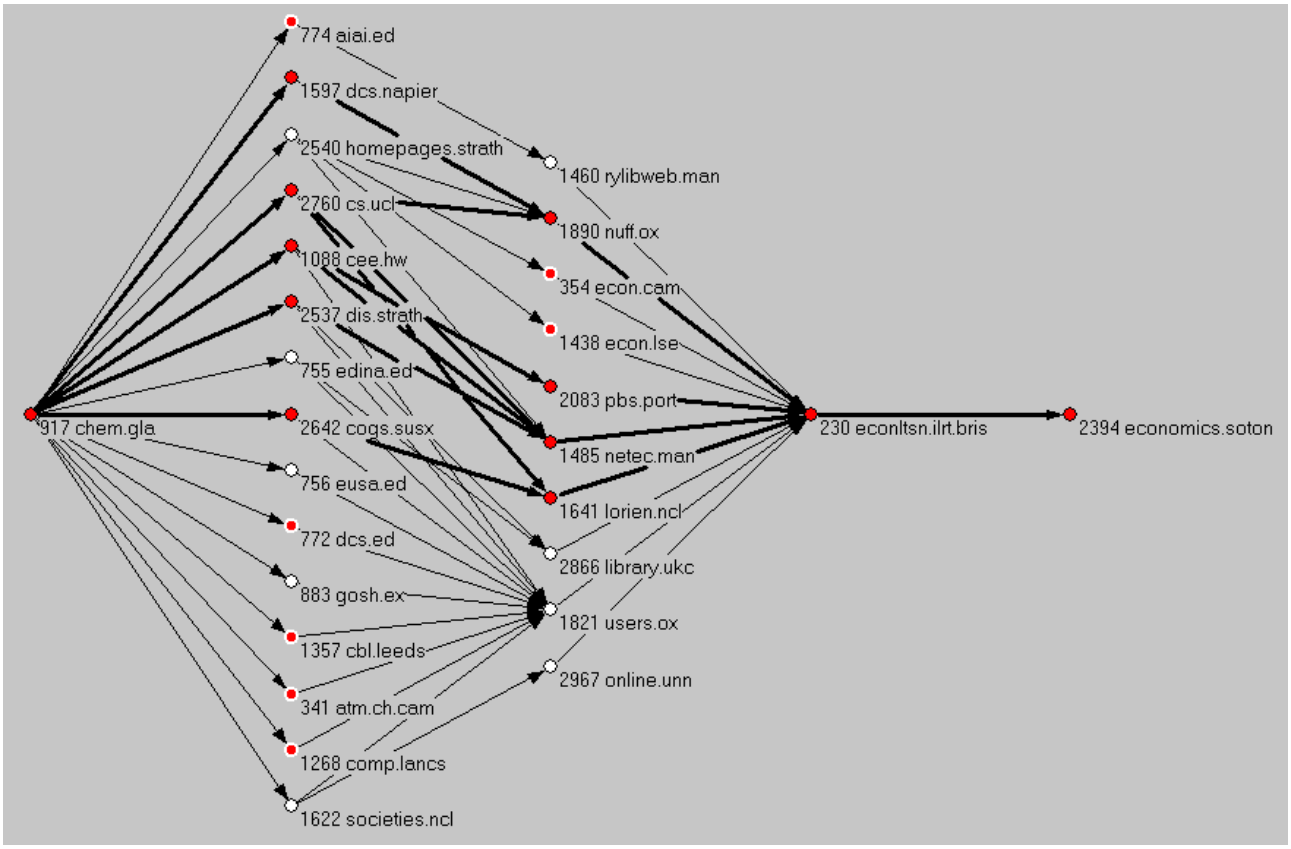


Figure 6.32.* Path net NH02. Six link paths in bold contain non-generic subsites only. Generic subsites are marked with white nodes. Excluded subsites on 'generic' link paths are marked with white-bordered red (dark) nodes. See Appendix 10 for affiliations. (*cf. color prints placed before appendices).

The number of link paths in a path net depends on the connectivity patterns between the nodes including the number of in-neighbors and out-neighbors for each node. According to Skiena (1996), there can be an exponential number of shortest link paths between two nodes.[59] Note there may be innumerous *longer* link paths between any node pairs connected by at least one shortest link path in a graph.

In Fig. 6.32 above, there are 27 different shortest link paths (path length 4) in path net NH02 connecting the start and end nodes. In other words, there are 27 different 'routes' of finding way along links between nodes 917 and 2394. The plentitude of shortest link paths makes such a subgraph less vulnerable to disconnection by removal of nodes. This aspect can be investigated by identifying so-called *node-independent paths*, paths not having any nodes in common except start and end node (White & Newman (2001). The minimum number of nodes that need to be removed to disconnect a pair of nodes (the so-called *connectivity* of the pair) in a graph is equal to the number of node-independent paths between the nodes (ibid.). Network robustness can thus be

[59] Roumeliotis (2002) describes an algorithm for finding the number of shortest paths between two nodes.

quantified by calculating the numbers of node-independent paths (ibid.). In path net NH02 in Fig. 6.32, there is one node-independent path, as it suffices to remove one node (node 230) to disconnect node 917 from node 2394.

Only six link paths marked with bold links do *not* pass any generic-type subsites in Fig. 6.32 and were thus included in the study. Twelve subsites were located on the six followed link paths. The remaining 21 link paths were thus excluded together with their subsites, including seven non-generic subsites marked with white-bordered nodes in the figure.

Table 6-16 shows the number of followed link paths, visited subsites and followed subsite level links in the 10 path nets (cf. Appendix 10).

path net	start node	end node	path length	# link paths	# followed link paths	% followed link paths	# subsites	# visited subsites	% visited subsites	# subsite level links	# followed subsite level links	% followed subsite level links
HN01	hum	atm	3	10	10	100.0	15	15	100.0	23	23	100.0
HN02	econ	chem	3	3	3	100.0	6	6	100.0	7	7	100.0
HN03	psy	math	4	6	6	100.0	8	8	100.0	12	12	100.0
HN04	speech	palaeo	3	1	1	100.0	4	4	100.0	3	3	100.0
HN05	geogr	eye	3	3	3	100.0	6	6	100.0	7	7	100.0
			3.2	*23*	*23*	*100.0*	*39*	*39*	*100.0*	*52*	*52*	*100.0*
NH01	atm	hum	3	12	5	41.7	15	8	53.3	25	11	44.0
NH02	chem	econ	4	27	8	29.6	28	12	42.9	53	18	34.0
NH03	math	psy	3	2	2	100.0	5	5	100.0	5	5	100.0
NH04	palaeo	speech	4	67	36	53.7	41	26	63.4	99	64	64.6
NH05	eye	geogr	4	7	7	100.0	13	13	100.0	18	18	100.0
			3.6	*115*	*58*	*50.4*	*102*	*64*	*62.7*	*200*	*116*	*58.0*
			3.4	**138**	**81**	**58.7**	**141**	**103**	**73.0**	**252**	**168**	**66.7**

Table 6-16. Followed link paths, visited subsites and followed subsite level links in the 10 path nets.

As will appear from Table 6-16 above, NH02 was the path net with the least number of link paths followed (29.6%) – due to many paths passing generic-type subsites. A total of 81 link paths (58.7% of all paths in the 10 path nets) were followed, passing 103 subsites (73.0%) and 168 subsite level links (66.7%). Not shown in the table is the detail that among the 104 *unique* subsite nodes (cf. Section 6.3.2) in the 10 path nets, 78 unique subsites were visited on the followed link paths.[60]

[60] Tables of affiliations in Appendix 10 show visited subsites (only subsites on link paths not passing generic-type subsite nodes were visited in the large path nets NH01, NH02 and NH04).

6.4.3 Internet Archive as 'web archaeological' tool

As stated by Wilkinson *et al.* (2003), a problem arises when trying to revisit web pages identified in earlier web crawls due to the dynamic nature of the Web. They report that in 86 cases of 550 (15.6%) revisited web pages in their study, "either the source or target page was no longer available or the source page no longer linked to the target page" (p. 51). These problems of removed pages and outdated so-called broken links are also noted by, e.g., Pitkow (1999), Koehler (1999b; 2002) and Day (2003). In an attempt to circumvent these problems, once more the Internet Archive was used as a 'web archaeological' tool, this time in order to recover as many as possible of the source and target pages on the followed paths in the 10 path nets.

The same procedure as described in Section 6.2 was conducted in the Internet Archive with regard to retrieving web pages indexed as close as possible to the time around July 2001 when the original link data set was collected. As before, this precaution was taken because the original full contents of the web pages had not been harvested by the web crawler in 2001 in order to reduce data storage. Because contents change on the dynamic Web, the Internet Archive could be used to access web page contents as close to the original ones that had their link data extracted by the web crawler.

The URLs of the source pages and target pages extracted from the original raw data files (cf. Section 6.4.1) were used to retrieve web pages from the Archive. As described above, all URLs were stored with their original domain names in the raw data files and not with the canonical name form. Furthermore, the URLs contained their original spelling with lower-case and upper-case letters. This turned out to be essential in the Internet Archive, because the UNIX system the Internet Archive runs on is case-sensitive[61]. For example, the Archive would give no results for a search on the URL *http://netec.mcc.ac.uk/econfaq/sc.html,* whereas the URL with correct upper-case letters would be found: *http://netec.mcc.ac.uk/EconFAQ/sc.html*. Fortunately, the Internet Archive had no problem retrieving the many URLs from the raw data set that had been stripped off the prefix *www*. The Archive treated such URLs as though they contained the prefix.

There were retrieved 530 web pages comprising 281 unique source pages and 249 unique target pages from the 81 followed link paths in the 10 path nets. It was not feasible to check all the 530 web pages in the present study to check how many were available on the current Web. However, the Internet Archive still has the advantage of providing page contents closer to a revisited 'old' web crawl than does the current Web.

45 source pages (16.0% of 281) and 30 target pages (12.0% of 249) were not available in the Internet Archive. The higher availability of target pages in the Archive may reflect that some of these pages belong to more *inlink-prone* page genres like institutional homepages as suggested in Section 6.4.5.1 below. Such inlink-prone pages

[61] 'Internet Archive Frequently Asked Questions': http://www.archive.org/about/faqs.php [visited 5.12.2002]

may have larger probability of being reached and retrieved by the Archive's web crawlers.

Nine of the 45 non-available source pages were due to robot's exclusion by site owners (cf. Section 4.2.4) preventing the Internet Archive from indexing the pages.[62] (Five of the 32 target pages were not available in the Archive due to this exclusion of web crawlers ('robots')). Of the non-available source pages in the Archive, 28 were retrieved from the current Web containing same contents of outlinks as identified in the original link data set. (12 target pages were found on the Web). The topics and genres of these pages were thus relatively unproblematic to classify. Five source pages were available on the Web but with changed contents, that is, they no longer contained links to the original targets. The remaining 12 source pages were not available on the Web. (18 target pages were not found on the Web). The topics and genres of some of the changed or non-available pages were identified by finding information on parent pages (the immediate superior pages in the web site file hierarchy) or sibling pages (other pages in the same file directory) found either in the Internet Archive or on the current Web. It turned out that a few pages stated as not archived by Internet Archive were in fact accessible via parent pages indexed in the Internet Archive.[63]

Terms embedded as parts of source URL and target URLs extracted from the raw data files could also provide some clues on probable topics and genres. Furthermore, one page was found by using the local site search at the university. The original URL *nuff.ox.ac.uk/library/gateways.shtml* had in this case been changed to a new file format, *nuff.ox.ac.uk/library/gateways.asp* – but still contained the identical set of outlinks as identified in the original data file. Google was used to search for URL fragments in order to find relocated pages because many directories are renamed, or pages are removed to new directories or new domain names, some pages also exist as copy versions on multiple web sites. For example, the original URL *earth.ox.ac.uk/internal/matlab/techdoc/ref/ gallery.html* was found neither in the Internet Archive nor on the Web. A Google search for: *inurl:matlab inurl:techdoc/ref/gallery.html* yielded several pages with similar contents, e.g., *www.nbs.ntu.edu.sg/userguide/MatLab6/help/techdoc/ref/gallery.html*. The page turned out to be a section of a large documentation of the mathematical software *MatLab*. This software documentation had thus been copied into many web sites, probably accompanying the download of the software.

There were some doubts on whether or not to include genres and topics identified on copy version pages. However, it was decided to include such pages in the analysis when all the original outlinks from the original data set source page were identically matched in both order and extent in the found copy version page.

---

[62] A robots.txt exclusion is backdated by the Internet Archive. Thus, if a web site posts a robots.txt ban today, than all previous copies of the site will be removed from the Archive.

[63] For example, Internet Archive states 'Sorry, no matches' when searched for the URL *www.geog.leeds.ac.uk/alumni/malumni/malumattend.html*. However, Internet Archive has archived the very same URL *http://web.archive.org/web/20010716113807/http://www.geog.leeds.ac.uk/alumni/malumni/malumattend.html* that turned out to be accessible from a parent page *http://web.archive.org/web/20010716113807/http://www.geog.leeds.ac.uk/alumni/malumni/*.

6.4.4 Retrieved pages and links

As will appear from Table 6-17 below, the retrieved 530 web pages, comprising 281 source pages and 249 target pages, implied that 57.5% of all the 922 web pages identified in the 10 path nets were visited.

path net	start node	end node	# subsites	# visited subsites	# source pages	# visited source pages	% visited source pages	# target pages	# visited target pages	% visited target pages
HN01	hum	atm	15	15	41	41	100.0	28	28	100.0
HN02	econ	chem	6	6	7	7	100.0	7	7	100.0
HN03	psy	math	8	8	29	29	100.0	28	28	100.0
HN04	speech	palaeo	4	4	4	4	100.0	3	3	100.0
HN05	geogr	eye	6	6	7	7	100.0	12	12	100.0
			39	*39*	*88*	*88*	*100.0*	*78*	*78*	*100.0*
NH01	atm	hum	15	8	68	26	38.2	47	23	48.9
NH02	chem	econ	28	12	134	37	27.6	114	37	32.4
NH03	math	psy	5	5	7	7	100.0	7	7	100.0
NH04	palaeo	speech	41	26	167	90	53.9	152	77	50.7
NH05	eye	geogr	13	13	33	33	100.0	27	27	100.0
			102	*64*	*409*	*193*	*47.2*	*348*	*172*	*49.4*
			141	**103**	**497**	**281**	**56.5**	**425**	**249**	**58.6**

Table 6-17. Retrieved source pages and target pages in the 10 path nets.

Table 6-18 further below shows that the 141 subsites in the 10 path nets were connected by 657 page level links. The 103 visited subsites on the 81 followed link paths were connected by 352 page level links. The followed 352 page level links thus comprised 53.6% of all page level links in the 10 path nets. When retrieving a source page from the Internet Archive or from the Web, the target URL from the original data set files was verified by searching for it in the source page HTML code.

Target pages were subsequently retrieved by simply clicking on the verified outlink on the archived source page. In a few cases, the Internet Archive could not always display the target pages. In such cases, most target URLs were retrieved through the search page in the Internet Archive. Otherwise, the target page was looked for on the current Web as described earlier.

Some problem links were encountered, here named *inactivated* links and *non-anchored* links. An *inactivated link* is not visible in a normal browser interface, but is still present in the underlying HTML code. By inserting an exclamation mark as the first character in an HTML tag, a web page creator may (perhaps just temporarily)

inactivate and hide some contents from the browser interface.[64] Two such inactivated links were identified among the 352 followed links.

A *non-anchored link* consists of a URL that has no clickable anchor text, possibly due to errors during the web page editing.[65] One non-anchored link was identified among the followed links. In an impact analysis of link counts, which was the original purpose for Dr Thelwall's web crawl in 2001, one could argue that such 'non-clickable' links, whether they are deliberately inactivated or erroneously non-anchored, reflect intentions by page authors and thus should be counted. However, in a connectivity analysis like the present study, one could also argue that all links should be traversable. However, even if a human web user cannot see a 'non-clickable' link, the hidden URL can be identified and traversed by a programmed web crawler. Based on the last argument, it was decided to include the three 'non-clickable' links in the study.

In order to facilitate the later identification of link types (e.g., personal or institutional; related to academic or non-academic activities), as well as target page genres and topics, the anchor texts and *anchor context*, that is, surrounding paragraph text and headings (cf. Brin & Page, 1998; Fürnkranz, 1998; Amitay, 2001) from source pages were manually extracted. Such anchor texts and contexts often contain useful information about the outlinked target page, sometimes more descriptive than contained in the target page itself.

As will appear from Table 6-18 below, the 497 source pages had on average 1.32 outlinks to other pages in the path nets (thus creating co-linked target pages). However, the 281 *visited* source pages had a slightly lower average of 1.25 *followed* outlinks. This lower average is partially due to the circumstance that not all outlinks were followed on visited pages, because they belonged to non-included link paths passing generic-type nodes. The 425 target pages received on average 1.55 inlinks from other pages in the path nets (thus reflecting co-linking source pages). As was the case above, the 249 *visited* target pages received a slightly lower average of 1.41 *followed* inlinks. The reasons parallel the above-mentioned.

A few pages occurred in more than one path net. For example, the institutional link list, *geog.le.ac.uk/cti/info.html*, 'Geo-Information Gateway', with links to world-wide earth science resources, made by the CTI (Computers in Teaching Initiative) Centre for Geography, Geology and Meteorology, University of Leicester) at node 1327 was a target page at node 1327 in both path net NH04 and NH05. Another example of multi-occurring pages is created by a researcher at node 917, the Department of Chemistry, University of Glasgow, with 8 pages in three path nets. He had a target page in path net HN02, a personal hobby page comprising a fan page for a football club, Partick Thistle FC. Furthermore, he had seven source pages (six in path net NH02 and one in NH05) as personal link lists mostly concerned with different football clubs.

The only page that was both a target page and source page in the same path net, was an institutional link list, *ma.hw.ac.uk/uk_maths.html*, at node 1089 path net HN03,

---

[64] For example, the HTML tag *<!-- <LI>ICMS <A HREF="http://www.ma.hw.ac.uk/icms/travel.html"> local transport information. -->* contains an inactivated link

[65] For example, the HTML code may contain this line: *"... by copying the file <A NAME="tex2html1" HREF="http://cbl.leeds.ac.uk/nikos/tex2html/doc/manual/labels.pl"></A>"*. Note the missing anchor text before the end tag </A>. The link is non-anchored.

the Department of Mathematics, Heriot-Watt University, with links to "UK Mathematics Departments, Centres and Institutes".

path net	start node	end node	# subsites	# visited subsites	# link paths	# followed link paths	# subsite level links	# followed subsite level links	# page level links	page level links / subsite level links	# followed page links	% followed page level links	page level links / source pages	followed page links / visited source pages	page level links / target pages	followed page links / visited target pages
HN01	hum	atm	15	15	10	10	23	23	48	2.09	48	100.0	1.17	1.17	1.71	1.71
HN02	econ	chem	6	6	3	3	7	7	7	1.00	7	100.0	1.00	1.00	1.00	1.00
HN03	psy	math	8	8	6	6	12	12	38	3.17	38	100.0	1.31	1.31	1.36	1.36
HN04	speech	palaeo	4	4	1	1	3	3	4	1.33	4	100.0	1.00	1.00	1.33	1.33
HN05	geogr	eye	6	6	3	3	7	7	12	1.71	12	100.0	1.71	1.71	1.00	1.00
			39	*39*	*23*	*23*	*52*	*52*	*109*	*2.10*	*109*	*100.0*	*1.24*	*1.24*	*1.40*	*1.40*
NH01	atm	hum	15	8	12	5	25	11	87	3.48	33	37.9	1.28	1.27	1.85	1.43
NH02	chem	econ	28	12	27	8	53	18	183	3.45	53	29.0	1.37	1.43	1.61	1.43
NH03	math	psy	5	5	2	2	5	5	8	1.60	8	100.0	1.14	1.14	1.14	1.14
NH04	palaeo	speech	41	26	67	36	99	64	232	2.34	111	47.8	1.39	1.23	1.53	1.44
NH05	eye	geogr	13	13	7	7	18	18	38	2.11	38	100.0	1.15	1.15	1.41	1.41
			102	*64*	*115*	*58*	*200*	*116*	*548*	*2.74*	*243*	*44.3*	*1.34*	*1.26*	*1.57*	*1.41*
			141	**103**	**138**	**81**	**252**	**168**	**657**	**2.61**	**352**	**53.6**	**1.32**	**1.25**	**1.55**	**1.41**

Table 6-18. Followed page level links in the 10 path nets.

In the next subsections, the path analysis of the 81 followed link paths will first deal with the genres of the visited source and target page, and then describe how the 352 followed links interconnected the genres.

6.4.5 Page genres

In order to enable the later identification of what types of web pages provide transversal links in an academic web space, a genre classification was made on all visited and extracted 530 source and target pages in the 10 path nets.

In media theory, literature studies and document theory, the concept of *genre* is used to cover the characteristics that differentiate texts from each other (Andersen, 2002). Freedman & Medway (1994) give an overview of genre studies with genres as 'types' or 'kinds' of discourse, characterized by similarities in content and form. In this context, genres may be defined as socially recognized regularities of form and purpose in documents (e.g., Roussinov *et al.*, 2001; Geisler *et al.*, 2001). In the dissertation, the term *genre* is used in a broad sense in accordance with contemporary web terminology for describing types of web sites as well as web pages (cf. e.g., Koehler, 1999a; Dillon & Gushrowski, 2000; Weare & Lin, 2000; Nilan, Pomerantz & Paling, 2001; Agatucci, 2001; Jackson-Sanborn *et al.*, 2002; Rehm, 2002). The Web is an interesting setting for studying genres, because "there are many communities meeting on the Web, bringing experiences with different genres and using the Web for many different purposes" (Roussinov et al., 2001). As stated in Section 6.2.2, genre classification on the Web is a very difficult task due to non-established, muddled and overlapping genres. Web pages

and aggregations of pages thus can "conform to existing print genres, merge multiple genres together, or create new ones" on the unregulated Web (Thelwall, 2002b). Hence, web genres and their defining conventions are "still in the process of becoming" (Agatucci, 2001).

Thelwall & Harries (2003) gives an excellent illustration of the diversity of web pages contained at a single academic web site that is "likely to contain information created by different types of authors (scholars, administrators, students), for different audiences (internal/external, prospective/current/past students, the public, other scholars), with differing content levels (academic papers, books, teaching notes, student assignments, job advertisements, hobby pages, photos or videos of family members), and in multiple recognizable and novel genres (lecture notes, link lists, frequently asked questions pages), and may even contain misinformation" (p. 595).

The 530 different web pages visited in the 10 path nets represented a wide diversity of genres reflecting the multitude of creators, purposes and audiences as illustrated in the above quotation. In order to obtain a clearer picture, the page genres were grouped into broader 'meta genres'. By an iterative process of revisiting, comparing and grouping similar pages, 17 meta genres crystallized as listed in Table 6-19 below. The meta genres were thus based on a kind of 'literary warrant' induced from the existing crop of 530 visited web pages. Again, the genre classification was conducted by the author alone, with the limitations this circumstance implies, as problematized earlier.

The genre categories partially overlap genre classifications in other studies. For example, Almind & Ingwersen (1997) classify web pages on the overall Web according to the function given them by their authors: personal home page, organizational home page, subject home page, pointer page, and resource pages. Rehm (2002) describes a web genre hierarchy on an academic web site in his attempt to build an automatic genre classification tool. Other web genre classifications that have been inspiring in the present study are made by Crowston & Williams (2000), Haas & Grams (1998, 2000), and Thelwall & Harries (2003). However, none of these genre classifications had sufficient specificity required in the present investigation of an academic web space.

During the course of close inspection of the web pages in order to classify their genres, an idea emerged to divide the visited web pages into two main categories, *personal* and *institutional* pages. The purpose was to get a picture of how the two categories appear in academic link structures, especially as providers of transversal links.

Personal web pages were defined as personally created pages used for personal academic or non-academic purposes. While primarily located in personal web site directories, some personal web pages are copied or moved into institutional directories. All 530 visited web pages, as well as their parent, sibling and child pages, were thoroughly examined, in order to identify such cross-locations. Page layout, text, links, and web page creator names were useful clues for such identification. Typical personal web pages are personal homepages, hobby pages, personal link lists, personal publications, software programs, and personal course pages including student's assignments. Among the visited pages were personal pages created by technical and administrative staff, researchers, PhD students and other students.

Institutional web pages are created for official, institutional and non-personal purposes, both academic and non-academic. The term *institutional* is used in the dissertation as a generic term to cover web pages at any academic unit level in the investigated web space.[66] Typical institutional web pages are thus homepages of schools, departments, centers, research groups, etc.; generic campus-wide service pages; institutional link lists pointing to research partners, institutions, research projects etc.; staff and publication lists; institutional reports, newsletters and other publications; conference and workshop pages.

The division into a personal and an institutional main category caused a similar division of meta genres across the two categories as shown in Table 6-19. Archives, homepages, link lists, teaching pages, software, publications and research projects were compiled into separate meta genres, either personal or institutional. Logically, some meta genres only had a institutional or personal counterpart, such as institutional conference pages or personal hobby pages.

Some pages were more difficult to identify as either institutional or personal. In such cases, their parent, sibling and child pages were visited to get the context. Such close inspection was also sometimes necessary in order to identify if a personal page creator was a researcher, technical staff, other staff, PhD student, undergraduate student, etc.

A special abbreviated notation was used, as shown in Table 6-19 below, with the prefix *i.* assigned to institutional page genres and *p.* for personal page genres.

[66] In the UK, the term *institution* usually refers to a university rather than, for example, a department (Thelwall, e-mail 17.8.2003).

meta genre	definition	examples
i.archive	providing access to collection of data in institutional archive or database	British National Corpus (linguistics), database search page on dinosaurs
i.conf	conferences, workshops, seminars and other researchers' meetings	homepages, programme, sessions, lists of delegates, presentations, abstracts[67]
i.generic	non-research-related, campus-wide service web pages	Web server statistics; campus-wide service pages, incl. studying abroad[68]
i.hp	homepages of academic institutions	homepages of international, national and university research institutions; other academic university institutions as libraries and museums
i.list	institutional pages with link lists as main content	institutional lists with site outlinks, institutional publication lists, staff lists, link-rich (text-sparse) resource guides[69]
i.proj	joint and single research groups, incl. specific group projects	homepages of joint research groups and single institutional research groups; research project descriptions,
i.publ	institutional publication (in a broad sense completed works presenting content-focused information for external users)	institutional research project reports, text-rich resource guides, journals
i.soft	institutional software programs, software manuals and other software descriptions	institutional software programs, download pages, demos, software documentation, manuals, FAQs[70], software tutorials
i.teach	institutional teaching-focused web pages	institutional courses, tutorials
p.archive	providing access to collection of data in personally administered archive or database	mailing list archive, discussion group archive[71]
p.hobby	personal (researcher, student) hobby webpage not related with the persons main academic activity	hobbies like football, running, sci-fi, private travel, Saxon shore forts, ancient Greek warship[72]
p.hp	personal homepage	personal homepage made by researcher, other staff, PhD student, other student, incl. CV, publication list
p.list	personal pages with link lists as main content	personal (researcher, other staff, PhD student, other student) pages with link lists, incl. bibliographies[73], research-related or teaching-related or leisure-related link lists; link-rich (text-sparse) resource guides
p.proj	personal (researcher, PhD student) research project pages	environmental studies, crystallography, cybergeography,
p.publ	personal publication (in a broad sense completed works presenting content-focused information for external users)	personal books, book reviews, reports, papers, conference presentations, posters, text-rich resource guides
p.soft	personal software programs, software manuals	personal software programs, download pages, demos, software documentation, manuals, FAQs, tutorials
p.teach	personal teaching-focused web pages	lecturer's course pages, tutorials[74]; students' assignments, course notes

Table 6-19. Typology of meta genres of academic web pages: 9 institutional genres (*i.*) and 8 personal (*p.*).

[67] *Full-text* conference papers and posters were placed in the *p.publ* genre.

[68] Web pages containing information on studies abroad were placed in *i.generic* and not in *i.teach* because they are university service for campus-wide users and not focused at specific teaching in particular topics.

[69] Resource guides often extensively cover a specific topic with links to web pages on other sites covering the same topic. Text-sparse resource guides resembling link lists were classified as *i.list* or *p.list*. Text-dense resource guides resembling papers or manuals were classified as *i.publ* or *p.publ*.

[70] Frequently Asked Questions (FAQ).

[71] In both examples, the archives were moderated by researchers on a private basis.

[72] Web page (*www.atm.ox.ac.uk/rowing/trireme*) devoted to a full size replica of an ancient Greek warship, the *Athenian Trireme*. The page is maintained by a researcher at Centre for Atmospheric, Oceanic and Planetary Physics, at the University of Oxford (node 1904 in path net HN01)

[73] Bibliographies classified as *p.list* if the entries are documents written by *others* than the page creator.

[74] Tutorials are placed in *p.teach* and not in *p.publ* because the intended readers are not external but internal users participating in a course.

The typology is not necessarily exhaustive for an academic web space. A larger sample of web pages would probably yield different and additional categories. Nor are the categories in the table mutually exclusive, but overlap with fluid borders. For example, institutional software documentation, manuals and tutorials were placed as *i.soft* and not as *i.publ* or *i.teach* because of the close relation to accompanying software programs.

In order to classify the pages in a consistent way, some rules of prioritized categorization order were necessary. Fig. 6.33 shows the decided prioritized order among the 9 institutional and 8 personal meta genres with the highest prioritized genres in front. For example, if a personal homepage also included a lengthy link list as part of the page, this page was not classified as a *p.list* but as *p.hp*, because *p.hp* 'overruled' *p.list* according to the prioritized categorization order shown in the figure. If a link list was included on a teaching-related page, e.g. a course page, the link list was classified not as *p.teach* but as a *p.list*.

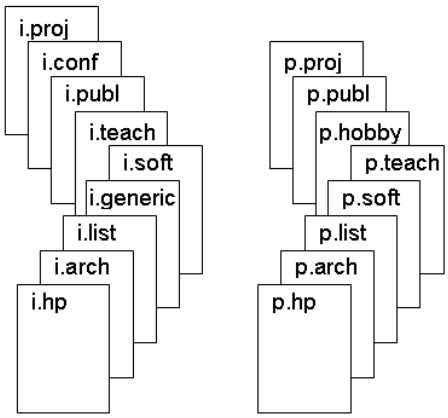


Figure 6.33. Prioritized order of genre classification of institutional and personal web pages. Highest priorities in front.

The selection of the prioritized order was based both on 'pre-coordinated' and 'post-coordinated' opinions by the present author. One 'pre-coordinated' opinion was that institutional and personal homepages were the top hierarchical pages in the corresponding institutional or personal web territories and thus should be prioritized highest in the genre order. Another 'pre-coordinated' opinion was that the link lists, either institutional or personal, were of special interest in this study because of their possible transversal links. The remaining prioritized order emerged in a 'post-coordinated' fashion when working with the page classifications and the definitions of the genres in Table 6-19. For example, the order of *i.soft* > *i.publ* > *i.proj* emerged when it turned out be most relevant to collate all web pages related to a software in the same category, because of the abovementioned close relation between software documentation and software program. A software manual page placed on a research project subsite was thus also placed in the *i.soft* category. The footnotes of Table 6-19 illustrate some more examples of the prioritized genre order.

Inevitably, higher prioritized genres will receive higher counts. For example, the *p.list* genre includes link lists related to *p.soft*, *p.teach*, *p.hobby*, *p.publ*, and *p.proj*. Consequently, lower prioritized genres will get lower counts.

6.4.5.1 Meta genres of visited pages

Table 6-20 below shows the meta genres of the visited 281 source pages and 249 target pages on the followed link paths in the 10 path nets.

meta genres	# visited source pages	% n=281	# visited source subsites	# visited target pages	% n=249	# visited target subsites
i.archive	0	0.0	0	4	1.6	4
i.conf	26	9.3	17	12	4.8	8
i.generic	6	2.1	3	10	4.0	7
i.hp	0	0.0	0	42	16.9	38
i.list	73	26.0	34	23	9.2	17
i.proj	16	5.7	14	24	9.6	17
i.publ	8	2.8	4	7	2.8	7
i.soft	6	2.1	4	9	3.6	5
i.teach	4	1.4	2	18	7.2	13
	139	*49.5*		*149*	*59.8*	
p.archive	0	0.0	0	2	0.8	2
p.hobby	2	0.7	2	12	4.8	7
p.hp	19	6.8	14	34	13.7	22
p.list	83	29.5	37	5	2.0	5
p.proj	0	0.0	0	4	1.6	4
p.publ	7	2.5	7	22	8.8	9
p.soft	14	5.0	4	14	5.6	8
p.teach	17	6.0	11	7	2.8	6
	142	*50.5*		*100*	*40.2*	
	281	**100.0**	**93**	**249**	**100.0**	**93**

Table 6-20. Meta genres of visited 281 source pages and 249 target pages.

It should be stressed that the findings of the 10 path nets cannot be generalized, but are indicative only, due to the small and non-random sample. The presentation in the following subsections should thus be viewed as an exploratory identification and conceptualization of elements in relation to how web page genres may be interconnected in an academic web space.

As shown in Table 6-20, the institutional and personal meta genres made up about 50% each of the visited source pages in the 10 path nets. This result may indicate a relatively large importance of personal web pages for providing site outlinks in an academic web space. The corresponding picture for target meta genres is somewhat different. About 60% of the visited target pages belonged to institutional meta genres, whereas 40% belonged to personal ones. Even if the share of personal meta genres thus is smaller with regard to the visited target pages, this result still may indicate a relatively large importance by personal web pages for receiving site inlinks in an academic web space.

Four meta genres were not represented among the visited source pages because they provided no outlinks in the path nets. Thus, there were no institutional or personal

archive pages, institutional homepages, or personal project pages among the visited source pages in the path nets. A larger sample size would probably have yielded outlinks from these meta genres. However, all meta genres were represented among the visited target pages.

A horizontal reading of Table 6-20 shows some differences between meta genres that may be *'site outlink-prone'* and *'site inlink-prone'* as suggested by this small sample. For example, the institutional homepages may not be outlink-prone but quite inlink-prone. As stated above, 0% of the visited source pages belonged to this meta genre, whereas 16.9% of the visited target pages did. However, the lacking site outlinks are quite comprehensible as the purpose of an institutional homepages is to function as access points to web pages *within* the institution, thus being 'site *selflink*-prone'. On the other hand, for example, personal link lists may be outlink-prone but less inlink-prone, again reflecting the purpose of the genre. However, it should be noted that it is not unusual that one link list points to another link list. More institutional link lists (9.2%) than personal link lists (2.0%) receive inlinks among the visited target pages, perhaps reflecting a larger authoritative quality of the institutional link lists.

Table 6-20 also shows the number of visited *subsites* containing the different page genres. For example, the 73 source pages with institutional link lists were located on 34 different subsites in the 10 path nets, reflecting that many subsites have more than one link list providing links to the followed link paths. These link lists may be very different and provide outlinks to different targets. However, some subsites also have source pages with very similar contents and outlinks but being located at different URLs in the file hierarchy of the subsite.

There were 103 visited subsites on the followed 81 link paths in the 10 path nets. However, the 10 subsites functioning as start nodes in the path nets cannot contain target pages, and they cannot contain source pages when they function as end nodes. There were thus a total of 93 visited *source subsites* and another 93 visited *target subsites* as shown in Table 6-20.

Tables 6-21 and 6-22 below give an overview of the most frequent source and target meta genres. The most frequent *source* meta genres were personal link lists (29.5%) and institutional link lists (26.0%) with a quite large 'jump' to the subsequent counts. Considering that the table lists the number of source pages providing links to subsites at other universities, this result is not surprising because the purpose of link lists is to provide such site outlinks. The third most frequent source meta genre was conference pages (9.3%) that typically had site outlinks to personal homepages of participators and to other conferences. The issue of which source and target meta genres are interlinked to each other is more elaborated in Section 6.4.5.4. The most frequent *target* meta genres were institutional homepages (16.9%), personal homepages (13.7%) and institutional research project pages (9.6%). This distribution was 'smoother' with no sharp 'jump'.

source meta genres	# visited source pages	%
p.list	83	29.5
i.list	73	26.0
i.conf	26	9.3
p.hp	19	6.8
p.teach	17	6.0
i.proj	16	5.7
p.soft	14	5.0
i.publ	8	2.8
p.publ	7	2.5
i.generic	6	2.1
i.soft	6	2.1
i.teach	4	1.4
p.hobby	2	0.7
i.archive	0	0.0
i.hp	0	0.0
p.archive	0	0.0
p.proj	0	0.0
	281	100.0

Table 6-21. Most frequent meta genres of visited *source* pages.

target meta genres	# visited target pages	%
i.hp	42	16.9
p.hp	34	13.7
i.proj	24	9.6
i.list	23	9.2
p.publ	22	8.8
i.teach	18	7.2
p.soft	14	5.6
i.conf	12	4.8
p.hobby	12	4.8
i.generic	10	4.0
i.soft	9	3.6
i.publ	7	2.8
p.teach	7	2.8
p.list	5	2.0
i.archive	4	1.6
p.proj	4	1.6
p.archive	2	0.8
	249	100.0

Table 6-22. Most frequent meta genres of visited *target* pages.

6.4.5.2 Source genres of followed links

As stated in Section 6.4.4, there were 352 links interconnecting the 281 source pages and 249 target pages on the followed link paths. The present subsection outlines how the 352 followed links interconnected the meta genres identified above.

As noted earlier, a source page may have outlinks to more than one target page in a path net as illustrated in Fig. 6.34 below. Furthermore, such target pages may belong to different meta genres. Correspondingly, a target page may have inlinks from different source pages. If a visited source page, for example, belongs to the meta genre *i.conf* (institutional conference pages) and has three outlinks to target pages belonging to, say, the meta genres *p.hp*, *i.hp*, and *i.proj*, the meta genre *i.conf* of this single source page will be counted three times, that is, one time for every followed outlink from the page. The target pages may be located in separate subsites as in Fig. 6.34, or may, for example, all be placed in the same subsite, perhaps as sibling pages belonging to the same web directory as in Fig. 6.35.

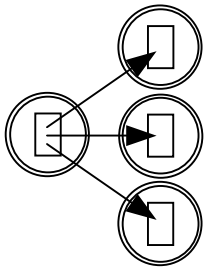


Figure 6.34. Source page with three co-linked target pages at different subsites.

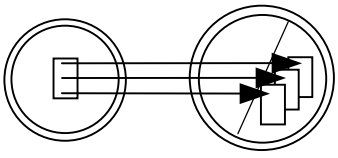


Figure 6.35. Source page with three co-linked target pages located in same subsite directory.

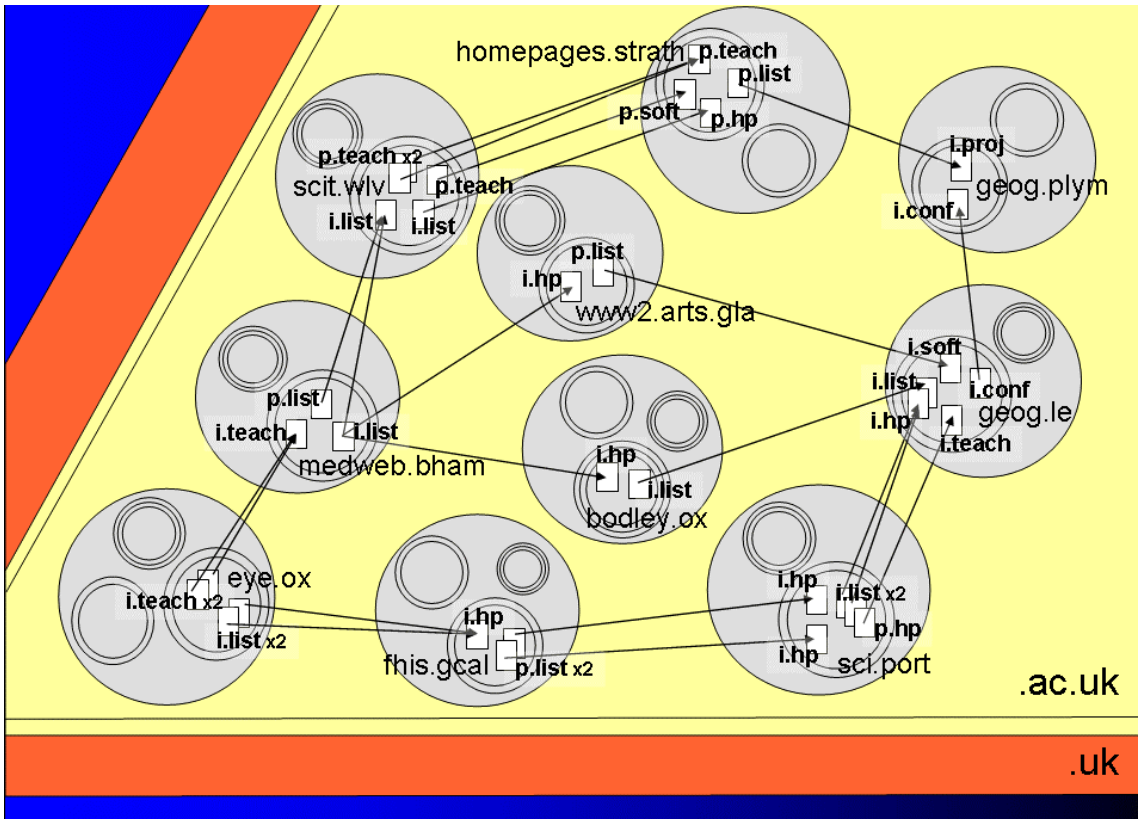


Figure 6.36.* Excerpt from path net NH05 with actual links between source pages and target pages.

Fig. 6.36 above (an extension of Fig. 6.31, Section 6.4.1) gives an example of the meta genres of source and target pages of all followed links in an excerpt of path net NH05. For example, an institutional link list (*medweb.bham.ac.uk/histmed/chmlinks.html*) at the Centre for the History of Medicine, School of Medicine, University of Birmingham has three outlinks in the path net. One is targeted to the earlier mentioned frequently outlinked institutional link list containing a clickable map of most UK universities and colleges (*scit.wlv.ac.uk/ukinfo/uk.map.html*, cf. Appendix 1), another is pointing to the homepage of the Centre for the History of Medicine, School of History & Archaeology, University of Glasgow (*www2.arts.gla.ac.uk/His/Med/*), and the third outlink is targeted to the homepage of the Bodleian Library, University of Oxford (*bodley.ox.ac.uk*).

As illustrated in Fig. 6.36, there is a large diversity of source and target meta genres of the followed links in the path nets. Table 6-23 below shows the distribution of followed links between source meta genres (divided into institutional and personal) and target meta genres sorted by frequency for each source meta genre.

institutional source meta genre	target meta genre	# links	sub total	personal source meta genre	target meta genre	# links	sub total
i.conf	p.hp	10	34	p.hobby	p.hp	3	3
i.conf	i.conf	9		p.hp	i.hp	8	22
i.conf	i.hp	4		p.hp	p.hp	4	
i.conf	i.proj	4		p.hp	i.publ	3	
i.conf	p.publ	3		p.hp	i.proj	2	
i.conf	i.generic	2		p.hp	i.teach	2	
i.conf	i.list	2		p.hp	p.soft	2	
i.generic	i.list	4	6	p.hp	i.soft	1	
i.generic	i.generic	2		p.list	p.publ	17	112
i.list	i.hp	33	87	p.list	i.hp	15	
i.list	i.list	14		p.list	p.hobby	11	
i.list	i.teach	10		p.list	p.hp	10	
i.list	i.archive	7		p.list	i.list	8	
i.list	i.proj	5		p.list	i.proj	8	
i.list	i.publ	5		p.list	i.conf	7	
i.list	p.hp	5		p.list	i.generic	7	
i.list	p.hobby	2		p.list	i.teach	7	
i.list	p.proj	2		p.list	i.soft	4	
i.list	p.teach	2		p.list	p.list	4	
i.list	p.publ	1		p.list	i.archive	3	
i.list	p.soft	1		p.list	p.soft	3	
i.proj	i.hp	9	20	p.list	p.teach	3	
i.proj	i.proj	3		p.list	i.publ	2	
i.proj	i.conf	2		p.list	p.proj	2	
i.proj	i.soft	2		p.list	p.archive	1	
i.proj	p.hp	2		p.publ	p.hp	2	7
i.proj	i.list	1		p.publ	i.archive	1	
i.proj	i.publ	1		p.publ	i.list	1	
i.publ	i.hp	7	8	p.publ	p.hobby	1	
i.publ	i.proj	1		p.publ	p.publ	1	
i.soft	i.conf	2	6	p.publ	p.teach	1	
i.soft	p.soft	2		p.soft	p.hp	9	21
i.soft	i.soft	1		p.soft	p.soft	5	
i.soft	p.hp	1		p.soft	i.hp	2	
i.teach	i.teach	4	4	p.soft	i.soft	2	
				p.soft	p.archive	2	
				p.soft	p.publ	1	
				p.teach	i.list	4	22
				p.teach	p.soft	4	
				p.teach	p.teach	4	
				p.teach	i.hp	2	
				p.teach	i.teach	2	
				p.teach	i.proj	1	
				p.teach	p.hobby	1	
				p.teach	p.hp	1	
				p.teach	p.list	1	
				p.teach	p.proj	1	
				p.teach	p.publ	1	
			165				**187**
							352

Table 6-23. Distribution of 352 followed links between all pairs of source meta genres (divided in institutional and personal) and target meta genres sorted by frequency for each source meta genre.

Table 6-23 shows that institutional homepages (*i.hp*) is the most frequent target meta genre for four source meta genres, *i.list, i.proj, i.publ*, and *p.hp* in the sample. This could, for instance, be a joint research project page pointing to homepages of partner

institutions, or a personal homepage pointing to the institutions of earlier jobs or studies.

Due to the small sample size, probably some pairs of meta genres are not represented in the table. For example, there are no links from *i.publ* to *i.publ*, or from *i.list* to *i.conf*. However, it should be recapitulated that the followed links are all between different universities and not *within* universities, thus excluding all links, for example, from one institutional publication to another publication at the same university. It is beyond the scope of this study to go into further details with the many different interlinked *genre pairs* in Table 6-23. In future large-scale studies, it would be interesting to investigate possible patterns of *genre connectivity* in academic web spaces using the methodologies and conceptual framework developed in the present study. In Section 6.4.5.4, some more general implications of genre connectivity are presented.

The distribution of most frequent source meta genres for the 352 followed links is close to the same distribution for visited source pages as appears from Table 6-24 below. The personal link lists provided almost a third (31.8%) of all followed links in the path nets and the institutional link lists about a quarter (24.7%). Again, there is a clear 'jump' from the second largest source meta genre to the third largest.

source meta genres	# followed links	% n=352	# visited source pages	% n=281
p.list	112	31.8	83	29.5
i.list	87	24.7	73	26.0
i.conf	34	9.7	26	9.3
p.hp	22	6.3	19	6.8
p.teach	22	6.3	17	6.0
p.soft	21	6.0	14	5.0
i.proj	20	5.7	16	5.7
i.publ	8	2.3	8	2.8
p.publ	7	2.0	7	2.5
i.generic	6	1.7	6	2.1
i.soft	6	1.7	6	2.1
i.teach	4	1.1	4	1.4
p.hobby	3	0.9	2	0.7
i.archive	0	0.0	0	0.0
i.hp	0	0.0	0	0.0
p.archive	0	0.0	0	0.0
p.proj	0	0.0	0	0.0
	352	**100.0**	**281**	**100.0**

Table 6-24. Most frequent source meta genres belonging to 352 followed links

Looking at the apportionment between institutional and personal source meta genres, Table 6-25 below shows that there are outlinks from personal link lists to all 17 target genres, whereas institutional link list genre has 12 *out-genres*, that is, target genres. These differences may reflect more diverse interests and purposes for creating personal link lists. However, again it should be stressed that these findings based on a small non-random sample should not be over-interpreted, as a randomized larger sample would probably yield other results. As noted earlier, the present presentation should thus be viewed as an exploratory identification and conceptualization of elements in academic web genre connectivity.

Table 6-25 shows the percentage of outlinks targeted to institutional pages for each source genre. For example, 54.5% of the followed links from personal lists were

targeted to institutional pages, whereas 85.1% of the followed links from the institutional lists had such targets. Again, this difference is perhaps not surprising because of the different purposes for the two list genres.

source meta genres	# target meta genres	# links	# links to inst. target pages	% n=352
p.list	17	112	61	54.5
i.list	12	87	74	85.1
p.teach	11	22	9	40.9
i.conf	7	34	21	61.8
i.proj	7	20	18	90.0
p.hp	7	22	16	72.7
p.publ	6	7	2	28.6
p.soft	6	21	4	19.0
i.soft	4	6	3	50.0
i.generic	2	6	6	100.0
i.publ	2	8	8	100.0
i.teach	1	4	4	100.0
p.hobby	1	3	0	0.0
		352	**226**	**64.2**

Table 6-25. Source meta genres with outlinks to most different target meta genres.

In Table 6-26 below, the distribution of institutional and personal meta genres is summed up. Reading the table horizontally, it appears that there are more (81.2%) outlinks from institutional source pages to institutional target pages, than (49.2%) from personal source pages to institutional target pages. Totally, 226 (64.2%) of all the 352 followed outlinks had institutional target pages.

links from/to	inst. target		pers. target			
inst. source	134	81.2%	31	18.8%	**165**	100.0%
pers. source	92	49.2%	95	50.8%	**187**	100.0%
	226	64.2%	126	35.8%	**352**	100.0%

Table 6-26. Distribution of followed outlinks between institutional or personal source pages or target pages.

6.4.5.3 Target genres of followed links

Reversing the spectator's perspective from the previous section in order to focus on the target page meta genres of the 352 followed links, Table 6-27 below shows the distribution of source meta genres on institutional and personal target meta genres.

source meta genre	institutional target meta genre	# links	sub total	source meta genre	personal target meta genre	# links	sub total
i.list	i.archive	7	11	p.soft	p.archive	2	3
p.list	i.archive	3		p.list	p.archive	1	
p.publ	i.archive	1		p.list	p.hobby	11	15
i.conf	i.conf	9	20	i.list	p.hobby	2	
p.list	i.conf	7		p.publ	p.hobby	1	
i.proj	i.conf	2		p.teach	p.hobby	1	
i.soft	i.conf	2		i.conf	p.hp	10	47
p.list	i.generic	7	11	p.list	p.hp	10	
i.conf	i.generic	2		p.soft	p.hp	9	
i.generic	i.generic	2		i.list	p.hp	5	
i.list	i.hp	33	80	p.hp	p.hp	4	
p.list	i.hp	15		p.hobby	p.hp	3	
i.proj	i.hp	9		i.proj	p.hp	2	
p.hp	i.hp	8		p.publ	p.hp	2	
i.publ	i.hp	7		i.soft	p.hp	1	
i.conf	i.hp	4		p.teach	p.hp	1	
p.soft	i.hp	2		p.list	p.list	4	5
p.teach	i.hp	2		p.teach	p.list	1	
i.list	i.list	14	34	i.list	p.proj	2	5
p.list	i.list	8		p.list	p.proj	2	
i.generic	i.list	4		p.teach	p.proj	1	
p.teach	i.list	4		p.list	p.publ	17	24
i.conf	i.list	2		i.conf	p.publ	3	
i.proj	i.list	1		i.list	p.publ	1	
p.publ	i.list	1		p.publ	p.publ	1	
p.list	i.proj	8	24	p.soft	p.publ	1	
i.list	i.proj	5		p.teach	p.publ	1	
i.conf	i.proj	4		p.soft	p.soft	5	17
i.proj	i.proj	3		p.teach	p.soft	4	
p.hp	i.proj	2		p.list	p.soft	3	
i.publ	i.proj	1		i.soft	p.soft	2	
p.teach	i.proj	1		p.hp	p.soft	2	
i.list	i.publ	5	11	i.list	p.soft	1	
p.hp	i.publ	3		p.teach	p.teach	4	10
p.list	i.publ	2		p.list	p.teach	3	
i.proj	i.publ	1		i.list	p.teach	2	
p.list	i.soft	4	10	p.publ	p.teach	1	
i.proj	i.soft	2					
p.soft	i.soft	2					
i.soft	i.soft	1					
p.hp	i.soft	1					
i.list	i.teach	10	25				
p.list	i.teach	7					
i.teach	i.teach	4					
p.hp	i.teach	2					
p.teach	i.teach	2					
			226				**126**
							352

Table 6-27. Distribution of source meta genres on institutional and personal target meta genres.

All 9 institutional meta genres and 8 personal meta genres were represented among the target pages. Institutional homepages was the genre receiving most of the followed links (22.7%) as shown in Table 6-28 below. At the same time, the institutional homepages comprised 16.9% of the visited target pages, reflecting the high number of inlinks per target page.

target meta genre	# followed links	% n=352	# visited target pages	% n=249
i.hp	80	22.7	42	16.9
p.hp	47	13.4	34	13.7
i.proj	24	6.8	24	9.6
i.list	34	9.7	23	9.2
p.publ	24	6.8	22	8.8
i.teach	25	7.1	18	7.2
p.soft	17	4.8	14	5.6
i.conf	20	5.7	12	4.8
p.hobby	15	4.3	12	4.8
i.generic	11	3.1	10	4.0
i.soft	10	2.8	9	3.6
i.publ	11	3.1	7	2.8
p.teach	10	2.8	7	2.8
p.list	5	1.4	5	2.0
i.archive	11	3.1	4	1.6
p.proj	5	1.4	4	1.6
p.archive	3	0.9	2	0.8
	352	**100.0**	**249**	**100.0**

Table 6-28. Most frequent target meta genres belonging to 352 followed links

Table 6-29 below shows that personal homepages had the highest diversity of inlinking source genres in this small case study, having 10 different *in-genres*, whereas institutional homepages had 8 different in-genres. Not surprisingly, only 38.3% of the inlinks to personal homepages originated from institutional source pages, whereas 66.3% of the inlinks to institutional homepages had such origin.

target meta genre	# source meta genres	# inlinks	# links from inst. source pages	% n=352
p.hp	10	47	18	38.3
i.hp	8	80	53	66.3
i.list	7	34	21	61.8
i.proj	7	24	13	54.2
p.publ	6	24	4	16.7
p.soft	6	17	3	17.6
i.teach	5	25	14	56.0
i.soft	5	10	3	30.0
i.conf	4	20	13	65.0
i.publ	4	11	6	54.5
p.teach	4	10	2	20.0
p.hobby	4	15	2	13.3
i.archive	3	11	7	63.6
i.generic	3	11	4	36.4
p.proj	3	5	2	40.0
p.list	2	5	0	0.0
p.archive	2	3	0	0.0
		352	**165**	**46.9**

Table 6-29. Target meta genres with inlinks from most different source meta genres.

Reading earlier Table 6-26 vertically, there are more (59.3%) inlinks to institutional target pages from institutional source pages, than (24.6%) to personal target pages from institutional source pages as shown in Table 6-30 below.

	inst. target		pers. target			
inst. source	134	59.3%	31	24.6%	165	46.9%
pers. source	92	40.7%	95	75.4%	187	53.1%
	226	100.0%	**126**	100.0%	**352**	100.0%

Table 6-30. Distribution of followed inlinks between institutional or personal source pages or target pages.

The two table readings may be summed by noting there are more links (187) *from personal* source pages, than (165) from institutional source pages, and more links (226) *to institutional* target pages, than (126) to personal target pages in this case study.

6.4.5.4 Web of genres

There were 83 different interlinked genre pairs in the case study of 10 path nets, that is, pairs of page meta genres connected by the 352 followed page level links. In Table 6-31, the most frequent genre pairs are listed (see full list in Appendix 15). The most frequent genre pairs were institutional link lists linking to institutional homepages (9.4%), personal link lists linking to personal publications (4.8%); personal link lists linking to institutional homepages (4.3%); and institutional link lists linking to other such lists (4.0%). Again, it should be noted that the pages are located at subsites belonging to different universities.

source meta genre	target meta genre	# links	% n=352
i.list	i.hp	33	9.4
p.list	p.publ	17	4.8
p.list	i.hp	15	4.3
i.list	i.list	14	4.0
p.list	p.hobby	11	3.1
i.conf	p.hp	10	2.8
i.list	i.teach	10	2.8
p.list	p.hp	10	2.8
i.conf	i.conf	9	2.6
i.proj	i.hp	9	2.6
p.soft	p.hp	9	2.6
p.hp	i.hp	8	2.3
p.list	i.list	8	2.3
p.list	i.proj	8	2.3
i.list	i.archive	7	2.0
i.publ	i.hp	7	2.0
p.list	i.conf	7	2.0
p.list	i.generic	7	2.0
p.list	i.teach	7	2.0
...	...	...	...

Table 6-31. Most frequently interlinked genre pairs (cut-off < 7 links).
Full list in Appendix 15.

An adjacency matrix was constructed in order to visualize which genre pairs were connected (see Appendix 15). Based on the matrix, the network analysis tool Pajek extracted the corresponding graph shown in Fig. 6.37. The width of the links reflects the

number of genre pairs. 'Selflinks', in this case between the same genres, are not shown graphically in Pajek. Some of the links are bi-directional showing that there were reciprocal links between the genres, for example, between *i.soft* and *p.soft*. In the figure, white nodes denote institutional meta genres, and red (dark) nodes personal.

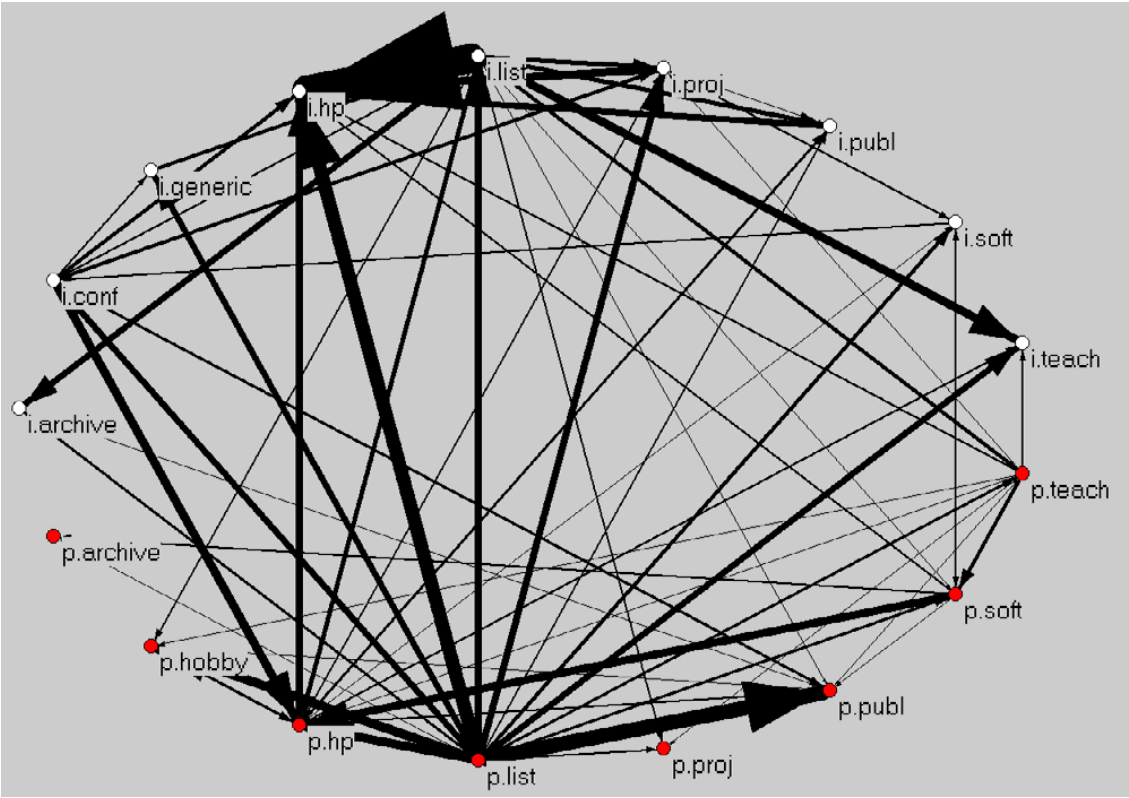


Figure 6.37.* A *web of genres.* Genre pairs among 352 followed links. Link width reflects link counts. Due to the Pajek software, thinner reciprocal links are concealed underneath thicker links. Genre selflinks are not shown. White nodes denote institutional meta genres and red personal.

On the real Web, it is possible to follow link paths connecting pages belonging to different genres as illustrated in Fig. 6.38 below. Fig. 6.37 above may give an impression of such *genre connectivity* and possible paths between genres, even if target pages have no links to source pages in the 10 path nets as discussed in Sections 4.2.2 and 6.4.1. In other words, in the 10 path nets, genres are only connected pair-wise in *inter-site* connections as shown in Fig. 6.39 and not *intra-site*.

Figure 6.38. Example of link path along page genres on the Web.

Figure 6.39. Example of link path along page genres on path nets in data set.

However, Fig. 6.39 above shows real genre pairs indicating the diversity of ways in which page genres link to each other. The figure gives an intuitive support to how the Web may be conceived as a *web of genres* with genres linked to other genres and with *genre drift*, that is, changes in page genres along link paths, as illustrated in Fig. 6.38 and 6.39 above.

The implications of genre connectivity and genre drift on small-world phenomena in document spaces on the Web will be further discussed in Chapter 7.

6.5 Transversal links in an academic web space

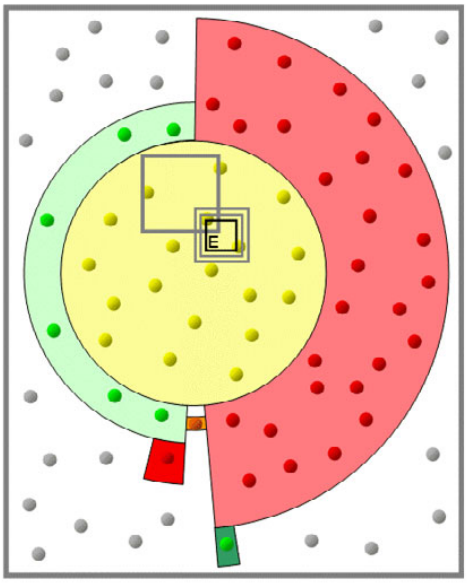


Figure 6.40. Step E in the five-step methodology: transversal links in an academic web space.

The methodology developed in the dissertation comprises five steps of 'zooming in' deeper and deeper into the data set. It is an exploration into more and more fine-grained web node levels.

Step A was concerned with identifying the graph components of the UK academic subweb space as of June-July 2001 represented by the delimited link data set containing only links between subsites located at different UK universities. 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 their overall topics and genres. The reason to focusing on subsites in the SCC was that all such subsites can reach each other through

link paths, thus enabling a sample of the shortest of these link paths between pairs of subsites. Such a sample was extracted in step C resulting in 10 path nets comprising all shortest paths in both directions between five pairs of subsites belonging to dissimilar topics. In step D, the genres and topics of the visited source and target pages along the followed link paths in the 10 path nets were classified.

The objective of all these developed methodological steps has been to lead up to the final step E in the present section, concerned with identifying what kind of links, web pages and web sites provide transversal shortcuts across dissimilar topical domains in a small-world academic web space – the main research question in the dissertation.

This section is organized as follows. First, transversal links and topic drift are discussed in Section 6.5.1. Different types of transversal links are outlined in Section 6.5.2, followed by a closer examination in Section 6.5.3 of link paths in the 10 path nets containing transversal links. Subsites with transversal links are characterized in Section 6.5.4, and finally, in Section 6.5.5, transversal page genres are identified.

6.5.1 Topic drift and transversal links

As noted in Section 2.3.1, most links within and between web sites connect web pages containing cognate topics (cf. Davison, 2000). This *topical propensity* leads to the emergence of topic-focused cluster-like formations in a web space. For example, in Fig. 5.14, Section 5.3.2, all the subsites in the neighborhood of node 945 were concerned with microbiology and biological sciences as was node 945. However, some links in a web neighborhood may break such topical linkage patterns and function as shortcuts between two dissimilar topical domains. In the conceptual framework presented in Section 2.3.1, the term *transversal* is used to denote such cross-topic links that may contribute to the formation of small-world properties in the shape of short link paths on the Web.

In many web experiments, so-called *topic drift* (Bharat & Henzinger, 1998) imposes a problem when the objective of the experiments is to conduct *focused* web crawls in order to identify interest communities and other topically focused areas on the Web. The topic drift problem is concerned with the change of topics when a human web surfer or a digital web crawler follows links from web page to web page. For example, a link path starting from a subsite on ophthalmology (eye research) may after a few steps have ended in a subsite concerned with geography. The 10 path nets in the present study all constitute examples of deliberately induced topic drift, in order to identify micro-level properties of transversal links affecting topic drift and small-world phenomena.

In order to identify topic drift and thus transversal links crossing topical domains, the overall topics of both source and target *pages* of the 352 followed links in the 10 path nets were compared, as well as the overall topics of the visited source and target *subsites* along the followed link paths. In theory, topic drift may be both *intra-site* and *inter-site*, that is, it may occur both *within* and *between* web sites, in this case subsites. Furthermore, topic drift may also be *inter*-page, that is, it may occur between web pages. Fig. 6.41 shows four different combinations of these different topic drifts, dependent on whether the web pages have the same topic as their subsite.

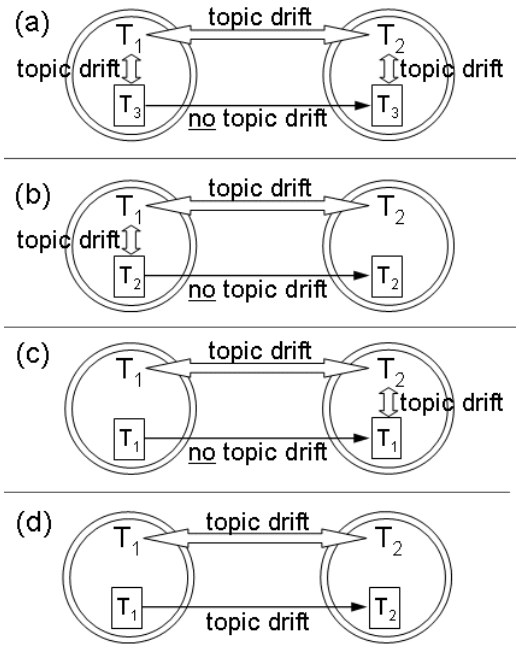


Figure 6.41. Four different instances of topic drift dependent on whether the web pages have the same topic T_x as the subsite they belong to.

In Fig. 6.41a-d, all pairs of subsites have the overall topics T_1 and T_2, respectively, with topic drift between the subsites. In Fig. 6.41a, there is *intra*-site topic drift within both the source and target subsite, and the source and target page has the same topic. An example from the present study is a researcher in chemistry (topic T_1 in Fig. 6.41a) with a personal link list about football (T_3) containing a link to a student's hobby page on the same football topic, located at a department in computer science and electrical engineering (T_2).

In Fig. 6.41b, there is no intra-site topic drift with regard to the target subsite and target page. For example, a researcher in astronomy (T_1) has a personal link list at a source subsite about cryptography legislation (T_2) with an outlink to a computer scientist's personal link list about the same topic. In this case, cryptography was classified as belonging to computer science, thus there was no topic drift within the target subsite. A similar example could illustrate Fig. 6.41c, however with no topic drift within the *source* subsite. In Fig. 6.41d, there is no intra-site topic drift within neither source nor target subsites. The page topics are thus similar to the subsite topics.

The four figures do not cover all possible combinations of topic drift. For instance, a variant of Fig. 6.41a could contain inter-page topic drift between two different page topics T_3 and T_4. Another combination could comprise pages where no clear overall topic could be identified, for example, some personal link lists pointing to a diversity of topics.

Because of the focus on links *between* subsites in this study, only links connecting subsites with different topics were considered. All four instances of topic drift in Fig. 6.41 thus include *inter-site* topic drift. However, as shown in Fig. 6.42 below, inter-*page* topic drift may occur even if there is no inter-*site* topic drift. Other combinations

of such inter-page topic drift not paralleled with inter-site topic drift will not be shown here.

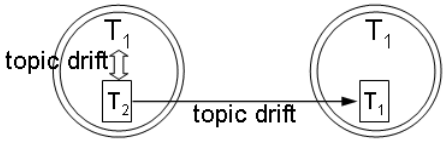


Figure 6.42. Example of inter-*page* topic drift not paralleled with inter-*site* topic drift.

6.5.2 Types of transversal links

There were identified 112 transversal links between dissimilar subsite topics among the 352 followed links in the 10 path nets. The 240 links not classified as transversal, comprised links connecting similar topics or links adjacent to generic-type subsites with no clear overall research topic. Table 6-32 shows the distribution of the 112 transversal links regarding intra-site topic drift in source and target subsites following the four types outlined in Fig. 6.41 above.

intra-site topic drift in source and target subsites	# transversal links	% n=112
in both source and target (type a)	17	15.2
in source only (type b)	17	15.2
in target only (type c)	11	9.8
in neither source nor target (type d)	67	59.8
	112	**100.0**

Table 6-32. Distribution of transversal links regarding intra-site topic drift in source and target subsites following the four types of Fig. 6.41.

The type (a) transversal links (intra-site topic drift in both source and target subsites) were typically between leisure-related web pages located in subsites with research topics. For example, a researcher in atmospheric, oceanic and planetary physics (node 1904, path net NH01), at Oxford, with a personal link list about running had a link to a researcher at the Department of Phonetics and Linguistics (node 2744), University College London, with a personal hobby page also about running, thus creating a transversal link between the two subsites. Note that the topic drift in this study thus is on the subsite level and not necessarily on the page level (both web pages dealt with the same topic, running, in the above example).

A type (b) transversal link (intra-site topic drift in the source subsite only) could be a researcher in electrical engineering (node 2615, path net NH01) having a web page about his hobby, Saxon shore forts, with a link to a personal homepage of a researcher in history at the Faculty of Arts (node 1451).

An example of a type (c) transversal link (intra-site topic drift in the target subsite only) is a link from an institutional link list on climate change at the Internet Biodiversity Service, University of East London (node 2858, path net NH04) targeted to an institutional research project on global climate change conducted at the Department

of Environmental and Geographical Sciences, Manchester Metropolitan University, but with web pages hosted at the Department of Computing and Mathematics (node 1572).

Table 6-32 shows that about 60% of the 112 transversal links belonged to the type (d) category with no intra-site topic drift in source or target subsites. In other words, the web pages had the same overall topics as each of their surrounding subsites. For example, a researcher in computer science (node 2760:) with a very extensive bibliography (*www.cs.ucl.ac.uk/staff/M.Sewell/papers.htm*)[75] including links to full text papers online, for instance a paper on computational economics at the NetEc subsite, an online clearinghouse in economics (node 1485: *netec.mcc.ac.uk*).

The question of 'transversality' and topic drift between and within subsites was determined by the author alone based on the affiliations and topical descriptions given by the visited subsites and web pages. Determining topic drift between subsites was not trivial because of the many interdisciplinary and multidisciplinary departments and other subunits at the UK universities, for example, as discussed further below in relation to geography and overlapping research areas in environmental studies and earth sciences. A pragmatic view on transversal links is employed in the dissertation, focusing on links crossing more clear-cut topical borders. For example, between humanities (node 337) and meteorology (node 1904) in link path HN01-04 in Table 6-36 further below (Section 6.5.4), with a transversal link connecting an institutional link list about ancient military history with an oceanographer's personal hobby web page about an ancient Greek warship.

More objective heuristics for deciding dissimilarity between topics were considered, for instance, employing measures of low co-word occurrence, low co-inlink and/or low co-outlink measures of the subsites. However, these heuristics were too manually time-consuming and impractical to implement without the necessary programming skills.

Of the 112 transversal links, 64 (57.1%) originated from a personal web page, whereas 48 (42.9%) came from an institutional web page as defined in Section 6.4.5.

As shown in Table 6-33 below, the *personal* links were subdivided into research-related, teaching-related, interest/leisure-related, and career-related links. The research-related and teaching-related links were directly related with the overall research and teaching activities, respectively, of the person as appearing from a close inspection of the person's local web territory including parent pages, sibling pages or child pages. The personal interest links were further subdivided into academic and non-academic target topics. There were 16 (14.3%) personal non-academic transversal links targeted to leisure-related topics, such as hobbies, charity, tourist information, and family relations (a link between two brothers studying European Studies (node 2099, path net HN01) and computer science (node 2387), respectively). The personal career-related link was created by a technical staff member at the Edinburgh Parallel Computing Centre (node 732, path net NH04) having a personal homepage with a link to a former employer, the Department of Phonetics and Linguistics, University College London (node 2744).

The *institutional* links were subdivided into generic, research-related and teaching-related links as shown in Table 6-33 (cf. Appendix 19). The generic links had

[75] Part of a very large personal web territory: *http://www.cs.ucl.ac.uk/staff/M.Sewell/*

broad target topics, for example, lists with UK academic web sites, and general advice on studies abroad for students. This category also comprised four automatic outlinks back to inlinking subsites created by a web server statistical program. These four links were judged neither academic nor non-academic due to their automated generation. Together with the abovementioned 16 personal non-academic transversal links, this left 92 (82.1%) of the 112 transversal links judged as academic (marked with yellow (grey) color in the table), comprising 48 personal and 44 institutional links. In other words, a quite large percentage of the transversal links were academic, either related to research or teaching. This point is further discussed in Section 7.1.1.

112	**TRANSVERSAL LINKS**			
64	PERSONAL			
	**20*	*personal research-related transversal link*		
		**3*	personal research-related (PhD student) transversal link	
		**17*	personal research-related (researcher) transversal link	
	**9*	*personal teaching-related transversal link*		
		**5*	personal teaching-related (researcher) transversal link	
		**4*	personal teaching-related (student) transversal link	
	34	*personal interest transversal link*		
		**18*	personal interest transversal link: academic	
			**2*	personal interest (adm.staff): lists with UK academic web sites
			**1*	personal interest (tech.staff): software
			**4*	personal interest (researcher): lists with UK academic web sites
			**3*	personal interest (researcher): legislation on information policy
			**4*	personal interest (researcher): research
			**2*	personal interest (researcher): teaching
			**1*	personal interest (PhD student): research
			**1*	personal interest (student): lists with UK academic web sites
		16	personal interest transversal link: non-academic leisure/hobby	
			9	personal interest (researcher): hobby (sport)
			2	personal interest (researcher): hobby (science fiction)
			1	personal interest (researcher): charity
			1	personal interest (researcher): tourist information
			2	personal interest (student): hobbies (backgammon, sport)
			1	personal familiar (student) relation: link to brother
	**1*	*personal career-related (tech.staff) transversal link: link to former employer*		
48	INSTITUTIONAL			
	14	*institutional generic transversal link*		
		**5*	studying abroad	
		**5*	lists with UK academic web sites	
		4	automatic outlink back to inlinking subsite: web server statistics	
	**31*	*institutional research-related transversal link*		
	**3*	*institutional teaching-related transversal link*		

Table 6-33. 112 transversal links subdivided into personal and institutional categories based on *source* page genre. Personal links are listed after profession. Counts of academic links are marked with yellow and an asterisk. Cf. Appendix 19.

6.5.3 Link paths with transversal links

The present section outlines how link paths containing transversal links were identified. As described in Section 6.4.2, all link paths in the smaller path nets were followed, whereas only link paths not passing generic-type subsite nodes were followed in the large path nets NH01, NH02 and NH04, giving a total of 81 followed link paths in the 10 path nets.

In Table 6-34, the overall topics of the *source subsites* providing the 352 links on the 81 followed link paths are listed. The overall topics are grouped after the 'hum/soc' and 'nat/tech' topic groups constructed when categorizing the 189 SCC subsites in Section 6.2.1. Logically, the same visited subsite may occur on several followed link paths in the same path net. Table 6-34 thus also shows counts of the topics of the 93 *unique* visited source subsites. As noted earlier, a start node in a path net cannot be the target of any links in the path net, and vice versa for an end node, thus the sum of 93 subsites (of 103 visited subsites) in the table.

For example, there were 3 medicine subsites providing 17 outlinks on the followed link paths. The abbreviations are also used in the link paths in Appendix 16.

source subsite topic	abbrev.	topic group	seed topic	unique visited source subsites	% n=93	followed outlinks	% n=352
generic	gen	-		10	10.8	19	5.4
generic: learning technology	gen/learn	-		2	2.2	5	1.4
multidisciplinary	multi-sci	-		1	1.1	3	0.9
arts & humanities	hum	A		5	5.4	10	2.8
humanities & social sciences	hum/soc	A	x	1	1.1	11	3.1
social sciences	soc	A		1	1.1	4	1.1
economics	econ	B	x	3	3.2	20	5.7
economics: learning tech.	econ/learn	B		1	1.1	3	0.9
linguistics	ling	C	x	4	4.3	8	2.3
archaeology	hum/archae	D		2	2.2	3	0.9
geography	geo	D	x	8	8.6	22	6.2
psychology	psych	E	x	1	1.1	1	0.3
earth sciences	earth	F	x	5	5.4	18	5.1
environmental studies	environ	F		2	2.2	7	2
medicine	med	G	x	3	3.2	17	4.8
chemistry	chem	H	x	3	3.2	12	3.4
astronomy	astro	I		1	1.1	4	1.1
meteorology	met	I	x	3	3.2	42	11.9
mathematics	ma	J	x	5	5.4	7	2
engineering	engineer	K		3	3.2	3	0.9
computer science	cs	L		14	15.1	73	20.7
comp.sci. & electr. engineering	cs/ee	L		8	8.6	34	9.7
comp.sci. & mathematics	cs/ma	L		4	4.3	19	5.4
comp.sci. & info.science/tech.	cs/is	L		2	2.2	6	1.7
comp.sci. & cognitive sciences	cs/cog	L		1	1.1	1	0.3
				93	**100.0**	**352**	**100.0**

Table 6-34. Topics of *all* visited source subsites with followed outlinks as well topics of *unique* visited source subsites. The topics are sorted by the topic groups of the 189 SCC subsites (Section 6.2.1): 'hum/soc' (A-E) and 'nat/tech' (F-L). The 10 seed subsite topics in the path nets are marked. The abbreviations are also used in the link paths in Table Appendix 16.

The line between 'hum/soc' and 'nat/tech' topics in Table 6-34 should not be drawn too sharply because of interdisciplinary overlaps. In the topical categorization of the sample of 189 SCC subsites in Section 6.2.1, geography was placed in the 'hum/soc' group following the RAE assessments (HERO 2001; cf. Appendix 9). This placing was justified by the interdisciplinary span encompassing both cultural and physical geography. However, when examining the followed links in the path nets it turned out to be natural in the present study to consider geography as closely related to earth sciences belonging to the 'nat/tech' group. For example, in link path NH04, an institutional link list at node 1343, the School of Earth Sciences, University of Leeds, (*earth.leeds.ac.uk/kennedy/ukfieldtrips.htm*) concerned with virtual field trips has an outlink to an institutional teaching page, 'the Virtual Field Course' (*www.geog.le.ac.uk/vfc/*) at the Department of Geography, University of Leicester. This link together with other links between earth sciences and geography was not considered a transversal link because of the interdisciplinary overlapping research areas of the disciplines.

Similarly, environmental studies and meteorology posed interdisciplinary overlaps with geography and earth sciences making it inconvenient to designate links between them as transversal. For instance, in path net NH04, a researcher in the Satellite Remote Sensing Team (research group in ocean circulation and climate) at the Southampton Oceanography Centre (node 2356) has made an outlink from his personal homepage *(www.soc.soton.ac.uk/JRD/SAT/pers/cipo.html*) to a FAQ (Frequently Asked Questions) on satellite imagery (*www.geog.nottingham.ac.uk/remote/satfaq.html*)[76] located at node 1709, the School of Geography, Faculty of Law and Social Sciences, University of Nottingham.

This example illustrates how topic drift may appear as sliding transitions along overlapping interdisciplinary topics.

In Fig. 6.43 below, subsites in path net HN03 belonging to the topical areas computer science (cs) and mathematics (math) are enclosed by thick yellow and thin white borders, respectively. Node 1089 is the School of Mathematical and Computer Sciences, Heriot-Watt University, that covers both topics. Transversal links crossing disciplinary borders between computer science and mathematics are marked with dashed bold links. A pragmatic view on transversal links is employed in the dissertation, incorporating links between computer science and all disciplines that use computer science as an auxiliary tool. Links between subsites in mathematics and all disciplines that make use of mathematical models were also treated as transversal links in the present study. Path net HN03 in Fig. 6.43 also contains a transversal link between the Department of Psychology, University of Manchester (node 1494) and the School of Computing and Information Technology, University of Wolverhampton (node 3020) – once again, the target was the highly inlinked map of UK universities, cf. Appendix 1.
The topics of the 10 seed subsite nodes that function as start and end nodes of the path nets naturally influence the number of subsites with similar topics in each path net. For example, in path net HN03 in Fig. 6.43 below, the end node topic is mathematics, as is the case with all the three in-neighbor nodes on level 3 in the path net.

[76] Available in the Internet Archive:
http://web.archive.org/web/19990506063212/http://www.geog.nottingham.ac.uk/remote/satfaq.html

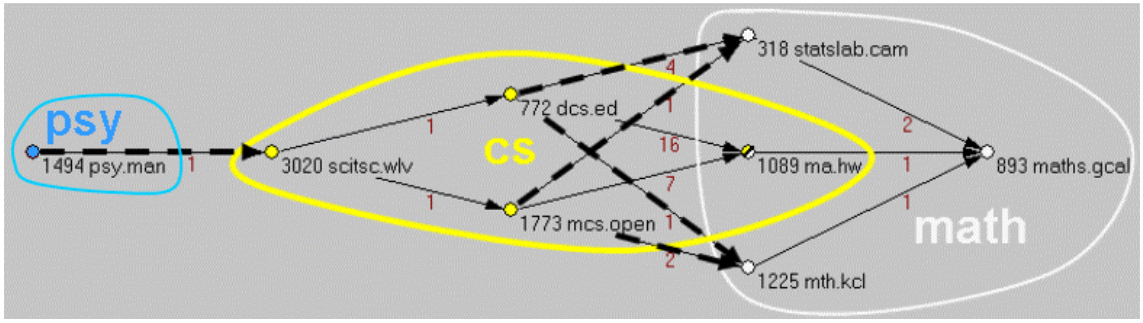


Figure 6.43.* Path net HN03 with the enclosed topical areas psychology (psy), computer science (cs) and mathematics (math). See Appendix 10 for affiliations. Transversal links crossing disciplinary borders are denoted in dashed bold. Counts of page level links are shown. (*cf. color prints placed before appendices).

A list of all 81 followed link paths in the 10 path nets as shown in Appendix 16 was used in the identification of where transversal links occurred on the link paths. Table 6-35 below gives an excerpt of the list showing the six different link paths between the start and end node in path net HN03 in Fig. 6.43 above.

path net	link path	level 0		level 1		level 2		level 3		level 4
HN03	HN03-01	1494psy	>	3020cs	-	772cs	**> #**	318ma	-	893ma
	HN03-02	1494psy	>	3020cs	-	772cs	-	1089cs/ma	-	893ma
	HN03-03	1494psy	>	3020cs	-	772cs	>	1225ma	-	893ma
	HN03-04	1494psy	>	3020cs	-	1773cs/ma	-	318ma	-	893ma
	HN03-05	1494psy	>	3020cs	-	1773cs/ma	-	1089cs/ma	-	893ma
	HN03-06	1494psy	>	3020cs	-	1773cs/ma	-	1225ma	-	893ma

Table 6-35. Excerpt of followed 81 link paths in 10 path nets (see Appendix 16 for full list). Subsites are denoted with id number and abbreviated topic. Bold right angle brackets (>) denote one or more research-related transversal links; hash sign (#) marks personal *non-academic* transversal links; non-bold right angle brackets (>) denote other transversal links; hyphens (-) mark non-transversal links.

As noted earlier, many links on the followed link paths connected subsites with similar topics, for example, a link between two mathematics-related subsites such as the link between node 1225 and node 893 in path net HN03 above. Links to and from campus-wide generic-type subsites were treated as non-transversal. For example, node 119 (*web.bham.ac.uk*) in path net HN01 (Fig. 6.44 below) was such a campus-wide subsite containing personal homepages for all the staff at the University of Birmingham. Node 335 (*cus.cam.ac.uk*) in the same path net is also a campus-wide generic-type subsite comprising the Central Unix Service for all university staff and students in Cambridge. There are five such generic-type subsites in path net HN01 as illustrated in Fig. 6.44 by not belonging to any of the four enclosed topical areas.

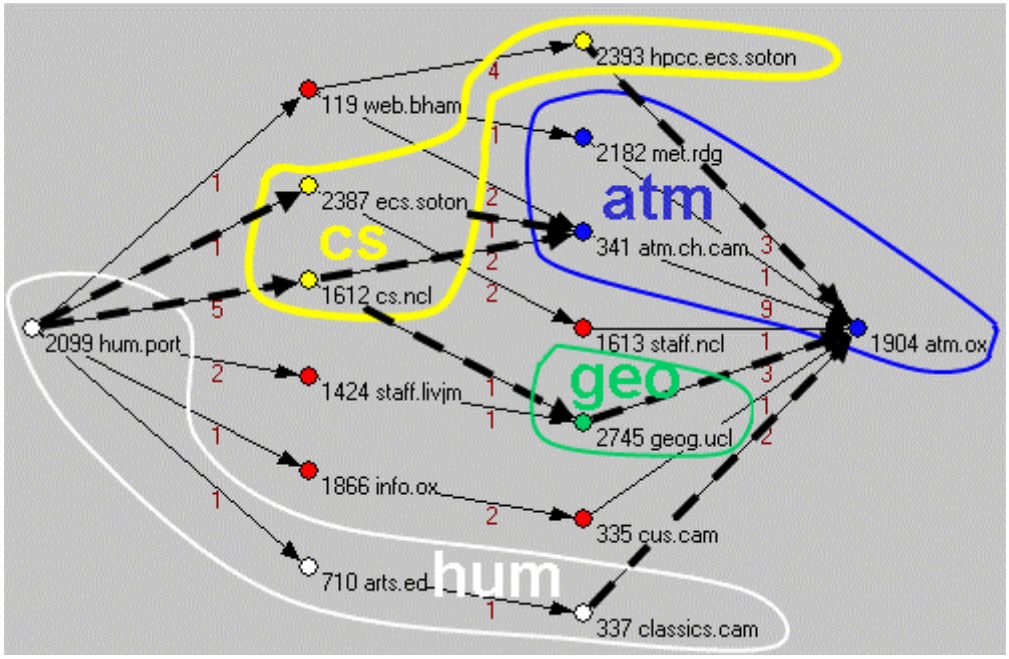


Figure 6.44.* Path net HN01 with enclosed topical areas humanities (hum), computer science (cs), geography (geo) and atmospheric sciences (atm). Non-enclosed nodes are generic-type. Transversal links are marked with dashed bold links. See Appendix 10 for affiliations.

As stated above, a pragmatic view on transversal links was employed, focusing on links crossing more clear-cut topical borders. For example, there is a transversal link between a humanities subsite (node 337) and a subsite in Atmospheric, Oceanic and Planetary Physics (node 1904) in path net HN01 in Fig. 6.44 above, connecting the earlier mentioned institutional link list about ancient military history with an oceanographer's personal hobby web page about an ancient Greek warship.

The present section has outlined how link paths containing transversal links were identified. Next section takes a closer look on what types of subsites provided transversal links.

6.5.4 Subsites with transversal links

Of the 81 followed link paths, 17 contained generic-type subsites and 58 (71.6%) contained computer-science-related subsites, including two link paths with both generic and computer science subsites. Only eight (9.9%) of the 81 followed link paths contained neither computer-science-related nor generic subsites as listed in Table 6-36 below.

path net	link path	level 0		level 1		level 2		level 3		level 4
HN01	HN01-04	2099hum	-	710hum	-	337hum	>	1904atm		
NH01	NH01-01	1904atm	-	1278environ	-	1451hum(incl. geo)	-	2099hum		
	NH01-03	1904atm	> #	2615ee	>	1451hum	-	2099hum		
	NH01-04	1904atm	#	2744ling	-	1451hum	-	2099hum		
	NH01-05	1904atm	#	2744ling	-	313ling	-	2099hum		
NH04	NH04-16	245earth	-	2228earth	>	213ee	#	2744ling	-	871ling
	NH04-08	245earth	-	1853hum/archaeo	-	337hum	-	2744ling	-	871ling
NH05	NH05-03	1885med	-	102med	>	922hum	>	1327geo	-	2068geo

Table 6-36. Only eight followed link paths contained *no* computer-science-related nor generic subsites. (See legend in Appendix 16).

In link path HN01-04, the transversal link between a humanities subsite (node 337) and a subsite in Atmospheric, Oceanic and Planetary Physics (node 1904) has been described in the previous section.

Link path NH01-01 in Table 6-36 contains no transversal links. Instead, the link path reflects an incremental topic drift comprising sliding transitions between overlapping topical areas. For example, the link between node 1904 (*atm.ox.ac.uk*) and node 1278 (*es.lancs.ac.uk*) connects a researcher at Atmospheric, Oceanic and Planetary Physics, Department of Physics, University of Oxford with a cognate research group in atmospheric chemistry at the Department of Environmental Science, Lancaster University. Furthermore, the link between node 1278 (*es.lancs.ac.uk*) and node 1451 (*art.man.ac.uk*) in the same link path connects a research group in hydrology with a cognate researcher engaged in the Royal Society Southeast Asia Rainforest Research Programme, at the School of Geography with web pages placed in a subdirectory at the Faculty of Arts, University of Manchester. The last stretch between node 1451(*art.man.ac.uk*) and node 2099 (*hum.port.ac.uk*) comprises two page level links connecting French Studies at the Faculty of Arts in Manchester with the School of Languages and Area Studies at the Faculty of Humanities and Social Sciences, University of Portsmouth.

Faculty subsites had not been excluded from the link paths because one of the seed nodes (2099 *hum.port.ac.uk*) was a faculty (Faculty of Humanities and Social Sciences, University of Portsmouth). However, the broadness of research topics covered by faculty subsites meant that only links to or from subsites with research topics *not* contained at the faculty were treated as transversal links in the study. For example, in the four link paths HN01-06/07/09/10 shown in Appendix 16, there are transversal links from the mentioned node 2099 to computer-science-related subsites (nodes 1612 and 2387).

Node 1853 (*.ashmol.ox.ac.uk*), The Ashmolean Museum, Museum of Art & Archaeology, University of Oxford, in link path NH04-08 in Table 6-36 also function as a topical 'overlapper' receiving an inlink from node 245 (*palaeo.gly.bris.ac.uk*), Palaeontology Research Group, Department of Earth Sciences, University of Bristol.

Table 6-37 below is an excerpt from Table 6-34 in the previous section. As shown in the table, about 31% of the unique visited source subsites were computer-science-related (hereafter CS). Further, 38% of all the 352 followed *outlinks* originated from these CS-related source subsites. This count covers subsites with more than one overall

topic, for example, subsites that in their affiliation combine computer science with electrical engineering, information science, cognitive science, or mathematics.

source subsite topic	unique visited source subsites	% n=93	followed outlinks	% n=352
cs	14	15.1	73	20.7
cs/ee	8	8.6	34	9.7
cs/ma	4	4.3	19	5.4
cs/is	2	2.2	6	1.7
cs/cog	1	1.1	1	0.3
	29	**31.2**	**133**	**37.8**

Table 6-37. *All* followed *outlinks* from computer-science-related *source* subsites.

The percentage of followed *inlinks* to CS subsites were slightly smaller (36.4%) as shown in Table 6-38 below. However, the percentage of visited CS-related target subsites was identical with the source subsites, because no CS subsites functioned as start nor end nodes in any path net.

target subsite topic	unique visited target subsites	% n=93	followed inlinks	% n=352
cs	14	15.1	52	14.8
cs+ee	8	8.6	44	12.5
cs+ma	4	4.3	29	8.2
cs/is	2	2.2	2	0.6
cs/cog	1	1.1	1	0.3
	29	**31.2**	**128**	**36.4**

Table 6-38. All followed *inlinks* to computer-science-related *target* subsites.

Among the transversal links, the share of CS-related links was somewhat higher, both regarding source and target subsites. Of the 112 transversal links, 46 (41.1%) originated from CS-related subsites, whereas 45 (40.2%) were received by CS-related subsites. Among the 48 source subsites for transversal links, 22 (45.8%) were CS-related. The corresponding count for the 47 transversal target subsites also was 22 (46.8%).

In the random sample of 189 SCC subsites in Section 6.2.1, 20 subsites[77] (10.6%) were judged as CS-related. Computer science thus constitutes a larger share among the visited subsites in the 10 path nets and an even larger share among the subsites connected by transversal links. Even if the sample of 10 path nets was small, this finding may indicate a special role of CS-related subsites as connectors on shortest link paths in an academic web space. For example, as shown in Table 6-35 and Fig. 6.43 in the previous section, the CS subsites function as a connector topic between psychology and mathematics on all six link paths in path net HN03. Another example, cf. Appendix 16, is how computer science connects chemistry and economics in path net NH02.

[77] In Table 6-2, Section 6.2.1, 15 subsites are directly affiliated with computer science. Furthermore, five of the subsites on informatics are very close to computer science, giving a total of 20 computer-science-related subsites in the sample.

The special role of computer science is supported by the circumstance that 15 of the 25 subsites with the highest betweenness centrality among the 7669 subsites were CS-related (Table 6-12, Section 6.3.2.4).

The role of CS-related subsites in academic link structures may reflect the auxiliary function of computer science in many scientific disciplines in natural sciences, humanities, and social sciences as mentioned above. Further, this may be combined with a more well-developed web presence and more experienced web literate behavior performed by CS-related persons and institutions, reflected by a larger number of created web pages and site outlinks. However, this latter hypothesis remains to be verified.

A number of subsites received *transversal inlinks* from more than one subsite in a path net. For example, Fig. 6.45 (identical with Fig. 6.44) below shows node 1904 (Atmospheric, Oceanic and Planetary Physics at Oxford) in path net NH04 that receives transversal inlinks from three subsites in High Performance Computing (node 2393), Geography (2745), and Faculty of Classics (337).

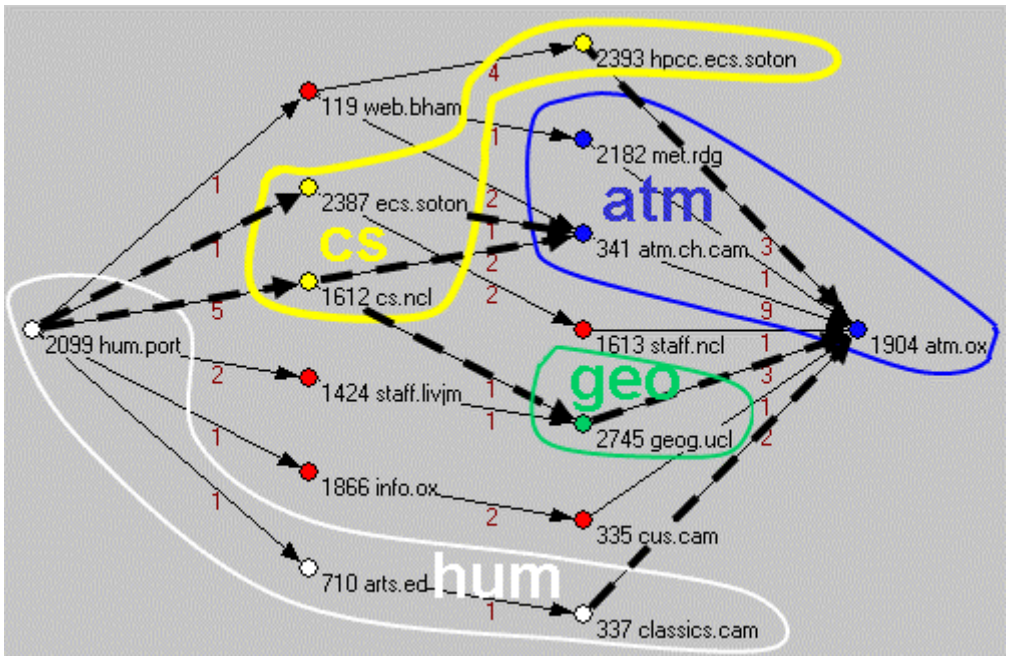


Figure 6.45. Three subsites in High Performance Computing (2393), Geography (2745), and Faculty of Classics (337) are *transversal in-neighbors* to node 1904 (Atmospheric, Oceanic and Planetary Physics at Oxford) in path net NH04. (See affiliations in Appendix 10). Cf. legend in Fig. 6.44.

In Table 6-39 below (based on Appendix 16), the subsites with more than one transversal in-neighbor are listed. Such *multi-transversal* subsites that provide many routes on shortest link paths may be of special interest as cross-topic connector nodes in an academic web space.

path net	id	path net level	short domain name	topic	# transversal *in-neighbors*	transversal in-neighbors
NH04	2744	3	phon.ucl	ling	6	213ee 732cs 1088cs/ee 1572cs/ma 2865cs 3017cs
HN01	1904	*3	atm.ox	atm	3	337hum 2393cs 2745geo
NH04	871	*4	speech.essex	ling	3	201cs 2372cs/ee 2387cs/ee
NH04	2387	3	ecs.soton	cs/ee	3	1327geo 1473ma 3060archaeo
HN02	917	*3	chem.gla	chem	3	1088cs/ee 1328cs 2865cs
NH02	1485	2	netec.man	econ	3	1088cs/ee 2537cs/is 2760cs
NH02	1641	2	lorien.ncl	chem	2	2642cs/cog 2760cs
NH04	732	2	epcc.ed	cs	2	1889earth 2356earth
NH04	1572	2	doc.mmu	cs/ma	2	1853hum/archaeo 2858environ
HN01	341	2	atm.ch.cam	atm	2	1612cs 2387cs
NH02	1890	2	nuff.ox	soc	2	1597cs 2760cs
NH04	2372	3	isis.ecs.soton	cs/ee	2	1327geo 1709geo
NH03	1494	*3	psy.man	psych	2	1268cs 2760cs

Table 6-39. Subsites with more than one *transversal in-neighbor*. An asterisk at the path net level denotes an end node. (See Appendix 16 for legend of topics of transversal in-neighbors).

Correspondingly, there were subsites providing *transversal outlinks* to more than one subsite in a path net. For example, node 917 (*chem.gla.ac.uk*: Department of Chemistry, University of Glasgow) in Fig. 6.46 further below provides transversal outlinks to five CS-related subsites in path net NH02. In Table 6-40 below, subsites with more than one transversal out-neighbor are listed.

path net	id	path net level	short domain name	topic	# transversal out-neighbors	transversal out-neighbors
NH02	917	*0	chem.gla	chem	5	1088cs/ee 1597cs 2537cs/is 2642cs/cog 2760cs
NH05	102	1	medweb.bham	med	3	917chem 922hum 3017cs
NH04	2356	1	soc.soton	earth	3	629cs/ee 732cs 1088cs/ee
NH04	1343	1	earth.leeds	earth	3	1619ma 1692cs/is 3017cs
NH02	2760	1	cs.ucl	cs	3	1485econ 1641chem 1890soc
NH04	1889	1	earth.ox	earth	2	732cs 1473ma
HN01	1612	1	cs.ncl	cs	2	341atm 2745geo
HN03	772	2	dcs.ed	cs	2	318ma 1225 ma
NH01	1904	*0	atm.ox	atm	2	2615ee 2744ling
NH04	1327	2	geog.le	geo	2	2372cs 2387cs/ee
HN01	2099	*0	hum.port	hum/soc	2	1612cs 2387cs/ee
NH04	2858	1	ibs.uel	environ	2	1572cs/ma 2865cs
NH02	1088	1	cee.hw	cs/ee	2	1485econ 2083econ
NH03	979	1	astro.gla	astro	2	1268cs 2760cs

Table 6-40. Subsites with more than one *transversal out-neighbor*. An asterisk at the path net level denotes a start node. (See Appendix 16 for legend of topics of transversal out-neighbors).

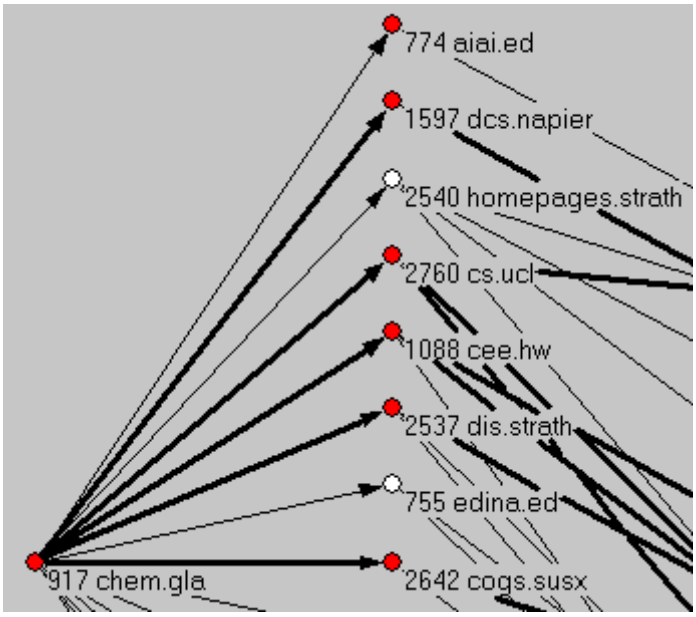


Figure 6.46. Five CS-related subsites (nodes 1088, 1597, 2537, 2642, 2760) are *transversal out-neighbors* to node 917 (Department of Chemistry, University of Glasgow) on followed link paths (in bold) in path net NH02. See affiliations in Appendix 10.

The present section has examined what types of subsites provide transversal links. Especially, there is an indication that computer-science-related subsites play an important role as cross-topic connectors in an academic web space. This point is further discussed in Section 7.1.3.

The next section is the final one in this lengthy empirical investigation into what types of web links, web pages and web sites function as transversal (cross-topic) connectors in small-world academic web spaces – the main research question in the dissertation. More specifically, the next section is concerned with what types of page genres provide transversal links.

6.5.5 Genres with transversal links

Looking at what web page genres provide and receive transversal links, a list similar to Table 6-31 in Section 6.4.5.4 was constructed. Table 6-41 below shows the 49 genre pairs (compared with 83 for all followed links) in relation to the 112 identified transversal links. The most frequent genre pairs were institutional link lists linking to institutional homepages (10.7%)(9.4% for all 352 followed links) and personal link lists linking to personal publications (8.0%)(4.8% for all followed links). The two most frequent genre pairs for the transversal links were thus identical with those of all the followed links, however with higher percentages for the two transversal genre pairs. The third most frequent transversal genre pair was personal link lists linking to personal hobby pages (7.1%). The third most frequent genre pair for all followed links was personal link lists linking to institutional homepages (4.3%).

source meta genre	target meta genre	# transversal links	%
i.list	i.hp	12	10.7
p.list	p.publ	9	8.0
p.list	p.hobby	8	7.1
i.list	i.generic	5	4.5
i.generic	i.list	4	3.6
p.list	i.generic	4	3.6
p.list	i.hp	4	3.6
p.list	i.list	4	3.6
i.list	i.proj	3	2.7
p.list	p.hp	3	2.7
p.list	p.list	3	2.7
p.teach	p.soft	3	2.7
i.conf	i.hp	2	1.8
i.conf	i.list	2	1.8
i.conf	p.hp	2	1.8
i.list	i.list	2	1.8
i.list	i.publ	2	1.8
i.list	p.hobby	2	1.8
i.proj	i.soft	2	1.8
p.hp	p.hp	2	1.8
p.list	i.proj	2	1.8
p.list	i.publ	2	1.8
p.list	i.soft	2	1.8
p.list	p.proj	2	1.8
p.soft	i.soft	2	1.8
i.conf	i.proj	1	0.9
i.conf	p.publ	1	0.9
i.list	i.teach	1	0.9
i.proj	i.conf	1	0.9
i.proj	i.hp	1	0.9
i.proj	p.hp	1	0.9
i.publ	i.proj	1	0.9
i.soft	i.soft	1	0.9
i.soft	p.hp	1	0.9
i.soft	p.soft	1	0.9
p.hobby	p.hp	1	0.9
p.hp	i.hp	1	0.9
p.hp	i.soft	1	0.9
p.hp	p.soft	1	0.9
p.list	p.archive	1	0.9
p.list	p.teach	1	0.9
p.publ	i.list	1	0.9
p.publ	p.publ	1	0.9
p.soft	i.hp	1	0.9
p.soft	p.soft	1	0.9
p.teach	i.list	1	0.9
p.teach	p.hobby	1	0.9
p.teach	p.list	1	0.9
p.teach	p.teach	1	0.9
		112	**100.0**

Table 6-41. Genre pairs for 112 transversal links sorted by frequency.

See Appendices 17 and 18 for genre pairs sorted by source genres and target genres.

6.5.5.1 Source genres of transversal links

According to Table 6-41 above, there are 9 transversal links between the genres of *p.list* and *p.publ*. However, 7 of these links originate from one single source page, the bibliography mentioned earlier[78] made by a researcher linking to 7 different papers at a single target site in economics. In order to gain a more balanced picture of the counts, Table 6-42 below gives a summary of more detailed counts of both links, pages and subsites for source genres and subgenres listed in Appendix 19 (also cf. Appendix 17). As will appear from Table 6-42, the most frequent genre providing transversal links was the personal link lists with 45 (40.2%) transversal outlinks originating from 34 web pages located on 18 different subsites. Several of the 48 subsites providing transversal outlinks contained pages with different genres. This circumstance explains why a simple sum of source subsites exceeds 48.[79] The table also shows that 12 of the 13 source genres present among all the followed links prevailed among the transversal links. Only the institutional teaching pages had no transversal outlinks, perhaps reflecting topically focused teaching.

source meta genre	# trans-versal outlinks	% n=112	unique source pages	% n=95	# source subsites
p.list	45	40.2	34	35.8	18
i.list	27	24.1	27	28.4	15
i.conf	8	7.1	6	6.3	6
p.teach	7	6.3	5	5.3	5
i.proj	5	4.5	4	4.2	4
p.hp	5	4.5	5	5.3	4
i.generic	4	3.6	4	4.2	1
p.soft	4	3.6	3	3.2	3
i.soft	3	2.7	3	3.2	2
p.publ	2	1.8	2	2.1	2
i.publ	1	0.9	1	1.1	1
p.hobby	1	0.9	1	1.1	1
i.teach	0	0.0	0	0.0	0
	112	**100.0**	**95**	**100.0**	**48**

Table 6-42. Most frequent meta genres of the 95 source pages providing 112 transversal links.

In Table 6-43 below, an excerpt from Appendix 19 (also cf. Table 6-33 in Section 6.5.2) shows some details of the *personal link lists* providing transversal links. The three numbers on each row are the counts of transversal links, pages, and subsites, respectively, for each category of personal link lists. Furthermore, the numbers of different persons that have created the lists are noted. For example, there are 23 persons behind the 34 link lists with the 45 transversal links. In other words, some persons have more than one link list in a path net. For example, a researcher at the Astronomy & Astrophysics Group, University of Glasgow (node 979, path net NH03) has made three bookmark lists on 'Cryptography and privacy', 'Cryptography legislation in the UK'

[78] *cs.ucl.ac.uk/staff/M.Sewell/papers.htm* - cf. Section 6.5.2.

[79] The simple sum of the subsites is 62, because several of the 48 subsites with transversal outlinks contained more than one page genre.

and 'Distance Education and Learning Technology: Organisations and link collections'. The transversal links from the two cryptography lists are targeted to a researcher in computer science (node 2760) also interested in this topic. The third list has a link to the Centre for the Study of Advanced Learning Technologies, Computing Department, Lancaster University (node 1268).[80]

45	**34**	**18**	**Personal link list (23 different persons)**						
		7	1	1	Bibliography (researcher)				
		2	1	1	Link list (adm. staff): bookmarks				
		30	26	15	Link list (researcher) (16 different persons incl. 1 multi-occurring in more than one path net)				
					4	4	3	Link list (researcher): research-related	
					1	1	1	Link list (researcher): personal academic interest (grammar)	
					3	2	2	Link list (researcher): UK academic sites	
					1	1	1	Link list (researcher): friends+scientists	
					1	1	1	Link list (researcher): leisure	
					2	2	1	Link list (researcher): sci-fi	
					8	7	3	Link list (researcher): sports	
					10	8	6	Link list (researcher): bookmarks (6 different persons)	
							1	1	1 Link list (researcher): bookmarks: research-related
							4	3	3 Link list (researcher): bookmarks: research-related + misc.
							5	4	2 Link list (researcher): bookmarks: personal academic interest
		3	3	3	Link list (PhD student) (3 different persons)				
					1	1	1	Link list (PhD student): research-related	
					2	2	2	Link list (PhD student): bookmarks	
							1	1	1 Link list (PhD student): bookmarks: research-related
							1	1	1 Link list (PhD student): bookmarks: research-related + misc.
		3	3	3	Link list (student) (3 different persons)				
					2	2	2	Link list (student): bookmarks	
							1	1	1 Link list (student): bookmarks: research-related + misc.
							1	1	1 Link list (student): bookmarks
					1	1	1	Link list (student): sports	

Table 6-43. Personal link lists providing transversal links sorted after type of link list and profession. The three numbers on each row are counts of transversal links, pages, and subsites, respectively. See full table of transversal source genres in Appendix 19.

The bookmark lists is an interesting subgenre within the personal link lists. These comprise long lists of bookmarks of visited web pages made by a person when using a browser to traverse the Web. The browser bookmarks can easily be converted into a web page and made publicly available as the ones investigated in the present study. The diversity of topics often present on such bookmark lists make them *transversality-prone*. The notation 'research-related + misc.' in Table 6-43 above covers a wide span of topical juxtapositions, for example, a computer scientist (node 2760, path net NH02) with a bookmark list containing a transversal link to the Department of Chemical and

[80] The three personal bookmark lists were retrieved from the Internet Archive:
http://web.archive.org/web/20010224022656/http://www.astro.gla.ac.uk/users/norman/bookmarks/lists-crypto.html
http://web.archive.org/web/20001209123200/http://www.astro.gla.ac.uk/users/norman/bookmarks/lists-S033.html
http://web.archive.org/web/20010224021317/http://www.astro.gla.ac.uk/users/norman/bookmarks/lists-S055.html

Process Engineering, Newcastle University (node 1641) regarding evolutionary computation and genetic programming. His bookmark list also includes general and leisure-related links as shown in Appendix 21.

Researchers' and students' bookmark lists are further discussed in Section 7.2.2 as providers of topical diversity and transversal links in small-world academic web spaces.

6.5.5.2 Target genres of transversal links

As was the case with Table 6-42 for the source genres, Table 6-44 below gives a summary of the more detailed counts of links, pages and subsites for *target* genres and subgenres listed in Appendix 20 (also cf. Appendix 18). Table 6-44 shows that the most frequent genre receiving transversal links was the institutional homepage with 21 (18.8%) transversal inlinks received by 16 web pages located on 15 different subsites. Almost all 17 target genres (except one: institutional archives) among the 352 followed links were represented among the transversal links.

target meta genre	# trans-versal inlinks	% n=112	# unique target pages	% n=94	# target subsites
i.hp	21	18.8	16	17.0	15
i.list	14	12.5	9	9.6	9
p.hobby	11	9.8	10	10.6	6
p.publ	11	9.8	11	11.7	5
p.hp	10	8.9	9	9.6	8
i.generic	9	8.0	5	5.3	3
i.soft	8	7.1	7	7.4	4
i.proj	7	6.3	7	7.4	5
p.soft	6	5.4	6	6.4	4
i.publ	4	3.6	3	3.2	3
p.list	4	3.6	4	4.3	4
p.proj	2	1.8	2	2.1	2
p.teach	2	1.8	2	2.1	1
i.conf	1	0.9	1	1.1	1
i.teach	1	0.9	1	1.1	1
p.archive	1	0.9	1	1.1	1
i.archive	0	0.0	0	0.0	0
	112	**100.0**	**94**	**100.0**	**47**

Table 6-44. Most frequent meta genres of 94 target pages receiving 112 transversal links.

6.5.5.3 Transversal links compared with all followed links

It would have been interesting to examine how transversal links differ from average site outlinks in the UK academic web space. However, it was too time-consuming to manually classify a random sample among the site outlinks in the original link data set. In the subsequent discussion (Section 7.1.1), comparisons are made with a study by Wilkinson *et al.* (2003). In the present study, the transversal links were compared with all the followed links in the 10 path nets, in order to identify possible differences.

The left side of Table 6-45 below contains the counts and percentages of institutional and personal page genres of *all* the 352 followed outlinks. The right side of the table contains the corresponding statistics for the 112 *transversal* outlinks.

	all followed outlinks					transversal outlinks				
source meta genre	followed outlinks	% n=352	visited source pages	% n=281	visited source subsites	trans-versal outlinks	% n=112	transv. source pages	% n=95	transv. source subsites
i.conf	34	9.7	26	9.3	17	8	7.1	6	6.3	6
i.generic	6	1.7	6	2.1	3	4	3.6	4	4.2	1
i.list	87	24.7	73	26.0	34	27	24.1	27	28.4	15
i.proj	20	5.7	16	5.7	14	5	4.5	4	4.2	4
i.publ	8	2.3	8	2.8	4	1	0.9	1	1.1	1
i.soft	6	1.7	6	2.1	4	3	2.7	3	3.2	2
i.teach	4	1.1	4	1.4	2	0	0.0	0	0.0	0
	165	*46.9*	*139*	*49.5*		*48*	*42.9*	*45*	*47.4*	
p.hobby	3	0.9	2	0.7	2	1	0.9	1	1.1	1
p.hp	22	6.3	19	6.8	14	5	4.5	5	5.3	4
p.list	112	31.8	83	29.5	37	45	40.2	34	35.8	18
p.publ	7	2.0	7	2.5	7	2	1.8	2	2.1	2
p.soft	21	6.0	14	5.0	4	4	3.6	3	3.2	3
p.teach	22	6.3	17	6.0	11	7	6.3	5	5.3	5
	187	*53.1*	*142*	*50.5*		*64*	*57.1*	*50*	*52.6*	
	352	**100.0**	**281**	**100.0**	**93**	**112**	**100.0**	**95**	**100.0**	**48**

Table 6-45. Comparison between genres of followed and transversal *outlinks* and *source* pages.

The small sample size of all the followed links including a smaller subset of transversal links gives small absolute counts of some genres in the table. Further, the inconsistent inclusion of generic subsites in small path nets but not in larger ones possibly have an influence on the distribution of genres. A comparison between the two genre distributions in the table should thus be cautious.

Immediately, the perhaps most apparent difference between the two sides of the table is the higher percentage of personal link lists among the transversal outlinks (40.2%) and source pages (35.8%) than among all the followed links (31.8%) and visited source pages (29.5%). Further, this difference also could explain the somewhat higher percentage, 57.1%, of transversal links originating from personal genres, taken all together, compared with the 53.1% of all followed links provided by personal genres. However, the H_0 hypothesis (no difference between the two proportions of personal link lists among all the followed outlinks (0.318) and the transversal links (0.402)) was rejected at the 90% level.[81] There was thus no evidence of a significant difference between the two proportions. It should be noted that the test violates prerequisites such as two *independent* random samples. Thus, even a positive test result would have had to be treated with caution. Future large-scale studies could cast light on whether personal link lists of researchers and students do have a special role for providing transversal links in academic web spaces.

[81] Reject H_0 if $|Z| > 1.645$; However, $|Z^*| = 1.592 < 1.645$; Thus insufficient evidence to reject H_0.

	all followed inlinks					transversal inlinks				
target meta genre	followed inlinks	% n=352	visited target pages	% n=249	visited target subsites	transversal inlinks	% n=112	transv. target pages	% n=94	transv. target subsites
i.archive	5	1.4	4	1.6	4	0	0.0	0	0.0	0
i.conf	20	5.7	12	4.8	8	1	0.9	1	1.1	1
i.generic	17	4.8	10	4.0	7	9	8.0	5	5.3	3
i.hp	80	22.7	42	16.9	38	21	18.8	16	17.0	15
i.list	34	9.7	23	9.2	17	14	12.5	9	9.6	9
i.proj	24	6.8	24	9.6	17	7	6.3	7	7.4	5
i.publ	11	3.1	7	2.8	7	4	3.6	3	3.2	3
i.soft	10	2.8	9	3.6	5	8	7.1	7	7.4	4
i.teach	25	7.1	18	7.2	13	1	0.9	1	1.1	1
	226	*64.2*	*149*	*59.8*		*65*	*58.0*	*49*	*52.1*	
p.archive	3	0.9	2	0.8	2	1	0.9	1	1.1	1
p.hobby	15	4.3	12	4.8	7	11	9.8	10	10.6	6
p.hp	47	13.4	34	13.7	22	10	8.9	9	9.6	8
p.list	5	1.4	5	2.0	5	4	3.6	4	4.3	4
p.proj	5	1.4	4	1.6	4	2	1.8	2	2.1	2
p.publ	24	6.8	22	8.8	9	11	9.8	11	11.7	5
p.soft	17	4.8	14	5.6	8	6	5.4	6	6.4	4
p.teach	10	2.8	7	2.8	6	2	1.8	2	2.1	1
	126	*35.8*	*100*	*40.2*		*47*	*42.0*	*45*	*47.9*	
	352	**100.0**	**249**	**100.0**	**93**	**112**	**100.0**	**94**	**100.0**	**47**

Table 6-46. Comparison between genres of followed and transversal *inlinks* and *target* pages.

The percentages of transversal links differ more with regard to the target genres than to the source genres. Some differences in the distribution of transversal inlinks are apparent, for example, the lower percentage (0.9%) of transversal links to conference pages compared with the 5.7% for all followed links. A similar lower percentage appears for the links to institutional teaching pages (0.9% compared with 7.1%). This may support the earlier observation with regard to the absence of transversal links from institutional teaching pages, as an indication of more topical focus on such pages. The same topical focus aspect could be in evidence for the conference pages. Other apparent differences in the table are the higher percentages of institutional generic-type pages (e.g., pages on studies abroad or a page with photos of Scotland) and institutional software pages as targets for transversal links. Interestingly, the share of transversal links to personal software pages is not remarkably changed. This could indicate a higher authoritative 'attractive force' of the institutional software pages compared with the personal ones with regard to attracting transversal links from outside a disciplinary neighborhood. Not surprisingly, the topically diversified personal hobby pages and personal link lists attracted more transversal links.

However, again caution must be exercised with regard to not over-interpret the data, taking into account the small sample sizes, and thus small absolute counts of some genres, as well as the non-random sampling of the investigated path nets. In the light of the discussion above on the validity of hypothesis testing in the present sample, no null hypotheses were tested on differences in the proportions of genres of transversal inlinks and all followed inlinks.

As noted in Section 6.3.1, it was evident that the five-step methodology could not yield any generalizable findings. Instead, the 10 path nets have been used as case studies for identification of phenomena and generation of concepts and hypotheses. In

Chapter 7, such generated hypotheses based on the indicative findings in the empirical chapters will be discussed.

6.6 Summary of findings

The following list gives a brief summary of the main findings disclosed in the investigation of small-world properties in the UK academic web space in the previous empirical chapters. As discussed above – and further elaborated in Chapter 8 – many of the findings are indicative only, especially due to the small and non-random sample of the 10 path nets as outlined above.

The findings are listed below together with the research question they belong to among the four research questions put forward in the dissertation. Within each listing, major findings are listed first. Brackets show the empirical sections concerned.

The first research question is concerned with the more general aspects of cohesion and interconnectivity:

1. *How cohesively interconnected are link structures in an academic web space?*

    - There was *sparse link connectivity* in the investigated UK academic web space. An average outlinking university web page in the UK data set had 11.6 outlinks comprising 10.1 *site selflinks* and 1.5 *site outlinks*. Of the site outlinks to all the Web, only 7.7% were targeted to the other 108 universities and their subsites in the *undelimited* data set (including links to and from university main sites). The delimited data set (only site outlinks between *subsites* at different universities) comprised 3.1% of all site outlinks at the 109 universities. The vast majority of site outlinks in the study thus were targeted to academic, commercial, and other targets *outside* the data set. (Section 5.4.2);
    - Detailed *'corona'* graph model depicting actual inter-component and intra-component adjacencies in UK academic subweb graph, including frequent links direct from the IN to the OUT component not shown in the traditional 'bow-tie'-model. (Section 5.1);
    - *Indicative ages of the graph components* in the UK academic web graph as indicated by first indexed dates in the Internet Archive. This showed, for example, that the OUT component in the UK academic web graph contained the oldest subsites, the IN component the youngest, and the SCC subsites were on average slightly younger than the OUT subsites. (Section 5.2).
    - *Power-law-like* distributions of in-neighbors/out-neighbors and inlinks/outlinks in the UK academic subweb as well as within the 10 path nets. This finding is in line with the concept of a *fractal 'self-similar' Web* (Dill *et al.*, 2001; Kumar *et al.*, 2002) with subsets of the Web displaying the same graph properties as the Web at large (Sections 5.4.3 and 6.3.2.2);
    - *Power-law-like* distribution of *betweenness centrality* in the investigated UK web space. (Section 6.3.2.4);

- Indication of close relation between Kleinberg's (1999a) concepts of *hubs and authorities* on the Web and the *betweenness centrality* measure. No literature has been found discussing such a relation. (Section 6.3.2.4);
- Low correlation measure indicates a *lack of 'assortative mixing'*: web nodes with high connectivity degrees (many in-neighbors and out-neighbors) do *not* tend to connect to other nodes with many connections. This finding yields indicative support to Newman's (2002) finding regarding no assortative mixing in networks on the Web (Section 6.3.2.3).

2. *In particular, to what extent can so-called small-world properties be identified in this web space?*

    - The characteristic path length and clustering coefficient of the investigated UK academic web meet the requirements for a *small-world network* as introduced by Watts & Strogatz (1998) (Section 5.3).

The last two research questions are listed together as they are logically connected regarding more specific factors that may contribute to small-world properties:

3. *If small-world link structures can be identified in this academic web space, which properties can be observed that contribute to such link structures?*
4. *Especially, what types of web links, web pages and web sites function as cross-topic connectors in small-world academic web spaces?*

    - Institutional and personal page genres made up about 50% each of the visited source pages in the 10 path nets. This result may indicate the relatively large influence *personal web pages* has for providing site outlinks in an academic web space. About 60% of the visited target pages belonged to institutional page genres, whereas 40% belonged to personal ones. This result also may indicate a relatively large influence by personal web pages for receiving site inlinks in an academic web space. (Section 6.4.5.1);
    - There were more followed *outlinks* (53%) from *personal* source pages, than from institutional source pages in the 10 path nets, perhaps reflecting more active link creations of personal web creators. On the other hand, there were more followed *inlinks* (64%) to *institutional* target pages, than to personal target pages, perhaps reflecting more relevant and authoritative contents of institutional pages. (Section 6.4.5.3);
    - There is a higher percentage of *personal link lists* among the *transversal links* (40%) and transversal source pages (36%) than among all the followed links (32%) and visited source pages (30%). This finding may indicate a special impact of personal web creators for the emergence of small-world phenomena across dissimilar topical web domains (Section 6.5.5.3);
    - About 31% of *all* visited subsites in the 10 path nets were *computer-science*-related (CS). However, about 46% of subsites providing or receiving *transversal* links were CS-related (cf. Section 6.5.4). Counting *links* instead of *subsites*, about 38% of all followed outlinks in the 10 path nets originated

from CS-related *subsites*. The percentage of followed inlinks to CS-related subsites was slightly smaller (36%). Of the transversal links, 41% originated from CS-related subsites, whereas 40% were received by CS-related subsites. In the random sample of 189 SCC subsites (cf. Section 6.2.1), about 11% were judged as CS-related. Computer science thus constitutes a larger share among the visited subsites in the 10 path nets and an even larger share among the subsites connected by transversal links. Even if the sample of 10 path nets was small, this finding may indicate a special role of CS-related subsites as cross-topic connectors on shortest link paths in an academic web space. (Section 6.5.4);

- *Rich diversity of genre pairs*. Links between the investigated academic subsites in the 10 path nets connect many different combinations of source and target page genres. (Section 6.4.5);
- Personal link lists provided almost a third (32%) of all followed outlinks in the 10 path nets and the institutional link lists about a quarter (25%). This suggests that such genres may be *'site outlink-prone'*. Institutional homepages and personal homepages received most followed inlinks in the path nets, 23% and 13%, respectively. Such genres may correspondingly be *'site inlink-prone'*. (Section 6.4.5).

The rich material from the investigated data set including the case studies of the 10 path nets gives rise to some additional interesting aspects to be further investigated in future large scale studies as proposed in the empirical sections. The brief list below gives a summary of the proposed future studies (brackets show sections concerned):

- *Longitudinal time series studies*, e.g., the years 2000-2003 of snapshots of the same UK population of academic web subsites, including how transversal cross-topic links change over the years. For example, do subsites get more interconnected over the years? Does the percentage of site outlinks grow? Does the SCC become larger? (Section 4.2.2);
- Investigate possible differences in distributions of *in-neighbors and out-neighbors* in different domains, reflecting different Web use between disciplines (cf. Kling & McKim, 2000; Jacobs, 2001). (Section 5.4.1);
- Compare *subsite topics and genres* of *different graph components*. In the study only subsite topics and genres in the SCC component were identified due to time-consuming manual examination. (Section 6.2);
- Test other *clustering* measures than *k-cores* on an academic web space, such as, for example, co-word occurrence and co-linkage (i.e. co-citation and bibliographic coupling). Co-linkage clusterization would benefit from topical data either manually or automatically identified (Section 6.3.2.4);
- Analyze possible patterns of *genre connectivity* in large-scale academic and non-academic web spaces using the methodologies and typologies developed in the present study. (Section 6.4.5.2);
- Employ more objective heuristics for determining *dissimilarity between topics*, possibly by including low co-inlink or co-outlink measures of web sites possibly combined with low co-word occurrences. (Section 6.5.2);

- Investigate if *computer-science-related* persons and institutions expose more well-developed web presence and more experienced web literate behavior reflected by a larger number of created web pages and site outlinks. (Section 6.5.4);
- Examine how *transversal links* differ from *average site outlinks* in an academic web space. (Section 6.5.5.3).

In the next chapter, aspects of identified phenomena and properties from the empirical chapters are discussed, for instance, with regard to the role of personal and institutional link creators for the emergence of small-world link structures in an academic document space. Further, hypothesized complementarities of topical uniformity and diversity in small-world link structures are discussed, as well as topic drift and genre drift in that context.

7 Discussion and perspectivation

The previous empirical chapters developed a five-step methodology for the sampling, identification and characterization of small-world properties in an academic web space. The methodology was developed to answer the overall research question of this dissertation as put forward in Chapter 1:

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

The overall research question was supplemented with three preceding questions regarding more general aspects of interconnectivity and small-world properties of link structures in an academic web space. All four research questions are recapitulated here:

1. *How cohesively interconnected are link structures in an academic web space?*
2. *In particular, to what extent can so-called small-world properties be identified in this web space?*
3. *If small-world link structures can be identified in this web space, which properties can be observed that contribute to such link structures?*
4. *Especially, what types of web links, web pages and web sites function as cross-topic connectors in small-world link structures in an academic web space?*

The four research questions reflect the overall objective of the dissertation as to yield a better understanding of what factors contribute to the formation of interconnected link topologies across an academic web space.

The employed data set constituted a 'frozen' snapshot picture of the publicly available link structure at the 109 UK universities as of June-July 2001. As noted in Section 4.2.2, this temporal delimitation does not capture the dynamics of the investigated link structures. Furthermore, the small and non-random sample of 10 path nets as well as the focus on UK university *subsites* after removal of main university sites imply there are no generalizable findings for the UK academic web space as a whole, less for academic web spaces world-wide. Academic web spaces may look different in other national university systems than in the investigated UK system. However, studying the academic pages within a single country is an acceptable scope for the dissertation. Even though the Web is world-wide, the pages or sites within a single country or within a single national university system form a conceptually coherent set that is therefore a valid object of study. The scope is thus valid, even if the results should not be seen as generalizable to the whole Web.

The indicative findings outlined in Section 6.6 are based on the rich material from the 'corona' model of the UK academic subweb, as well as the closely investigated case studies of the 10 path nets. Especially, the case studies enabled identification of phenomena and revealed interesting indicative aspects of the phenomena that may generate more general hypotheses and perspectives, some of which will be discussed in this chapter, for example, regarding topical uniformity, diversity and genre drift.

Furthermore, the developed five-step methodology necessitated generation of concepts that could describe aspects of, for example, path nets and genre connectivity.

The present chapter is divided into three sections comprising three 'cross-sections' or 'transversal trails' that cut across the identified phenomena and properties from the previous chapters:

1. *Personal and institutional connectors in small-world webs*. The role of personal and institutional link creators for the emergence of cohesive small-world link structures in a distributed and 'non-engineered' academic document space.
2. *Complementarities in the formation of small-world webs*. Hypothesized complementarities of topical uniformity and diversity in small-world link structures. Furthermore, hypothesized complementarities of topic drift and genre drift in small-world connectivity of a distributed document space.
3. *Exploratory capabilities in a small world*. The exploitation of the above complementarities in the possibilities to explore a distributed document space.

The first two sections are especially concerned with the *third* research question dealing with what properties may contribute to the emergence of small-world link structures in an academic web space.

The chapter ends with some reflections on what overall implications may be drawn for overall library and information science frameworks based on the preceding discussions.

7.1 Personal and institutional connectors in small-world webs

In this section, roles of personal and institutional link creators for the emergence of small-world link structures in a self-organized academic web space are discussed, including the roles of especially computer-science subsites and persons as 'betweenness-central' connectors in this small-world web space. Mayor discussion points are marked with bullets.

The close examination of the 10 path nets provided a good understanding of how academic web creators connect documents, topics, genres, and sites on the Web; and how an academic document space thus becomes interconnected across web sites, including how small worlds emerge in this space. One of the most interesting indicative findings was on the suggested impact of personal link creators on the emergence of small-world phenomena in the investigated web space.

From a library and information science perspective, with accessibility and navigability in information systems as core issues, the role of personal link creators is especially interesting in order to understand how the non-engineered and self-organized architecture of a distributed information space – in this case, the UK academic web of subsites – provide access points, traversal options, and small-world phenomena in the shape of short link distances both within topics and across topics.

The examined path nets literally provided *cross-sections* through the investigated web space. Paraphrasing Bush (1945) who envisioned personal hypertext-like information systems, where a scientist could *"build a trail of his interest through the*

maze of materials available to him", one might say that the shortest link paths and thus also the investigated *path nets* comprise *trails* through the maze of human interests and knowledge as reflected in the investigated academic web space.

As noted in Section 6.4.5.1, the institutional and personal page genres made up about 50% each of the visited *source* pages in the 10 path nets. This finding indicates a relatively high importance of personal web pages for providing site outlinks in an academic web space. The corresponding picture for *target* genres was somewhat different. About 60% of the visited target pages belonged to institutional genres, whereas 40% belonged to personal ones. This circumstance may reflect a larger authoritative quality and 'attractive force' of institutional page genres than of personal ones. However, even if the share of personal genres thus is smaller with regard to the visited target pages, this result still indicates a relatively large influence by personal web pages for receiving site inlinks in an academic web space, hence:

- **personal web pages may be important providers and receivers of links across sites in an academic web space**

Counting links instead of pages as above, there were more followed site outlinks (53%) from *personal* source pages, than from *institutional* source pages in the 10 path nets, perhaps reflecting more active link creations by personal web creators. On the other hand, there were more followed site inlinks (64%) to institutional *target* pages, than to personal target pages, perhaps reflecting more relevant and authoritative contents of institutional pages (cf. Section 6.4.5.3). Yet again, personal and institutional page genres may also be considered as complementary by the circumstance that they provide a diversity of link sources and targets for each other, hence:

- **more site *outlinks* derive from *personal* source web pages**
- **more site *inlinks* given to *institutional* target web pages**
- **personal and institutional page genres complement each other**

Inspired by the terminology of Bray (1996), one could say that *personal* web pages in an academic web space yield high *luminosity* providing 'illumination' to outlinked target pages and sites. Correspondingly, *institutional* web pages gain high *visibility* receiving 'illumination' from inlinking source pages and sites. Visibility and luminosity are thus concerned with mutual *linking empowerment* in web spaces, of large importance today; as search engines like Google utilize the strength of interlinkage when ranking search results (cf. Brin & Page, 1998; Walker, 2002).

The Web has provided individuals inside and outside academia with an unprecedented tool for obtaining visibility and luminosity by using personal homepages for self presentation including pointers to personal interests (cf. Miller, 1995; Wynn & Katz, 1997; Bates & Lu, 1997; Chandler, 1998; Miller & Arnold, 2001). As stated by Björneborn & Ingwersen (2001):

> *"the breakthrough for everybody to express themselves, practically without control from authorities, to become visible world-wide, also by linking to what pages one wants to link to, to assume credibility by being "there", and to obtain access to data, information, values and knowledge in many shapes and degrees of*

> *truth, has generated a reality of freedom of information - also in regions and countries otherwise poor of infrastructure."* (p. 69)

The personal homepage may be considered the first uniquely new genre on the Web (e.g. Erickson, 1996; Dillon & Gushrowski, 2000), cf. Section 6.4.5. According to Erickson (ibid.), the possibilities to blend professional and personal appearances on personal homepages are one of the main keys to why the Web is becoming a fundamentally different and perhaps more attractive information system than preceding ones:

> *"Hobbies, research interests, pets, professional publications, children, politics, friends, colleagues, all are grist for the personal page."* (p. 15)

In this context, Thelwall (2002d) describes how the Web provides a "*public unrefereed creative space that is used for informal research, teaching and recreational information, for example in personal home pages.*" (p. 563). According to Wynn & Katz (1997), linking capability provides home page creators with unique means to *"build social context around themselves"* (p. 310) by pointing to friends, colleagues, etc., and thus disclose more *social embeddedness* (ibid.) than perhaps revealed in everyday life.

The blending of personal and professional presentations on web pages located under a university's auspices has made many UK universities introduce explicit codes of practice and guidelines with regard to the setting up and use of web servers and web pages, especially what contents and outlinks are acceptable on personal web pages of university staff and students (cf. Charlesworth, 1996). For example, the University of Newcastle upon Tyne makes this disclaimer on a subsite with students' clubs and societies (*www.societies.ncl.ac.uk*):

> *"The content of this server is the sole responsibility of the individual maintainers. The University of Newcastle upon Tyne exercises no editorial control but reserves the right to remove material which breaches copyright, is offensive, obscene, defamatory, inaccurate, or otherwise brings the University into disrepute. The University will take immediate action if problems of this nature are brought to its attention."*

Another university, the University of Strathclyde, makes a similar disclaimer on their subsite containing personal homepages for staff and students (*homepages.strath.ac.uk*) in order to prevent the university from being associated with the contents of the personal homepages:

> *"The pages on this server carry the views of individual members of staff or students of the University of Strathclyde. The information here should not be understood as representing the policy of the University of Strathclyde or be an accurate representation of the University. The University of Strathclyde should not be associated with the views of individual members of the University which are expressed here."*

Some of the subsites, as represented in the two examples above, thus have an intentional web policy from the university's part in order to obtain a clear separation between institutional and personal web pages.

7.1.1 Academic vs. non-academic links

The close examination of all the source and target pages along the followed link paths in the 10 path nets revealed a striking richness of impressively extensive, well-organized and well-maintained *personal web territories* comprising, for instance, researcher's personal curriculum vitae, publication lists, full-text preprints, research interests, ongoing or finished research projects, teaching curricula, bookmark lists, hobby interests, etc. Many researchers thus use the Web as a *visibility* and *luminosity* tool – as outlined in the previous section – for personal knowledge organization.

As noted in Section 2.4.1, the Web is thus increasingly used in both formal and informal scholarly communication and collaboration (e.g., Cronin *et al.*, 1998, Zhang, 2001; Thelwall & Wilkinson, 2003; Wilkinson *et al.*, 2003). Webometrics thus offers potentials of tracking aspects of scientific endeavor traditionally more hidden for bibliometric or scientometric studies, such as the use of research results in teaching and by the general public (Björneborn & Ingwersen, 2001; Thelwall & Wilkinson, 2003a) or the actual use of scientific web pages. The arrival of researchers' personal web pages and web territories – including transversal links – has opened new possibilities to investigate facets of previously concealed types of informal – both academic and non-academic – activities conducted by scholars.

As noted in Section 6.5.5.3, it was not feasible to examine how transversal links differ from average site outlinks in the UK academic web space due to the time-consuming manual classification of such a sample. Instead, the transversal links were compared with all the followed links in the 10 path nets in order to identify possible differences. However, a study by Wilkinson *et al.* (2003) investigated motivations for academic web site interlinking by examining a random sample of 414 site outlinks between UK university web sites from the same original data set (including links to and from university main sites) as the present study is based upon. Potential link motivations were classified by the four researchers in the Wilkinson study. In spite of inter-indexer difficulties to handle multiple potential link motivations and fluid web page genres, it was clear that the vast majority, over 90%, of the site outlinks was created for broadly scholarly reasons, including teaching, whereas 7.5% were classified as non-academic links targeting recreational and tourist information topics.

As outlined in Section 6.5.2 regarding the identified 112 transversal links that connected topically dissimilar subsites in the 10 path nets, 92 (82%) transversal links were judged as academic, comprising 48 personal and 44 institutional links. The *academic* links comprised links related directly to research and teaching done by a person or an institution, as well as links to academic targets of more general or peripheral interest to a person. There were 16 (14%) personal *non-academic* transversal links targeted to leisure-related topics, such as hobbies, charity, tourist information, and family relations. Four automatic outlinks created by a web server statistical program (that targeted back to inlinking subsites) were judged neither academic nor non-academic due to their automated generation.

The percentage of non-academic transversal links (14%) in the 10 path nets was thus almost twice as high as the 7.5% non-academic inter-site links found in the random sample in the Wilkinson study. Even if the 10 path nets do not comprise a representative sample, this result may indicate an

- **important role of *non-academic* links for enabling *cross-topic* connections in an academic web space**

However, as indicated by the high percentage (82%) of identified *academic* transversal links:

- **main factor behind cross-topic connections are *academic* transversal links**

There were about 57% personal and 43% institutional links among the identified transversal links in the 10 path nets. However, the links in the Wilkinson study were not divided into personal and institutional categories. It is thus not possible to determine whether personal links constitute a larger share of transversal links than of overall site outlinks in an academic web space.

The entanglement of scholarly and non-scholarly contents in an academic web space clearly makes it more difficult to develop criteria for differentiation of scientific web pages from other types of contents on the Web in order to identify and retrieve such scholarly contents (cf. Jepsen *et al.*, 2002). However, the methodologies in the present study could perhaps be developed to *avoid non-academic content and links* in an academic web space, for example, by identifying and filtering away personal page genres with dominance of non-academic contents and links, such as some of the personal link lists and hobby pages. On the other hand, too rigid avoidance of such pages could limit detection and exploitation of well-developed personal link lists that provide extensive and quality pointers collated by persons being experts in scientific domains.

However, even non-academic transversal links may be exploited in an academic web space because they contract link distances between academic web sites just like academic transversal links do. Short link distances in an academic web space may be of importance for how exhaustive harvests that automated web crawlers can conduct when they traverse the Web by following links from page to page. Non-academic transversal links may thus benefit *more exhaustive web crawls* by providing direct access to parts of academic web sites that would not otherwise be visited and harvested by web crawlers sent out, for instance, by webometric surveys, by search engines like Google, or by web archiving projects like the Internet Archive.

7.1.2 Transversal links and weak ties

As outlined in Section 6.3, the path nets with topically juxtaposed seed nodes were constructed tools for identifying and locating transversal links. The deliberate juxtaposition of pairs of topically dissimilar subsites enabled confined and thus investigable *'mini small worlds'* in the shape of path nets comprising subgraphs of all shortest link paths between the juxtaposed web nodes in an academic web space. However, it was not possible to deduce the frequency of transversal links between subsites in the overall UK delimited data set (only links between subsites at different universities). Instead, the identified 112 transversal links comprised about a third (32%)

of the 352 followed links in the 10 path nets. However, this share is not generalizable to the 207,865 inter-site links in the UK data set. First, the sample of 10 path nets is small and non-random, and second, and more importantly, a path net represents a special *cross-section* of a web space that only includes those links fitting into the shortest link paths between the start node and end node of the path net. The single path nets thus do *not* represent 'typical' link structures in a web space.

An interesting study with regard to how frequently site outlinks connect *dissimilar* topics compared to site outlinks connecting *similar* topics is given by Thelwall & Wilkinson (forthcoming). They investigated topical similarities between 500 random sites and subsites in the UK academic web space as of 2002 with regard to three possible linkage types: direct links, co-inlinks (co-citation) and co-outlinks (bibliographical coupling). Using human assessment for topical similarity between the sites, the authors found that sites connected by a combination of *all* three linkage types (cf. Fig. 7.1 below) not surprisingly had the highest probability (43%) of topical similarity. However, and perhaps more surprisingly, the two co-linkage types only yielded marginal improvement over using direct links alone. Thus, 42% of sites connected by only direct links were topically similar. This specific result thus does not support the information retrieval theory of *poly-representation* (Ingwersen, 1992; 1994) that is concerned with how a multi-evidence of cognitive representations (in this case, a combination of links and co-links) may be exploited for retrieving more relevant documents. This lack of support may be due to more muddled motivations for making links in an academic web space compared to citation motivations in scientific literature.

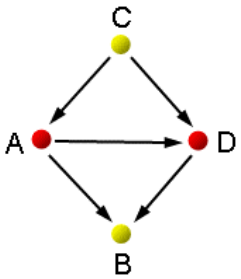


Figure 7.1. Direct link between A and D, that further have co-inlinks from C (analogous to co-citation), and co-outlinks to B (bibliographic coupling)

However, an interesting fact in the context of this dissertation focus on topical *dissimilarities* between web sites, the figures found by Thelwall & Wilkinson (ibid.) indicate the extent of *topic drift* in an academic web space even if this issue is not discussed by the authors. Thus, when 42% of sites connected by direct links were topically similar, this means there was topic drift in 58% of these cases of inter-site links.[82] This percentage is thus higher than the abovementioned percentage of identified transversal links (32%) in the 10 path nets. Different human assessments of topical similarity may explain some of this deviation. However, as emphasized above, single path nets do *not* represent 'typical' link structures in a web space, which may explain

[82] Not surprisingly, the topic drift was much larger when two web sites share co-inlinks only (90% such pairs of sites had topic drift) or co-outlinks only (78%).

the lower percentage of transversal links in the investigated small sample of 10 path nets. Future large-scale studies of path nets and transversal links in academic web spaces may cast light on this issue.

As noted in Section 5.4.2, there was *sparse link connectivity* in the investigated academic web space. This was reflected in the delimited data set (site outlinks between subsites at different universities) that only comprised 3.1% of all site outlinks at the 109 universities.[83] Furthermore, even many SCC nodes in the data set had only low connectivity degrees (cf. Section 6.1). Such sparse connectivity suggests how vulnerably close to isolation a web node can be even if it belongs to the strongest connected component in a web space. Every site outlink and inlink, including transversal ones, may thus be important for providing cohesion and preventing isolation of nodes and disintegration of a web network. This finding of sparse connectivity in the investigated UK subweb graph is in line with the suggestion made by Watts & Strogatz (1998) on that small-world properties "might be common in sparse networks with many vertices, as even a tiny fraction of short cuts would suffice".

As noted earlier in Section 3.5, the Web resembles a social network (cf. e.g., Wellman, 2001; Kumar *et al.*, 2002; Adamic & Adar, 2003). According to Kumar *et al.* (2002), the aforementioned fractal self-similarity with subsets of the Web that display the same power-law-like connectivity distributions as the Web at large (cf. Section 6.6) is also pervasive in social networks.

In this context, it would also be interesting to pursue a hypothesis that transversal links may function as *weak ties* using a social network analytic term (Granovetter, 1973; 1982) for explaining macro-level social cohesion and possibilities for rapid diffusion of ideas and epidemics across social boundaries through peripheral social contacts, so-called 'weak ties', cf. Section 3.1. Such transversal 'weak ties' may also be seen as to how academic authors often cite a few sources outside their own scientific domains, so-called *'boundary crossings'* (Klein, 1996a, 1996b; Pierce, 1999). Transversal links may thus function as weak ties and boundary crossings between heterogeneous web clusters. In that respect, transversal links may be conceived as generating rich *transitional areas* (cf. Turner, Davidson-Hunt & O'Flaherty, 2003) where diversified topical areas meet on the Web. For example, such transitional areas occur on the boundaries between the topical areas in the large path net in Fig. 7.3 shown in Section 7.2.2 further below.

Of course, many transversal links reflect idiosyncrasy, for example, personal non-scientific hobbies as discussed above. But then again, other transversal links on scientists' web pages may reflect emerging 'research fronts' in scientific domains, or cross-disciplinary 'invisible colleges' (cf. below). Revealing such 'hidden' connections could render useful information about new directions in the evolving interconnectedness of science, discovering new relationships and patterns. Transversal links crossing scientific boundaries could provide creative insights, thus adding a new connotation to the earlier mentioned notion of *'the strength of weak ties'* from social network analysis (Granovetter, 1973; 1982). This hypothesis was not possible to test in the dissertation, because it was not feasible to identify topical clusters (cf. Section 6.3.2.4) and thus strong ties were not identified – less weak ties.

[83] Cf. finding by Thelwall (2003c) that 97% of *all* pages in a sample of the UK academic web space were neither the source nor target of any inter-site link (p.12).

As the Web includes more and more informal self-presentations and link creations by academics, the sociology of science may employ small-world approaches including measures of betweenness centrality for identifying *'invisible colleges'* in the shape of informal scholarly communication networks (Price, 1961; Crane, 1972) and central *'gatekeepers'* across link structures in academic web spaces reflecting networked knowledge creation and diffusion. As noted in Section 3.2, Garfield (1979) early envisaged such small-world approaches to identify 'gatekeepers' and 'invisible colleges' in informal scholarly communication networks. As stressed earlier in Section 6.3.2.4, the betweenness centrality of gatekeepers on the Web is not concerned with control of information transfers as usually is the case in social networks. On the Web, betweenness centrality reflects how gatekeeper nodes allow human web surfers and digital web crawlers to access and traverse large parts of the Web graph. For example, a biochemical researcher (*www.chem.gla.ac.uk/~johnm*) who occurred in two of the 10 investigated path nets (NH02 and NH05) with non-academic (football) transversal links may perhaps function as such a gatekeeper contributing to the cohesion of the UK academic web space, however idiosyncratic such non-academic links may seem to be.

Measuring the betweenness centrality of nodes in, for example, an academic web space where extensive link connectivity data are available – as in the present study – may be developed for *automatic detection* of personal and institutional *betweenness-central gatekeepers* in an academic web space.

Identification of betweenness-central personal gatekeepers may be important in order to facilitate so-called *social navigation* on the Web (cf. Dourish & Chalmers, 1994; Erickson, 1996; Dieberger, 1997; Kautz *et al.*, 1997). Social navigation is concerned with finding and accessing information, e.g., on the Web, through the pointers of *persons* with special interests and expertise regarding the concerned information or topic. Publicized bookmark lists and other pointer pages on the Web free for everyone to use are examples of access points for social navigation.

7.1.3 Institutional connectors

The 10 randomly selected *seed subsite nodes* in the investigated path nets belonged to topics in humanities & social sciences, economics, psychology, linguistics, geography, atmospheric physics, chemistry, mathematics, palaeontology, and ophthalmology (eye research). By chance, none of the seed nodes were computer-science-related (hereafter CS-related). Given the small sample size, this turned out to be an advantage as CS-related seed nodes would have blurred the indicative finding that CS-related subsites play an essential role as cross-topic connectors in academic web spaces. CS-related subsites thus occurred on the shortest paths connecting seed nodes in six of the 10 path nets (cf. Appendix 16). In the remaining four path nets, typically generic-type subsites functioned as intermediate connectors between topically dissimilar subsites.

About 31% of *all* visited *subsite* nodes in the 10 path nets were computer-science-related (CS). However, as many as about 46% of the subsites providing or receiving *transversal* links were CS-related (cf. Section 6.5.4). (Counting links instead of subsites, 41% transversal links originated from CS-related subsites, whereas 40% were received by CS-related subsites). In the random sample of 189 SCC subsites (cf. Section

6.2.1) only about 11% were judged as CS-related. Computer science thus constituted a larger share among the visited subsites in the 10 path nets and an even larger share among the subsites connected by transversal links. Even if the sample of 10 path nets was small, this finding may

- **indicate an important role of CS-related subsites as cross-topic connectors in an academic web space**

As noted earlier in Section 6.5.4, the role of CS-related subsites in academic link structures most likely reflects the auxiliary function of computer science in many scientific disciplines in natural sciences, technology, humanities, and social sciences. Computer programming thus forms part of a multitude of scientific disciplines, for example, in linguistics, economics, geography, etc. This auxiliary function may be combined with a more well-developed and unconstrained web presence and more experienced web literate behavior performed by CS-related persons and institutions, reflected by a relatively larger number of created web pages and site outlinks. However, this latter hypothesis remains to be verified. A historical fact though, is the circumstance that university computer-science departments especially in the USA have been key players in the development of the Internet since the early beginning in the 1960s (e.g., Abbate, 1999).

The essential role of CS-related subsites for small-world connectivity patterns in an academic web space was also indicated by the circumstance that 15 CS-related subsites were placed among the 25 subsites with the highest *betweenness centrality* – being top *hub and authority* sites – among all 7669 subsites in the data set (cf. Table 6-12**,** Section 6.3.2.4). This indication of computer-science subsites and thus indirectly CS-related persons as 'betweenness-central' connectors in a small-world academic web space is supported by the finding by Thelwall (2002c) that the most highly targeted pages in the UK academic web space were predominantly university home pages and computing-related departments and initiatives, as investigated in a survey based on the same 2001 data set as the present study.

According to Kling & McKim (2000), there are – and will likely continue to be – large differences between different scientific fields in the way electronic media, including the Web, are implemented and utilized, cf. Section 2.4.1. In some fields, there are *'open flow fields'* where researchers freely share unrefereed preprints (ibid., p. 1315). Other fields are more *'restricted flow fields'*. In this context, it is symptomatic that the original World-Wide Web project was designed to facilitate such open information flow among nuclear physicists at CERN (cf. Berners-Lee & Cailliau, 1990). As noted by Thelwall (2003a), a researcher in a high Web use area, such as computing, is more likely to be familiar with external web pages and make links to them in her own pages. Conversely, the same will be true for her peers and so bigger inlink counts could also be expected.

Wilkinson, Thelwall & Li (forthcoming) note that disciplinary variations mean that some disciplines interlink more than others. Research by Tang & Thelwall (fortchoming) indicates that in the US web spaces, the hard sciences probably interlink more than social sciences, whereas the humanities hardly interlink at all.

Looking at academic web spaces at large, Menczer (2001) states that academic Web pages are better connected to each other than commercial pages in that they are more prone of pointing to other similar pages. In other words, according to Menczer (ibid.) it is easier to find related pages browsing through academic pages than through commercial pages. This reflects the different goals of the two communities, where commercial web sites are not eager to make site outlinks allowing potential customers to slip away.

So far, the discussion has dealt with possible roles of personal and institutional web creators for understanding the emergence of small-world link structures in an academic web space – from a *micro* level perspective. The next section gives a more *macro* level perspective on small-world link structures by incorporating overall conceptualizations, hypothesized complementarities and metaphors.

7.2 Complementarities in the formation of small-world webs

From a library and information science perspective it is interesting to understand how a distributed document space becomes interconnected and navigable in spite of no central coordination or control of *what* document collections or documents are included in this document space; *where* they are located; and *how* they are interconnected. Furthermore, it is intriguing to understand how small-world phenomena in the shape of *both* short local *and* short global link distances emerge and affect accessibility and navigability in this non-engineered document space.

In this section, some hypothesized complementarities in the formation of small world link structures in a distributed academic web space will be introduced and discussed. Section 7.2.1 recapitulates relations between *scale-free* network features and small-world properties (cf. Section 3.5) before discussing some findings. Section 7.2.2 is concerned with possible complementarities of *topical uniformity* and *diversity* in small-world academic web spaces. Section 7.2.3 deals with the perceived *web of genres* including hypothesized complementarities of *topic drift* and *genre drift* in small-world connectivity. Finally, Section 7.2.4 gives an intuitive visualization how page genres and links may shorten distances in *crumpled-up* web spaces.

7.2.1 Scale-free networks and small-world properties

As noted earlier in Section 3.5, the occurrence of so-called *scale-free* network features (Barabási & Albert, 1999; Barabási Albert & Jeong, 2000) is essential to an understanding of connectivity and cohesion on the Web – including small-world link structures. In a scale-free network, there is no 'typical' node, that is, no characteristic 'scale' to the degree of connectivity. Scale-free distributions of inlinks and outlinks show long *power-law* tails (cf. Section 5.4), implying that only a small share of web nodes receive or provide many links, whereas the bulk of nodes has quite few inlinks or outlinks each. As listed in Section 3.5, a range of power-law distributions have been identified on the Web. In the present study, power-law like distributions were found for

in-neighbors/out-neighbors and *inlinks/outlinks* in the UK academic subweb as well as within the 10 path nets (Sections 5.4 and 6.3.2.2). Also the measure of *betweenness centrality* (Section 6.3.2.4) in the investigated UK web space showed power-law-like distributions.

According to Barabási & Albert (1999), scale-free link distributions are rooted in two generic mechanisms of many real-world networks: *continuous growth* and *preferential attachment* ("rich-get-richer"). In this framework, the Web is an open self-organizing system that grows by the continuous addition of new nodes and links where the probability of connecting to a node depends on the number of links already attached to the node. As noted earlier, this significance of preferential attachment for power-law distributions is well known in bibliometrics as *'the Matthew effect'* (Merton, 1968) and *'cumulative advantage'* (Price, 1976).

As also noted in Section 3.5, small-world properties on the Web may be explained by scale-free link distributions, in which a relatively small number of well-connected nodes serve as hubs (e.g., Steyvers & Tenenbaum, 2001). Strongly connected hubs and authorities (Kleinberg, 1999a) may thus create interlinked 'backbones' in a web network onto which more 'peripheral' nodes can be attached. This explanation is supported by the indication (cf. Section 6.3.2.4) in the investigated UK data set, that subsite nodes with high *betweenness centrality* may tend to link to other nodes also with high betweenness centrality, so-called *assortative mixing* (Newman, 2002). Furthermore, there is an indication that subsites in the UK data set that have high betweenness centrality – and thus function as connectors on many shortest paths between nodes in the UK academic web space – also tend to be *hubs* and *authorities* in a Kleinberg (1999a) sense (cf. Section 6.3.2.4).

As outlined in Sections 3.3 and 5.3, small-world phenomena in a network may be defined by two parameters: high overall *clustering coefficient* and low characteristic *path length* (Watts & Strogatz, 1998). The overall clustering coefficient of a network (cf. Section 5.3.2) is the average of the local clustering coefficients for all the nodes in the network. Each local clustering coefficient reflects how densely connected the neighborhood is around each network node, that is, reflects how short *local* link distances are. On the other hand, the characteristic path length of a network (cf. Section 5.3.1) reflects how short *global* link distances in the network are. Small-world phenomena thus reflect the complementary co-existence of *short local* distances and *short global* distances in a network.

In the library and information science approach employed in this dissertation, *preferential attachment* in the investigated small-world web space may be seen reflected in *topical uniformity* in link structures. In the next section is presented a hypothesis regarding how complementarities of topical uniformity and topical diversity may contribute to the emergence of small-world link structures.

7.2.2 Topical uniformity and diversity

Contrary to libraries, the Web is a distributed information system (Berners-Lee, 1989/1990) without centrally engineered construction and maintenance of the system (cf. Section 1.1). The Web may be conceived as constructed by *'collaborative weaving'*

by millions of link creators interlinking their web pages with web pages of others distributed on millions of servers. In this globally interwoven *'patchwork'* of many thousands of topics, the local link creators attach their links for a multitude of socio-cognitive reasons reflecting a diversity of personal interests, social preferences, power structures, topical relations, navigational aids, etc. (cf. Cronin *et al.*, 1998; Kim, 2000; Walker, 2002; Park, 2002; Park *et al.*, 2002; Thelwall, 2003d).

However, there exists no '*cybercartographic'* maps covering vast and dynamically changing link topologies of the global Web. When selecting the link targets, the millions of link creators make use of their limited knowledge of available web pages and sites. Well-established and well-known web sites within different topical domains both in academic and non-academic web spaces will thus inevitably attract more inlinks due to the prevalence of *preferential attachment* as mentioned in Section 7.2.1.

The preferential attachment implies that most links within and between web sites connect web pages containing cognate topics (cf. Davison, 2000). This dominance of *intra-topic* links, here called *topical propensity*, leads to the emergence of topic-focused cluster-like formations in a web space. A *topic cluster* may consist of web pages and web sites of researchers and their institutions making links to other researchers, institutions, projects and papers *within* their own scientific discipline or community. A community or cluster on the Web can be defined as a collection of web pages which have more links between them, i.e. have greater *link density*, than with the rest of the Web (cf. Kleinberg & Lawrence, 2001).

The case studies of the 10 path nets have given a visual indication of the *topic topology* across an academic web space, that is, how topics are distributed and interconnected in this web space.[84] In order to discuss how small-world phenomena emerge in a distributed document network, it may be fruitful to use the terms *intra-topic* and *inter-topic* links, as well as the related terms *topical uniformity* and *topical diversity*.[85]

Intra-topic links thus connect web sites belonging to the same overall topic as already indicated above. In this context, the term *topic cluster* will broadly be used to designate structural formations on the Web of such intra-topic links. Topic clusters are characterized by *topical uniformity*, that is, homogeneity of overall topics within the cluster.

Correspondingly, *inter*-topic links connect web sites belonging to dissimilar overall topics. The terms *inter-topic link* and *transversal link* are thus synonymous. *Topical diversity*, that is, heterogeneity of topics within a web site may provide such inter-topic links between different topic clusters.

As mentioned earlier, it was not feasible to identify topic clusters in the data set, apart from the so-called *k-cores* in Section 6.3.2.4 comprising groups of subsites interconnected by at least *k* links (with the *53-core* dominated by computer-science-related subsites). In future studies, it would interesting to include other clustering

[84] The terms *topic* and *topology* are both derived from the Greek term *topos* meaning place. According to Bolter (1991, p.106), the Greek term was used in ancient rhetoric to refer to commonplaces, conventional units and methods of thoughts. In the Renaissance, *topics* became headings that could be used to organize any field of knowledge.

[85] The terms *intra-topic* and *inter-topic* are also used by, e.g., Papadimitriou *et al.* (1998).

measures on an academic web space, for instance, based on frequencies of direct inter-site links, co-inlinks (co-citation) and/or co-outlinks (bibliographic coupling) (cf. 6.3.2.5). However, link data alone will not be sufficient in order to filter out topic clusters because of topic drift along link paths in link structures (cf. Section 6.5.1). Topical data either identified manually (as in the present study) or automatically will be necessary for identifying topic clusters, for instance, by including metadata and frequent content words.

In this context, the term *topic* has been used in a pragmatic common-sense way throughout the dissertation when identifying the overall topic of a subsite (cf. Sections 6.2.1 and 6.5.2). Future studies would benefit from more objective heuristics to determine topics – and *similarities* between topics – by using, for example, high-frequent co-word occurrences (including metadata) between the investigated web site (or web page) and the sites (or pages) with which it is most strongly co-linked (and/or co-linking) (cf. e.g., Menczer, 2001; Haveliwala *et al.*, 2002; Jacobs, 2002).

Correspondingly, *dissimilarities* between topics – and thus also the determination of what links are transversal – could be more objectively determined by lack of high-frequent co-word occurrence (cf. Section 6.3.2.5). However, problems of diversified terminology within a topical domain may blur such co-word measuring (cf. Leydesdorff, 1997). Furthermore, as proposed in Section 6.3.2.5, lack of direct co-linkage between two web sites in the strongly connected component (SCC) of a relatively small academic web space may indicate topical dissimilarity between the sites. The hypothesis is based on a non-verified assumption that two sites on similar topics would probably either be co-linked or co-linking in the SCC of a relatively small academic web space.

Much web research is concerned with automatic categorization of web contents. Among the pioneers in this area were Pirolli *et al.* (1996) who explored web analysis techniques for automatic categorization utilizing both link structures, text content, metadata similarity, and usage data. Later, work by Bharat & Henzinger (1998) and others have focused on combining content analysis and connectivity analysis for "topic distillation" to achieve more focused web crawls and information retrieval on the Web.

Naturally, large-scale academic web spaces, let alone the whole Web, contain more complex and muddled topical structures than outlined by the dichotomy of the 'ideal type' terms 'topical uniformity' and 'diversity'. One may imagine web sites, for example, the generic-type and multidisciplinary subsites in the 10 path nets, having *'multiple memberships'* (using a term from White, 2003) belonging simultaneously to more than one topic cluster, because of no overall site topic. Furthermore, in academic web spaces, there are probably overlapping clusters, due to sliding transitions between overlapping interdisciplinary topics, for example, as described in Section 6.5.3 regarding path net NH04, where environmental studies and meteorology posed interdisciplinary overlaps with geography and earth sciences making it inconvenient to designate links between them as inter-topic, that is, transversal. The boundaries of a topic cluster will furthermore depend on the employed threshold values whether frequencies of direct links, co-links, words, or combinations of these are used. There may thus be several non-overlapping clusters on the same topic in a web space.

In this context, it should be noted that the rich topical diversity encountered in the investigated UK academic web space may reflect a deliberate government policy to

promote diversity of institutional missions in higher education (Dearing, 1997; Thelwall & Wilkinson, forthcoming). Furthermore, many disciplines are interdisciplinary and cross-institution collaboration is increasingly encouraged in UK universities (Chen *et al.,* 1998).

As stated earlier in Section 7.1.2, it would be interesting in future large-scale studies to investigate a random sample of academic site outlinks in order to investigate how frequent are transversal links where the target site topic is dissimilar from the source site topic compared with the frequency of site outlinks connecting sites of similar topics, thus forming topic web clusters. In other words, how frequent are *inter-topic* (transversal) inter-site links compared with *intra-topic* inter-site links on the Web?

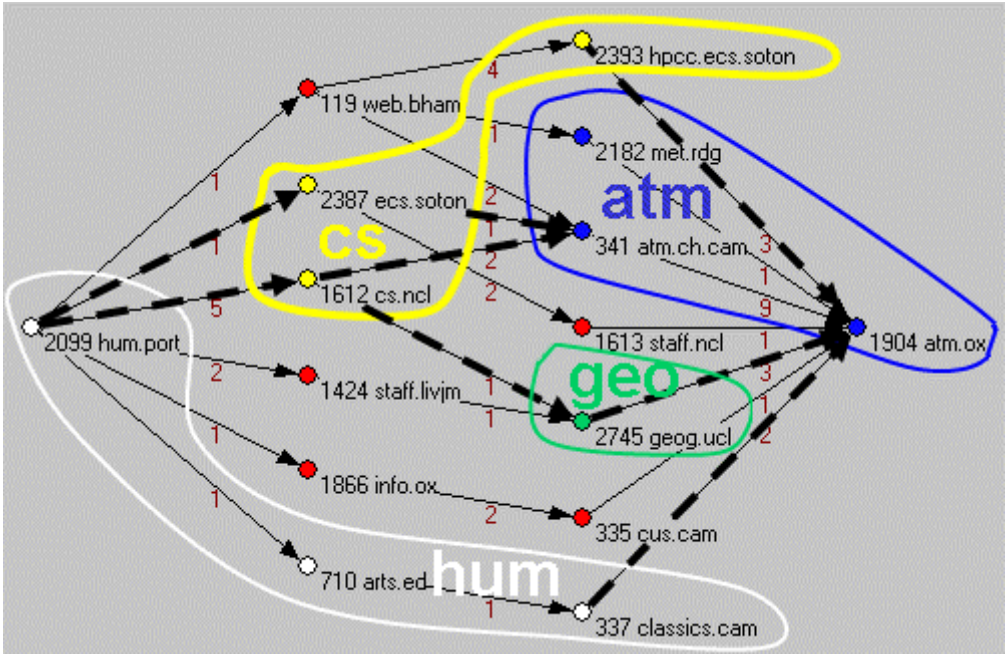


Figure 7.2.* Path net HN01 with enclosed topical areas humanities (hum), computer science (cs), geography (geo) and atmospheric sciences (atm). Non-enclosed nodes are generic-type. Transversal links are marked with dashed bold links. See Appendix 10 for affiliations.

The terms inter-topic, intra-topic, topical uniformity and topical diversity can be further exemplified by Fig. 7.2 of path net HN01 (cf. Section 6.5.3) showing all the shortest link paths between node 2099 (*hum.port.ac.uk*), Faculty of Humanities and Social Sciences, University of Portsmouth, and node 1904 (*atm.ox.ac.uk*), Atmospheric, Oceanic and Planetary Physics, Department of Physics, University of Oxford.

In the figure, one of the link paths between start node 2099 and end node 1904 passes the two subsite nodes 710 and 337 representing topics in the humanities (see affiliations in Appendix 10) just like the start node. The enclosed area 'hum' in the figure reflects an overall topical uniformity between the three humanities subsites connected by two intra-topic site level links. Correspondingly, the topical area 'atm' in the figure reflects an overall topical uniformity between the three atmospheric science subsites connected by two intra-topic site level links (equaling 10 intra-topic *page level*

links). The dashed bold links in the figure show transversal links, that is, inter-topic site level links reflecting topical diversity.

None of the 10 path nets exceeds path length 4 (cf. Table 6-7, Section 6.3.2). The topical areas in these path nets thus can have only relatively short intra-topic link paths, like the one comprising two intra-topic links in the humanities in Fig. 7.2 above. However, the UK data set of 7669 subsites also contained longer path nets. For instance, there were three path nets with path length 10 – the *longest* of the shortest link paths in the whole data set (cf. Table 5-3, Section 5.3.1). Fig. 7.3 below shows one of these three path nets with all the shortest link paths between SCC-node 438 (*www-hcl.phy.cam.ac.uk*) Hitachi Cambridge Laboratory, Department of Physics, University of Cambridge and OUT-node 3128 (*asian-mgt.abs.aston.ac.uk*), The Aston Centre for Asian Business and Management, University of Aston (cf. Fig. 5.12, Section 5.3.1).[86]

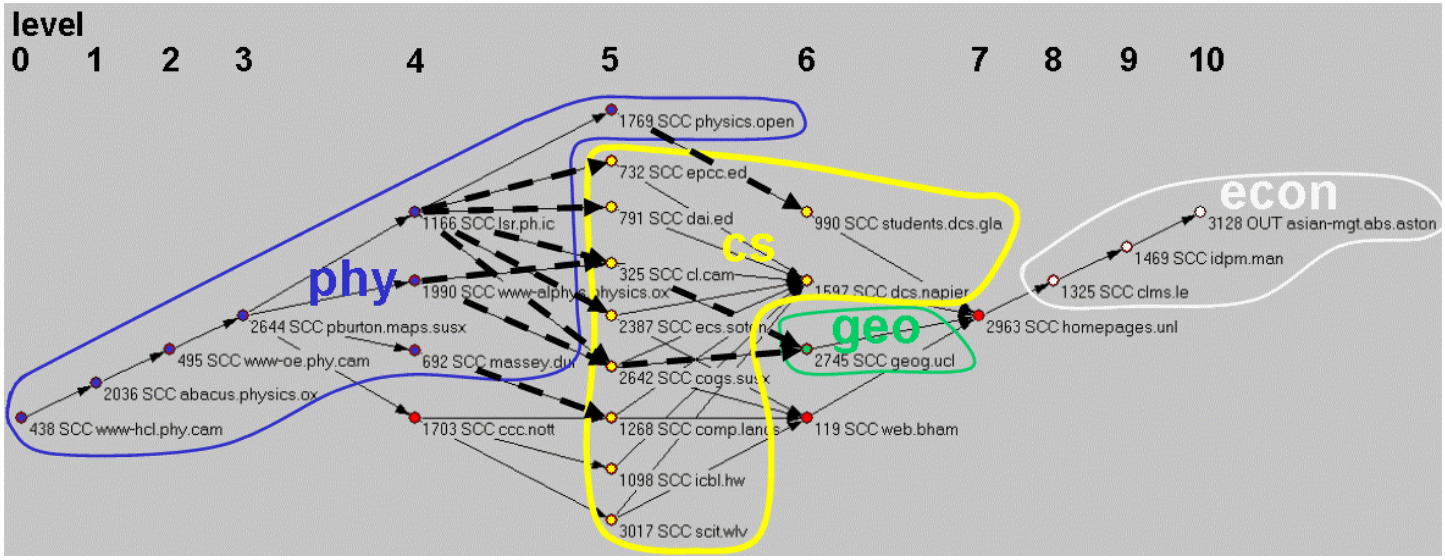


Figure 7.3.* Path net containing all shortest link paths of length 10 between node 438 (*www-hcl.phy.cam.ac.uk*) and node 3128 (*asian-mgt.abs.aston.ac.uk*) with enclosed topical areas physics (phy), computer science (cs), geography (geo) and economics/management (econ). Non-enclosed nodes are generic-type. Transversal links are marked with dashed bold links. Levels show link distances from start node. Due to limited space, initial and final nodes in the path net are drawn together. See Appendix 6 for affiliations of nodes in the path net.

The topical area 'phy' covers subsites in quantum physics, optical physics and general physics (cf. Appendix 6). An initial single 'path-thread' from path net level 0 to 3 furcates into four directions at level 3 with three nodes at level 4 and one node at level 5 still in the physics area. The topical area 'phy' is thus quite large and long, probably resembling how path nets would look in web spaces larger than the delimited and

[86] There were two other pairs of subsites separated by shortest link paths of length 10. One pair was – yet again – SCC-node 438 (*www-hcl.phy.cam.ac.uk*), Hitachi Cambridge Laboratory, Cambridge, and OUT-node 3439 (*cbshiva-2.mrc-lmb.cam.ac.uk*), a subsite (neither available in the Internet Archive nor on the Web) at the MRC Laboratory of Molecular Biology, Cambridge. The other pair was IN-node 1655 (*petrus.ncl.ac.uk*), The Whitaker Lab, School of Cell and Molecular Biosciences, University of Newcastle, and – yet again – OUT-node 3128 (*asian-mgt.abs.aston.ac.uk*), The Aston Centre for Asian Business and Management, University of Aston. Node 438 was thus start node in two of the three path nets with shortest path length 10 and node 3128 was end node also in two path nets.

relatively small UK academic web space. However, this conjecture remains to be verified in large-scale studies of link structures in academic web spaces possibly cross-national and combined with other types of web spaces, for example, commercial ones.

The long path net in Fig. 7.3 above also ends with a single path-thread from level 7 to 10. The last three subsites are concerned with labor market studies and management studies enclosed within the overall topical area 'economics/management'. As was the case in the sampled 10 path nets, computer science functioned as connector nodes on the link paths in the long path net as illustrated with the large topical area 'cs' in the figure.

The three path nets with path length 10 all had single path-threads starting the path net and ending it. Subsites in these path-threads had only few inlinks and outlinks. Between the two threads is the characteristic diamond-shaped middle section of more well-connected subsites. A start-thread thus provides a link path leading from a low-connected start node to a well-connected node that can reach out to wider parts of a web space, whereas an end-thread provides a link path from a well-connected node in to a low-connected end node.

Probably, a topic cluster on the Web will have peripheral nodes as well as core nodes (for instance, in the shape of pages, sites or other aggregations, depending on the employed research scope) all reachable along link paths from each other, like nodes in the strongest connected component (SCC) of a web graph. As noted earlier in Sections 5.1 and 6.6, subsets of the Web display the same characteristics as the Web at large. This *fractal 'self-similarity'* in the Web (Dill *et al.,* 2001; Kumar *et al.*, 2002) was evident in the present study in the 'corona' model of the UK academic subweb containing the same graph components as the Web at large, as well as the power-law like distributions of in-neighbors/out-neighbors and inlinks/outlinks. As noted in Section 3.5, Thelwall & Wilkinson (2003b) and Baeza-Yates & Castillo (2001) found graph components in national academic web spaces and countries, respectively, similar to graph components of the whole Web.

According to Dill *et al.* (2001), the self-similarity in the Web also means that so-called *TUC*s, "thematically unified clusters" (similar to the notion of topic clusters in the present approach) show 'bow-tie' model structures (cf. Section 5.1) containing the same graph components as larger web graphs.

Fig. 7.4 below incorporates graph components in a hypothetical example of a shortest link path between web node A in topic cluster T and web node H in topic cluster V. In the figure, each topic cluster comprises a 'corona' subgraph with the same component adjacencies as outlined in Section 5.1. Nodes A and B are thus peripheral nodes in cluster T and belong to the IN component of this cluster. One could hypothesize that if a link path starts in such a peripheral cluster-component and the link path shall end in another topic cluster, the link path must pass more well-connected nodes with sufficiently diversified range of target topics. Naturally, peripheral and low-connected nodes with only few in-neighbors/out-neighbors may nevertheless be connected to different topics. But the probability of having *transversal* (i.e. topically dissimilar) neighbors will most likely rise for nodes with *higher connectivity degree* (i.e. having more neighbors) and *higher betweenness centrality*. However, the latter assumptions have not been verified in the present study.

Some of the physics subsite nodes in the initial path thread in Fig. 7.3 above, may belong to an IN component of a topic cluster in physics in the investigated UK academic subweb. Correspondingly, some of the last subsite nodes in the final path thread in the same figure may belong to an OUT component of an economics/management cluster. (Recapitulating Section 5.1, nodes in the IN component can reach the SCC through directed link paths but cannot in turn be reached from the SCC. Correspondingly, pages in the OUT component can be reached through directed link paths from the SCC but cannot reach back.) In future studies of academic web spaces, it would be interesting to verify such graph components in topic clusters.

In the hypothetical example in Fig. 7.4 below, the link path has reached node C belonging to the SCC of the cluster. Node C has no out-neighbors in topical directions that can reach node H. However, node C is connected with node D also in the cluster SCC, and D has a transversal out-neighbor E belonging to topic cluster U. Through node F and yet a transversal link to node G, the link path can reach node H in cluster V.

The graph components only apply to the corresponding topic cluster. This means that node H belonging to the OUT component of topic cluster V may very well have outlinks to nodes *outside* the cluster even if it has no out-neighbors *within* the cluster.

On the real Web, the picture of topical clusters is more complex than schematically shown by the hypothetical and simplified 'ideal type' example in Fig. 7.4. As noted earlier in this section, topic clusters may overlap, for instance, due to sliding transitions between interdisciplinary topics. The peripheral 'tail' of one topical domain may thus be part of the core of another one.

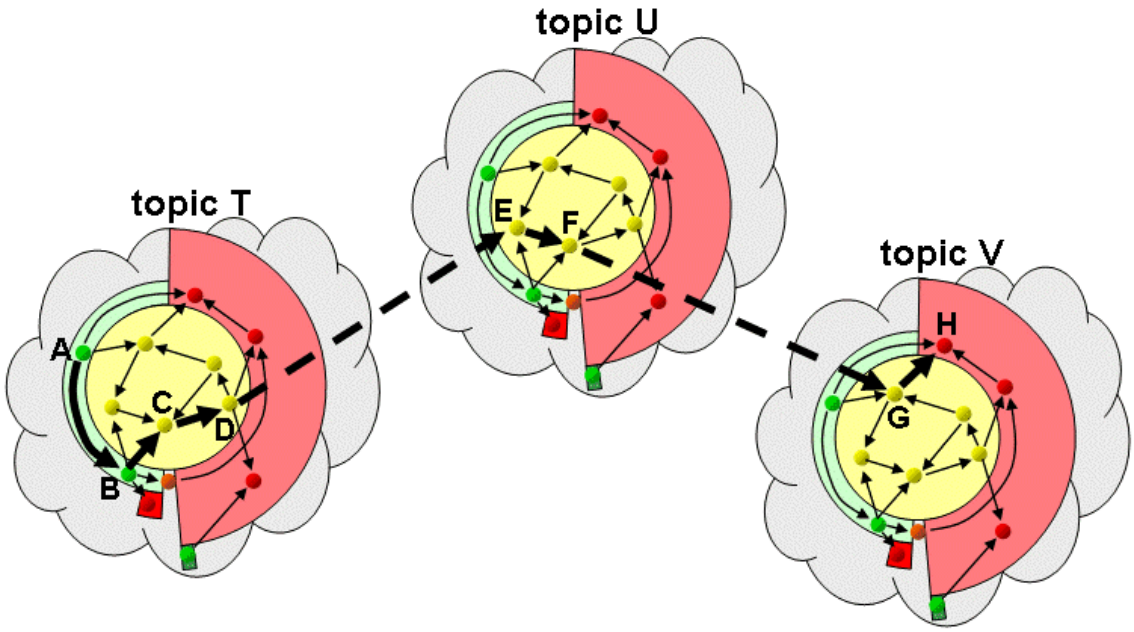


Figure 7.4.* Example of shortest link path (bold links) between nodes A and H crossing three topic clusters, each of which with 'corona'-like graph components. Transversal (inter-topic) links are marked with dashed bold links.

Fig. 7.5 below shows a more simplified version of Fig. 7.4 above with only the SCC of each topic cluster marked for sake of simplicity. The figure illustrates how *topical uniformity*, for example, of the four nodes A-D, within a topic cluster can provide an

intra-cluster link path from a low-connected node A to a more well-connected node D that has sufficient *topical diversity* in order to lead the link path in an *inter-cluster* direction where the end node H can be reached with the least possible link steps.

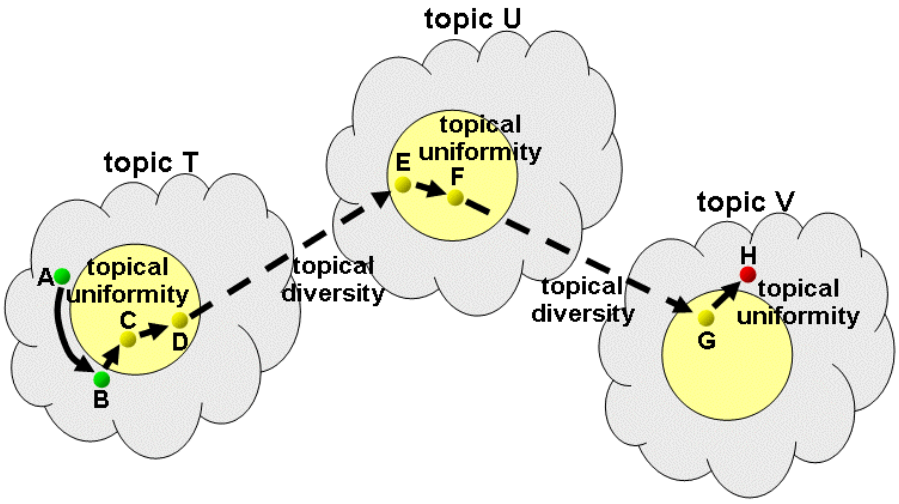


Figure 7.5.* Simplified version of Fig. 7.4 with a shortest link path comprising steps of topical uniformity and diversity to reach from node A to H crossing three topic clusters, each of which with a strongly connected component (SCC) denoted by an inner circle. Transversal (inter-topic) links are marked with dashed links.

As stated earlier, large-scale web spaces naturally contain more complex and muddled topical structures than indicated by the dichotomy of the 'ideal type' terms 'topical uniformity' and 'topical diversity' shown in Fig. 7.4 and 7.5, for example, because web sites may belong to more than one topic cluster due to no overall site topic, or clusters may overlap due to interdisciplinary topics. Nevertheless, the simplified dichotomy may clarify some essential mechanisms in how small-world phenomena emerge in interweaved document spaces on the Web. Small-world phenomena on the Web *may* thus be hypothesized as emerging through complementarities of topical uniformity and topical diversity as suggested above. The arrows in the following bulleted and distilled expressions should be read as *'may contribute to'* as the relations are hypothesized only:

- **topical uniformity + diversity → small world**

As argued in connection with Fig. 7.5 above, topical uniformity may support short distances *within* topic clusters by enabling low-connected nodes to reach well-connected nodes within a cluster. Correspondingly, topical diversity may support short distances *between* topic clusters by means of transversal links. The distributed knowledge organization of document spaces on the Web means that these hypothesized complementarities between intra-topic/intra-cluster *and* inter-topic/inter-cluster link structures emerge in non-engineered and self-organized ways without any central control or coordination. The co-existence of short local and short global distances in small-world web spaces may thus be brought about by *non-intentional 'collaborative weaving'* by a multitude of local link creators:

- **non-engineered intra-topic + inter-topic link structures → intra-cluster + inter-cluster structures → small world**

As stated earlier, some topic clusters can be research-related, for instance, comprising interlinked web pages and web sites of researchers and their projects, papers and institutions within a scientific domain. Other web clusters and interest communities can be in the shape of topic-specific web portals, subject gateways and resource guides. In the approach presented in this paper, such communities and clusters on the Web reflect an important tendency: the human urge to create local 'islands' of topic-focused uniformity on the Web. A conflicting tendency is the urge to expose a diversity of topical interests, for example, on personal homepages (cf. Chandler, 1998; Dillon & Gushrowski, 2000). For instance, a researcher or a layman may have converted bookmarks made in their browser for web pages visited on the Web and published them as long link lists on web pages with links to favorite web sites related to a wide range of work and leisure interests in what could be called *diversified resource collation*. Researchers may have made bookmarks to web pages of related researchers, institutions, projects and articles (cf. Abrams *et al.*, 1998; Sørensen *et al.*, 2001; Gottlieb & Dilevko, 2001). Furthermore, as indicated in the investigated 10 path nets, some bookmarks may be transversal links made to more peripheral scientific interests, or to leisure-related web sites. Bookmark lists thus reflect *trails* (cf. Bush, 1945) of researchers' diverse interests, preferences and actions on the Web, and constitute an obvious area for scientometric and webometric investigation.

From a library and information science perspective, it would be interesting to investigate further aspects behind this apparent human urge to *both* create topical uniformity *and* expose topical diversity as manifested in the small-world link structures of the investigated academic web space.

In a PhD course paper (Björneborn, 2002), the author pursues a hypothesis that the complementarities of topical uniformity and diversity for the emergence of small-world phenomena on the Web may be traced back to two major socio-technical *'control revolutions'*. These are concerned with new information technologies and practices as tools for regulation and centralized control in the 19th and 20th century industrial society (Beniger, 1986), and deregulation and decentralized control in the late 20th century information society (Shapiro, 1999).

Universal classification systems and centrally controlled world-encompassing information repositories and bibliographies, as suggested by the Belgian documentalist and internationalist Paul Otlet and others at the turn of the 20th century (cf. Rayward, 1994), as well as mainframe computers, electronic databases and the field of information science emerging from the 1950s, may be regarded as logical offshoots of the first-mentioned control revolution above. Furthermore, libraries, whether public or scientific, may be viewed as instruments to control the access to the knowledge universe through means of acquisition, classification and indexing, etc.[87] The advanced industrial revolution in the second half of the 19th century was accompanied and fueled by an

[87] In this library context, it may be interesting to note that classification is concerned with to *merge* together similar objects, whilst indexing is to *discriminate* between objects.

unprecedented growth in scientific specialization and publications which called for the need of library solutions. Thus, it is logical to include libraries in the framework of Beniger's (1986) control revolution even if the author does not make this inclusion.

Contrary to this, the Internet, the Web, and personal computers are vehicles for the latter control revolution for the decentralization and individualization of control in the construction and use of information systems as argued by Shapiro (1999). Vannevar Bush (1945) may be viewed as an initiator of this new control revolution with his seminal vision of a hypertext-like association-based and personalized information system, the *Memex* (cf. the Prelude).

However, the current Web may be conceived as embracing both the contrasting but complementary socio-technical regulatory and deregulatory trends traceable to the two control revolutions outlined above, as reflected in the topic-focused uniformity (due to preferential attachment, including global initiatives for metadata standardization, semantic web projects, subject gateways, etc.) *and* the topic-scattered diversity (reflecting human curiosity and broad interests of link creators including blends of personal and professional endeavors, bookmark lists, etc.) manifested in, for instance, small-world academic web spaces:

- **contrasting but complementary socio-technical trends**
  → regulatory + deregulatory tendencies → small-world

7.2.3 Web of genres

Fig. 7.6 below (identical with Fig. 6.37 in Section 6.4.5.4) shows pairs of page genres identified in the close examination of the 10 path nets revealing how links between the investigated academic subsites connect many different combinations of source and target page genres. Contrary to scientific literature where references interconnect a small number of different genre types, such as scientific papers, monographies, etc., an academic web space contains a much larger diversity of genres many of which not connected by cross-references in traditional scientific document networks known in paper media, for example, research project descriptions, teaching syllabi, and researchers' self-presentations. The term *genre* is here used in a broad sense in accordance with contemporary web terminology as outlined in Section 6.4.5.

The investigated path nets embraced a rich diversity in interconnected genre pairs. In large-scale academic web spaces, let alone the Web at large, there will inevitably be additional page genres as well as richer diversity of page genre combinations. The rich diversity of genre pairs may reflect a corresponding diversity of link motivations, including motivations related to teaching, research, institutional 'credit links' (Thelwall (2002c), leisure interests, social contacts, etc. A good overview of link motivations in an academic web space is given in Thelwall (2003d).

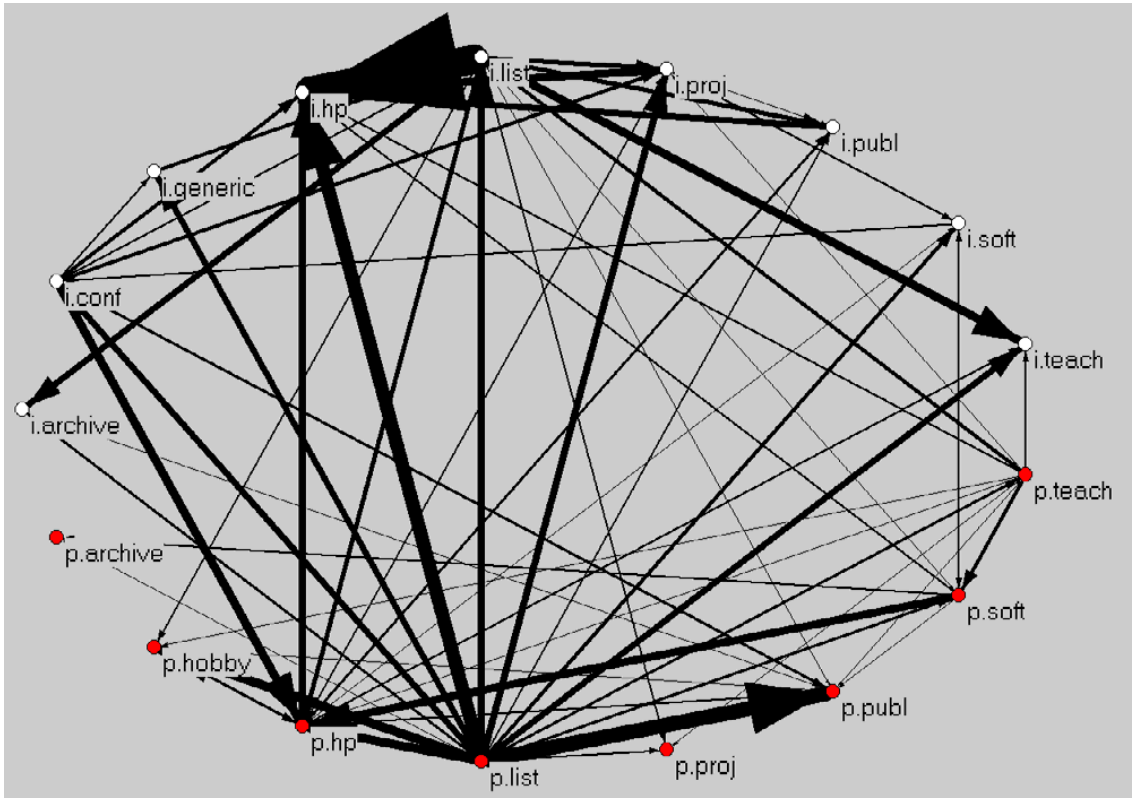


Figure 7.6. A *web of genres.* Pairs of page genres among 352 followed links in the 10 path nets. Link width reflects link counts. Due to the Pajek software, thinner reciprocal links are concealed underneath thicker links. Genre selflinks are not shown. White nodes denote institutional genres and red (dark) personal.

As noted in Section 6.4.5.4, Fig. 7.6 gives an impression of the diverse *genre connectivity* and possible link paths between genres in an academic web space. Furthermore, the figure gives an intuitive support to how the Web may be conceived as a *web of genres* with page genres linked to other genres and with *genre drift*, that is, changes in genres of pages along link paths. No studies have been found discussing how academic or other web page genres are interconnected in large web spaces and how the Web may be conceived as a web of genres. A diverse web of genres may affect possibilities for small-world link structures. In this context, the concepts of *genre drift* and *topic drift* may be combined with those of *topical uniformity* and *diversity* outlined in the previous section in order to understand how small-world properties emerge in document spaces on the Web, viewed from a library and information science perspective.

The concept of *topical diversity* is closely related to that of *topic drift* (cf. Section 6.5.1). As outlined by Bharat & Henzinger (1998), topic drift is concerned with the change of topics when a human web surfer or a digital web crawler follows links from web page to web page. The topic drift problem thus negatively affects possibilities for topically *focused* web crawls. However, in context of the dissertation, understanding how topic drift emerges on the Web is important in order to understand how small-world properties emerge in the shape of short distances across topical domains on the Web.

Using the constructed genre matrix (Appendix 15) behind Fig. 7.6 above, some possible shortest link paths between genres may be illustrated as in Fig. 7.7a-c below, showing how genre drift hypothetically may create shortest link paths in an academic web space. For example, a link path starting on a personal hobby page in Fig. 7.7b may end on a conference page after passing a personal homepage linking to an institutional research project page.[88]

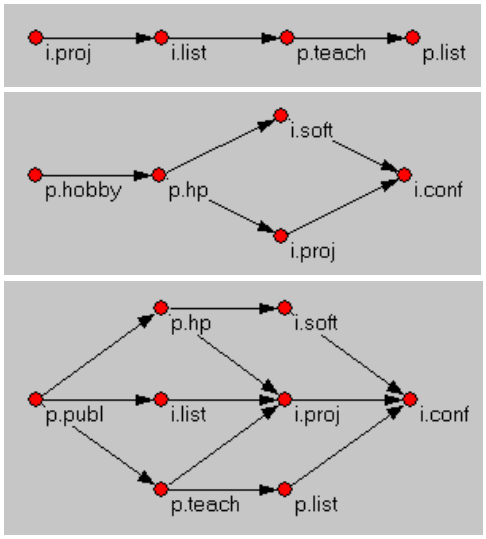


Figure 7.7a-c. Genre drift along link paths may hypothetically create shortest paths in an academic web space (based on genre matrix from 10 path nets).

The three figures may be transformed into the example in Fig. 7.8 below that illustrates complementarities of genre drift *and* topic drift along shortest link paths across topic clusters in an academic web space. Where *genre drift* is concerned with the change of genres along link paths when a human web surfer or a digital web crawler follows links from web page to web page, *topic drift* (cf. Section 6.5.1) is concerned with the change of topics when links are followed from page to page. As stated above, the 10 path nets in the dissertation all constitute examples of deliberately induced topic drift in order to construct investigable 'mini' small-world link structures.

[88] NB. In the 10 path nets the page genres are only connected *pair-wise*, cf. Fig. 6.39, Section 6.4.5.4. In the path nets, a *target* genre at a node is thus not directly linked with a *source* genre at the same node, because the focus in the dissertation is on *inter*-site links – not *intra*-site.

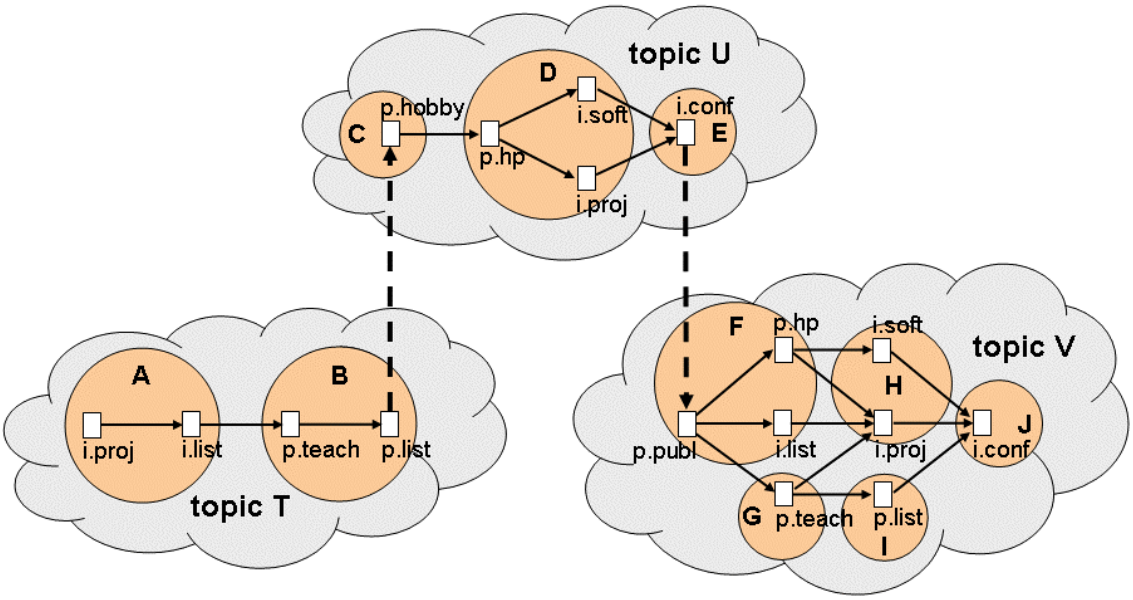


Figure 7.8.* Intra-cluster genre drift and inter-cluster topic drift along shortest link paths from web site A in topic cluster T to site J in topic cluster V. Transversal inter-topic links are denoted with dashed bold links.

In web site A in Fig. 7.8, a research project page (*i.proj*) in, say, astronomy could have a site *selflink* to an institutional link list on the same topic. The link list in turn has a site *outlink* to a researcher's course page on astronomy (*p.teach*) at site B having a site selflink to a colleague astronomer's personal bookmark list on the topic. The bookmark list also contains some transversal links to the colleague's peripheral research interests as well as leisure interests including, say, football. The target of the transversal link is a hobby page about a football club, the page made by a researcher in chemistry. The bookmark list thus creates a connection to topic cluster U comprising chemistry web sites. As the figure reproduces link paths based on a constructed genre matrix as noted earlier, only links between *different* page genres are represented in the figure. On the real Web, links would naturally connect similar genres as well, for example, two institutional link lists located on different sites.

In the dissertation, focus is on the topics of interconnected web *sites*. Focus in this discussion section is thus on inter-*site* topic drift and not on inter-*page* topic drift in accordance with the discussion in Section 6.5.1 on this issue. However, instead of topic clusters comprising sites, one could operate with clusters comprising pages or any other level of web node aggregation chosen as appropriate in a research design. In the above case, a topic cluster of football *pages* would cover web sites in many domains, both inside and outside academic web spaces, including sites in astronomy and chemistry as exemplified above.

In Fig. 7.4 and 7.5 in Section 7.2.2 was shown how *topical uniformity* within a cluster could provide short distances from a low-connected web node to a more well-connected node enabling short link paths from the low-connected node to a node in another topic cluster. Fig. 7.8 above gives an intuitive hint about some possible underlying mechanisms that should be taken into account in order to understand how short link distances emerge in web spaces.

As simplistically visualized in Fig. 7.8, web sites and topic clusters consist of diverse genres, for instance, in an academic web space it could be institutional homepages, link lists, course pages, conference pages, etc., as well as personal homepages, link lists, etc. *Genre drift* within web sites and within a topic cluster containing many web sites may thus enable short link paths from a web page with no transversal links to a web page containing transversal links. Such a transversal link in turn creates topic drift by reaching out to other topic clusters hence causing short link distances between topic clusters. Small-world phenomena on the Web may thus hypothetically be conceived as entailed by complementarities of genre drift and topic drift associated with distributed web page and link creations:

- **distributed page/link creations → genre drift + topic drift → small world**

As suggested above, topic clusters with genre diversity entail genre drift along intra-cluster links. At the same time, some web page genres, like institutional or personal link lists, are more diversity-prone containing transversal links. Such genres with topical diversity thus function as providers of topic drift along inter-cluster links. The combination of *intra-cluster genre drift* and *inter-cluster topic drift* may thus contribute to the emergence of small-world properties of academic web spaces, and possibly in other areas of the Web as well:

- **topic clusters with genre diversity → intra-cluster genre drift**
- **genres with topical diversity → inter-cluster topic drift**
- **intra-cluster genre drift + inter-cluster topic drift → small world**

Once again it should be noted that the arrows in the above bulleted and distilled expressions should be read as *'may contribute to'* as the relations are hypothesized only.

7.2.4 Crumpled-up web spaces

In continuation of the discussion about the Web as a web of genres with inherent genre drift and topic drift, this section will introduce and discuss an intuitive visualization how page genres and links may shorten distances on the Web.

In Section 6.4.5.1, the notion of *outlink-prone* and *inlink-prone* page genres was introduced. Outlink-prone genres like personal link lists provide many outlinks to other genres, whereas inlink-prone page genres such as institutional homepages correspondingly receive many inlinks from other genres. In this context, a metaphor of *hooks* and *lugs* can be used to visualize how genres may contribute to short distances across web spaces by 'crumpling up' these web spaces. Fig. 7.11 below shows a *hook* (Fig. 7.9) grasping a so-called *lug* (Fig. 7.10).

Figure 7.9. Hook.

Figure 7.10. Lug.

Figure 7.11. Hook grasping lug.

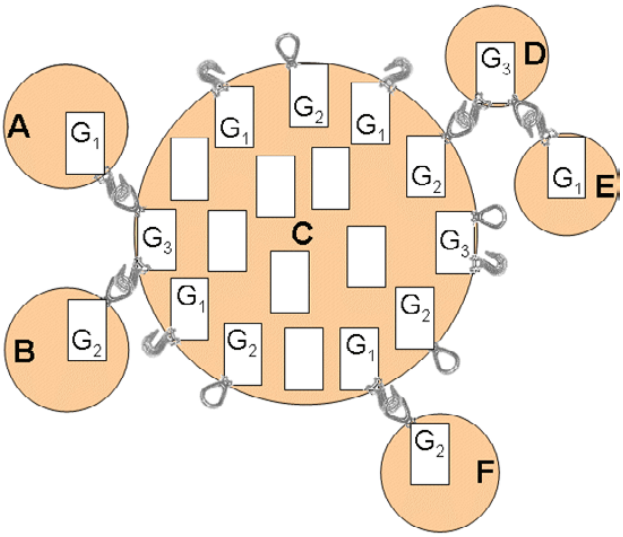


Figure 7.12.* Some web page genres may function as outlink-prone hook genres (G_1), inlink-prone lug genres (G_2), or combined hook&lug genres (G_3), here pulling web sites A-F close together.

In Fig. 7.12 above, hooks and lugs have been mounted symbolically on web pages. Pages with just hooks represent outlink-prone page genres (G_1), like the abovementioned personal link lists providing many outlinks to other genres. Pages with only lugs represent inlink-prone page genres (G_2) such as institutional homepages receiving many inlinks from other genres. Pages with both lugs and hooks represent in-/outlink-prone page genres (G_3), like institutional link lists both receiving and providing many links. As visualized in the figure, such 'hook&lug' genres may create short distances directly along interlinked web *pages* placed on different web sites, for example, the inlinking web site A and outlinked site B. Naturally, real web pages can contain more outlinks and inlinks than represented by the maximum one hook and lug included for sake of simplicity in the figure. On the Web, pages can thus pull together many other pages, with page genres from academic and non-academic web spaces providing a diversity of link sources and targets for each other.

Fig. 7.12 illustrates how 'hook genres', 'lug genres' and 'hook&lug genres' pull web sites A-F close to each other thus contracting link distances on the Web – and 'crumpling up' web topologies.

The intuition of 'crumpled-up' web topologies (cf. Björneborn, 2001a) may be illustrated in a very simplified way by taking a piece of chequered paper and crumpling it up by pulling together points from opposite parts of the paper. Paths along chequered

lines between two selected points on the paper get shorter the more new contact points and merged line paths are created between the folded layers in the crumpled-up paper, cf. Fig. 7.13.

Figure 7.13. Crumpled-up paper.

However, at a certain point it is not possible to keep on crumpling up the paper because of the inflexible three-dimensionality of the paper. Contrary to this, possibilities of 'crumpling up' web spaces are unlimited because the virtual space of the Web is billion-dimensional. Every new outlink on the Web may interconnect with any existing and accessible web page. Every new outlink may thus cut across the existing billions of links in the Web space. In accordance with the above metaphor of 'hook genres' and 'lug genres', every new outlink may furthermore be conceived as a 'hook' capable of contracting and reshaping existing web spaces. Using this spatial hook metaphor, *transversal* links may be considered as hooks that pull 'distant' web nodes close to each other and thereby also pull together the neighborhoods and clusters which these nodes already belong to – and the neighborhoods of the neighborhoods, etc. – thus possibly reinforcing small-world properties of the crumpled-up Web.

- **transversal links = 'hooks' → crumpled-up small-world web spaces**

Fig. 7.14 below shows a 3D visualization of the same path net as in Fig. 7.3 in Section 7.2.2 with all shortest link paths (path length 10) between the blue node 438 (physics subsite) and the white node 3128 (economics/management). The figure may give an intuition of how transversal links pull topical areas close to each other in a web space. For example, transversal links pull together the encircled nodes 1168, 325 and 2745, belonging to physics, computer science and geography, respectively. As a path net does not show all neighbors of the nodes, but only those neighbors located on the shortest link paths, it should be noted that all neighbors of the above three subsites will also be pulled close together by the transversal links, thus crumpling up this local web space. Transversal links thus not only pull separate web pages and web sites close to each other, but also pull entire web neighborhoods close together. The latter aspect parallels

what happens when also the surroundings of two points on the above piece of crumpled-up paper come close to each other when the two points are pulled together.

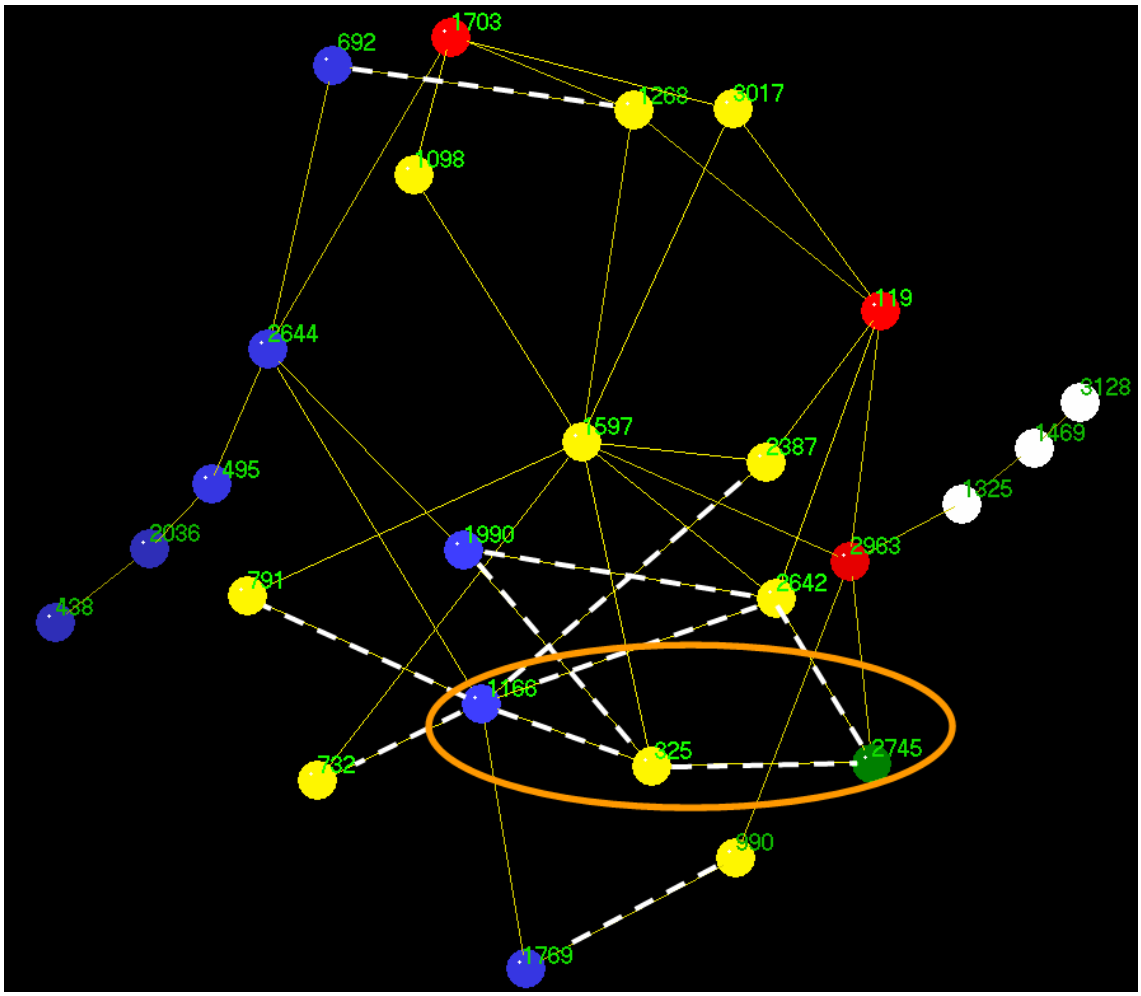


Figure 7.14.* 3D visualization made in network software tool *Kinemage* of the same path net as in Fig. 7.3 (Section 7.2.2) containing all shortest link paths (length 10) between node 438 (*www-hcl.phy.cam.ac.uk*) and node 3128 (*asian-mgt.abs.aston.ac.uk*). Blue nodes denote physics subsites, yellow computer science, green geography, white economics/management, and red generic-type subsites (* cf. color print placed before appendices). See Appendix 6 for affiliations of subsites nodes. Encircled nodes 1168, 325 and 2745 exemplify topical domains pulled together by transversal links. Transversal links are marked with dashed bold lines.

The crumpled-up link structures on the Web imply that diverse topics and page genres may be separated by one or just a few links. Web genres and topics thus may co-exist literally side by side in crumpled-up web topologies due to genre drift and topic drift as discussed in Section 7.2.3.

The spatial metaphor of crumpled-up web topologies may be extended to intuitively grasp the complementarities of topical uniformity and diversity for creating both short local and short global distances in a web space as discussed in Section 7.2.2. Each topic cluster may thus be conceived as a crumpled-up subset of a web space where a strongly connected component of the cluster may yield short local link distances

within the topically uniform cluster. As suggested above, transversal links spanning topical diversity yield short global distances by crumpling up the overall web space, that is, by pulling topic clusters close together. Fig. 7.15 below illustrates three crumpled-up topic clusters close to each other within an embracing crumpled-up web space. As the figure mimics three dimensions only, the figure can just give a dim impression of how topic clusters crumple up in a pervasive billion-dimensional web space.

Figure 7.15.* Crumpled-up web space with three crumpled-up topic clusters.

7.3 Exploratory capabilities in a small world

> *"The raison d'être of hypertext is movement in innumerable directions"*
> (Davenport & Cronin, 1990, p. 189)

The indicated influence of personal link creators and computer science subsites for the emergence of small-world link structures in the investigated academic web space have been discussed in the previous sections, as well as the hypothesized complementarities of topical uniformity and diversity, including genre drift and topic drift, for the emergence of such small-world web spaces.

Summing up some of the preceding discussions, the distributed knowledge organization of a vast academic web space may lead to small-world properties of this web space. This is because topical uniformity *and* diversity in the local link creations of personal and institutional web constructors may affect the emergence of topic clusters and transversal links that in turn may enable small-world properties in the shape of both short local and global link distances in the web space. These hypothesized relations may be condensed in the following way:

- **distributed knowledge organization → non-engineered link creation → topical uniformity + diversity → topic clusters + transversal links → short local + short global link distances → small-world properties**

Small-world knowledge organization comprising *'collaborative weaving'* by millions of link creators may thus be conceived as an important non-engineered *organizing principle* for handling a vast distributed document space by the emergence of topic clusters and transversal links enabling the co-existence of both short local and global link distances in this vast document space.

Logically, the *construction* of an information system affects the *functionality* of the system: *Knowledge organization* affects options for users' *information behavior*, as also noted in Section 1.3. In other words, the ways resources are organized and interconnected in an information system influence the ways in which the system can be navigated and exploited by users that access information by goal-directed searching, browsing, serendipitous encountering or combinations of these different information behaviors (cf. e.g., Cove & Walsh, 1988; Catledge & Pitkow, 1995; Erdelez, 1995). For instance, the open shelves in modern physical libraries allow patrons to browse along the shelves and, for instance, serendipitously encounter an exposed book on a topic not searched for, but nevertheless of interest for the patron.

Hypertextual link structures similarly affect accessibility and navigability on the Web. This point is underlined by Berners-Lee & Cailliau (1990) in their proposal for a new hypertext project called WorldWideWeb: "*HyperText is a way to link and access information of various kinds as a web of nodes in which the user can browse at will*." The Web was thus envisaged as a tool to facilitate easy access to networked information sharing and browsing as noted in Chapter 1. Early on, pre-Web hypertext researchers as, for instance, Yankelovich, Meyrowitz & van Dam (1985) were also concerned with this functionality of hypertextual document systems that should *"help scholars both create connections and follow those made by others."* (p. 16). As mentioned earlier, this focus on hypertextual link creations and link traversals conducted by scholars can be traced back to Bush (1945). As noted in Section 2.4.1, Davenport & Cronin (1990) who were early LIS recognizers of hypertext potentials for the conduct and creativity of science, stress how hypertext entails freedom of movement within and between texts – as quoted in the opening quote of this section.

In the present study, this hypertextual freedom of movement includes accessibility and navigability within and between topical domains in academic web spaces. In continuation of this line of discussion, a natural follow-up question is: How does the distributed knowledge organization leading to small-world properties of web spaces affect *exploratory capabilities* across topical domains in such vast document networks? This question is addressed in the next sub-sections on *small-world explorations* and *small-world serendipity*. The sub-sections should be conceived as perspectivations pointing to future studies.

7.3.1 Small-world explorations

In the dissertation, the term *'exploratory capabilities'* covers possibilities for users to navigate and access information in an information system. Exploratory capabilities in the shape of accessibility and navigability in information systems is thus a core issue in library and information science, as noted earlier. The use of the term is inspired by Doyle (1963) who used 'exploratory capability' to supplement the traditional criterion of 'relevance' for evaluating performance of an information retrieval (IR) system. Other researchers in library and information science, like Bates (1986; 1989) and Pejtersen (1991) have followed up on the importance of exploratory capabilities with regard to online information retrieval.

Supposedly, small-world knowledge organization should affect exploratory capabilities in a web space because it should be easier to explore this space if both local and global distances are short. Furthermore, as small-world knowledge organization comprises both topical uniformity and diversity it may perhaps facilitate both *convergent* (goal-directed) searching and *divergent* (serendipitous) browsing. (The concepts of 'convergent' and 'divergent' are further elaborated in Section 7.3.2).

However, there are no such straightforward causal relations between small-world knowledge organization and easy exploratory capabilities. As stated by several researchers (e.g., Walsh, 1998; Kleinberg, 2000; Adamic *et al.*, 2001; Vázquez, 2001; Watts, Dodds, & Newman, 2002; Menczer, 2002; White & Houseman, 2003) it is difficult to identify shortest link paths across a web space if only information about *local* link topologies is available, as is the case for human web surfers or digital web crawlers following links one at a time on the Web having no maps over larger web areas. In the present study, the transversal links contributing to short link distances between topics could be identified because information about the *overall* link topologies of the *whole* investigated academic subweb was at hand. However, special so-called *decentralized algorithms* have been developed that utilize local (*'decentral'*) connectivity information for identifying short paths through a network if no global (*'central'*) link data are vacant (e.g., Kleinberg, 2000; Adamic *et al.*, 2001; Menczer, 2002). In particular, hub-like nodes with many outlinks may be exploited (e.g., Kleinberg, 2000; Adamic *et al.*, 2001) in such decentralized algorithms.

The indication of *transversal-prone* personal link lists and computer-science sites in the present study may contribute to the development of such decentralized algorithms for finding short paths that reach 'distant' nodes across the Web. Furthermore, as already suggested in Section 7.1.1, transversal-prone pages and sites may be exploited for more *exhaustive* web crawls by search engines that could use personal link lists (especially bookmark lists, easy identifiable by the high link density) or computer-science sites as starting points for the web crawlers allowing traversals along transversal links in order to reach a wider range of topical areas. On the other hand, as suggested earlier, if identified (perhaps by techniques based on the present study), transversal links may also be *avoided* allowing more topically *focused* web crawls.

Future studies may reveal more aspects and applications of relations between small-world knowledge organization and exploratory capabilities, including potentials for serendipity as put into perspective in the next sub-section.

7.3.2 Small-world serendipity

Contrary to the previous sections that discuss findings and hypotheses derived from the empirical investigations, this section is concerned with more analytical spin-offs.

In future studies it would be interesting to follow up on the original underlying idea behind the dissertation concerned with how small-world link structures – including small-world *co-linkage* and other topological factors – may affect potentials of *serendipity*, computer-supported *knowledge discovery* ('data mining') and *creativity stimulation* in information spaces (cf. e.g., Swanson, 1986; Bawden, 1986; O'Connor, 1988; Davies, 1989; Van Andel, 1994; de Jong & Rip, 1997; Ford, 1999; Björneborn & Ingwersen, 2001)

Small-world knowledge organization may thus create a *possibility space* (using a term from Perkins, 1992; 1995) with enhanced chances for serendipitous information encountering and computer-supported knowledge discovery when traversing a document space with short distances between topics. The rationale behind this hypothesis is that the shorter link distances are between web pages belonging to different topical domains in crumpled-up and small-world web spaces, the larger the probability is to encounter and discover something unexpected while traversing these link structures.

Serendipity is *"the art of making an unsought finding"* (van Andel, 1994). Such unsought findings may belong to the same topical area that is already under exploration – or perhaps more frequently, when encountering unexpected information belonging to a completely different topical area. In the following discussion, such unexpected cross-topic findings are in focus.

According to Rice (1988), potentials for serendipity are proportional to the number of access points in an information system. In this context, Celoria (1969) states that '*loopholes'* in library structures may allow 'higher browsing', a kind of deliberately serendipity-stimulating behavior by a patron browsing across boundaries of topically heterogeneous domains (also cf. Liestman, 1992). Transversal links may function as such unpredictable 'loopholes' and subject access points in an information system as the Web.

In general, hyperlinks are subject access points and guiding tools that may help users to discover optional and alternative directions in order to encounter information in a hypertextual information system, e.g., in the Web. *Inconsistencies* assigned by authors, indexers, etc., may create unexpected access points and 'loopholes' between information nodes in an information system. In other words, potentials for serendipity in an information system depend on the *lack of total order and uniformity*, that is, the *presence of diversities.*

Serendipity typically occurs when users engage in exploratory, browsing information behavior (cf. e.g., Bates, 1986, 1989; Chang & Rice, 1993; Erdelez, 1995, 1997, 2000; Williamson, 1998; Toms, 1998, 2000). Such behavior may help researchers

and others discover information, which they did not know in advance they needed, or which they did not know existed – and thus could not make a request about to an information system. In other words, serendipity may occur when the *interest space* of an user, i.e. the multitude of tasks, problems and interests, which are more or less urgent or latent in the user's life and related to work, leisure, etc., are triggered when the user traverses an *information space* (broadly defined: e.g., a city, a library, the Web) and encounters contents, options and pointers offered by this information space.

A human surfer or digital crawler exploring the Web by following links from web page to web page has the possibility to move from one topic cluster to another topically 'distant' cluster using a single transversal link as a shortcut. For example, a researcher in information science could have made a transversal link on a web page with a bookmark list targeted to another of his interest fields, creativity stimulation, thus contributing to small-world crumpled-up link structures by connecting two dissimilar topics.

'Convergent' (topic-focused) and *'divergent'* (topic-scattered) link structures, with topic clusters corresponding to the former type and transversal links to the latter, may support exploration of the Web conducted by a combination of convergent (goal-directed) and divergent (serendipitous) ways. Convergent goal-directed search behavior may thus identify a central point of information that subsequently may function as point of departure for divergent serendipitous browsing. Correspondingly, information that is unexpectedly encountered while browsing may lead to a need for more focused search strategies. Users moving through an information space may change direction and behavior several times as their information needs and interests perhaps develop or get triggered dependent on options and opportunities occurring on the way through this space. For example, goal-directed searchers may find desired information in topic-focused web clusters. Then again, goal-directed searchers (being sufficiently curious and 'non-blinkered') as well as more intuitively browsing surfers may be led astray by transversal links and other topic-scattering links. This *topic drift* may give rise to new goal-directed search behavior in a new topic area. Thus, convergent and divergent information behavior may complement each other supported by complementary convergent and divergent link structures.

Complementarities of convergent and divergent information behavior are also stressed by Van Andel (1994) who notes that *"directed (re)search and serendipity do not exclude each other, but conversely, they complement and even reinforce each other"* (p. 644).

The use of the terms *convergent* and *divergent* is inspired by Ford (1999). In his article 'Information retrieval and creativity: towards support of the original thinker', Ford (ibid.) describes how traditional information retrieval (IR) systems use *convergent* information processing in order to create as exact matches as possible between representations of requests and representations of documents. Ford advocates for the need of IR systems *also* capable of *divergent* information processing by stimulating creative, divergent thinking of users of IR systems. This approach is in accordance with related work of Bawden (1986), O'Connor (1988) and others on the stimulation of creativity in information systems.

Understanding complementarities of convergence and divergence on the Web – both in link structures that are explored and in the behavior of human or digital agents

that are exploratory – could be important in order to develop creative methods of computer-supported knowledge discovery on the Web (cf. Björneborn & Ingwersen, 2001).

Discussing how serendipity might be induced, facilitated or at least not inhibited in an information system, different approaches have been taken (cf. e.g., Bawden, 1986; Twidale *et al.*, 1997; Toms, 1998). For example, Toms (1998) warns against negative side effects of the intensive focus on information filtering, intelligent agents and user profiles on the Web that may limit the exposure to unanticipated information and narrow the need for browsing. As stated by Toms (ibid.),

> *"these techniques may eliminate or seriously handicap the ability of a user to experience a chance encounter with digital information [...] as users are boxed into or presented with narrower and narrower semantic content segments which are based on perceived interest or system-based profiles of user actions, ultimately destroying social, educational, political and intellectual diversity*."

This caution is analogous to the '*balkanization'* of science cautioned by Van Alstyne & Brynjolfsson (1996) in their concern that *"the intellectual cross-pollination of ideas can suffer"* if *intra*-disciplinary interactions substitute for *inter*-disciplinary ones leading to scholarly *"tunnel vision and peripheral blindness"* (p. 1480). However, the small-world properties of the UK academic web space reflect short link distances between a rich diversity of scientific domains and thus also indicate that all scholarly link creators are *not* necessarily 'balkanized' into an insularity of subpopulations. As stated by Van Alstyne & Brynjolfsson (ibid.), balkanization can be avoided if scientists use information technologies to seek, select and encourage both diverse contacts and specialized ones. The earlier noted (Section 7.2.2) rich topical diversity encountered in the investigated UK academic web space may reflect a deliberate government policy to promote diversity of institutional missions in higher education (Dearing, 1997; Thelwall & Wilkinson, forthcoming), thus countering risks of scientifically 'balkanization'.

In order to stimulate serendipitous information encountering on the Web, Campos & Figueiredo (2001) suggest software agents that browse the Web in a simulation of typical human browsing behavior in order to uncover useful and not sought for information that may stimulate curiosity and serendipitous insights by providing new entry points to users. Further, the heterogeneity of the Web can be a fertile source for making discoveries. As stated by de Jong & Rip (1997), discoveries often result from "making *unexpected* combinations of heterogeneous resources" (italics in original), implying that it is not possible in advance to tell what resources are required.

The small-world approach developed in the present study with transversal links identified in path nets induced by randomly juxtaposed topics could perhaps be incorporated in such computer-supported *serendipity facilitation* as suggested by Campos & Figueiredo (2001) above.

7.4 Summary of generated hypotheses

The following list gives a brief summary of the hypotheses that have been generated but not verified from the close investigation of the case studies of the 10 path nets as discussed in the previous sections (brackets show sections concerned). As noted at the beginning of chapter 7, the generated hypotheses are especially concerned with the third research question dealing with what properties may contribute to the emergence of small-world link structures in an academic web space:

If small-world link structures can be identified in this academic web space, which properties can be observed that contribute to such link structures?

- *Complementarities* between topical *uniformity* (topic clusters) and *diversity* (topic drift, i.e., transversal links) in a web space may support the emergence of small-world properties in the web space (Section 7.2.2);
- *Topic clusters* on the Web may consist of *graph components* that correspond to the 'corona' or 'bow-tie' model of larger web graphs (Section 7.2.2);
- *Genre drift* may create short paths both within and between topic clusters in a web space. Genre drift combined with complementarities between *topic clusters* and *topic drift* may support the emergence of small-world web spaces:
  - topic clusters with genre diversity → intra-cluster genre drift
  - genres with topical diversity → inter-cluster topic drift
  - intra-cluster genre drift + inter-cluster topic drift → small world

  (Section 7.2.3);
- Small-world link structures with topic clusters, topic drift (transversal links) and genre drift may reflect non-engineered *distributed knowledge organization* of a vast information space enabling both short local and global distances along link paths (Section 7.2.3);
- Small-world knowledge organization comprising *'collaborative weaving'* by a multitude of link creators may be an important *non-engineered organizing principle* for handling a vast distributed document space by enabling both short local and global link distances (Section 7.3);
- Small-world knowledge organization may affect *exploratory capabilities* within and across topical domains in the information space because exploration is facilitated by co-existence of short local and global link distances (Section 7.3);
- Small-world knowledge organization may create a *possibility space* with enhanced chances for serendipitous information encountering and computer-supported knowledge discovery when traversing a document space with short distances between topics (Section 7.3.1);
- Small-world knowledge organization comprising both topical uniformity and diversity may facilitate both *convergent* (goal-directed) searching and *divergent* (serendipitous) browsing (Section 7.3.2).

Future studies are needed in order to test these hypotheses.

7.5 Implications for library and information science

This chapter ends with some reflections on what overall implications may be drawn for library and information science based on the discussions unfolded above.

Research in library and information science (LIS) has traditionally been concerned with how information systems shall be constructed in order to support goal-directed, convergent information retrieval, '*recovery'*, and not so much concerned with how information systems *also* may facilitate divergent information behavior related to serendipity, creativity stimulation and information '*discovery'*. This dichotomy of 'recovery' and 'discovery' is inspired by Garfield (1986). The traditional focus in LIS, encompassing system-driven and user-driven approaches (cf. Ingwersen, 1992), on supporting goal-directed information *recovery*, is reflected by Belkin's (1978) influential definition of the overall aim of LIS:

> "*Facilitating the effective communication of desired information between human generator and human user"* (p. 58).

Important keywords here are *effective communication* and *desired information*. This LIS approach with its focus on system control, effective IR (information retrieval) algorithms, rational search behavior, explicit user needs, and task-based relevance assessments, etc., and with its 'top-down' regulatory approach of optimizing information systems is still a core mainstream research area in LIS. Logically, this is an important and necessary research area, since there is a big need for supporting convergent, task-driven, goal-directed information behavior in information systems. However, this approach may no longer be sufficient to cope with issues necessitated by the rise of the Internet and the Web. Instead, the traditional LIS approach could be *complementarily extended* to *also* encompass issues concerned with understanding the construction and use of information systems *without* centralized control. For example, issues concerned with distributed knowledge organization, topical diversity, small-world phenomena, serendipity, knowledge discovery and creativity stimulation as discussed above. As stated by Kumar *et al.* (2002):

> *"The global distribution and diversity of Web content creation is in sharp contrast to the more controlled, homogeneous domains that fostered classical information retrieval."*

Thus, there is a need for redefining the overall aim and explanatory framework of LIS research, so it encompasses *both* convergent and divergent information behavior conducted by users in *both* 'top-down'-constructed information systems (e.g., traditional libraries and bibliographic databases) *and* in 'bottom-up'-constructed distributed information systems such as the Web. A redefined and broadly encompassing LIS aim should include how users of both *engineered* ('top-down') and *non-engineered* ('bottom-up') information systems may be supported in order to *explore and exploit options embedded* in these information systems.

Such a redefined LIS approach could raise questions like: *How easy is it to explore an information terrain and how lucid are options presented in this terrain?* In the framework of this dissertation, such *exploratory capabilities* of an information terrain are concerned with (1) *navigability* or freedom of movement for users traversing

the information terrain, and with (2) *terrain lucidity*; how easy it is to see the information terrain and the options embedded in it. For example, a web surfer can only follow already existing links between web pages. Thus, navigability and freedom of movement depends on *where* and *whereto* millions of web constructors have placed and targeted hyperlinks. Terrain lucidity in information systems is concerned with traditional core LIS concepts such as classification systems, index terms, metadata, display formats, etc. If the content of a web page, or a search engine result, etc., is not displayed in a lucid and understandable way, users conducting convergent or divergent information behavior may fail to see information details or pointers, which otherwise could have been used for renewed information behavior. Furthermore, it is easier to discover a useful navigable '*loophole'* in an information system when there is a *contrasting background* of lucid order. For instance, the topic hinted by the clickable anchor text of a transversal link may be interesting to follow on the contrasting background of the overall topic displayed in a web cluster.

Traditional classification systems and thesauri may provide options for navigation and browsing in an information system, as well as for refining and redefining requests. However, the claim of this dissertation is that a centrally constructed and controlled information system may be too rigid in order to provide the same up-to-date and flexible options as the complementarities of uniformity and diversity in the distributed knowledge organization of small-world web spaces.

Local order on the Web may emerge through both *coordinated* and spontaneous *uncoordinated* actions of web constructors. A web-based library OPAC using a classification system and thesaurus is an example of pre-coordinated local order on the Web. Furthermore, parts of a web cluster may be created by coordinated actions by web constructors, for instance, researchers participating in cross-institutional projects. Another example of pre-coordinated link structures are so-called *web rings* consisting of web sites linking to topic-related sites in pre-arranged ring-like structures for members in an interest community.[89] However, large parts of a topical web cluster may also be created when web constructors uncoordinatedly *'hook on'* and attach their own web territories by making links to existing topically related nodes in a cluster, perhaps unaware of other territories in the same cluster. Thus, the global distribution of links on the Web affecting the cohesiveness of link structures depends on local self-organizing, coordinated and uncoordinated activities reflecting a *'division of labor'* between millions of web constructors.

Furthermore, both the *construction* and *traversal* of link structures on the Web depend on millions of individuals' different limits of *cognitive capacity*. There are individual limits as to how many outlinks different web constructors are capable and interested to put into their web pages depending on purposes and desired levels of maintenance of the web pages. There are also individual limits as to how many outlinks different web users are capable and interested to embrace and follow on a web page (cf. Khan & Locatis, 1998). Thus, the activities of millions of link creators and link followers are manifested in self-organizing and *self-regulative* ways, as in an ecological system, affected by the abovementioned limited individual cognitive capacities. In other words, due to the distributed, diverse and dynamic nature of the Web, it is not possible

[89] Cf. http://www.webring.org

centrally to control the *optimal* number of links on web pages. Instead, the number of constructed and traversed links on the Web is brought about by the abovementioned distributed and self-regulative activities.

Our limited cognitive capacities combined with the unpredictable construction of crumpled-up web spaces imply that we can grasp only limited segments of the complex link structures. It is thus easy to experience information overload and to become disorientated and 'lost in space' unable to grasp the overall topological structure of the space (cf., e.g., Conklin, 1987; Darken & Sibert, 1996). However, a counter-argument against the notion of disorientation in hypertext being solely a negative thing, could be that getting lost may enable us to serendipitously discover new places, which we might not otherwise have stumbled across (Twidale *et al.*, 1997).

However, even if the Web – like other complex and self-organizing systems, for instance, the biosphere – cannot be controlled in details, it is worthwhile discussing how we can create regions and clusters of local order on the Web in the shape of subject gateways, resource guides, web portals, metadata, etc. In this context, the ambitions of librarians and others to create order on the Web should not be ridiculed, however unrealistic these ambitions may sometimes seem. Instead, ambitions of creating *global* order on the Web, e.g., through global metadata standards, may realistically and beneficially lead to improved *local* order, e.g., clusters of academic institutions implementing Dublin Core metadata.

Traditionally, information systems such as libraries, bibliographic databases, etc., have functioned as 'memory institutions' (cf. Hjørland, 2000). Even Vannevar Bush (1945) used this metaphor for his *Memex*-machine: 'memory extender'. On the Web, new memory institutions have emerged, for example, scientific preprint archives (cf. e.g., Harnad & Carr, 2000; Dalgaard, 2001; Zhang, 2001) and the Internet Archive (Kahle, 1997), accumulating and archiving information resources and thus supporting the *collective memory* of mankind.

However, due to its dynamic and adaptive nature, the Web may also support *collective intelligence* (cf. de Beer, 1996; Lévy, 1997; de Kerckhove, 1998; Heylighen, 1999), because of the easy, rapid ways to publish and adapt web resources as well as to create and follow hyperlink pointers to other resources. Furthermore, many link creators also are link followers - and vice versa. The Web may thus speed up the diffusion of knowledge both within and across communities of researchers, laymen, etc. In this context, Bollen (1995) states that "[t]he evolution of knowledge and ideas on the WWW takes place by the continuous process of changing and replacing of links between related nodes". Indeed, Bollen and other researchers like Mayer-Kress & Barczys (1995), Heylighen & Bollen (1996) and Heylighen (1999) conceive a *global brain* as an emergent structure from the world-wide computing networks. Another propagator of such a global brain is Abraham (1996), who used the term *webometry* (cf. Section 2.1) for measuring the complexity of the Web. He states that "[t]he explosive growth of the WWW may be viewed as the neurogenesis phase in the embryogenesis of a new planetary civilization" (ibid.). Interestingly, Manfred Kochen who was a pioneer in small-world theorizing (cf. Section 3.2) envisaged similar brain-like macro structures of world-encompassing information systems (Kochen, 1972; 1975b), drawing on ideas from H.G. Wells in the 1930s on a *World Brain* (Wells, 1938; Price, 1975; Garfield, 1975; Rayward, 1999).

Small-world phenomena – and thus complementarities of topic clusters and transversal links – may have an important role in the diffusion of knowledge on the Web. Traditionally, collective intelligence has been manifested in the collaborative activities known as scientific research, publishing and referencing, where scientists build on each other's research results. The easy explorative, publishing and referencing options on the Web may support collective intelligence for researchers as well as for other web actors. Further, if web resources also become more extensively archived – and dynamic link connectivity structures get mapped and visualized, for example, in search engine displays, the Web may be able to support *both* collective memory *and* collective intelligence in the future.

8 Conclusion

> *"'Would you tell me, please, which way I ought to go from here?'*
> *'That depends a good deal on where you want to get to,' said the Cat.*
> *'I don't much care where--' said Alice.*
> *'Then it doesn't matter which way you go,' said the Cat.*
> *'--so long as I get somewhere,' Alice added as an explanation.*
> *'Oh, you're sure to do that,' said the Cat, 'if you only walk long enough."*
>
> (Lewis Carroll, *Alice's Adventures in Wonderland,* 1865. Chapter VI)

The dissertation ends where it began, by quoting Alice and the Cheshire Cat; apropos finding one's way through a PhD project – walking long enough along transversal trails.

Much scientific progress depends on generating new trails of describing and conceptualizing phenomena. Research is thus characterized by the fact that you cannot know in advance exactly where you are heading when you begin the long walks into the new territory to be explored. Hence, at the beginning of this PhD project three years ago, this mastodon-sized dissertation was not anticipated, as also noted in Chapter 1. However, the scope of the overall objective, the four research questions, and not least the richness of aspects discovered in the 'corona' model of the UK academic subweb and in the case studies of the 10 path nets resulted in this behemoth. The rich material from the 'corona' model of the UK academic subweb and from the case studies of the 10 path nets thus enabled identification of phenomena and revealed interesting indicative aspects of the phenomena that could generate more general hypotheses, conceptualizations and perspectives as elaborated and discussed in the previous chapter.

The overall objective of the dissertation has been to yield a better understanding of what factors contribute to the formation of an interconnected topology of link structures across an academic web space, especially in the shape of small-world topologies in this space. In this context, the dissertation has been concerned with answering the four research questions as put forward in the dissertation:

1. *How cohesively interconnected are link structures in an academic web space?*
2. *In particular, to what extent can so-called small-world properties be identified in this web space?*
3. *If small-world link structures can be identified in this academic web space, which properties can be observed that contribute to such link structures?*
4. *Especially, what types of web links, web pages and web sites function as cross-topic connectors in small-world academic web spaces?*

The first three research questions regard more general aspects of interconnectivity and small-world properties in an academic web space leading up to the fourth and main

research question regarding types of links, pages and sites functioning as cross-topic connectors in small-world academic web spaces.

As stated earlier, the small and non-random sample of 10 path nets; the subsite level delimitation of the original data set from a national university system; as well as the temporal delimitation giving a 'frozen' snapshot picture that does not capture the dynamics of the investigated link structures: these delimitations imply that there are no generalizable answers to the above research questions for academic web spaces world-wide. However, the findings may be indicative and fruitful for future large-scale investigations. Moreover, the developed conceptual framework and methodologies may contribute to more fully appreciate and investigate the link structures and other webometric characteristics of the Web.

The dissertation thus may contribute to library and information science including webometrics by providing elements of a novel conceptual framework, as well as some methodologies, models, findings, and hypotheses that may be useful for understanding small-world properties in academic web spaces as listed in the following overview:

- **novel conceptual framework**
  - diagram *defining* webometrics in a bibliometric LIS framework
  - consistent *link terminology* and web *node diagrams*
  - *transversal* links connecting dissimilar topics
  - *genre connectivity* and *genre drift* in a *web of genres*
  - *crumpled-up web* with links as contracting *hooks*

- **novel methodologies**
  - five-step methodology for sampling, identification and characterization of small-world properties in an academic web space
    - web node diagrams for zooming into web node levels
    - detailed considerations, delimitations and techniques for circumventing inherent data problems and general web problems
  - confined and investigable *'mini small worlds'* in the shape of *path nets* by juxtaposition of pairs of dissimilar topical seed nodes
    - *path analysis* of followed shortest link paths in path nets
    - identifying *transversal* links connecting dissimilar topics
  - *genre connectivity analysis* using genre typology divided into personal and institutional genres with consistent prioritized genre categorization
  - Internet Archive used as *'web archaeological'* tool
    - to validate domain names
    - to indicate ages of web graph components
    - to classify *subsite* topics and genres of 'old' web data
    - to classify *page* topics and genres of 'old' web data

- **novel findings**
  - UK academic web contains *small-world* properties
    - high clustering coefficient and low characteristic path length
    - indicative relation hubs/authorities and betweenness centrality

        - power-law-like distributions: in-neighbors/out-neighbors, and inlinks/outlinks
        - supports notion of a fractal 'self-similar' Web
    - detailed '*corona'*-model of graph components in UK academic subweb, including inter-component and intra-component link structures
        - links direct from OUT to IN components (not in 'bow-tie'-model)
        - indicative ages of graph components based on Internet Archive
    - *web of genres* with *genre drift*
        - rich diversity of genre pairs
        - outlink-prone *'hook genres'* and inlink-prone *'lug genres'*
    - indicative finding of links, page genres and subsites functioning as small-world providers in an academic web space
        - computer science subsites may be important cross-topic connectors
        - personal link creators may be important connectors across sites and topics in academic web space
        - personal link lists provide about 40% of transversal outlinks in 10 path nets
        - over 80% of transversal links in 10 path nets are academic

- **novel hypotheses**
    - *small-world knowledge organization*
        - may comprise *complementarities*: topical uniformity + diversity → clustering + cross-clustering → small world
        - topic clusters with genre diversity → intra-cluster genre drift
        - genres with topical diversity → inter-cluster topic drift
        - intra-cluster *genre drift* + inter-cluster *topic drift* → small world
    - small-world knowledge organization
        - may comprise *'collaborative weaving'* by many link constructors
        - non-engineered *organizing principle* to handle vast document space by enabling short local *and* short global link distances in document space
    - small-world knowledge organization → *exploratory capabilities*
        - exploring information space with short local and global distances
        - short distances between topics → *possibility space* for serendipity and computer-supported knowledge discovery
        - small-world → topical uniformity + diversity → → *convergent* (goal-directed) searching + *divergent* (serendipitous) browsing

The next two subsections give a more detailed overview of these contributions with respect to conceptualizations and methodologies used to answer the four research questions, as well as the indicative answers found to the questions.

8.1 Research questions 1 and 2

1. *How cohesively interconnected are link structures in an academic web space?*
2. *In particular, to what extent can so-called small-world properties be identified in this web space?*

In order to answer these two more overall research questions and handle the different types of link structures and web node levels in the investigated web space it was necessary first to develop a consistent *conceptual framework* concerned with the following aspects (brackets show sections concerned):

- *Basic link terminology* with, for example, inlinks, outlinks, selflinks, reciprocal links, transversal links, link paths, co-inlinks, and co-outlinks (Section 2.3.1);
- *Basic web node diagrams* for illustrating and handling different web node levels denoted with simple geometrical figures: *quadrangles* (web pages), *diagonal lines* (web directories), *circles* (web sites) and *triangles* (top level domains, TLDs) (Section 2.3.2);
- *Advanced link terminology and diagrams* for webometric studies of different web node levels employing *micro*, *meso* and *macro level* perspectives (Section 2.3.3).

The developed link terminology and node diagrams have been supportive with regard to describing and illustrating different structural aspects in the different 'zooming-in' steps of analysis in the dissertation; both at the *macro* level concerned with the 'corona'-model of the UK web graph, the *meso* level of path nets among subsites, and at the *micro* level of closely examined source and target pages and links.

Subsequently, a range of *methods* were developed and measures applied regarding the concerned research questions:

- Detailed *methodological considerations*, *delimitations and techniques* for handling and circumventing inherent data problems with the given data set (computational limits, typos, etc.) and with general web problems (elusiveness of web pages and sites, etc.) (Chapters 4, 5, and 6).
- Detailed *'corona'* graph model depicting actual inter-component and intra-component adjacencies in a web graph (Section 5.1);
- Wide range of *graph measures* (including characteristic path lengths; clustering coefficients; distributions of in-neighbors/out-neighbors, inlinks/outlinks, and in-distance/out-distance; assortative mixing; betweenness centrality; cores; hubs & authorities; co-linkage) were applied to investigate the *macro-level* (UK academic web) as well as *medium-level* (10 path nets) connectivity patterns in the investigated web space (Chapter 5 & Section 6.3.2).

The wide range of graph measures were especially applied to get a more complete picture of how cohesively interconnected are link structures in the UK web graph, as

addressed in the first research question. The application of these graph measures led to the following *findings* regarding research questions 1 and 2 concerned with graph theoretic connectivity aspects of the investigated web graph. Major findings are listed first:

- The characteristic path length and clustering coefficient of the investigated UK academic web meet the requirements for a *small-world network* (Watts & Strogatz, 1998), thus answering research question 2. The characteristic path length was 3.5 and the diameter (maximum link distance) was 10 between reachable subsites (Section 5.3);
- *Sparse link connectivity* in the investigated academic web space. An average outlinking university web page in the UK data set had 11.6 outlinks comprising 10.1 *site selflinks* and 1.5 *site outlinks*. Not surprisingly, most links thus point to other pages *within* the same site. Of the site outlinks to all the Web, only 7.7% were targeted to the other 108 universities and their subsites in the *undelimited* data set (including links to and from university main sites). The delimited data set (only site outlinks between *subsites* at different universities) comprised 3.1% of all site outlinks at the 109 universities. The vast majority of outlinks in the study thus were targeted to academic, commercial, and other targets *outside* the data set. (Section 5.4.2);
- Detailed *'corona'* graph model depicting actual inter-component and intra-component adjacencies in UK academic subweb graph, including frequent links direct from the IN to the OUT component not shown in the traditional 'bow-tie'-model. (Section 5.1);
- *Indicative ages of the graph components* in an academic web graph. Average ages of subsites as indicated by first indexed dates in the Internet Archive showed, for example, that the OUT component in the investigated UK academic web graph contained the oldest subsites, the IN component the youngest, and the Strongest Connected Component (SCC) subsites were on average slightly younger than the OUT subsites. (Section 5.2);
- *Power-law-like* distributions of in-neighbors/out-neighbors and inlinks/outlinks in the UK academic subweb as well as within the 10 path nets. This finding is in line with the concept of a *fractal 'self-similar' Web* (Dill *et al.*, 2001; Kumar *et al.*, 2002) with subsets of the Web displaying the same graph properties as the Web at large (Sections 5.4.3 & 6.3.3.2);
- *Power-law-like* distribution of *betweenness centrality* in the investigated academic web space. (Section 6.3.2.4);
- Indication of close relation between Kleinberg's (1999a) concepts of *hubs and authorities* on the Web and *betweenness centrality*. No literature has been found discussing such a relation. (Section 6.3.2.4);
- Low correlation measure indicates a *lack of 'assortative mixing'*: web nodes with high connectivity degrees (many in-neighbors and out-neighbors) do *not* tend to connect to other nodes with many connections. This finding yields indicative support to Newman's (2002) finding regarding the lack of assortative mixing in networks on the Web (Section 6.3.2.3).

8.2 Research questions 3 and 4

3. *If small-world link structures can be identified in this academic web space, which properties can be observed that contribute to such link structures?*
4. *Especially, what types of web links, web pages and web sites function as cross-topic connectors in small-world academic web spaces?*

These two research questions concerned with more specific small-world properties of the investigated academic web space were answered using the following *methods* and *conceptualizations* (brackets show sections concerned):

- *Five-step methodology* for sampling, identification and characterization of small-world properties in an academic web space, especially, identifying what types of links, web pages and web sites provide transversal shortcuts across dissimilar topical domains in an academic web space (Chapter 6);
- Focus on academic *subsites* as carriers of presumed topically focused contents. Exclusion of university main web sites because of multi-topicality. This focus enabled more clear-cut sampling and identification of transversal link structures. (Section 4.2.1);
- *'Web archaeological'* use of the *Internet Archive* for the retrieval, verification, and characterization of older links, web pages and web sites that may have changed or disappeared from the dynamic Web. Even if the Internet Archive does not cover the entire Web, a remarkably high percentage – over 90% – of the investigated UK academic subsites (minus 53 obvious domain name typos) were indexed in the Archive. Furthermore, 99% of the SCC subsites were indexed in the Archive. In the 10 path nets, 84% of the visited source pages and 88% of the visited target pages were available in the Archive. The Internet Archive is thus an excellent *web archaeological tool* – at least for investigating the UK academic web. (Sections 4.2.4, 5.2, 6.2, & 6.4.3);
- Novel concept of so-called *transversal links* as cross-topic shortcuts between dissimilar topical web domains (Sections 2.3.1 & 6.5);
- Novel metaphor of a *'crumpled-up'* Web with links as *hooks* reshaping and 'crumpling-up' web spaces by pulling web pages, sites and neighborhoods close together (Section 7.2.4). This metaphor has been helpful for conceiving the intricacy of the investigated topologies.

The five-step methodology provided a stepwise 'zooming-in' into more and more fine-grained web node levels in the investigated UK academic web space; starting with the *macro* level 'corona'-model, followed by the *meso* level path nets of subsites, and ending with the closely examined *micro* level of pages and links. An important part of the five-step methodology was concerned with *path nets* in order to construct investigable small-world link structures – *'mini small worlds'*:

- Deliberate *juxtaposition* of pairs of topically dissimilar web nodes (intentionally induced *topic drift*) in order to construct confined and thus investigable small-world link structures in the shape of *path nets* comprising subgraphs of all

shortest link paths between the juxtaposed web nodes in an academic web space. (Section 6.3);
- Elaborated *diagrams* and *terminology* for *path nets* in the shape of all shortest link paths between single pairs of web nodes (Section 6.3);
- Detailed *path analysis* of subsites, web pages and links along followed shortest link paths in the case studies of 10 path nets (Sections 6.3, 6.4, & 6.5).

This method of constructing investigable 'mini small worlds' in the shape of path nets connecting topically dissimilar subsites provided interesting indicative findings as listed further below. Another essential part of the five-step methodology dealt with enabling the identification of what kind of page genres provided transversal links in the path nets. For this purpose, especially the following conceptual and methodological aspects were important:

- Novel approach of conceiving the Web as a *'web of genres'*, that is, as a web of interconnected web page genres, including outlink-prone *'hook genres'* and inlink-prone *'lug genres'* (Sections 6.4.5.4 & 7.2.4);
- Novel approach of including so-called *genre drift* together with *topic drift* for explaining small-world phenomena on the Web (Sections 6.4.5.4 & 7.2.3).
- Consistent *page genre typology* of personal and institutional web page genres for investigating the *genre connectivity* in an academic web space (Section 6.4.5);
- Rules of *prioritized genre categorization order* in order to use developed genre typology to classify page genres in a consistent way (Section 6.4.5);

These concepts of *genre connectivity* and *genre drift* are perhaps some of the most fruitful outcomes of the dissertation as useful elements in a LIS approach for explaining the emergence of small-world link structures, as hypothesized in Section 7.2.3. As noted in Section 6.4.5, the term *genre* is used in the dissertation in a broad sense in accordance with contemporary web terminology for describing types of web pages as well as types of aggregated web pages, for example, in the shape of web sites).

The following *indicative findings* were made, especially, regarding research question 4 concerned with what types of web links, web pages and web sites function as cross-topic connectors in an academic web space. The first set of findings deals with *personal link creators* such as researchers and students as important connectors across sites and topics in an academic web space. Personal web pages thus provide about 53% of all followed site outlinks and 35% of site inlinks in the 10 path nets. Further, personal web pages provide about 57% of transversal outlinks and 42% of transversal inlinks in the 10 path nets. Personal link lists was the largest cross-topic page genre providing about 40% of transversal outlinks in the 10 path nets. Over 80% of the identified transversal links in the 10 path nets were related to academic activities like research or teaching:

- Institutional and personal page genres made up about 50% each of the visited source pages in the 10 path nets. This finding may indicate a relatively large influence of *personal* web pages for providing site outlinks in an academic web

space. About 60% of the visited target pages belonged to institutional page genres, whereas 40% belonged to personal ones. This result also may indicate the relatively large influence by personal web pages for receiving site inlinks in an academic web space. (Section 6.4.5.1);

- There were more followed *outlinks* (53%) from *personal* source pages, than from institutional source pages in the 10 path nets, perhaps reflecting more active link creations of personal web creators. On the other hand, there were more followed *inlinks* (64%) to *institutional* target pages, than to personal target pages, perhaps reflecting more relevant and authoritative contents of institutional pages. (Section 6.4.5.3);
- Personal web pages provide about 53% (as noted above) of *all* followed site outlinks and 36% of site inlinks in the 10 path nets, whereas personal web pages provide about 57% of identified *transversal outlinks* and 42% of *transversal inlinks* in the 10 path nets. This yields yet an indication that personal link creators may be important connectors across sites and topics in academic web spaces (Sections 6.4.5 and 6.5.5);
- There is a higher percentage of *personal link lists* among the *transversal links* (40%) and transversal source pages (36%) than among all the followed links (32%) and visited source pages (30%). This finding may indicate a special impact of personal link lists for the emergence of *small-world* phenomena across dissimilar topical web domains (Section 6.5.5.3).

The next set of important indicative findings deals with *computer science subsites* as important cross-topic connectors in an academic web space:

- About 31% of *all* visited subsites in the 10 path nets were *computer-science*-related (CS). However, about 46% of subsites providing or receiving *transversal* links were CS-related (cf. Section 6.5.4). Counting *links* instead of *subsites*, about 38% of all followed outlinks in the 10 path nets originated from CS-related *subsites*. The percentage of followed inlinks to CS-related subsites were slightly smaller (36%). Of the transversal links, 41% originated from CS-related subsites, whereas 40% were received by CS-related subsites. In the random sample of 189 SCC subsites (cf. Section 6.2.1), about 11% were judged as CS-related. Computer science thus constitutes a larger share among the visited subsites in the 10 path nets and an even larger share among the subsites connected by transversal links. Even if the sample of 10 path nets was small, this finding may indicate a special role of CS-related subsites as cross-topic connectors on shortest link paths in an academic web space. (Section 6.5.4);

The role of CS-related subsites in academic link structures most likely reflects the auxiliary function of computer science in many scientific disciplines in natural sciences, technology, humanities, and social sciences. This auxiliary function may be combined with a more well-developed and unconstrained web presence and more experienced web literate behavior performed by CS-related persons and institutions. Future qualitative studies of academic link creation practices may reveal more details on this issue.

The last set of indicative findings is concerned with more general aspects of *page genres* in the investigated path nets:

- *Rich diversity of genre pairs*. Links between the investigated academic subsites in the 10 path nets connect many different combinations of source and target page genres. This may reflect a corresponding diversity of link motivations (Section 6.4.5);
- Personal link lists provided almost a third (32%) of all followed site outlinks in the 10 path nets and the institutional link lists about a quarter (25%). Such genres are thus *'site outlink-prone' ('hook genres'*). Institutional homepages and personal homepages received most followed site inlinks in the path nets, 23% and 13%, respectively. Such genres are correspondingly *'site inlink-prone' ('lug genres')*. (Section 6.4.5).

As stated above, the listed indicative findings are especially concerned with research question 4 regarding what types of web links, web pages and web sites function as cross-topic connectors in an academic web space. Research question 3 was concerned with a more abstract level concerned with which properties may contribute to small-world link structures in academic web spaces. Especially the discussion in Section 7.2 dealt with such aspects, for instance, regarding the hypothesized complementarities of topical uniformity and diversity for the formation of small-world webs, including factors like topic drift and genre drift. The summary in Section 7.4 gives an overview of this discussion.

8.3 Final remarks

> "*Whether we are building a central monopoly-controlled consensus information system versus a decentralized, pluralistic, free-access communication system makes all the difference in the world."* (Parker, 1975, p. 22)

> *"... a vision encompassing the decentralised, organic growth of ideas, technology, and society. The vision I have for the Web is about anything being potentially connected with anything. It is a vision that provides us with new freedom, and allows us to grow faster than we ever could when we were fettered by the hierarchical classification systems into which we bound ourselves."* [...]
> *"The ultimate goal of the Web is to support and improve our weblike existence in the world."* (Berners-Lee, 1999, pp. 1 & 133)

The dissertation has aimed to provide a better understanding – from a library and information science perspective – of elements affecting the construction and use of a self-organizing information system such as the Web. In spite of many attempts to control and restrict free access to information and freedom of expression on the Web (cf. e.g., Hamilton, 2002), the Web as conceived by Berners-Lee above has to a high degree become that *"decentralized, pluralistic, free-access communication system"* as

envisioned by Parker (1975) above in his article on 'Who should control society's information resources?' The article was symptomatically published in an anthology edited by Manfred Kochen (1975a) who was central in small-world theorizing (cf. Section 3.2).

The dissertation has focused on small-world properties in academic web spaces. As outlined in the previous sections, the dissertation has contributed to library and information science including webometrics with regard to novel conceptualizations, methodologies, findings, models and hypotheses for understanding small-world academic web spaces.

However, one could ask to what extent was the small-world idea a 'useful' one to explore the Web, or the best one? In retrospect, one could thus ask if it would have been more directly useful or easier to try another approach to get the kind of results found in the dissertation. As stated earlier in Section 7.2.2, clustering techniques have not been applied in the dissertation for various reasons. Indeed, it would have been interesting to apply clustering techniques on the investigated academic web space, for instance, based on frequencies of direct inter-site links, co-inlinks and/or co-outlinks (cf., e.g., Thelwall & Wilkinson, forthcoming). However, a big problem with such an approach would be that link data alone would not be sufficient in order to filter out topic clusters because of the inherent topic drift in link structures. Topical data either identified manually and time-consumingly (as in the present study) or automatically, would be necessary for identifying topic clusters, for instance, by including metadata and content words. It is not possible to say whether such approaches of clustering web pages by topic or genre and then looking for cross-topic or cross-genre links would have yielded more interesting or 'useful' findings than the present study. Future studies are needed to test this issue.

Nevertheless, the approach undertaken in the present study by the five-step methodology for sampling, identification and characterization of small-world properties in an academic web space, especially, deliberately juxtaposing pairs of topically dissimilar web nodes in order to construct investigable *'mini small worlds'* in the shape of path nets, and subsequently 'zooming' into subsites and page genres providing transversal links; this approach has been fruitful as demonstrated by the earlier listed findings. The steps of the primarily manually executed data extractions and categorizations were advantageous in one important respect. The time-consuming pace of steps enabled detection of dimensions of web pages and links perhaps otherwise neglected. The close manual inspection of the web pages and links revealed interesting findings, for instance, the discovered genre connectivity.

In addition to the future studies of topic clusters in academic web spaces suggested above, other future studies are also needed with respect to whether small-world web spaces actually have a likely impact on users' browsing behavior; or is it unlikely that users will find or follow transversal links on small-world link paths? Some of these aspects were discussed in Section 7.3.1 with respect to the lack of straightforward causal relations between small-world knowledge organization and easy exploratory capabilities.

One important outcome of this dissertation has been a realization that the traditional LIS approach with its focus on system control, effective IR algorithms, rational search behavior, explicit user needs, etc., may no longer be sufficient to cope

with issues necessitated by the rise of the Internet and the Web, that are information systems *without* centralized control, as argued in Section 7.5. Thus, there is a need for redefining the overall aim and explanatory framework of LIS research, so it encompasses *both* 'top-down'-constructed information systems (e.g., traditional libraries and bibliographic databases) *and* 'bottom-up'-constructed distributed information systems such as the Web, in order to cope with issues concerned with distributed knowledge organization, small-world phenomena, topical diversity, genre connectivity, serendipity, knowledge discovery and creativity stimulation as discussed in the dissertation.

As discussed in Section 7.3.1, small-world phenomena in the shape of short link paths across the Web may affect exploratory capabilities including better possibilities for serendipity, knowledge diffusion and creativity stimulation. Thus, the unpredictability of link structures on the Web may be an advantage if the Web is used as a *possibility space* for cross-topic exploration and discoveries. However, the unpredictability of crumpled-up web spaces also implies that users can grasp only limited segments of the complex link structures. As noted in Section 7.5, this may result in disorientation and information overload. Fortunately, searchers and surfers on the Web can be helped by local and global projects aiming at making the Web more ordered and navigable, for example, quality-assessed resource guides, topic-specific web portals, semantic web and metadata standards. Thus, information specialists, librarians and other *web territory organizers* – including researchers and students as revealed in the close examination of the visited web pages in the 10 path nets – have an important role as *distributed knowledge organizers*, i.e., as providers of local order and navigation aids in a global self-organizing small-world Web.

However, the Web, like other complex self-organizing systems such as the biosphere, cannot be controlled in details. Such control may not even be desirable, since we risk ruining small-world *'loopholes'* and other intricate but beneficial structures facilitating exploratory capabilities on the Web. In the future, navigation tools and visualization tools of search engines and browsers will hopefully get more sophisticated, for instance, with zoomable maps of link structures showing optional convergent and divergent web page neighborhoods, etc. Whether we are link creators or link followers, we may thus develop better skills in exploiting the complementarities between topical uniformity and diversity on the Web.

As noted earlier, the prefix *hyper* in hypertext derives from a Greek term meaning *over, beyond, transcendent*; thus very appropriate in this context of small-world *transversal* interlinkages on the Web. As argued in the dissertation, the self-organized small-world architecture of large regions on the Web may be considered as an important non-engineered organizing principle for structuring a vast information space by enabling both short local and global link distances. Webometric approaches may reveal characteristics of how such complex bottom-up-aggregated web spaces provide traversal options and access points to information. In this context, *academic* web spaces are of special interest, because science may be considered as a complex, largely self-organizing socio-cognitive system (e.g., van Raan, 2000; Leydesdorff, 2001; Sandstrom, 2001). The Web with its capability to host self-organizing activities in the shape of 'collaborative weaving' may thus be seen as a natural environment and playing ground for the self-organizing system of scholars.

In addition to such applications of small-world approaches, the findings and hypotheses in the dissertation may give rise to other possible applications for utilizing and exploiting small-world web spaces as outlined in the following list:

- *Sociology of science*: As the Web includes more and more informal self-presentations and link creations by scholars, the sociology of science may employ small-world approaches including measures of betweenness centrality for automatic detection of 'invisible colleges' and central 'gatekeepers' across link structures in academic web spaces reflecting networked knowledge creation and diffusion, and providing options for so-called *social navigation* strategies in the abovementioned self-organizing system of scholars (cf. Section 7.1.2).
- *Web mining*: Computer-supported knowledge discovery on the Web, so-called *web mining*, could include small-world approaches for identifying fertile areas for cross-disciplinary exploration. Transversal links may give hints to identify *'undiscovered public knowledge'* (cf. Swanson, 1986; and Section 3.4).
- *Social network analysis*: On the Web at large, including both academic and non-academic web spaces, small-world link structures connecting apparently dissimilar topics may reflect emerging cultural and social formations across web communities, including cross-disciplinary contacts in science. Such formations may be examined in social network analysis.
- *Library and information science*: Analyzing small-world web spaces may help library and information science understand how distributed knowledge organization by a multitude of link creators ('collaborative weaving') contribute to the formation of topic clusters, transversal links, and thus small-world document networks on the Web. For example, LIS may utilize small-world approaches for facilitating exploratory capabilities across topics in *digital libraries*, making them as serendipity-prone as physical libraries are for people browsing shelves.
- *Search engines*: Small-world approaches may be exploited for more *exhaustive* web crawls, for instance, by using personal link lists (especially bookmark lists, easy identifiable by the high link density) or computer-science sites as starting points for the search engine web crawlers allowing traversals along transversal links in order to reach a wider range of topical areas. If identified (perhaps by techniques based on the present study), transversal links may also be *avoided* allowing more topically *focused* web crawls (cf. Sections 7.1.1 and 7.3.1).
- *Browsers*: Mapping and visualization of small-world web neighborhoods in browser windows may stimulate serendipitous browsing across topics by showing serendipity-prone 'jump spots' for human web surfers (cf. Section 7.5).

The dissertation has amalgamated webometrics and small-world theory originally deriving from social network analysis. As noted earlier in Section 3.2, innovative information scientists and bibliometric pioneers like Manfred Kochen and Eugene Garfield early on envisaged potentials of small-world theory for bibliometrics and scientometrics, for example, for identifying gatekeepers and 'invisible colleges' in informal scholarly communication networks. Thus, the ring now closes as small-world approaches again enter the domains of bibliometrics and scientometrics in the shape of

webometric studies of small-world link structures in academic web spaces as in this dissertation.

The dissertation has defined webometrics in a bibliometric and LIS framework. This framework as well as the link terminology and diagram notation proposals should be seen as conceptual foundations and building blocks by which future discoveries and perspectives of the Web and webometrics hopefully will thrive. The distinction between web node levels, its terminological impact, and the proposal of a consistent diagram notation is necessary if the intricate structures of the Web shall be understood and analyzed. There exists a constant possibility of loosing the point of perspective in such analysis, in particular if terminological rigor is lacking.

As stated earlier in connection with the diagram of bibliometrics embracing webometrics (Fig. 2.1 in Section 2.2), the inclusion of webometrics expands the field of bibliometrics, as webometrics inevitably will contribute with further methodological developments of web-specific approaches. As ideas rooted in bibliometrics, scientometrics and informetrics contributed to the emergence of webometrics, ideas in webometrics might now contribute to the development of these embracing fields.

Traditional bibliometrics and scientometrics have had many years to develop quite precise methodologies for data extraction and analysis. Contrary to this, webometrics is a new research field now passing through a necessary tentative and exploratory phase. Webometrics is thus still in its infancy, and, furthermore, must handle data of a much more messy, non-standardized, diverse and dynamic nature (as demonstrated in the empirical chapters in this dissertation) than traditional bibliographic data used in bibliometrics and scientometrics.

Furthermore, the Web has only existed well over 10 years and its complex topologies, functionalities and potentials are far from fully explored or exploited. We are thus only on the threshold of understanding the complexities of the Web. Moreover, the Web itself, including its uses and technologies, is continuously changing and developing, and will thus keep up providing new intriguing challenges and opportunities for webometrics.

The Web is thus still at its inception as a *possibility space*, the potentials of which for human information sharing and innovation still wait for exploration and development. In this connection, library and information science has an important research role to play, to which end the dissertation hopefully can contribute.

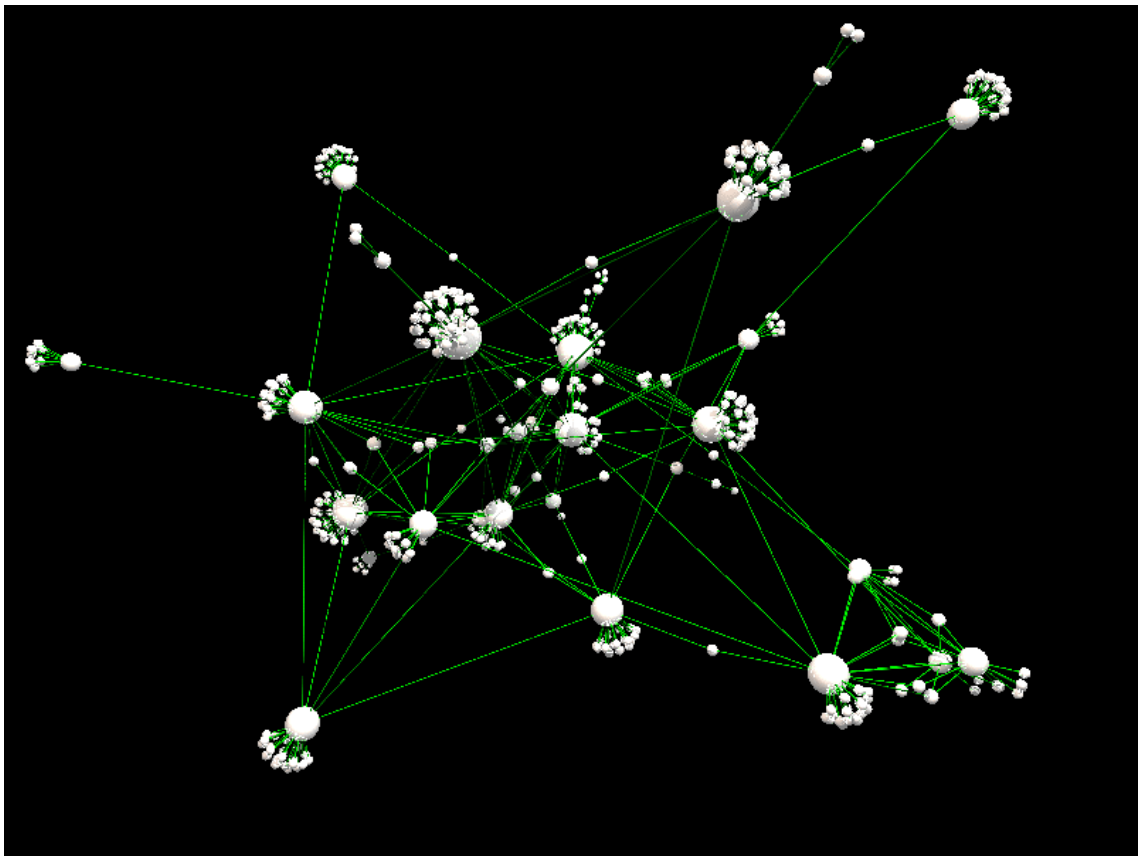
The dissertation has dealt with the rich diversity of human activity and creativity as reflected and manifested in the distributed and self-organizing knowledge organization of the Web, especially, in the small-world link structures of an academic web space. In this context, the dissertation has contributed to a better understanding of how the topology of the Web behaves like a living organism *threading trails* across topical domains and genres. The hypothesized exploratory capabilities across such a crumpled-up and short-distanced web organism, that constantly grow and adapt by including new topics and links, may be described by paraphrasing the famous five laws of library science by S.R. Ranganathan (1931)[90] in the following five 'laws' for *web connectivity*:

[90] Ranganathan (1931): *The five laws of library science*: "Books are for use. Every reader his or her book. Every book its reader. Save the time of the reader. The Library is a growing organism."

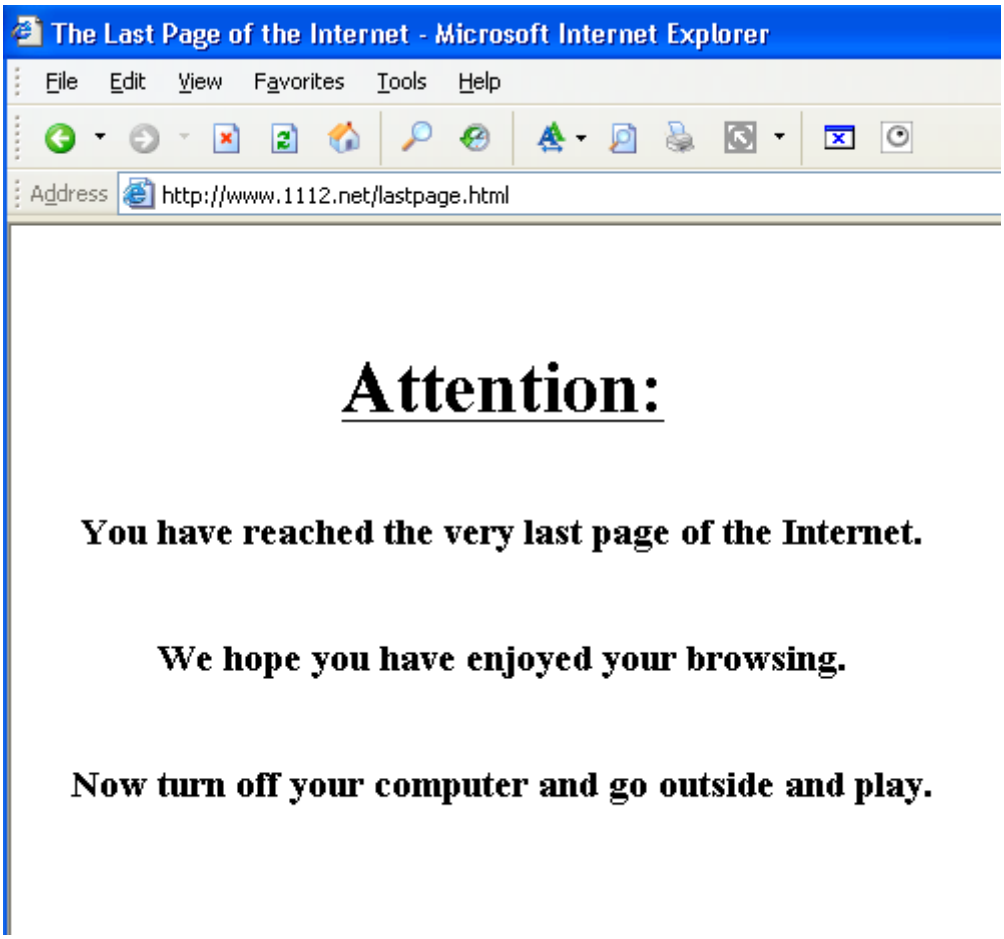
- *Links are for use* – the very essence of hypertext;
- *Every surfer his or her link* – the rich diversity of links across topics and genres;
- *Every link its surfer* – ditto;
- *Save the time of the surfer* – visualizing web clusters and small-world shortcuts;
- *The Web is a growing organism*.

Postlude

"Connecto ergo sum." (Björneborn, 1998)



(Wood *et al.,* 1995)


The Last Page of the Internet - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://www.1112.net/lastpage.html
Attention:
You have reached the very last page of the Internet.
We hope you have enjoyed your browsing.
Now turn off your computer and go outside and play.

References

All URLs are validated 30.10.2003.

Aarseth, Espen (1997). *Cybertext : perspectives on ergodic literature*. Baltimore: The Johns Hopkins University Press.

Abbate, Janet (1999). *Inventing the Internet*. Cambridge, Mass.: The MIT Press.

Abraham, Ralph H. (1996). 'Webometry: measuring the complexity of the World Wide Web'. Visual Math Institute, University of California at Santa Cruz. Available: http://www.ralph-abraham.org/vita/redwood/vienna.html

Abrams, David; **Baecker**, Ron & **Chignell**, Mark (1998). 'Information archiving with bookmarks: personal web space construction and organization'. *Proceedings of Human Factors in Computing Systems, CHI 98*. ACM Press. pp. 41-48.

Adamic, Lada A. (1999). 'The small world web'. In: Abitetoul, Serge & Vercoustre, Anne-Marie (eds.). *Proceedings of the 3rd European Conference on Digital Libraries, Paris.* Berlin: Springer. pp. 443-452. (Lecture Notes in Computer Science; 1696).

Adamic, Lada A. & **Adar**, Eytan (2003). 'Friends and neighbors on the Web'. *Social Networks*, 25(3): 211-230.

Adamic, Lada A. & **Huberman**, Bernardo A. (2000). 'Power-law distribution of the World Wide Web'. *Science*, 287: 2115a

Adamic, Lada A. & **Huberman**, Bernardo A. (2001). 'The Web's hidden order'. *Communications of the ACM,* 44(9): 55-59.

Adamic, Lada A.; **Lukose**, Rajan M.; **Puniyani**, Amit R. & **Huberman**, Bernardo A. (2001). 'Search in power-law networks'. *Physical Review E*, 64: 46135.

Agatucci, Cora (2001). 'Cyber rhetoric (3): web genres & purposes'. Central Oregon Community College. Available: http://www.cocc.edu/hum299/lessons/rhet3.html

Aguillo, Isidro F. (2002). 'Cybermetrics : definitions and methods for an emerging discipline'. *Séminaires de l'ADEST*, Paris, 14 February, 2002. Available: http://www.upmf-grenoble.fr/adest/seminaires/ISIDRO/Cybermetrics.ppt

Ahmed, E. & **Abdusalam**, H.A. (2000). 'On social percolation and small world network'. *The European Physical Journal B*, 16(3): 569-571.

Albert, Réka & **Barabási**, Albert-László (2002). 'Statistical mechanics of complex networks'. *Reviews of Modern Physics*, 74(1): 47-97.

Albert, Réka; **Jeong**, Hawoong & **Barabási**, Albert-László (1999). 'Diameter of the World-Wide Web'. *Nature*, 401(Sept. 9): 130-131.

Albert, Réka; **Jeong**, Hawoong & **Barabási**, Albert-László (2000). 'Error and attack tolerance of complex networks'. *Nature*, 406(July 27): 378-381.

Almind, Tomas C. & **Ingwersen**, Peter (1996). 'Informetric analysis on the World Wide Web: A methodological approach to 'internetometrics''. Centre for Informetric Studies, Royal School of Library and Information Science. (CIS Report 2).

Almind, Tomas C. & **Ingwersen**, Peter (1997). 'Informetric analyses on the World Wide Web: methodological approaches to 'webometrics''. *Journal of Documentation*, 53(4): 404-426.

Amaral, L.A.N.; **Scala**, A; **Barthélémy**, M. & **Stanley**, H.E. (2000). 'Classes of small-world networks'. *Proceedings of the National Academy of Sciences*, Oct. 10, 2000, 97(21): 11149-11152.

Amitay, Einat (2001). *What lays in the layout: using anchor-paragraph arrangements to extract descriptions of Web documents*. PhD Thesis. Macquarie University, Australia. Available: http://www.ics.mq.edu.au/~einat/thesis/final.pdf

Andersen, Jack (2002). 'The concept of genre: when, how, and why?' *Knowledge Organization*, 28(4): 203-204.

Andersen, Svein S. (1997). *Case-studier og generalisering : forskningsstrategi og design.* Bergen-Sandviken: Fagbokforlaget.

Baeza-Yates, Ricardo & **Castillo**, Carlos (2001). 'Relating web characteristics with link based web page ranking'. *Proceedings of SPIRE 2001*. Laguna San Rafael, Chile: IEEE CS Press. pp. 21-32.

Bagnoli, Franco & **Bezzi**, Michele (2001). 'Small world effects in evolution'. Available: http://arxiv.org/PS_cache/cond-mat/pdf/0007/0007458.pdf

Barabási, Albert-László (2001). 'The physics of the Web'. *Physics World*, July 2001. Available: http://physicsweb.org/article/world/14/7/9

Barabási, Albert-László (2002). *Linked : the new science of networks*. Cambridge, Mass.: Perseus Publishing.

Barabási, Albert-László & **Albert**, Réka (1999). 'Emergence of scaling in random networks'. *Science*, 286(5439, Oct 15): 509-512.

Barabási, Albert-László; **Albert**, Réka & **Jeong**, Hawoong (1999). 'Mean-field theory for scale-free random networks'. *Physica A*, 272: 173-187.

Barabási, Albert-László; **Albert**, Réka & **Jeong**, Hawoong (2000). 'Scale-free characteristics of random networks: the topology of the world-wide web'. *Physica A*, 281: 69-77.

Barabási, Albert-László; **Jeong**, Hawoong; **Neda**, Zoltan; **Ravasz**, Erzsebet; **Schubert**, Andras & **Vicsek**, Tamás (2002). 'Evolution of the social network of scientific collaborations'. *Physica A*, 311(3-4): 590-614.

Bar-Ilan, Judit (1997). 'The "Mad Cow disease", usenet newsgroups and bibliometric laws'. *Scientometrics*, 39(1): 29-55.

Bar-Ilan, Judit (1998). 'The mathematician, Paul Erdos (1913-1996) in the eyes of the Internet'. *Scientometrics*, 43(2): 257-267.

Bar-Ilan, Judit (2001). 'Data collection methods on the Web for informetric purposes : a review and analysis'. *Scientometrics*, 50(1): 7-32.

Bar-Ilan, Judit (2002). 'Methods for measuring search engine performance over time'. *Journal of the American Society for Information Science and Technology*, 53(4): 308-319.

Bar-Ilan, Judit (forthcoming). 'The use of Web search engines in information science research'. *Annual Review of Information Science and Technology*, 38.

Bar-Ilan, Judit & **Peritz**, Bluma C. (2002). 'Informetric theories and methods for exploring the Internet: an analytical survey of recent research literature'. *Library Trends*, 50(3): 371-392.

Barrat, A. & **Weigt**, M. (2000). 'On the properties of small-world network models'. *The European Physical Journal* B, 13(3): 547-560.

Batagelj, Vladimir & **Mrvar**, Andrej (2000). 'Some analyses of Erdös collaboration graph'. *Social Networks*, 22: 173-186.
Preprint available: http://vlado.fmf.uni-lj.si/pub/networks/doc/erdos/erdos.pdf

Bates, Marcia J. (1986). 'An exploratory paradigm for online information retrieval'. In: Brookes, B.C. (ed.). *Intelligent information systems for the Information Society. Proceedings of the 6th International Research Forum in Information Science (IRFIS 6), Frascati, Italy, September 16-18, 1985.* Amsterdam: North-Holland. pp. 91-99.

Bates, Marcia J. (1989). 'The design of browsing and berrypicking techniques for the online search interface'. *Online Review,* 13(5): 407-424.

Bates, Marcia J. & **Lu**, Shaojun (1997). 'An exploratory profile of personal home pages: content, design, metaphors'. *Online & CDROM Review*, 21(6): 331-340.

Bavelas, Alex (1948). 'A mathematical model for group structure'. *Applied Anthropology*, 7: 16-30.

Bawden, David (1986). 'Information systems and the stimulation of creativity'. *Journal of Information Science,* 12: 203-216.

Belkin, N.J. (1978). 'Information concepts for information science'. *Journal of Documentation*, 34(1): 55-85.

Beniger, James R. (1986). *The control revolution : technological and economic origins of the information society.* Cambridge, Mass.: Harvard University Press.

Bergman, Michael K. (2001). 'The Deep Web: Surfacing hidden value'. *Journal of Electronic Publishing*, 7(1). Available: http://www.press.umich.edu/jep/07-01/bergman.html

Berners-Lee, Tim (1989/1990). 'Information management: a proposal'. Available: http://www.w3.org/History/1989/Proposal.html

Berners-Lee, Tim (1997). 'Realising the full potential of the Web'. Available: http://www.w3.org/1998/02/Potential.html

Berners-Lee, Tim; with **Fischetti**, Mark (1999). *Weaving the Web : the past, present and future of the World Wide Web by its inventor*. London: Orion Business Books.

Berners-Lee, Tim & **Cailliau**, Robert (1990). 'WorldWideWeb: proposal for a hypertext project'. Available: http://www.w3.org/Proposal.html

Bernstein, Mark (1998). 'Patterns of hypertext'. *Proceedings of ACM Hypertext '98*. Available: http://www.eastgate.com/patterns/Print.html

Bharat, Krishna; **Broder**, Andrei; **Henzinger**, Monika; **Kumar**, Puneet & **Venkatasubramanian**, Suresh (1998). 'The Connectivity server: fast access to linkage information on the Web'. *Proceedings of the 7th International World Wide Web Conference (WWW7)*. pp. 469-477.

Bharat, Krishna & **Henzinger**, Monika R. (1998). 'Improved algorithms for topic distillation in a hyperlinked environment'. In: Croft, W. Bruce et al. (eds.). *Proceedings of the 21st annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, 1998. pp. 104-111.

Bianconi, G. & **Barabási**, A.-L. (2001). 'Competition and multiscaling in evolving networks'. *Europhysics Letters*, 54: 436-442. Preprint available: http://www.nd.edu/~networks/PDF/Competition%202001.pdf

Bilke, Sven & **Peterson**, Carsten (2001). 'Topological properties of citation and metabolic networks'. *Physical Review* E64, 036106. Preprint available: http://arxiv.org/PS_cache/cond-mat/pdf/0103/0103361.pdf

Björk, Bo-Christer & **Turk,** Ziga (2000). 'How scientists retrieve publications: an empirical study of how the Internet is overtaking paper media'. *The Journal of Electronic Publishing*, 6(2). Available: http://www.press.umich.edu/jep/06-02/bjork.html

Björneborn, Lennart (1998). *Connecto ergo sum : synliggørelse - sammenkobling - synergi - webside-fællesskaber : fænomener omkring forskeres brug af personlige websider og links. [Connecto ergo sum : visibilization, connectization, communitization and innovativization : phenomena regarding researchers' use of personal web pages and links]*. BSc Thesis. Copenhagen: Royal School of Library and Information Science.

Björneborn, Lennart (2000). *Verdensvævet som 'small-world'-netværk og mulighedsrum : omridset af en forståelsesmodel for transversale links på World Wide Web. ['Small-World' Web and Possibility Space : outlining a conceptual framework for transversal links on the Web.]* MSc Thesis. Copenhagen: Royal School of Library and Information Science.

Björneborn, Lennart (2001a). 'Small-world linkage and co-linkage'. *Proceedings of the 12th ACM Conference on Hypertext and Hypermedia*, *Århus, Denmark*. New York: ACM Press. pp. 133-134.

Björneborn, Lennart (2001b). 'Necessary data filtering and editing in webometric link structure analysis'. Working paper. Royal School of Library and Information Science.

Björneborn, Lennart (2002). 'Two control revolutions and a small-world Web - complementary aspects of uniformity and diversity in an information system'. Paper for PhD course in LIS theory, Swedish School of Library and Information Studies, Borås, Sweden.

Björneborn, Lennart & **Ingwersen**, Peter (2001). 'Perspectives of webometrics'. *Scientometrics*, 50(1): 65-82.

Björneborn, Lennart & **Ingwersen**, Peter (forthcoming). 'Towards a basic framework for webometrics'. *Journal of the American Society for Information Science and Technology.* Special Issue on Webometrics.

Bøgh Andersen, Peter (1998). 'WWW as self-organizing system'. *Cybernetics & Human Knowing,* 5(2): 5-41.

Bohland, J.W. & **Minai**, A.A. (2001). 'Efficient associative memory using small-world architecture'. *Neurocomputing*, 38-40: 489-496.

Bollen, Johan (1995). 'Algorithms for the evolution and development of knowledge networks that use common semantics'. *Principia Cybernetica Project Symposium 'The Evolution of Complexity : Evolutionary and cybernetic foundations for transdisciplinary integration'*, Free University of Brussels. Available: http://pespmc1.vub.ac.be/Einmag_Abstr/JBollen.html

Bollobás, Béla (1998). *Modern graph theory*. New York: Springer.

Bolter, Jay David (1991). 'Topographic writing: hypertext and the electronic writing space'. In: Delany, Paul & Landow, George P. (eds.) (1991). *Hypermedia and literary studies*. Cambridge, Mass.: The MIT Press. pp. 105-118.

Bopp, T. & **Hampel**, T. (2001). 'Magellan, the Paderborn approach to distributed knowledge organization'. In: Montgomerie, C. & Viteli, J. (eds.): *Proceedings of ED-MEDIA 2001*, Charlottesville (Va.): Association for the Advancement of Computing in Education 2001, pp. 649–655.

Borgman, Christine L. & **Furner**, Jonathan (2002). 'Scholarly communication and bibliometrics'. *Annual Review of Information Science and Technology*, 36: 3-72.

Bossy, Marcia J. (1995). 'The last of the litter: "Netometrics". *Solaris*, 2 ('Les sciences de l'information : bibliométrie, scientométrie, infométrie'). Presses Universitaires de Rennes.
Available: http://biblio-fr.info.unicaen.fr/bnum/jelec/Solaris/d02/2bossy.html

Botafogo, Rodrigo A., **Rivlin**, Ehud, **Shneiderman**, Ben (1992). 'Structural analysis of hypertexts: identifying hierarchies and useful metrics'. *ACM Transactions on Information Systems*, 10(2): 142-180.

Botafogo, Rodrigo A. & **Shneiderman**, Ben (1991). 'Identifying aggregates in hypertext'. *Proceedings of the 3rd ACM Conference on hypertext.* pp. 63-74.

Bradford, Samuel Clements (1934). 'Sources of information on specific subjects'. *British Journal of Engineering*, 137: 85-86

Bray, Tim (1996). 'Measuring the Web'. *WWW5 Conference*. Available: http://www5conf.inria.fr/fich_html/papers/P9/Overview.html

Brin, Sergey & **Page**, Lawrence (1998). 'The anatomy of a large-scale hypertextual Web search engine'. *Computer Networks and ISDN Systems,* 30: 1-7.

Broder, Andrei; **Kumar**, Ravi; **Maghoul**, Farzin; **Raghavan**, Prabhakar; **Rajagopalan**, Sridhar; **Stata**, Raymie; **Tomkins**, Andrew & **Wiener**, Janet (2000). 'Graph structure in the Web'. *Proceedings of the 9th WWW Conference*. Also published in: *Computer Networks*, 33(1-6): 309-320. Available online: http://www.almaden.ibm.com/cs/k53/www9.final

Brookes, B.C. (1990). 'Biblio-, sciento-, infor-metrics??? What are we talking about?' In: Egghe, Leo & Rousseau, Ronald (eds.). *Informetrics 89/90 : selection of papers submitted for the Second International Conference on Bibliometrics, Scientometrics and Informetrics. London, Ontario, Canada, 5-7 July 1989*. Amsterdam: Elsevier Science Publishers. pp.31-43.

Burden, Peter (2001). 'UK Universities and Colleges.' Available at the Internet Archive:
http://web.archive.org/web/20010707114102/http://www.scit.wlv.ac.uk/ukinfo/alpha.html

Burden, Peter (2003). 'UK Sensitive Map : Universities : Version 5'. University of Wolverhampton. Available: http://www.scit.wlv.ac.uk/ukinfo/uk.map.html

Burnett, Gary; **Besant**, Michele & **Chatman**, Elfreda A. (2001). 'Small worlds: normative behavior in virtual communities and feminist bookselling'. *Journal of the American Society for Information Science and Technology*, 52(7): 536-547.

Bush, Vannevar (1945). 'As we may think'. In: Nyce, James M. & Kahn, Paul (eds.) (1991). *From Memex to hypertext : Vannevar Bush and the Mind's Machine.*

Boston: Academic Press. pp. 85-110. Note: Originally published in *The Atlantic Monthly*, July, 1945, 176(1): 101-108.

Cailliau, Robert (1995). 'A little history of the World Wide Web : from 1945 to 1995'. World Wide Web Consortium. Available: http://www.w3.org/History.html

Campos, Jose & **de Figueiredo**, A. Dias (2001). 'Searching the unsearchable: inducing serendipitous insights'. *Proceedings of the Workshop Program at the Fourth International Conference on Case-Based Reasoning, ICCBR 2001, Technical Note AIC-01-003*. Washington, DC: Naval Research Laboratory, Navy Center for Applied Research in Artificial Intelligence. pp. 159-164. Available: http://citeseer.nj.nec.com/campos01searching.html

Castells, Manuel (1996). *The rise of the network society*. Oxford: Blackwell Publishers. (The information age : economy, society and culture; 1)

Castells, Manuel (2001). *The Internet galaxy : reflections on the Internet, business, and society.* New York: Oxford University Press.

Catledge, Lara D. & **Pitkow**, James E. (1995). 'Characterizing browsing strategies in the World-Wide Web.' *Computer Networks and ISDN Systems*, 26(6): 1065-1073.

Celoria, Francis (1969). 'The archaeology of Serendip'. *Library Journal,* 94:1846-1848.

Chakrabarti, Soumen; **Dom**, Byron E.; **Kumar**, S. Ravi; **Raghavan**, Prabhakar; **Rajagopalan**, Sridhar; **Tomkins**, Andrew; **Gibson**, David & **Kleinberg**, Jon (1999). 'Mining the Web's link structure'. *IEEE Computer*, 32(8): 60-67.

Chakrabarti, Soumen; **Joshi**, Mukul M.; **Punera**, Kunal & **Pennock**, David M. (2002). 'The structure of broad topics on the Web'. *WWW2002 Conference*. Available: http://www2002.org/CDROM/refereed/338/

Chandler, Daniel (1998). 'Personal home pages and the construction of identities on the Web'. Available: http://www.aber.ac.uk/media/Documents/short/webident.html

Chang, Shan-Ju & **Rice**, Ronald E. (1993). 'Browsing : a multidimensional framework'. *Annual Review of Information Science and Technology,* 28: 231-276.

Charlesworth, Andrew (1996). 'Legal issues on the WWW - draft code of practice'. ILTU, University of Hull. Available: http://www.agocg.ac.uk/reports/mmedia/legal/legal.pdf

Chen, Chaomei; **Newman**, J.; **Newman**, R. & **Rada**, R. (1998). 'How did university departments interweave the Web: a study of connectivity and underlying factors'. *Interacting with Computers,* 10: 353-373.

Chi, Ed H.; **Pitkow**, James; **Mackinlay**, Jock; **Pirolli**, Peter; **Gossweiler**, Rich & **Card**, Stuart K. (1998). 'Visualizing the evolution of Web ecologies'. *Proceedings of Human Factors in Computing Systems, CHI 98*. ACM Press. pp. 400-407.

Chatman, Elfreda A. (1991). 'Life in a small world: applicability of gratification theory to information-seeking behavior'. *Journal of the American Society for Information Science*, 42(6): 438-449.

Chu, Heting; **He**, Shaoyi & **Thelwall**, Mike (2002). 'Library and information science schools in Canada and USA: a webometric perspective'. *Journal of Education for Library and Information Science* 43(2): 110-125.

Clever Project, Members of the (1999). 'Hypersearching the Web'. *Scientific American*, 280(6) June: 54-60.

Collins, James J. & **Chow**, Carson C. (1998). 'It's a small world'. *Nature*, 393(June 4): 409-410.

Conklin, Jeff (1987). 'Hypertext : an introduction and survey'. *IEEE Computer,* 20(9): 17-41.

Cooley, R.; **Mobasher**, B. & **Srivastava**, J. (1997). 'Web mining: information and pattern discovery on the World Wide Web'. *Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97).* Available: http://citeseer.nj.nec.com/cooley97web.html

Cooper, W.S. (1969). 'Is interindexer consistency a hobgoblin?' *American Documentation*, 20: 268–278.

Cossentino, M. & **Lo Faso**, U. (2001). 'Workgroup Hypermedia Editor: a tool to support a strategy for co-operative hypermedia production'. Paper at *the European Conference on Computer-Supported Collaborative Learning, Euro-CSCL 2001*. Available: http://www.mmi.unimaas.nl/euro-cscl/Papers/31.doc

Cove, J.F. & **Walsh**, B.C. (1988). 'Online text retrieval via browsing'. *Information Processing & Management*, 24(1): 31-37.

Crane, Diana (1972). *Invisible colleges : diffusion of knowledge in scientific communities.* Chicago: University of Chicago Press.

Cronin, Blaise (2001). 'Bibliometrics and beyond: some thoughts on web-based citation analysis'. *Journal of Information Science*, 27(1): 1-7.

Cronin, Blaise & **McKim**, Geoffrey (1996). 'Science and scholarship on the World Wide Web: a North American perspective'. *Journal of Documentation*, 52(2): 163-171.

Cronin, Blaise; **Snyder**, Herbert W.; **Rosenbaum**, Howard; **Martinson**, Anna & **Callahan**, Ewa (1998). 'Invoked on the Web'. *Journal of the American Society for Information Science,* 49(14): 1319-1328.

Crowston, Kevin & **Williams**, Marie (2000). 'Reproduced and emergent genres of communication on the World Wide Web'. *The Information Society*, 16: 201-215.

Cui, Lei (1999). 'Rating health web sites using the principles of citation analysis: a bibliometric approach'. *Journal of Medical Internet Research*, 1(1):e4. Available: http://www.jmir.org/1999/1/e4/

Dalgaard, Rune (2001). 'Hypertext and the scholarly archive: intertexts, paratexts and metatexts at work'. *Proceedings of the 12th ACM Conference on Hypertext and Hypermedia*. New York: ACM Press. pp. 175-184.

Darken, Rudolph P. & **Sibert**, John L. (1996). 'Wayfinding strategies and behaviors in large virtual worlds'. *CHI 96 Electronic Proceedings.* Available: http://web.archive.org/web/20020219095518/http://www.acm.org/sigchi/chi96/proceedings/papers/Darken/Rpd_txt.htm

Davenport, Elisabeth & **Cronin**, Blaise (1990). 'Hypertext and the conduct of science'. *Journal of Documentation*, 46(3): 175-192.

Davenport, Elisabeth & **Cronin**, Blaise (2000). 'The citation network as a prototype for representing trust in virtual environments'. In: Cronin, Blaise & Atkins, Helen Barsky (eds.) (2000). *The web of knowledge : a festschrift in honor of Eugene Garfield*. Medford, N.J.: Information Today. pp. 517-534.

Davies, Roy (1989). 'The creation of new knowledge by information retrieval and classification'. *Journal of Documentation,* 45(4): 273-301.

Davis, Gerald F.; **Yoo**, Mina & **Baker**, Wayne E. (2002). 'The network topography of the American corporate elite, 1982-2001'. University of Michigan Business School. Available: http://experiments.gsia.cmu.edu/speakers/Davis.pdf

Davison, Brian D. (2000). 'Topical locality in the Web'. In: Belkin, Nicholas J. *et al.* (eds.). *Proceedings of the 23rd annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press. pp. 272-279.

Day, Michael (2003). 'Collecting and preserving the World Wide Web : A feasibility study undertaken for the JISC and Wellcome Trust'. UKOLN, University of Bath. Version 1.0 - 25 February 2003. Available: http://www.jisc.ac.uk/uploaded_documents/archiving_feasibility.pdf

de Beer, Fanie (1996). 'The anthropology of Cyberspace'. In: Ingwersen, Peter & Pors, Niels Ole (eds.). *Integration in perspective : Proceedings : CoLIS 2 : Second International Conference on Conceptions of Library and Information Science, October 13-16, 1996.* Copenhagen: Royal School of Librarianship. pp. 117-133.

de Bono, Edward (1967). *The use of lateral thinking*. London: Jonathan Cape.

de Bra, Paul (2000). 'Using hypertext metrics to measure research output levels'. *Scientometrics*, 47(2): 227-236.

de Castro, Rodrigo & **Grossman**, Jerrold W. (1999). 'Famous trails to Paul Erdös'. *The Mathematical Intelligencer*, 21(3): 51-63. Preprint available: http://www.oakland.edu/~grossman/trails.pdf

de Jong, Hidde & **Rip**, Arie (1997). 'The computer revolution in science: steps towards the realization of computer-supported discovery environments'. *Artificial Intelligence*, 91: 225-256.

de Kerckhove, Derrick (1998). *Connected intelligence : the arrival of the web society*. London: Kogan Page.

Dearing, Ron (1997). *Higher education in the learning society.* National Committee of Inquiry into Higher Education, NCIHE. Available: http://www.leeds.ac.uk/educol/ncihe/

Deo, Narsingh & **Gupta**, Pankaj (2001). 'World Wide Web: a graph-theoretic perspective'. Computer Science Technical report CS-TR-01-001. University of Central Florida. Available: http://www.cs.ucf.edu/~pgupta/www_tech.pdf

Deutsch, Karl W. (1989). 'The small world problem: the growth of a research idea'. In: Kochen, Manfred (ed.). *The small world*. Norwood, N.J.: Ablex Publishing Corporation, 1989. pp. xv-xxi.

Dieberger, Andreas (1997). 'Supporting social navigation on the World Wide Web'. *International Journal of Human-Computer Studies*, 46(6): 805-825.

Dill, Stephen; **Kumar**, S. Ravi; **McCurley**, Kevin; **Rajagopalan**, Sridhar; **Sivakumar**, D. & **Tomkins**, Andrew (2001). 'Self-similarity in the Web'. *Proceedings of the 27th International Conference on Very Large Data Bases*, pp. 69-78.

Dillon, Andrew & **Gushrowski**, Barbara A. (2000). 'Genres and the Web: is the personal home page the first uniquely digital genre?'. *Journal of the American Society for Information Science,* 51(2): 202-205.

Ding, Chris; **Zha**, Hongyuan; **He**, Xiaofeng; **Husbands**, Parry & **Simon**, Horst (2002). 'Link analysis: hubs and authorities on the World Wide Web'. LBNL Tech Report 47847. May 7, 2001 Available: http://www.nersc.gov/research/SCG/cding/papers_ps/hits3.ps

Dodge, Martin (1999a). 'Journey to the centre of the Web'. In: Staple, G.C. (ed.). *TeleGeography 1999: Global Telecommunications Traffic Statistics & Commentary*. Washington, DC: TeleGeography, Inc. pp. 150-154. Preprint available: http://www.casa.ucl.ac.uk/martin/telegeography_webx.pdf

Dodge, Martin (1999b). 'The geography of Cyberspace'. CASA Working Paper 8. Centre for Advanced Spatial Analysis, University College London. Available: http://www.casa.ucl.ac.uk/cyberspace.pdf

Dodge, Martin & **Kitchin**, Rob (2001). *Mapping cyberspace*. London: Routledge.

Dodge, Martin & **Kitchin**, Rob (2002). 'New cartographies to chart Cyberspace'. *GeoInformatics*, 5(April/May): 38-41. Available: http://www.casa.ucl.ac.uk/martin/geoinformatics_article.pdf

Doreian, Patrick & **Fararo**, Thomas J. (1985). 'Structural equivalence in a journal network'. *Journal of the American Society for Information Science,* 36(1): 28-37.

Dorogovtsev, S.N. & **Mendes**, J.F.F. (2000). 'Evolution of reference networks with aging'. *Physical Review E,* 62(2): 1842-1845. Preprint available: http://arxiv.org/PS_cache/cond-mat/pdf/0001/0001419.pdf

Dorogovtsev, S.N. & **Mendes**, J.F.F. (2002). 'Evolution of networks'. *Advances in Physics*, 51(4): 1079-1187. Preprint available: http://arxiv.org/PS_cache/cond-mat/pdf/0106/0106144.pdf

Dorogovtsev, S. N.; **Mendes**, J. F. F. & **Samukhin**, A.N. (2000). 'Structure of growing networks with preferential linking'. *Physical Review Letters*, 85(21): 4633-4636.

Dourish, Paul & **Chalmers**, Matthew (1994). 'Running out of space: models of information navigation'. *Proceedings of HCI'94*. Available: http://www.dcs.gla.ac.uk/~matthew/papers/hci94.pdf

Downie, J. Stephen (1996). 'Informetrics and the World Wide Web: A case study and discussion'. *Proceedings of the 24th Annual Conference of the Canadian Association for Information Science*, 2-3 June 1996, Toronto, Ontario. pp. 130-141.

Dubé, Line & **Paré**, Guy (2001). 'Case research in information systems : current practices, trends, and recommendations'. École des Hautes Études Commerciales de Montréal, Canada. Available: http://gresi.hec.ca/SHAPS/cp/gescah/formajout/ajout/test/uploaded/cahier0112.doc

Ebel, Holger; **Mielsch**, Lutz-Ingo & **Bornholdt**, Stefan (2002). 'Scale-free topology of e-mail networks'. *Physical Review E*, 66: 035103(R).

Efe, Kemal; **Raghavan**, Vijay; **Chu**, C. Henry; **Broadwater**, Adrienne L.; **Bolelli**, Levent & **Ertekin**, Seyda (2000). 'The shape of the Web and its implications for searching the Web'. *Proceedings of the International Conference on the Advances in Infrastructure for Electronic Business, Science, and Education on the Internet*, L'Aquila, Italy, July 31-Aug 6, 2000. Available: http://citeseer.nj.nec.com/317732.html

Egghe, L. (2000). 'New informetric aspects of the Internet: some reflections - many problems'. *Journal of Information Science*, 26(5): 329-335.

Egghe, Leo & **Rousseau**, Ronald (1990). *Introduction to informetrics : quantitative methods in library, documentation and information science.* Amsterdam: Elsevier.

Egghe, Leo & **Rousseau**, Ronald (2002). 'Co-citation, bibliographic coupling and a characterization of lattice citation networks'. *Scientometrics*, 55(3): 349-361.

Egghe, Leo & **Rousseau**, Ronald (2003a). 'A measure for the cohesion of weighted networks'. *Journal of the American Society for Information Science and Technology*, 54(3): 193-202.

Egghe, Leo & **Rousseau**, Ronald (2003b). 'BRS-compactness in networks: theoretical considerations related to cohesion in citation graphs, collaboration networks and the Internet'. *Mathematical and Computer Modelling*, 37(7-8): 879-899 .

Eppstein, David & **Wang**, Joseph (2001). 'Fast approximation of centrality'. *Proceedings of the 12th Symposium on Discrete Algorithms, ACM and SIAM*. pp. 228-229.

Erdelez, Sanda (1995). *Information encountering: an exploration beyond information seeking*. Doctoral dissertation. New York: Syracuse University.

Erdelez, Sanda (1997). 'Information encountering: a conceptual framework for accidental information discovery'. In: Vakkari, Savolainen & Dervin (eds.). *Information seeking in context: proceedings of an international conference on research in information needs, seeking and use in different contexts. 14-16 August, 1996, Tampere, Finland.* London: Taylor Graham. pp. 412-421.

Erdelez, Sanda (2000). 'Towards understanding information encountering on the Web'. In: *Proceedings of the 63rd ASIS Annual Meeting,* 37. Medford, NJ.: Information Today. pp. 363-371.

Erickson, Thomas (1996). 'The World Wide Web as social hypertext'. *Communications of the ACM*, 39(1): 15-17.

Etzioni, Oren (1996). 'The World-Wide Web: quagmire or gold mine?' *Communications of the ACM*, 39(11): 65-68.

Faloutsos, Michalis; **Faloutsos**, Petros & **Faloutsos**, Christos (1999). 'On power-law relationships of the Internet topology'. *SIGCOMM 1999 Conference*. Available: http://www.acm.org/sigcomm/sigcomm99/papers/session7-2.html
Note: Also published in *Computer Communication Review*, 1999, 29: 251-262.

Fang, Yong & **Rousseau**, Ronald (2001). 'Lattices in citation networks: an investigation into the structure of citation graphs'. *Scientometrics*, 50(2): 273-287.

Ferrer i Cancho, Ramon; **Janssen**, C. & **Solé**, Ricard V. (2001). 'Topology of technology graphs: small world patterns in electronic circuits'. *Physical Review E*, 64: 046119.

Ferrer i Cancho, Ramon & **Solé**, Ricard V. (2001). 'The small-world of human language'. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, 268(1482): 2261-2265. Preprint available: http://www.santafe.edu/sfi/publications/Working-Papers/01-03-016.pdf

Finholt, Thomas A. (2002). 'Collaboratories'. *Annual Review of Information Science and Technology*, 36: 73-107.

Flake, Gary William; **Lawrence**, Steve; **Giles**, C. Lee & **Coetzee**, Frans M. (2002). 'Self-organization and identification of web communities'. *IEEE Computer*, 35(3): 66-71.

Ford, Nigel (1999). 'Information retrieval and creativity : towards support for the original thinker'. *Journal of Documentation*, 55(5): 528-542.

Freedman, Aviva & **Medway**, Peter (1994). 'Locating genre studies: antedecents and prospects'. pp. 1-20. In: Freedman & Medway (eds.). *Genre and the New Rhetoric*. London: Taylor & Francis.

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

Furner, Jonathan; **Ellis**, David & **Willett**, Peter (1996). 'The representation and comparison of hypertext structures using graphs'. In: Agosti, Maristella & Smeaton, Alan F. (eds.). *Information retrieval and hypertext.* Boston: Kluwer Academic Publishers. pp. 75-96.

Fürnkranz, Johannes (1998). 'Using links for classifying web-pages : Technical Report OEFAI-TR-98-29'. Austrian Research Institute for Artificial Intelligence. Available: http://citeseer.nj.nec.com/153148.html

Garfield, Eugene (1955). 'Citation indexes for science : a new dimension in documentation through association of ideas'. *Science*, 122(July 15): 108-111.

Garfield, Eugene (1975). 'The World Brain as seen by an information entrepreneur'. In: Kochen, Manfred (ed.). (1975). *Information for action : from knowledge to wisdom*. New York: Academic Press. pp. 155-160.

Garfield, Eugene (1979). 'It's a small world after all'. *Current contents*, 43(October 22): 5-10. Also in: *Essays of an Information Scientist*, 1979-80, 4: 299-304. Available: http://www.garfield.library.upenn.edu/essays/v4p299y1979-80.pdf

Garfield, Eugene (1986). 'The metaphor-science connection. *Current Comments*, 42(October 20): 3-10.

Garfield, Eugene (1989). 'Manfred Kochen: in memory of an information scientist pioneer qua World Brain-ist'. *Current Contents*, 25(June 19): 3-14. Also in: *Essays of an Information Scientist,* 1989, 12: 166-169. Available: http://www.garfield.library.upenn.edu/essays/v12p166y1989.pdf

Garfield, Eugene (1994). 'Linking literatures: An intriguing use of the citation index', *Current Contents,* 21(May 23): 3-5.

Garner, Ralph (1967). 'A computer oriented, graph theoretic analysis of citation index structures'. In: Flood, Barbara (ed.) (1967). *Three Drexel information science research studies*. Drexel Press. pp. 3-46. Available: http://www.garfield.library.upenn.edu/rgarner.pdf

Garrido, Maria & **Halavais**, Alexander (2003). 'Mapping networks of support for the Zapatista movement: applying social-networks analysis to study contemporary social movements'. In: McCaughey, Martha & Ayers, Michael D. (eds.). *Cyberactivism: online activism in theory and practice*. London: Routledge.

Garton, Laura; **Haythornthwaite**, Caroline & **Wellman**, Barry (1999). 'Studying online social networks.' In: Jones, Steve (ed.). *Doing Internet research : critical issues and methods for examining the Net*. Thousand Oaks, Cal.: SAGE Publications. pp. 75-105.

Geisler, Cheryl; **Bazerman**, Charles; **Doheny-Farina**, Stephen; **Gurak**, Laura; **Haas**, Christina; **Johnson-Eilola**, Johndan; **Kaufer**, David S.; **Lunsford**, Andrea; **Miller**, Carolyn R.; **Winsor**, Dorothy & **Yates**, Joanne (2001). 'IText : future directions for research on the relationship between information technology and writing'. *Journal of Business and Technical Communication*, 15(3): 269-308.

Gibson, David; **Kleinberg**, Jon & **Raghavan**, Prabhakar (1998). 'Inferring web communities from link topology'. *Proceedings of the 9th ACM Conference on Hypertext and Hypermedia*. New York: ACM Press. pp. 225-234.

Girardin, Luc (1995). *Cyberspace geography visualization : mapping the World-Wide Web to help people find their way in cyberspace*. The Graduate Institute of International Studies, Geneva. Available: http://www.girardin.org/luc/cgv/report/report.pdf

Girardin, Luc (1996). 'Mapping the virtual geography of the World-Wide Web'. *Proceedings of the 5th WWW Conference*. Available: http://www.girardin.org/luc//cgv/www5/index.html

Girvan, Michelle & **Newman**, M.E.J. (2002). 'Community structure in social and biological networks'. *Proceedings of the National Academy of Sciences*, 99, 8271-8276. Preprint available: http://arxiv.org/PS_cache/cond-mat/pdf/0112/0112110.pdf

Glänzel, Wolfgang (2001). 'National characteristics in international scientific co-authorship relations'. *Scientometrics*, 51(1): 69-115.

Glänzel, Wolfgang & **Schubert**, Andras (2003). 'A new classification scheme of science fields and subfields designed for scientometric evaluation purposes'. *Scientometrics*, 56(3): 357-367.

Goodrum, Abby A., **McCain**, Katherine W., **Lawrence**, Steve & **Giles**, C. Lee (2001). 'Scholarly publishing in the Internet age: a citation analysis of computer science literature'. *Information Processing & Management*, 37: 661-675.

Gottlieb, Lisa & **Dilevko**, Juris (2001). 'User preferences in the classification of electronic bookmarks: implications for a shared system'. *Journal of the American Society for Information Science and Technology*, 52(7): 517-535.

Granic, Isabela & **Lamey**, Alex V. (2000). 'The self-organization of the Internet and changing modes of thought'. *New Ideas in Psychology*, 18(1): 93-107.

Granovetter, Mark S. (1973). 'The strength of weak ties'. *American Journal of Sociology*, 78(6): 1360-1380.

Granovetter, Mark S. (1982). 'The strength of weak ties : a network theory revisited'. In: Jones, Steve (ed.). *Social structure and network analysis.* Beverly Hills, Cal.: SAGE Publications. pp. 105-130.

Gross, Jonathan & **Yellen**, Jay (1999). *Graph theory and its applications.* Boca Raton, Flor.: CRC Press.

Grossman, Jerrold W. (2003). 'The Erdös number project'. Available: http://www.acs.oakland.edu/~grossman/erdoshp.html

Grossman, Jerrold.W. & **Ion**, Patrick D.F. (1995). 'On a portion of the well-known collaboration graph'. *Congressus Numerantium*, 108: 129-131.

Guare, John (1990). *Six degrees of separation : a play*. New York: Vintage.

Guice, Jon (1998). 'Looking backward and forward at the Internet'. *The Information Society*, 14: 201-211.

Haas, Stephanie W. & **Grams**, Erika S. (1998). 'Page and link classifications: connecting diverse resources'. *Proceedings of the 3rd ACM Conference on Digital Libraries*. pp. 99-107.

Haas, Stephanie W. & **Grams**, Erika S. (2000). 'Readers, authors, and page structure: a discussion of four questions arising from a content analysis of web pages'. *Journal of the American Society for Information Science*, 51(2):181–192.

Hamilton, Stuart (2002). 'An overview of global Internet access barriers'. pp. 15-30. In: *IFLA/FAIFE Summary Report 2002: Libraries, conflicts and the Internet.* Copenhagen: IFLA/FAIFE.

Hardy, Christine (2001). 'Self-organization, self-reference and inter-influences in multilevel webs: beyond causality and determinism'. *Cybernetics & Human Knowing*, 8(3): 35-59.

Harnad, Stevan & **Carr**, Leslie (2000). 'Integrating, navigating and analyzing eprint archives through open citation linking (the OpCit Project). *Current Science,* 79(5): 629-638.

Harter, Stephen P. & **Ford**, Charlotte E. (2000). 'Web-based analyses of E-journal impact: approaches, problems, and issues'. *Journal of the American Society for Information Science,* 51(13): 1159-1176.

Haveliwala, Taher H.; **Gionis**, Aristides; **Klein**, Dan & **Indyk**, Piotr (2002). 'Evaluating strategies for similarity search on the Web'. *WWW2002 Conference*. Available: http://www2002.org/CDROM/refereed/75/

Hayes, Brian (2000a). 'Graph theory in practice: part I'. *American Scientist*, 88(1): 9-13. Available: http://www.americanscientist.org/template/AssetDetail/assetid/14708

Hayes, Brian (2000b). 'Graph theory in practice: part II'. *American Scientist*, 88(2): 104-109. Available: http://www.americanscientist.org/template/AssetDetail/assetid/14717

Henzinger, Monika (2001). 'Hyperlink analysis for the Web'. *IEEE Internet Computing*, 5(1): 45-50.

Hernández-Borges, Angel A.; **Pareras**, Luis G. & **Jiménez**, Alejandro (1997). 'Comparative analysis of pediatric mailing lists on the internet'. *Pediatrics*, 100(2):e8. Available: http://www.pediatrics.org/cgi/content/full/100/2/e8

HERO (2001). '2001 Research Assessment Exercise: The Outcome : RAE 4/01'. *Higher Education & Research Opportunities in the UK*. Available: http://www.hero.ac.uk/rae/Pubs/4_01/

Herring, Susan C. (2002). 'Computer-mediated communication on the Internet'. *Annual Review of Information Science and Technology*, 36: 109-168.

Hertzel, Dorothy H. (1987). 'Bibliometrics, History of the development of ideas in'. *Encyclopedia of Library and Information Science*, vol. 42, supplement 7, pp. 144-219. New York: Marcel Dekker.

Heydon, A. & **Najork**, M. (1999). 'Mercator: a scalable, extensible Web crawler'. *World Wide Web*, 2: 219-29.

Heylighen, Francis (1999). 'Collective intelligence and its implementation on the Web: algorithms to develop a collective mental map'. *Computational & Mathematical Organization Theory*, 5(3): 253-280.

Heylighen, Francis & **Bollen**, Johan (1996). 'The World-Wide Web as a super-brain: from metaphor to model'. Principia Cybernetica Project. Available: http://pespmc1.vub.ac.be/papers/WWWSuperBRAIN.html

Hjortgaard Christensen, Finn; **Ingwersen**, Peter & **Wormell**, Irene (1997). 'Online determination of the journal impact factor and its international properties'. *Scientometrics*, 40(3): 529-540.

Hjørland, Birger (2000). 'Documents, memory institutions and information science'. *Journal of Documentation*, 56(1): 27-41.

Huberman, Bernardo A. (2001). *The laws of the Web : patterns in the ecology of information.* Cambridge, Mass.: The MIT Press.

Huberman, Bernardo A. & **Adamic**, Lada (1999). 'Growth dynamics of the World-Wide Web'. *Nature*, 401 (September 9): 131.

Huberman, Bernardo A.; **Pirolli**, Peter; **Pitkow**, James E. & **Lukose**, Rajan M. (1998). 'Strong regularities in World Wide Web surfing'. *Science,* 280(April): 95-97.

Hummon, Normann P. & **Doreian**, Patrick (1989). 'Connectivity in a citation network: The development of DNA theory'. *Social Networks*, 11: 39-63. Available: http://www.garfield.library.upenn.edu/papers/hummondoreian1989.pdf

Huotari, Maija-Leena & **Chatman**, Elfreda (2001). 'Using everyday life information seeking to explain organizational behavior'. *Library & Information Science Research*, 23: 351-366.

Hurd, Julie M. (2000). 'The transformation of scientific communication: a model for 2020'. *Journal of the American Society for Information Science,* 51(14): 1279-1283.

Ingwersen, Peter (1992). *Information retrieval interaction*. London: Taylor Graham.

Ingwersen, Peter (1994). 'Polyrepresentation of information needs and semantic entities : elements of a cognitive theory for information retrieval interaction'. In: Croft, W. B. and van Rijsbergen, C. J. (eds.). *SIGIR '94 : Proceedings of the seventeenth annual international ACM-SIGIR conference on research and development in information retrieval*, 3-6 July 1994, Dublin, Ireland. London: Springer-Verlag. pp. 101-110.

Ingwersen, Peter (1998). 'The calculation of Web impact factors'. *Journal of Documentation*, 54 (2): 236-243.

Jackson, Michele H. (1997). 'Assessing the structure of communication on the World Wide Web.' *Journal of Computer Mediated Communication*, 3(1). Available: http://www.ascusc.org/jcmc/vol3/issue1/jackson.html

Jackson-Sanborn, Emily; **Odess-Harnish**, Kerri & **Warren**, Nikki (2002). 'Web site accessibility: a study of six genres'. *Library Hi Tech*, 20(3): 308-317.

Jacobs, Neil (2001). 'Information technology and interests in scholarly communication: A discourse analysis'. *Journal of the American Society for Information Science and Technology* , 52(13): 1122-1133.

Jacobs, Neil (2002). 'Co-term network analysis as a means of describing the information landscapes of knowledge communities across sectors'. *Journal of Documentation*, 58(5): 548-562.

Jepsen, Erik Thorlund; **Seiden**, Piet; **Björneborn**, Lennart; **Lund**, Haakon & **Ingwersen**, Peter (2002). 'WebTAPIR - scientific information retrieval on the World Wide Web'. In: Fidel, R., Bruce, H., Ingwersen, P. and Vakkari, P. (eds.) *Proceedings of the 4th International Conference on Conceptions of Library and Information Science, CoLIS4, Seattle, July, 2002*. Greenwood Village, Co.: Libraries Unlimited. pp. 309-312.

Jespersen, S.; **Sokolov**, I.M. & **Blumen**, A. (2000). 'Small-world rouse networks as models of cross-linked polymers'. *Journal of Chemical Physics*, 113(17): 7652-7655.

Jin, Shudong & **Bestavros**, Azer (2002). 'Small-world internet topologies : possible causes and implications on scalability of end-system multicast'. Technical Report BUCS-TR-2002-004. Computer Science Department, Boston University. Available: http://www.cs.bu.edu/techreports/pdf/2002-004-internet-topology-smallworld-sources.pdf

Kahle, Brewster (1997). 'Preserving the Internet'. *Scientific American*, March, 276(3):82-83. Preprint available: http://www.archive.org/sciam_article.html

Kalorkoti, Kyriakos (2003). 'CS2 Algorithms and Data Structures Note 11: Breadth-First Search and Shortest Paths'. School of Informatics, University of Edinburgh. Available: http://www.informatics.ed.ac.uk/teaching/classes/cs2/LectureNotes/CS2Bh/ADS/ads11.pdf [visited 20.6.2003; not available 30.10.2003]

Kautz, Henry; **Selman**, Bart & **Shah**, Mehul (1997). 'Referral Web: combining social networks and collaborative filtering'. *Communications of the ACM*, 40(3): 63-65.

Kelly, Brian (1995). 'History of WWW developments at Leeds University'. Available: http://www.leeds.ac.uk/ucs/WWW/www_history.html

Kelly, Brian & **Peacock**, Ian (1999). *WebWatching UK Web communities: final report for the WebWatch Project*. *British Library Research and Innovation Report 146*. Available: http://www.ukoln.ac.uk/web-focus/webwatch/reports/final/report.pdf

Kessler, M.M. (1963). 'Bibliographic coupling between scientific papers'. *American Documentation*, 14(1): 10-25.

Khan, Kushal & **Locatis**, Craig (1998). 'Searching through Cyberspace: the effects of link display and link density on information retrieval from hypertext on the World Wide Web'. *Journal of the American Society for Information Science,* 49(2): 176-182.

Kim, Hak Joon (2000). 'Motivations for hyperlinking in scholarly electronic articles: a qualitative study'. *Journal of the American Society for Information Science,* 51(10): 887-899.

Kinouchi, O.; **Martinez**, A. S.; **Lima**, G. F.; **Lourenço**, G. M. & **Risau-Gusman**, S. (2002). 'Deterministic walks in random networks: an application to thesaurus graphs'. Available: http://arxiv.org/PS_cache/cond-mat/pdf/0110/0110217.pdf

Kirstein, Peter T. (1999). 'Early experiences with ARPANET and INTERNET in the UK'. *IEEE Annals of the History of Computing*, 21(1): 38-44. Available: http://www.cs.utexas.edu/users/chris/sigcomm/t1/kirstein.arpahistory.material.prn.pdf

Kleczkowski, Adam & **Grenfell**, Bryan T. (1999). 'Mean-field-type equations for spread of epidemics: the 'small world' model'. *Physica A*, 274(1-2): 355-360.

Klein, Julie Thompson (1996a). *Crossing boundaries : knowledge, disciplinarities, and interdisciplinarities.* Charlottesville, Virg.: University Press of Virginia.

Klein, Julie Thompson (1996b). 'Interdisciplinary needs: the current context'. *Library Trends*, 45(2): 134-154.

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

Kleinberg, Jon M. (1999b). 'The small-world phenomenon: an algorithmic perspective'. Available: http://www.cs.cornell.edu/home/kleinber/swn.d/swn.html

Kleinberg, Jon M. (2000). 'Navigation in a small world'. *Nature*, 406 (August 24): 845.

Kleinberg, Jon M; **Kumar**, Ravi; **Raghavan**, Prabhakar; **Rajagopalan**, Sridhar & **Tomkins**, Andrew (1999). 'The Web as a graph: measurements, models, and methods'. *Proceedings of the 5th International Computing and Combinatorics Conference.* Also in: *Lecture Notes in Computer Science*, 1627: 1-18. Available: http://citeseer.nj.nec.com/kleinberg99web.html

Kleinberg, Jon & **Lawrence**, Steve (2001). 'The structure of the Web'. *Science*, 294(Nov. 30): 1849-1850.

Kleinfeld, Judith (2000). 'History of the small-world problem'. Columbia University. Available: http://web.archive.org/web/20020205080249/http://smallworld.sociology.columbia.edu/history.html

Kling, Rob & **Callahan**, Ewa (forthcoming). 'Electronic journals, the Internet, and scholarly communication'. *Annual Review of Information Science and Technology*, 37, forthcoming.

Kling, Rob & **McKim**, Geoffrey (2000). 'Not just a matter of time: field differences and the shaping of electronic media in supporting scientific communication'. *Journal of the American Society for Information Science*, 51(14):1306-1320.

Knoke, David & **Kuklinski**, James H. (1982). *Network analysis*. Beverly Hills, Cal.: SAGE Publications.

Kochen, Manfred (ed.)(1967). *The growth of knowledge : readings on organization and retrieval of information*. New York: John Wiley & Sons.

Kochen, Manfred (1972). 'WISE: a World Information Synthesis and Encyclopedia'. *Journal of Documentation*, 28(4): 322-343.

Kochen, Manfred (ed.). (1975a). *Information for action : from knowledge to wisdom*. New York: Academic Press.

Kochen, Manfred (1975b). 'Evolution of brainlike social organs'. In: Kochen, Manfred (ed.). (1975). *Information for action : from knowledge to wisdom*. New York: Academic Press. pp. 1-18.

Kochen, Manfred (ed.) (1989). *The small world*. Norwood, N.J.: Ablex Publishing Corporation.

Koehler, Wallace (1999a). '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.

Koehler, Wallace (1999b). 'An analysis of web page and web site constancy and permanence'. *Journal of the American Society for Information Science,* 50(2): 162-180.

Koehler, Wallace (2002). 'Web page change and persistence - A four-year longitudinal study'. *Journal of the American Society for Information Science and Technology*, 53(2): 162-171.

Korfhage, Robert R.; **Bhat**, U. Narayan & **Nance**, Richard E. (1972). 'Graph models for library information systems'. *The Library Quarterly*, 42(1): 31-42.

Kosala, Raymond & **Blockeel**, Hendrick (2000). 'Web mining research: a survey'. *SIGKDD Explorations,* 2(1): 1-15. Available: http://citeseer.nj.nec.com/kosala00web.html

Kumar, Ravi; **Raghavan**, Prabhakar; **Rajagopalan**, Sridhar & **Tomkins**, Andrew (1999). 'Trawling the web for emerging cyber-communities'. *Proceedings of the 8th International World Wide Web Conference*, pp. 403-415. Amsterdam: Elsevier.

Kumar, Ravi; **Raghavan**, Prabhakar; **Rajagopalan**, Sridhar & **Tomkins**, Andrew (2002). 'The Web and social networks'. *IEEE Computer*, 35(11): 32-36.

Kuster, Richard J. (1996). 'A bibliometric study of the remote hypertext links in public library World Wide Web sites'. *Proceedings of the ASIS Mid-Year Meeting, San Diego, California*. Medford, NJ: Information Today. pp. 338-343.

Larson, Ray R. (1996). 'Bibliometrics of the World Wide Web: an exploratory analysis of the intellectual structure of Cyberspace'. In: Hardin, Steve (ed.). *Global complexity : information, chaos, and control. Proceedings of the 59th ASIS Annual Meeting, Baltimore, Maryland.* Medford, NJ: Learned Information Inc./ASIS. pp. 71-78.

Lawrence, Steve (2001). 'Free online availability substantially increases a paper's impact'. *Nature*, 411(May 31): 521.

Lawrence, Steve & **Giles**, C. Lee (1998). 'Searching the World Wide'. *Science,* April: 98-100.

Lawrence, Steve & **Giles**, C. Lee (1999). 'Accessibility of information on the Web'. *Nature*, 400, July 8: 107-109.

Leazer, Gregory H. & **Furner**, Jonathan (1999). 'Topological indices of textual identity networks'. *Proceedings of the 62nd ASIS Annual Meeting,* 36. Medford, NJ.: Information Today. pp. 345-358.

Levene, Mark & **Poulovassilis**, Alexandra (2001). 'Web dynamics'. *Software Focus*, 2(2): 60-67.

Lévy, Pierre (1997). *Collective intelligence : mankind's emerging world in cyberspace*. New York: Plenum.

Leydesdorff, Loet (1997). 'Why words and co-words cannot map the development of the sciences'. *Journal of the American Society for Information Science*, 48(5): 418-427.

Leydesdorff, Loet (2001). *The challenge of scientometrics : the development, measurement, and self-organization of scientific communication*. 2. ed. Universal Publishers (uPUBLISH.com).

Leydesdorff, Loet & **Curran**, Michael (2000). 'Mapping university-industry-government relations on the internet: the construction of indicators for a knowledge-based economy'. *Cybermetrics*, 4(1). Available: http://www.cindoc.csic.es/cybermetrics/articles/v4i1p2.html

Li, Xuemei; **Thelwall**, Mike; **Musgrove**, Peter & **Wilkinson**, David (2003). 'The relationship between the WIFs or inlinks of Computer Science Departments in UK and their RAE ratings or research productivities in 2001'. *Scientometrics*, 57(2): 239-255.

Liestman, Daniel (1992). 'Chance in the midst of design: approaches to library research serendipity'. *Reference Quarterly,* 31(4): 524-532.

Lotka, Alfred J. (1926) 'The frequency distribution of scientific productivity.' *Journal of the Washington Academy of Sciences*, 16(12, June 19): 317-323.

Lyman, Peter & **Varian**, Hal R. (2000). 'How big is the information explosion'. *iMP, Information Impacts Magazine*, Nov. 2000. Available:

http://web.archive.org/web/20020220072409/http://cisp.org/imp/november_2000/11_00lyman.htm

Marchiori, M. & **Latora**, V. (2000). 'Harmony in the small-world'. *Physica A*, 285(3-4): 539-546.

Mathias, Nisha & **Gopal**, Venkatesh (2000). 'Small-worlds: how and why'. Available: http://arxiv.org/PS_cache/cond-mat/pdf/0002/0002076.pdf

Matzat, Uwe (1998). 'Informal academic communication and scientific usage of internet discussion groups'. *Proceedings IRISS '98 International Conference, 25-27 March 1998, Bristol, UK*. Available: http://sosig.ac.uk/iriss/papers/paper19.htm

Mayer-Kress, Gottfried & **Barczys**, Cathleen (1995). 'The global brain as an emergent structure from the worldwide computing network, and its implications for modeling'. *The Information Society*, 11: 1-27.

McKiernan, Gerry (1996). 'CitedSites(sm) : citation indexing of Web resources'. Available: http://www.public.iastate.edu/~CYBERSTACKS/Cited.htm

Medina, Alberto; **Matta**, Ibrahim & **Byers**, John (2000). 'On the origin of power laws in Internet topologies'. *ACM Computer Communications Review*, 30(2): 18-28.

Menczer, Filippo (2001). 'Links tell us about lexical and semantic Web content'. Technical Report Computer Science Abstract CS.IR/0108004. Available: http://arxiv.org/PS_cache/cs/pdf/0108/0108004.pdf

Menczer, Filippo (2002). 'Growing and navigating the small world Web by local content'. *Proceedings of the National Academy of Sciences*, 99(22): 14014-14019. (October 29, 2002).

Merton, Robert K. (1968). 'The Matthew Effect in science'. *Science*, 159(3810, January 5): 56-63. Available: http://www.garfield.library.upenn.edu/merton/matthew1.pdf

Mettrop, Wouter & **Nieuwenhuysen**, Paul (2001). 'Internet search engines - fluctuations in document accessibility'. *Journal of Documentation*, 57(5): 623-651.

Meyer, Eric K. (2000). 'Web metrics: too much data, too little analysis'. In: Nicholas, D. & Rowlands, I. (eds.). *The Internet: its impact and evaluation. Proceedings of an international forum held at Cumberland Lodge, Windsor Park, 16-18 July 1999*. London : Aslib/IMI. pp. 131-144.

Middleton, Iain; **McConnell**, Mike & **Davidson**, Grant (1999). 'Presenting a model for the structure and content of a university World Wide Web site'. *Journal of Information Science*, 25(3): 219-227.

Miles-Board, Timothy; **Kampa**, Simon; **Carr**, Leslie & **Hall**, Wendy (2001). 'Hypertext in the Semantic Web'. *Proceedings of the 12th ACM Conference on Hypertext and Hypermedia*. New York: ACM Press. pp. 237-238.

Milgram, Stanley (1967). 'The small-world problem'. *Psychology Today*, 1(1): 60-67.

Miller, Hugh (1995). 'The presentation of self in electronic life: Goffman on the Internet'. The Nottingham Trent University, Department of Social Sciences. Available: http://ess.ntu.ac.uk/miller/cyberpsych/goffman.htm

Miller, Hugh & **Arnold**, Jill (2001). 'Self in web home pages: gender, identity and power in Cyberspace'. pp. 73-94. In: Riva, Giuseppe & Galimberti, Carlo (eds.).

Towards cyberpsychology: mind, cognitions and society in the Internet age. Amsterdam: IOS Press.

Milo, R.; **Shen-Orr**, S.; **Itzkovitz**, S.; **Kashtan**, N.; **Chklovskii**, D. & **Alon**, U. (2002).'Network motifs: simple building blocks of complex networks'. *Science*, 298(5594, Oct 25): 824-827.

Molyneux, Robert E. & **Williams**, Robert V. (2000). 'Measuring the Internet'. *Annual Review of Information Science and Technology*, 34: 287-339.

Montoya, Jose M. & **Solé**, Ricard V. (2002). 'Small world patterns in food webs'. *Journal of Theoretical Biology*, 214(3): 405-412.

Moore, Cristopher & **Newman**, M.E.J. (2000). 'Epidemics and percolation in small-world networks'. *Physical Review E*, 61: 5678-5682.

Moukarzel, Cristian F. (1999). 'Spreading and shortest paths in systems with sparse long-range connections'. *Physical Review E*, 60: R6263-R6266.

Moulthrop, Stuart (1994). 'Rhizome and resistance: hypertext and the dreams of a new culture'. In: Landow, George P. (ed.). *Hyper/text/theory*. Baltimore: The Johns Hopkins University Press. pp. 299-319.

Moulthrop, Stuart & **Kaplan**, Nancy (1995). 'Citescapes : supporting knowledge construction on the Web'. Poster at *WWW4 Conference*. Available: http://iat.ubalt.edu/moulthrop/essays/citescapes/citescapes.html

Nance, Richard E.; **Korfhage**, Robert R. & **Bhat**, U. Narayan (1972). 'Information networks: definitions and message transfer models'. *Journal of the American Society for Information Science*, 23(4): 237-247.

Nelson, Ted (1967). 'Getting it out of our system'. In: Schecter, George (ed.). *Information retrieval : a critical view.* Washington, D.C.: Thompson Book Company. pp. 191-210.

Newman, M.E.J. (2000). 'Models of the small world'. *Journal of Statistical Physics*, 101(3-4): 819-841.

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.

Newman, M.E.J. (2002). 'Assortative mixing in networks'. *Physical Review Letters*, 89: 208701. Preprint available: http://arxiv.org/pdf/cond-mat/0205405

Nilan, Michael Sanford; **Pomerantz**, Jeffrey; **Paling**, Stephen (2001). 'Genres from the bottom up: what has the Web brought us?' *Proceedings of the 64th Annual Meeting of the American Society for Information Science and Technology*. 38: 330-339.

O'Connor, Brian (1988). 'Fostering creativity : enhancing the browsing environment'. *International Journal of Information Management*, 8: 203-210.

Otte, Evelien & **Rousseau**, Ronald (2002). 'Social network analysis: a powerful strategy, also for the information sciences'. *Journal of Information Science*, 28(6): 441-454.

Papadimitriou, Christos H.; **Raghavan**, Prabhakar; **Tamaki**, Hisao & **Vempala**, Santosh (1998). 'Latent semantic indexing: a probabilistic analysis'. In: *Proceedings of the ACM Conference on Principles of Database Systems (PODS)*. Preprint available: http://www.cs.berkeley.edu/~christos/ir.ps

Park, Han Woo (2002). 'Examining the determinants of who is hyperlinked to whom: a survey of webmasters in Korea'. *First Monday*, 7(11). Available: http://firstmonday.org/issues/issue7_11/park/

Park, Han Woo; **Barnett**, George A. & **Nam**, In-Yong (2002). 'Hyperlink-affiliation network structure of top web sites: Examining affiliates with hyperlink in Korea'. *Journal of the American Society for Information Science and Technology*, 53(7): 592-601.

Park, Han Woo & **Thelwall**, Mike (2003). 'Hyperlink analyses of the World Wide Web: A review'. *Journal of Computer-Mediated Communication*, 8(4). Available: http://www.ascusc.org/jcmc/vol8/issue4/park.html

Parker, Edwin B. (1975). 'Who should control society's information resources?'. In: Kochen, Manfred (ed.). (1975). *Information for action : from knowledge to wisdom*. New York: Academic Press. pp. 21-31.

Pejtersen, Annelise Mark (1991). *Interfaces based on associative semantics for browsing in information retrieval*. Risø National Library. (Risø-M-2883).

Pendleton, Victoria E.M. & **Chatman**, Elfreda A. (1998). 'Small world lives: implications for the public library'. *Library Trends*, 46(4): 732-752.

Pennock, David M.; **Flake**, Gary W.; **Lawrence**, Steve; **Glover**, Eric J. & **Giles**, C. Lee (2002). 'Winners don't take all: characterizing the competition for links on the web'. *Proceedings of the National Academy of Sciences*, 99(8): 5207-5211. Preprint available: http://modelingtheweb.com/modelingtheweb.pdf

Perkins, David N. (1992). 'The topography of invention'. In: Weber, Robert J. & Perkins, David N. (eds.). *Inventive Minds : creativity in technology*. New York: Oxford University Press. pp. 238-250.

Perkins, David N. (1995). 'Insight in minds and genes'. In: Sternberg, Robert J. & Davidson, Janet E. (eds.). *The nature of insight*. Cambridge, Mass.: The MIT Press. pp. 495-533.

Persson, Olle (1994). 'The intellectual base and research fronts of JASIS 1986-1990'. *Journal of the American Society for Information Science*, 45(1): 31-38.

Pierce, Sydney J. (1999). 'Boundary crossing in research literatures as a means of interdisciplinary information transfer.' *Journal of the American Society for Information Science,* 50(3): 271-279.

Pinski, G. & **Narin**, F. (1976). 'Citation influences for journal aggregates of scientific publications: theory, with applications to the literature of physics'. *Information Processing and Management*, 12: 297-312.

Pirolli, Peter; **Pitkow**, James & **Rao**, Ramana (1996). 'Silk from a sow's ear: extracting usable structures from the Web'. *CHI 96 Electronic Proceedings.* Available: http://www.acm.org/sigchi/chi96/proceedings/papers/Pirolli_2/pp2.html

Pirolli, Peter L.T. & **Pitkow**, James E. (1999). 'Distributions of surfers' paths through the World Wide Web'. *World Wide Web*, 2: 29-45.

Pitkow, James (1997). *Characterizing World Wide Web ecologies.* Doctoral dissertation. Georgia Institute of Technology. Available: http://www.pitkow.com/docs/1997-Pitkow-Dissertation.pdf

Pitkow, James (1999). 'Summary of WWW characterizations'. *World Wide Web*, 2: 3-13.

Pitkow, James & **Pirolli**, Peter (1997). 'Life, death, and lawfulness on the electronic frontier'. *CHI 97 Electronic Publications.* Available: http://www.acm.org/sigchi/chi97/proceedings/paper/jp-www.htm

Polanco, Xavier; **Boudourides**, Moses A.; **Besagni**, Dominique & **Roche**, Ivana (2001). 'Clustering and mapping web sites : for displaying implicit associations and visualizing networks'. Working paper v.1.2. Available: http://www.math.upatras.gr/~mboudour/articles/web_clustering&mapping.pdf

Pool, Ithiel de Sola & **Kochen**, Manfred (1978/1979). 'Contacts and influence'. In: Kochen, Manfred (ed.). *The small world*. Norwood, N.J.: Ablex Publishing Corporation, 1989. pp. 3-51.
Note: Originally published in *Social Networks*, 1978/79, 1:5-51.

Postman, Neil (1993). *Technopoly : the surrender of culture to technology*. Vintage Books.

Price, Derek de Solla (1961). *Science since Babylon.* Yale University Press.

Price, Derek de Solla (1965). 'Networks of scientific papers'. In: Kochen, Manfred (1967)(ed.). *The growth of knowledge : readings on organization and retrieval of information*. New York: John Wiley & Sons. pp. 145-155.
Note: Originally published in *Science*, 149(July 30, 1965): 510-515.

Price, Derek de Solla (1970). 'Citation measures of hard science, soft science, technology and nonscience'. In: Nelson, C.E. & Pollock, D.K. (eds.). *Communication among scientists and engineers*. Lexington, Mass.: Heath Lexington Books. pp. 3-22.

Price, Derek de Solla (1975). 'Some aspects of "World Brain" notions'. In: Kochen, Manfred (ed.). (1975). *Information for action : from knowledge to wisdom*. New York: Academic Press. pp. 177-192.

Price, Derek de Solla (1976). 'A general theory of bibliometric and other cumulative advantage processes'. *Journal of the American Society for Information Science*, 27(5): 292-306.

Prime, C., **Bassecoulard**, E. & **Zitt**, M. (2002). 'Co-citations and co-sitations: a cautionary view on an analogy'. *Scientometrics*, 54(2):291-308.

Pritchard, Alan (1984). *On the structure of information transfer networks*. M. Phil. thesis. School of Librarianship, Polytechnic of North London.

Qin, Jian & **Norton**, M. Jay (eds.)(1999). 'Introduction' (In issue: Knowledge Discovery in Bibliographic Databases). *Library Trends*, 48(1): 1-8.

Ranganathan, S. R. (1931). *The five laws of library science*. Madras: The Madras Library Association.

Rayward, W. Boyd (1994). 'Visions of Xanadu: Paul Otlet (1868-1944) and hypertext'. *Journal of the American Society for Information Science,* 45(4): 235-250.

Rayward, W. Boyd (1999). 'H.G. Wells's idea of a World Brain: a critical reassessment'. *Journal of the American Society for Information Science,* 50(7): 557-573.

Redner, S. (1998). 'How popular is your paper? An empirical study of the citation distribution'. *European Physical Journal B*, 4: 131-134. Preprint available: http://arxiv.org/PS_cache/cond-mat/pdf/9804/9804163.pdf

Rehm, Georg (2002). 'Towards automatic web genre identification : a corpus-based approach in the domain of academia by example of the academic's personal

homepage'. *Proceedings of the 35th Hawaii International Conference on System Sciences*. Available: http://www.uni-giessen.de/~g91063/pdf/HICSS35-rehm.pdf

Rice, James (1988). 'Serendipity and holism: the beauty of OPACs'. *Library Journal,* 113 (Feb. 15): 138-141.

Rodriguez i Gairin, J.M. (1997). 'Valorando el impacto de la informacion en Internet: Alta-vista, el "Citation Index" de la Red'. *Revista Espanola de Documentacion Scientifica*, 20(2): 175-181.

Rogers, Everett M. (1995). *Diffusion of innovations*. 4.ed. New York: The Free Press.

Roumeliotis, John (2002). 'Algorithmic graph theory'. Victoria University, Australia. Available: http://www.staff.vu.edu.au/johnr/scm2711/graph1.pdf

Rousseau, Ronald (1997). 'Sitations: an exploratory study'. *Cybermetrics,* 1(1). Available: http://www.cindoc.csic.es/cybermetrics/articles/v1i1p1.html

Rousseau, Ronald (1998/1999). 'Daily time series of common single word searches in AltaVista and NorthernLight'. *Cybermetrics,* 2/3(1). Available: http://www.cindoc.csic.es/cybermetrics/articles/v2i1p2.html

Rousseau, Brendan & **Rousseau**, Ronald (2000). 'LOTKA: A program to fit a power law distribution to observed frequency data'. *Cybermetrics*, 4(1): paper 4. Available: http://www.cindoc.csic.es/cybermetrics/articles/v4i1p4.html

Roussinov, Dmitri; **Crowston**, Kevin; **Nilan**, Mike; **Kwasnik**, Barbara; **Cai**, Jin; **Liu**, Xiaoyong (2001). 'Genre based navigation on the Web'. *Proceedings of the 34th Hawaii International Conference on System Sciences (HICSS-34)*. Available: http://wpcarey.asu.edu/fac/droussinov/hicss.doc

Rowlands, Ian (1999). 'Who can count the dust of Jacob? From bibliometrics to cybermetrics'. In: Nicholas, D. & Rowlands, I. (eds.). *The Internet: its impact and evaluation*. pp. 114-130.

Sandbothe, Mike (1996). *Interactivity-hypertextuality-transversality : a media-philosophical analysis of the Internet.* Available: http://www.sandbothe.net/241.html
Note: Later published in *Hermes. Journal of Linguistics*, 24(February 2000): 81-108.

Sandstrom, Pamela Effrein (2001). 'Scholarly communication as a socioecological system'. *Scientometrics*, 51(3): 573-605.

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

Schwartz, Michael F. & **Wood**, David C.M. (1993). 'Discovering shared interests using graph analysis'. *Communications of the ACM*, 36(8): 78-89.

Scott, John (2000). *Social network analysis : a handbook*. 2. ed. Thousand Oaks, Cal.: SAGE Publications.

Shapiro, Andrew L. (1999). *The control revolution : how the Internet is putting individuals in charge and changing the world we know*. New York: Public Affairs.

Shepherd, M.A.; **Watters**, C.R. & **Cai**, Yao (1990). 'Transient hypergraphs for citation networks.' *Information Processing & Management*, 26(3): 395-412.

Sigman, Mariano & **Cecchi**, Guillermo A. (2002). *Proceedings of the National Academy of Sciences*, 99(3): 1742-1747. Preprint available: http://arxiv.org/ftp/cond-mat/papers/0106/0106509.pdf

Simpson, Rosemary; **Renear**, Allen; **Mylonas**, Elli & **van Dam**, Andries (1996). '50 years after 'As we may think': The Brown/MIT Vannevar Bush Symposium'. *Interactions of the ACM*, 3(2): 47-67. Preprint available: http://www.cs.brown.edu/memex/Bush_Symposium_Interact_2.html

Skiena, Steven (1996). 'Lecture 18 - shortest path algorithms'. Course CSE 373/548 - Analysis of Algorithms. Spring 1996. Department of Computer Science, SUNY Stony Brook. Available: http://www.cs.sunysb.edu/~algorith/lectures-good/node18.html

Skvoretz, John & **Fararo**, Thomas J. (1989). 'Connectivity and the small world problem'. In: Kochen, Manfred (ed.). *The small world*. Norwood, N.J.: Ablex Publishing Corporation, 1989. pp. 296-326.

Small, Henry (1973). 'Co-citation in the scientific literature: a new measure of the relationship between two documents'. *Journal of the American Society for Information Science,* 24(4): 265-269.

Small, Henry (1999). 'A passage through science: crossing disciplinary boundaries'. *Library Trends*, 48(1): 72-108.

Small, Henry (2000). 'Charting pathways through science: exploring Garfield's vision of a unified index to science'. In: Cronin, Blaise & Atkins, Helen Barsky (eds.) (2000). *The web of knowledge : a festschrift in honor of Eugene Garfield*. Medford, N.J.: Information Today. pp. 449-473.

Smeaton, Alan F. (1995). 'Building hypertexts under the influence of topology metrics'. Presented at *IWHD'95, International Workshop on Hypermedia Design*, Montpellier, France, June 1995. Available: http://www.compapp.dcu.ie/~asmeaton/pubs/IWHD-ext-abs.ps

Smith, Alastair (1999). 'A tale of two web spaces: comparing sites using web impact factors'. *Journal of Documentation*, 55(5): 577-592.

Smith, Alastair & **Thelwall**, Mike (2001). 'Web impact factors and university research links'. *Proceedings of the 8th International Conference on Scientometrics and Informetrics*, *ISSI-2001*. pp. 657-664.

Smith, Alastair & **Thelwall**, Mike (2002). 'Web Impact Factors for Australasian universities'. Scientometrics, 54(3): 363-380.

Snyder, Herbert & **Rosenbaum**, Howard (1999). 'Can search engines be used as tools for web-link analysis? A critical view'. *Journal of Documentation*, 55(4): 375-384.

Sørensen, C.; **Macklin**, D. & **Beaumont**, T. (2001). 'Navigating the World Wide Web: bookmark maintenance structures'. *Interacting with Computers*, 13: 375-400.

Souma, Wataru; **Fujiwara**, Yoshi & **Aoyama**, Hideaki (2001). 'Small-world effects in wealth distribution'. Available: http://arxiv.org/PS_cache/cond-mat/pdf/0108/0108482.pdf

Spertus, Ellen (1997). 'ParaSite: mining structural information on the Web'. *WWW6 Conference*. Available: http://decweb.ethz.ch/WWW6/Technical/Paper206/Paper206.html

Sporns, Olaf (2003). 'Network analysis, complexity, and brain function'. *Complexity*, 8(1): 56-60.

Steyvers, Mark & **Tenenbaum**, Joshua B. (2001). 'The large-scale structure of semantic networks: statistical analyses and a model for semantic growth'. Available: http://arxiv.org/pdf/cond-mat/0110012

Strogatz, Steven H. (2001). 'Exploring complex networks'. *Nature*, 410(March 8): 268-276.

Sun, Kai & **Ouyang**, Qi (2001). 'Distance distribution and reliability of small-world networks'. *Chinese Physics Letters*, 18(3): 452-454. Available: http://mail.phy.pku.edu.cn/~nl/undergraduate/w418.pdf

Swanson, Don R. (1986). 'Undiscovered public knowledge'. *Library Quarterly,* 56(2): 103-118.

Swanson, Don R. & **Smalheiser**, Neil R. (1997). 'An interactive system for finding complementary literatures: a stimulus to scientific discovery'. *Artificial Intelligence*, 91: 183-203.

Swanson, Don R. & **Smalheiser**, Neil R. (1999). 'Implicit text linkages between Medline records: using Arrowsmith as an aid to scientific discovery'. *Library Trends*, 48(1): 48-59.

Tague-Sutcliffe, Jean (1992). 'An introduction to informetrics'. *Information Processing & Management*, 28(1): 1-3.

Tang, Rong & **Thelwall**, Mike (forthcoming). 'Disciplinary differences in US academic departmental web site interlinking'. *Library and Information Science Research*.

Tassier, T. & **Menczer**, F. (2001). 'Emerging small-world referral networks in evolutionary labor markets'. *IEEE Transactions on Evolutionary Computation*, 5(5): 482-492.

Thelwall, Mike (2000). 'Web impact factors and search engine coverage'. *Journal of Documentation*, 56(2): 185-189.

Thelwall, Mike (2001a). 'Extracting macroscopic information from web links'. *Journal of the American Society for Information Science and Technology*, 52(13): 1157-1168.

Thelwall, Mike (2001b). 'A publicly accessible database of UK university website links and a discussion of the need for human intervention in web crawling'. Available: http://www.scit.wlv.ac.uk/~cm1993/papers/a_publicly_accessible_database.pdf

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

Thelwall, Mike (2001d). 'An initial exploration of the link relationships between UK university web sites'. *ASLIB Proceedings*, 54(2): 118-126.

Thelwall, Mike (2001e). 'Results from a web impact factor crawler'. *Journal of Documentation*, 57(2): 177-191.

Thelwall, Mike (2002a). 'A comparison of sources of links for academic Web impact factor calculations'. *Journal of Documentation*, 58(1): 66-78.

Thelwall, Mike (2002b). 'Conceptualizing documentation on the Web: an evaluation of different heuristic-based models for counting links between university web sites'. *Journal of the American Society for Information Science and Technology,* 53(12): 995-1005.

Thelwall, Mike (2002c). 'The top 100 linked-to pages on UK university web sites: high inlink counts are not usually associated with quality scholarly content'. *Journal of Information Science*, 28(6): 483-492.

Thelwall, Mike (2002d). 'Evidence for the existence of geographic trends in university web site interlinking'. *Journal of Documentation*, 58(5): 563-574.

Thelwall, Mike (2002e). 'A research and institutional size based model for national university Web site interlinking'. *Journal of Documentation*, 58(6): 683-694.

Thelwall, Mike (2002f). 'Methodologies for crawler based Web surveys'. *Internet Research*, 12(2): 124-138.

Thelwall, Mike (2003a). 'Web use and peer interconnectivity metrics for academic Web sites'. *Journal of Information Science*, 29(1): 1-10.

Thelwall, Mike. (2003b). 'Can Google's PageRank be used to find the most important academic Web pages?' *Journal of Documentation*, 59(2): 205-217.

Thelwall, Mike. (2003c). 'A layered approach for investigating the topological structure of communities in the Web'. *Journal of Documentation*, 59(4): 410-429.

Thelwall, Mike (2003d). 'What is this link doing here? Beginning a fine-grained process of identifying reasons for academic hyperlink creation'. *Information Research*, 8(3): paper no. 151. Available: http://informationr.net/ir/8-3/paper151.html

Thelwall, Mike (forthcoming). 'Methods for reporting on the targets of links from national systems of university web sites'. *Information Processing & Management*.

Thelwall, Mike & **Harries**, Gareth (2003). 'The connection between the research of a university and counts of links to its web pages: an investigation based upon a classification of the relationships of pages to the research of the host university'. *Journal of the American Society for Information Science and Technology*, 54(7): 594-602.

Thelwall, Mike & **Harries**, Gareth (forthcoming). 'Do better scholars' web publications have significantly higher online impact?' *Journal of the American Society for Information Science and Technology.*

Thelwall, Mike & **Smith**, Alastair (2002). 'Interlinking between Asia-Pacific University Web sites'. *Scientometrics*, 55(3): 363-376.

Thelwall, Mike & **Tang**, Rong (2003). 'Disciplinary and linguistic considerations for academic Web linking: An exploratory hyperlink mediated study with Mainland China and Taiwan'. *Scientometrics*, 58(1): 153-179.

Thelwall, Mike; **Vaughan**, Liwen & **Björneborn**, Lennart (forthcoming). 'Webometrics'. *Annual Review of Information Science and Technology*, 39, forthcoming.

Thelwall, Mike & **Wilkinson**, David (2003a). 'Three target document range metrics for university Web sites'. *Journal of the American Society for Information Science and Technology*, 54(6): 490-497.

Thelwall, Mike & **Wilkinson**, David (2003b). 'Graph structure in three national academic Webs: power laws with anomalies'. *Journal of the American Society for Information Science and Technology*, 54(8): 706-712.

Thelwall, Mike & **Wilkinson**, David (forthcoming). 'Finding similar academic Web sites with links, bibliometric couplings and colinks'. *Information Processing and Management.* forthcoming.

THES (2001). 'League Tables 2001'. *Times Higher Education* Supplement, May 18, 2001. pp. T1-T4.

Thomas, Owen & **Willett**, Peter (2000). 'Webometric analysis of departments of librarianship and information science'. *Journal of Information Science*, 26(6): 421-428.

Toms, Elaine G. (1998). 'Information exploration of the third kind: the concept of chance encounters'. A position paper for the *CHI98 Workshop on Information Exploration*.

Toms, Elaine G. (2000). 'Serendipitous information retrieval'. In: *Proceedings of the First DELOS Network of Excellence Workshop on Information Seeking, Searching and Querying in Digital Libraries, Zurich, Switzerland*. European Research Consortium for Informatics and Mathematics. Available: http://www.ercim.org/publication/ws-proceedings/DelNoe01/3_Toms.pdf

Tsikrika, Theodora & **Lalmas**, Mounia (2002). 'Combining web document representations in a bayesian inference network model using link and content-based evidence'. *Proceedings of the Twenty-Fourth European Colloquium on Information Retrieval Research (ECIR '02)*, Glasgow, March 2002 : *Lecture Notes in Computer Science,* 2291, pp. 53-72.

Turner, Nancy J.; **Davidson-Hunt**, Iain J. & **O'Flaherty**, Michael (2003). 'Living on the edge: ecological and cultural edges as sources of diversity for social-ecological resilience'. *Human Ecology*, 31(3): 439-461.

Twidale, Michael B.; **Nichols**, David M. & **Paice**, Chris D. (1997). 'Browsing is a collaborative process'. *Information Processing & Management*, 33(6): 761-783.

Valverde, Sergi; **Ferrer i Cancho**, Ramon & **Solé**, Ricard V. (2002). 'Scale-free networks from optimal design'. Preprint available: http://arxiv.org/abs/cond-mat/0204344

Van Alstyne, Marshall & **Brynjolfsson**, Erik (1996). 'Could the Internet balkanize science?' *Science,* 274(November 29): 1479-1480.

van Andel, Pek (1994). 'Anatomy of the unsought finding : serendipity: origin, history, domains, traditions, appearances, patterns and programmability. *British Journal for the Philosophy of Science*, 45(2): 631-648.

van Raan, Anthony F.J. (2000). 'On growth, ageing, and fractal differentiation of science'. *Scientometrics*, 47(2): 347-362.

van Raan, Anthony F.J. (2001). 'Bibliometrics and internet: some observations and expectations'. *Scientometrics*, 50(1): 59-63.

Vaughan, Liwen & **Shaw**, Debora (forthcoming). 'Bibliographic and web citations: What is the difference?' *Journal of the American Society for Information Science and Technology*.

Vaughan, Liwen & **Thelwall**, Mike (2003). 'Scholarly use of the Web: what are the key inducers of links to journal Web sites?' *Journal of the American Society for Information Science and Technology*, 54(1): 29-38.

Vázquez, Alexei (2001). 'Knowing a network by walking on it: emergence of scaling'. *Europhysics Letters*, 54: 430-435. Preprint available: http://arxiv.org/PS_cache/cond-mat/pdf/0006/0006132.pdf

Venkatraman, Mahadevan; **Bin**,Yu & **Singh**, Munindar P. (2000). 'Trust and reputation management in a small-world network'. *Proceedings of the 4th*

International Conference on MultiAgent Systems, 2000. IEEE Computer Society. pp. 449-450.

Walker, Jill (2002). 'Links and power: the political economy of linking on the Web'. *Proceedings of Hypertext 2002*. Baltimore: ACM Press. pp. 78-79.

Walsh, Toby (1998). 'Search in a small world'. *Proceedings of IJCAI-99*. Available: http://www-users.cs.york.ac.uk/~tw/Papers/wijcai99.pdf

Wasserman, Stanley & **Faust**, Katherine (1994). *Social network analysis : methods and applications*. Cambridge: Cambridge University Press.

Watts, Duncan J. (1999a). 'Networks, dynamics, and the small-world phenomenon'. *American Journal of Sociology*, 105(2): 493-527.

Watts, Duncan J. (1999b). *Small worlds : the dynamics of networks between order and randomness*. Princeton, N.J.: Princeton University Press. (Princeton Studies in Complexity)

Watts, Duncan J. (1999c). 'The Internet, the small world, and the nature of distance'. Messages, Museum of Science, Boston. Available: http://aries.mos.org/internet/essay.html

Watts, Duncan J.; **Dodds**, Peter Sheridan & **Newman**, M. E. J. (2002). 'Identity and search in social networks'. *Science*, 296(5571, May 17): 1302-1305.

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

Weare, Christopher & **Lin**, Wan-Ying (2000). 'Content analysis of the World Wide Web : opportunities and challenges'. *Social Science Computer Review*, 18(3): 272-292.

Weiss, Ron; **Vélez**, Bienvenido; **Sheldon**, Mark A.; **Namprempre**, Chanathip; **Szilagyi**, Peter; **Duda**, Andrzej & **Gifford**, David K. (1996). 'HyPursuit: a hierarchical network search engine that exploits content-link hypertext clustering'. *Proceedings of the 7th ACM Conference on Hypertext and Hypermedia*. ACM Press. pp. 180-193. Available: http://www.psrg.lcs.mit.edu/publications/Papers/hypertabs.htm

Wellman, Barry (2001). 'Computer networks as social networks'. *Science*, 293(Sep 14): 2031-2034.

Wells, H.G. (1938). 'World encyclopaedia'. In: Kochen, Manfred (ed.). *The growth of knowledge : readings on organization and retrieval of information*. New York: John Wiley & Sons. pp. 11-22.

White, Douglas R. (2003). 'Ties, Weak and Strong'. In: Karen Cristensen & David Levinson (eds.). *Encyclopedia of Community*. Thousand Oaks, CA: Sage. Preprint: http://eclectic.ss.uci.edu/~drwhite/pw/EncyclopediaofCommunity.pdf

White, Douglas R. & **Newman**, M. E. J. (2001). 'Fast approximation algorithms for finding node-independent paths in networks'. Available: http://www.santafe.edu/sfi/publications/Working-Papers/01-07-035.pdf

White, Douglas R. & **Houseman**, Michael (2003). 'The navigability of strong ties: small worlds, tie strength, and network topology : self-organization in strong-tie small worlds'. *Complexity*, 8(1): 72-81.

White, Howard D. & **McCain**, Katherine W. (1989). 'Bibliometrics'. *Annual Review of Information Science and Technology*, 24: 119-186.

Wilhite, Allen (2001). 'Bilateral trade and `small-world' networks'. *Computational Economics*, 18(1): 49-64.

Wilkinson, David; **Harries**, Gareth; **Thelwall**; Mike & **Price**; Liz (2003). 'Motivations for academic web site interlinking: evidence for the Web as a novel source of information on informal scholarly communication'. *Journal of Information Science,* 29(1): 49-56.

Wilkinson, David; **Thelwall**, Mike & **Li**, Xuemei (forthcoming). 'Exploiting hyperlinks to study academic Web use'. *Social Science Computer Review.*

Williamson, Kirsty (1998). 'Discovered by chance: the role of incidental information acquisition in an ecological model of information use'. *Library & Information Science Research,* 20(1): 23-40.

Wilson, Robin J. & **Watkins**, John J. (1990). *Graphs : an introductory approach*. New York: John Wiley & Sons.

WISER (2001). 'Web indicators for scientific, technological and innovation research - WISER: part B: proposal description'. *The WISER project: Web Indicators for Science, Technology & Innovation Research.* Available: http://www.webindicators.org/

Wood, Andrew; **Drew**, Nick; **Beale**, Russell & **Hendley**, Bob (1995). 'HyperSpace: Web browsing with visualisation'. *Proceedings of WWW Conference, 1995*. Available: http://www.igd.fhg.de/archive/1995_www95/proceedings/posters/35/

World Wide Web Consortium (1999). 'HTML 4.01 Specification : W3C Recommendation 24 December 1999 : Appendix B : Performance, Implementation, and Design Notes : The robots.txt file'. Available: http://www.w3.org/TR/html4/appendix/notes.html#h-B.4.1.1

World Wide Web Consortium (2002). 'Naming and Addressing: URIs, URLs, ...'. Available: http://www.w3.org/Addressing/

Wynn, Eleanor & **Katz**, James A. (1997). 'Hyperbole over cyberspace: self-presentation and social boundaries in Internet home pages and discourse'. *The Information Society*, 13: 297-327.

Yankelovich, Nicole; **Meyrowitz**, Norman & **van Dam**, Andries (1985). 'Reading and writing the electronic book'. *IEEE Computer*, 18(10): 15-30.

Yao, Y.Y.; **Zhong**, Ning; **Liu**, Jiming & **Ohsuga**, Setsuo (2001). 'Web intelligence (WI) : research challenges and trends in the new information age'. *Lecture Notes in Artificial Intelligence*, 2198: 1-17.

Yin, Robert K. (1994). *Case Study research, Design and Methods*. 2 ed. Beverly Hills, Ca.: Sage.

Yook, Soon-Hyung; **Jeong**, Hawoong & **Barabási**, Albert-László (2002). 'Modeling the Internet's large-scale topology'. *Proceedings of the National Academy of Sciences*, 99(21): 13382-13386.

Zanette, Damian H. (2001). 'Dynamics of rumor propagation on small-world networks'. Available: http://arxiv.org/PS_cache/cond-mat/pdf/0110/0110324.pdf

Zhang, Yin (2001). 'Scholarly use of Internet-based electronic resources'. *Journal of the American Society for Information Science and Technology*, 52(8): 628-654.

Zipf, George Kingsley (1949). *Human behavior and the principle of least effort: an introduction to human ecology*. Cambridge, MA, Addison-Wesley.

Color prints

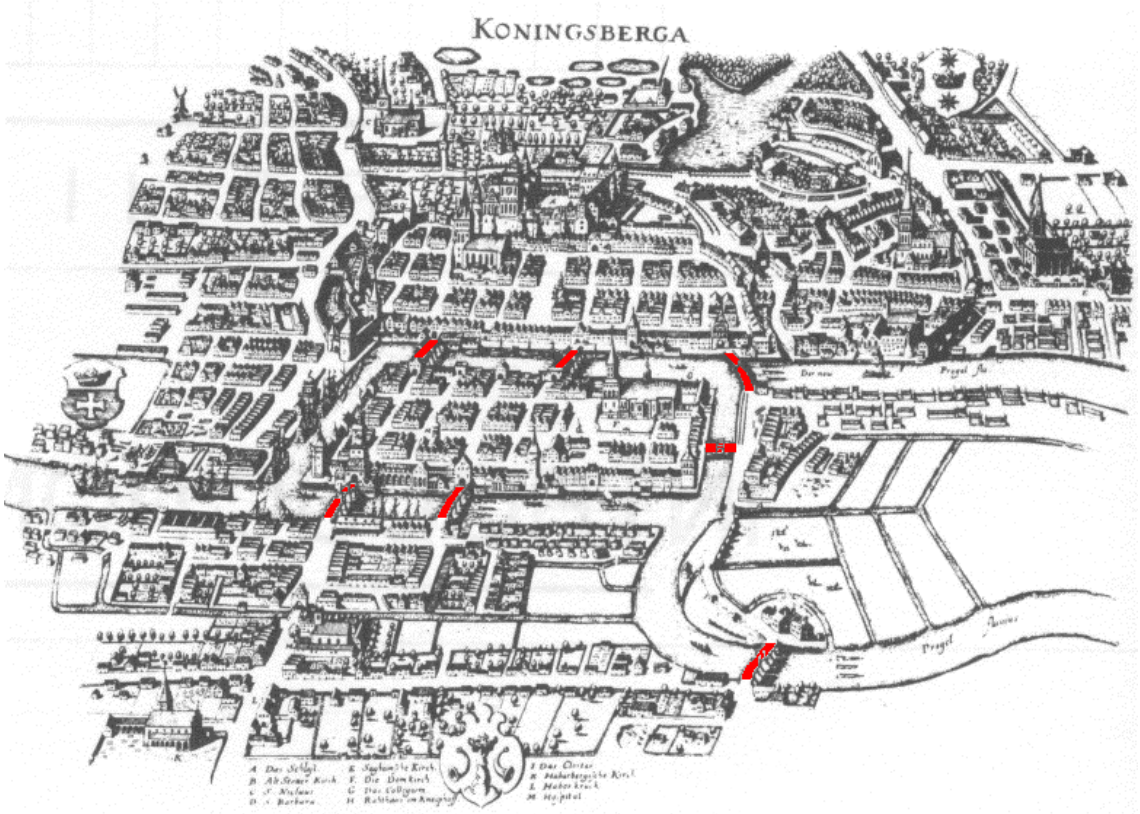


Figure 3.1. The seven bridges of Königsberg (map from Wilson & Watkins, 1990). Colored bridges added by present author.

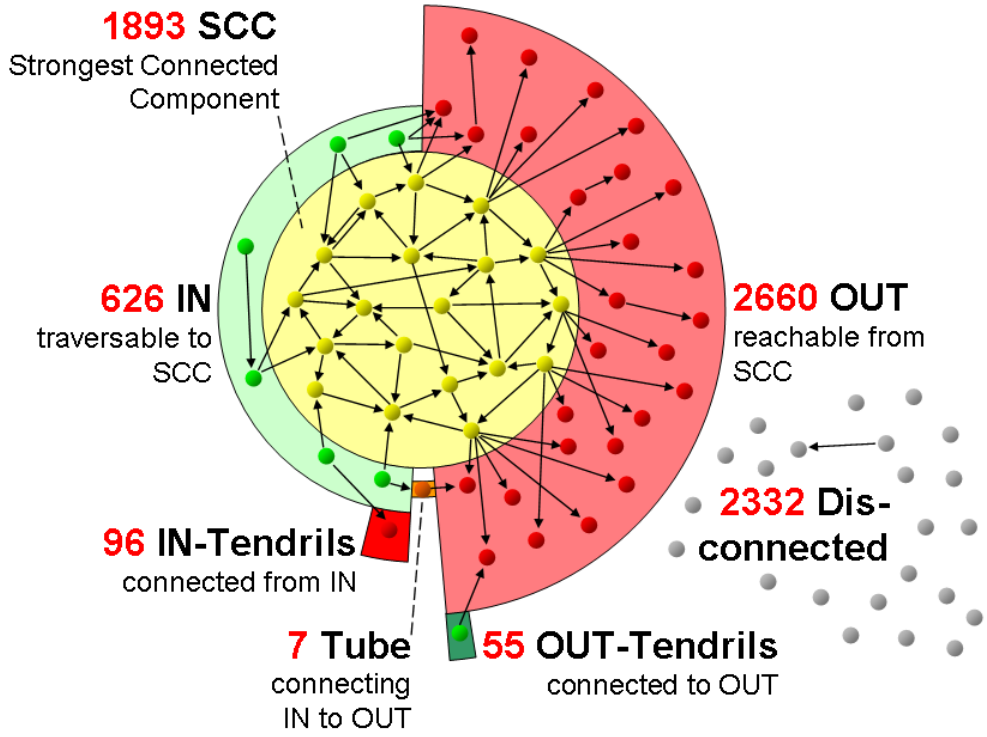


Figure 5.4. A 'corona' model of the Web graph structure of 7669 UK university subsites as of 2001. The number of nodes and sizes of components in the figure roughly represent the actual numbers and sizes. Green and red graph colors symbolize where link paths may start and stop, respectively.

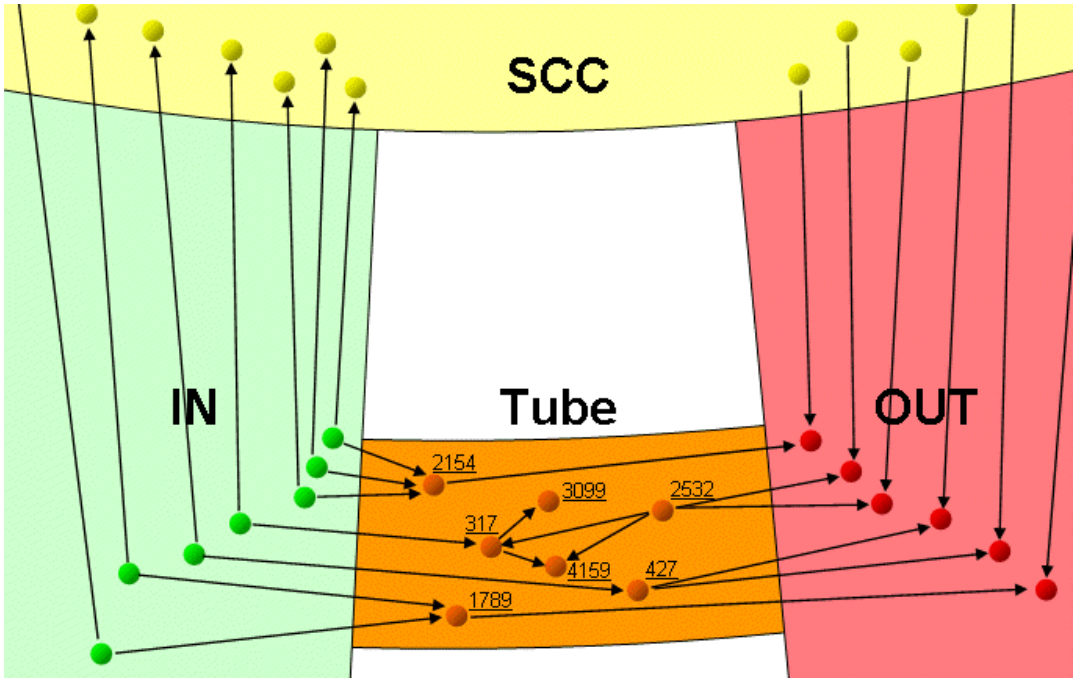


Figure 5.5. Link structures within the Tube component of the 'corona' graph model of the UK academic subweb 2001. The seven Tube nodes have id numbers assigned the 7669 subsite nodes.

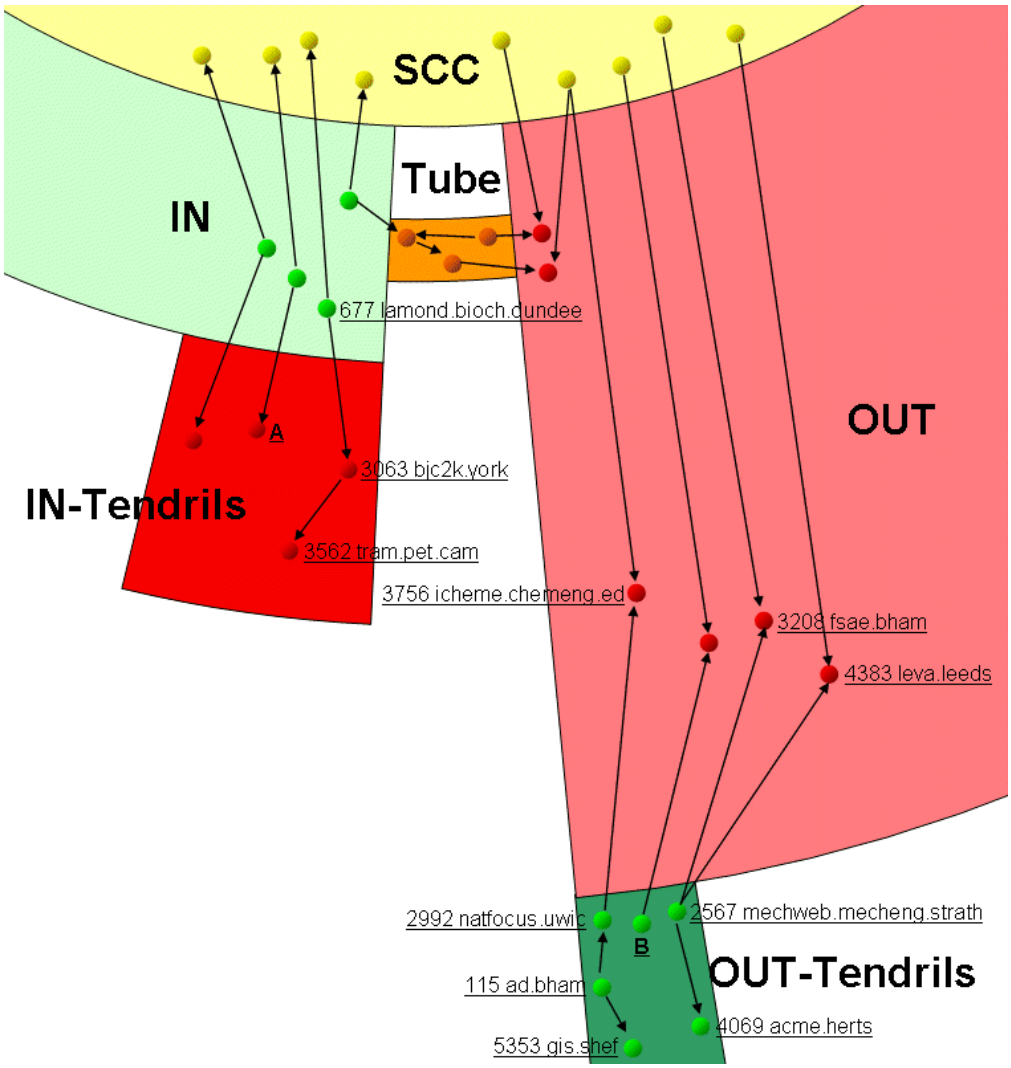


Figure 5.7. The actual link structures of nodes in the IN-Tendrils and OUT-Tendrils with intra-component links. Nodes A and B represent the majority of Tendril nodes with no intra-component links.

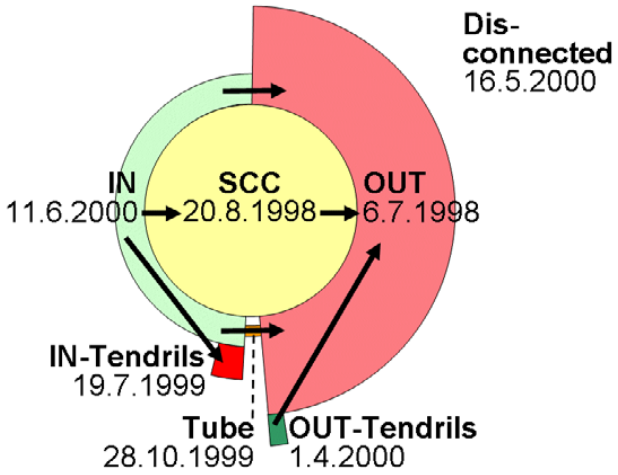


Figure 5.9. Indicative ages of graph components based on average first time indexing in the Internet Archive of 6868 subsites (cf. Table 5-2).

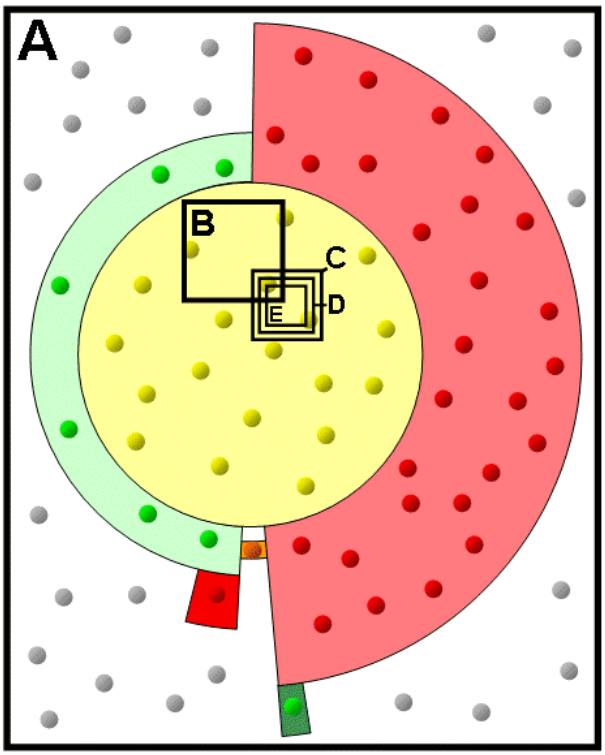


Figure 6.1. Five-step methodology (A-E) for sampling, identifying and characterizing transversal links.

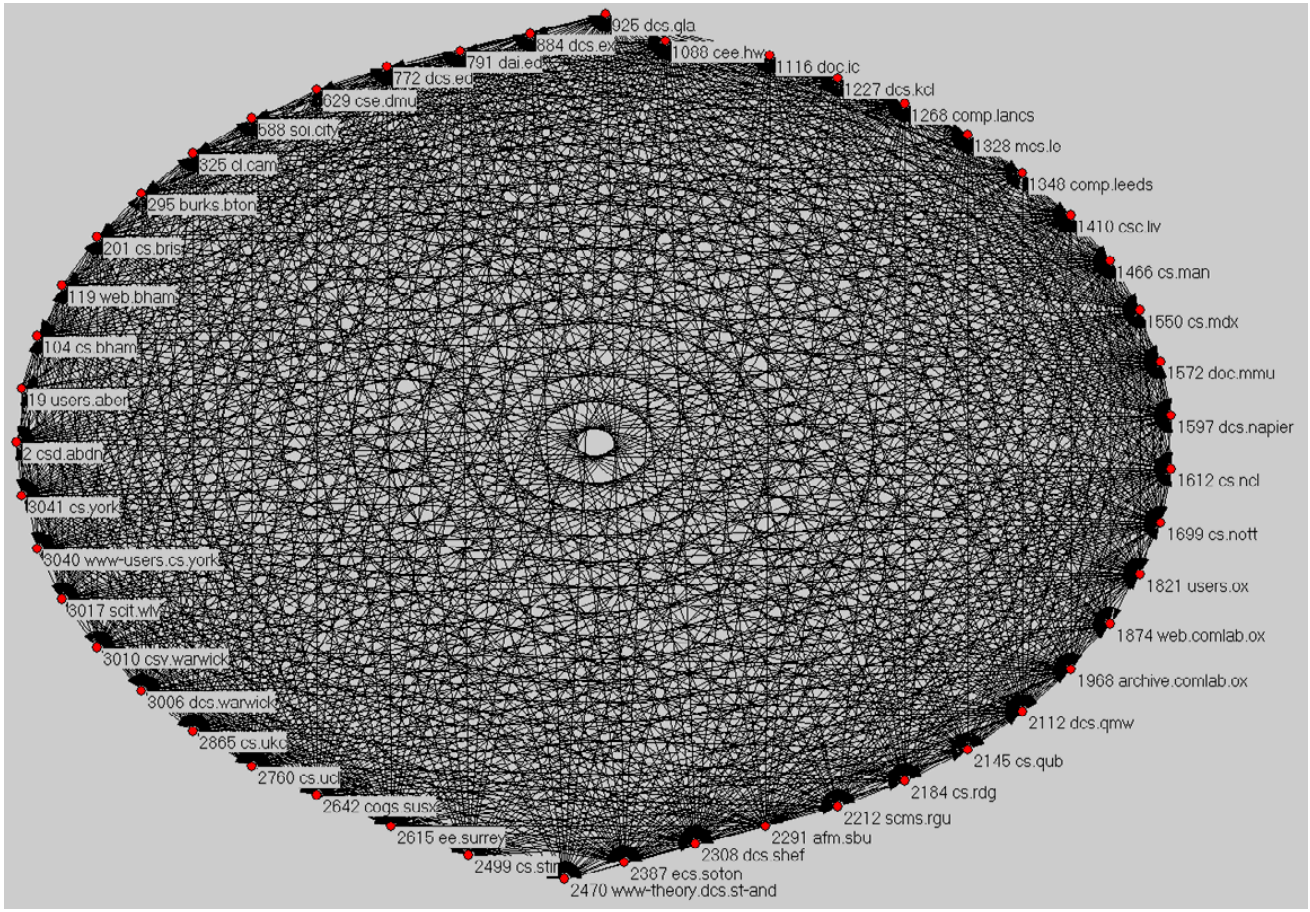


Figure 6.20. ***Out-23-core*** **containing 47 subsite nodes with at least** ***23 outlinks*** **to other core nodes.**

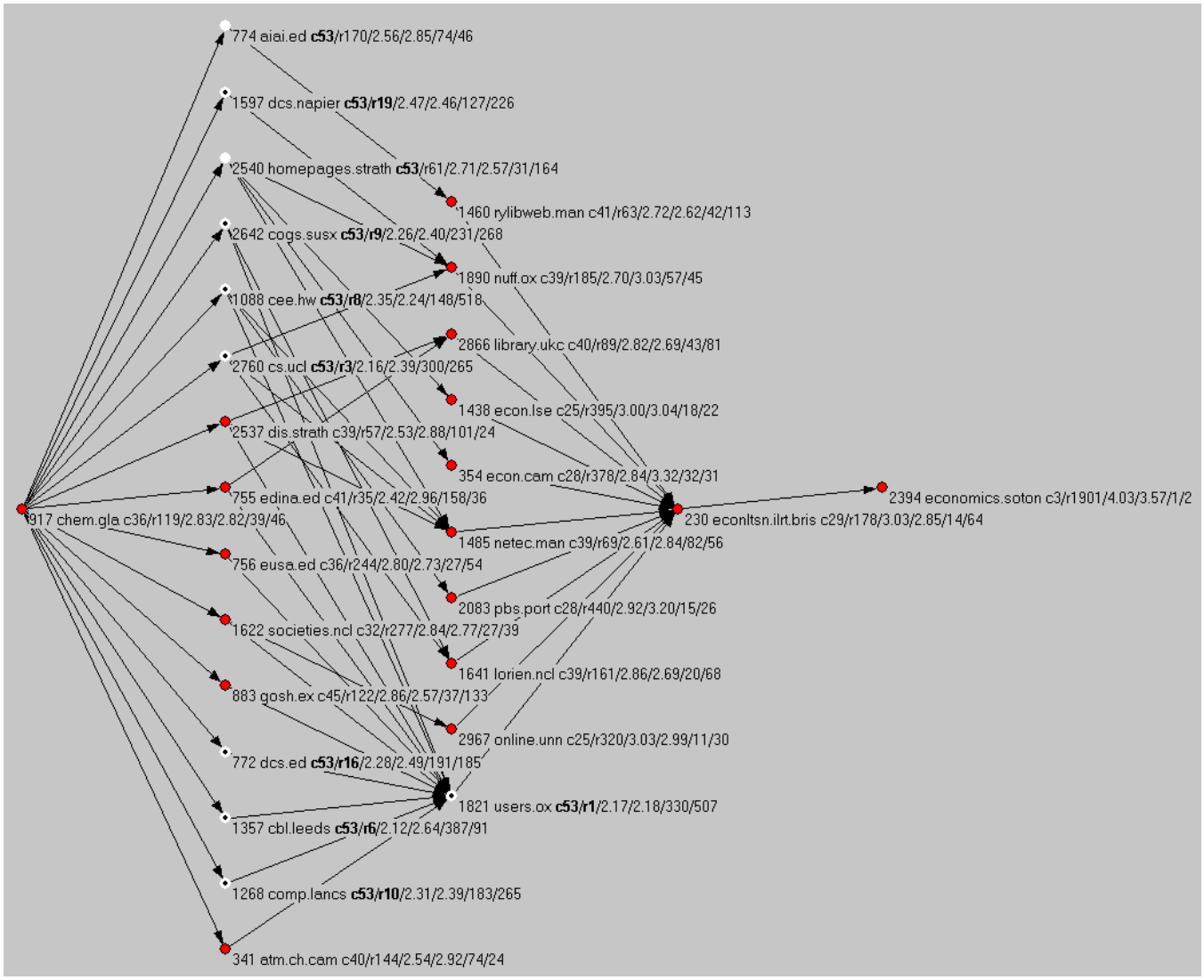


Figure 6.21. Path net NH02 with graph measures in a string at each subsite node showing core (c), betweenness centrality rank (r), average in-distance/out-distance, and number of in-neighbors/out-neighbors. Subsites belonging to the 53-core are marked in white, and subsites with bc rank < 25 are marked in white with a black spot.

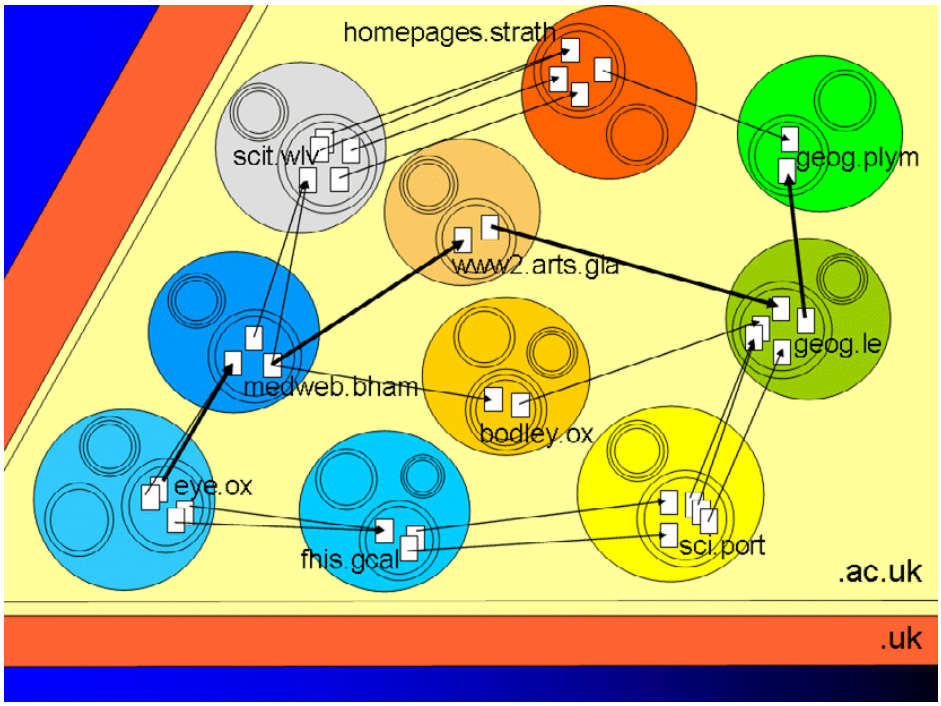


Figure 6.31. Node diagram with link path visualization. Excerpt from path net NH05 with actual source pages and target pages. All links belong to shortest link paths (path length 4) between start node *eye.ox.ac.uk* and end node *geog.plym.ac.uk*. Bold links show one example of such a link path.

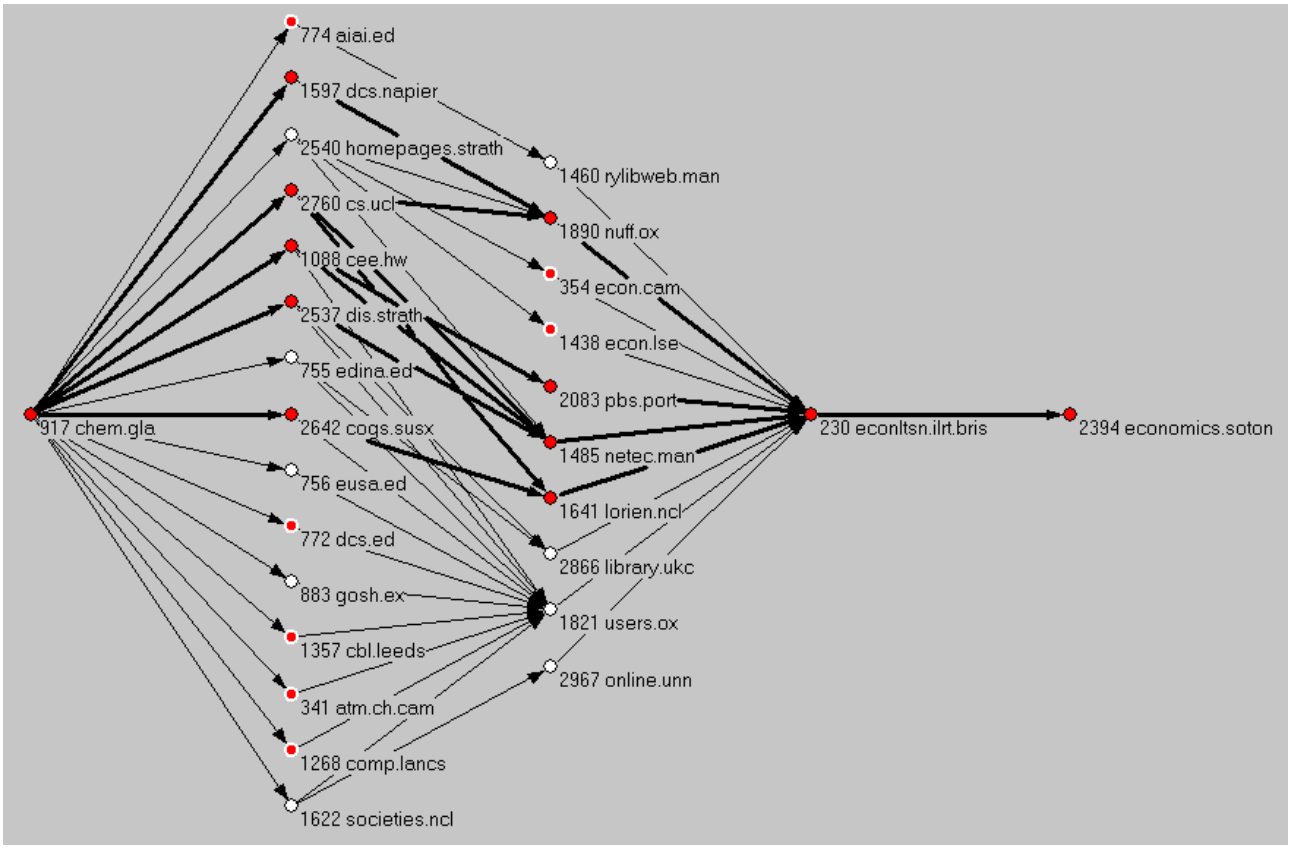


Figure 6.32. Path net NH02. Six link paths in bold contain non-generic subsites only. Generic subsites are marked with white nodes. Excluded subsites on 'generic' link paths are marked with white-bordered red nodes.

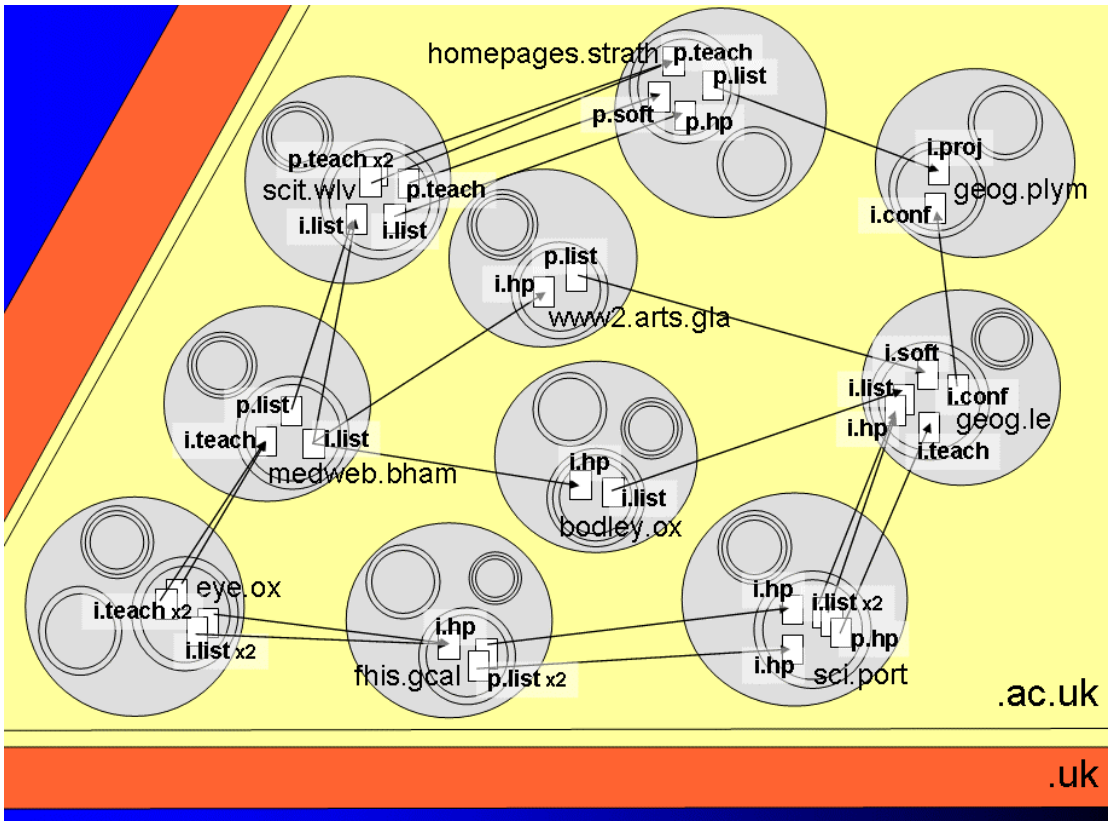


Figure 6.36. Excerpt from path net NH05 with actual links between source pages and target pages.

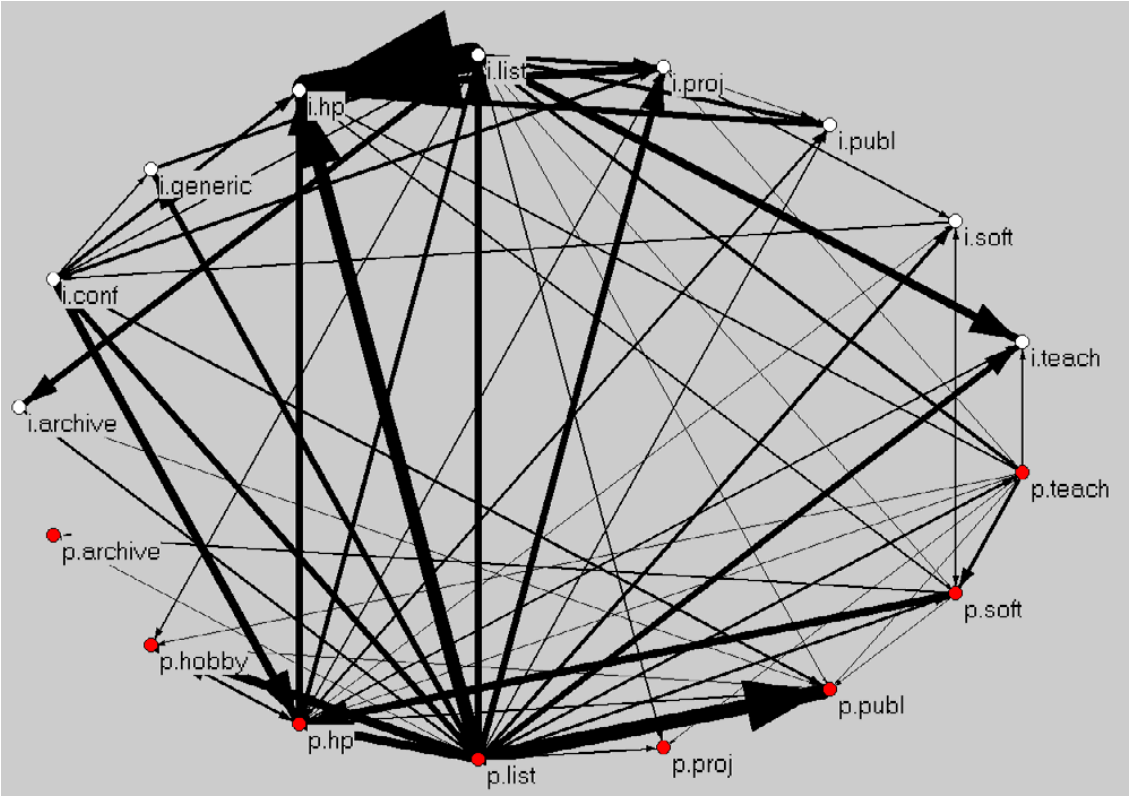


Figure 6.37. A *web of genres.* Genre pairs among 352 followed links. Link width reflects link counts. Due to the Pajek software, thinner reciprocal links are concealed underneath thicker links. Genre selflinks are not shown. White nodes denote institutional meta genres and red personal.

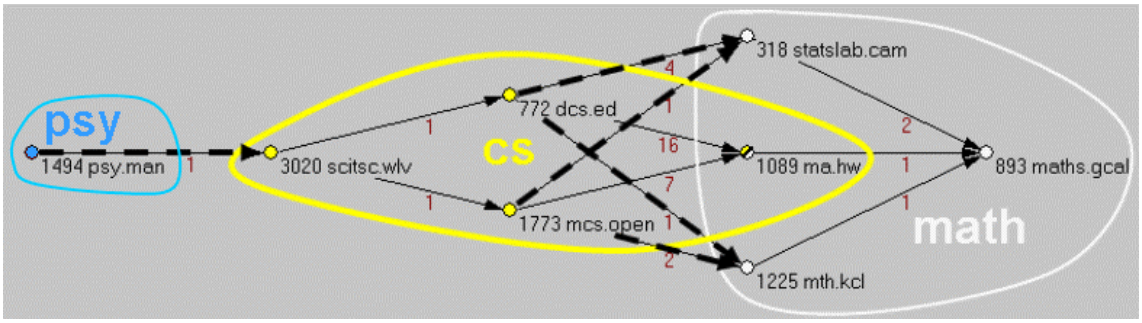


Figure 6.43. Path net HN03 with the enclosed topical areas psychology (psy), computer science (cs) and mathematics (math). Transversal links crossing disciplinary borders are denoted in dashed bold. Counts of page level links are shown.

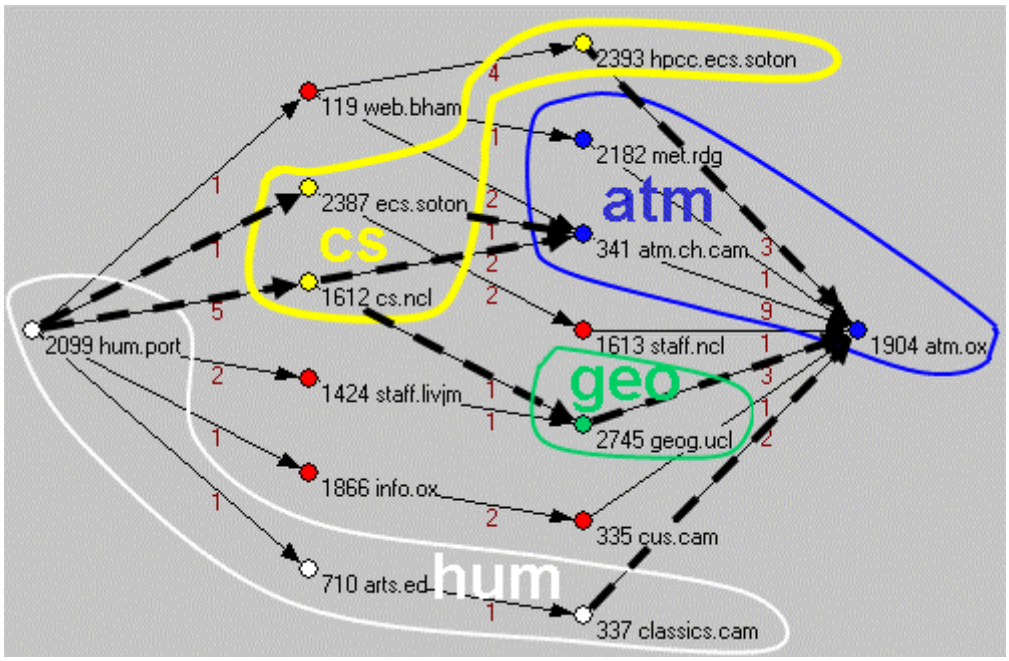


Figure 6.44 & 7.2. Path net HN01 with enclosed topical areas humanities (hum), computer science (cs), geography (geo) and atmospheric sciences (atm). Non-enclosed nodes are generic-type. Transversal links are marked with dashed bold links.

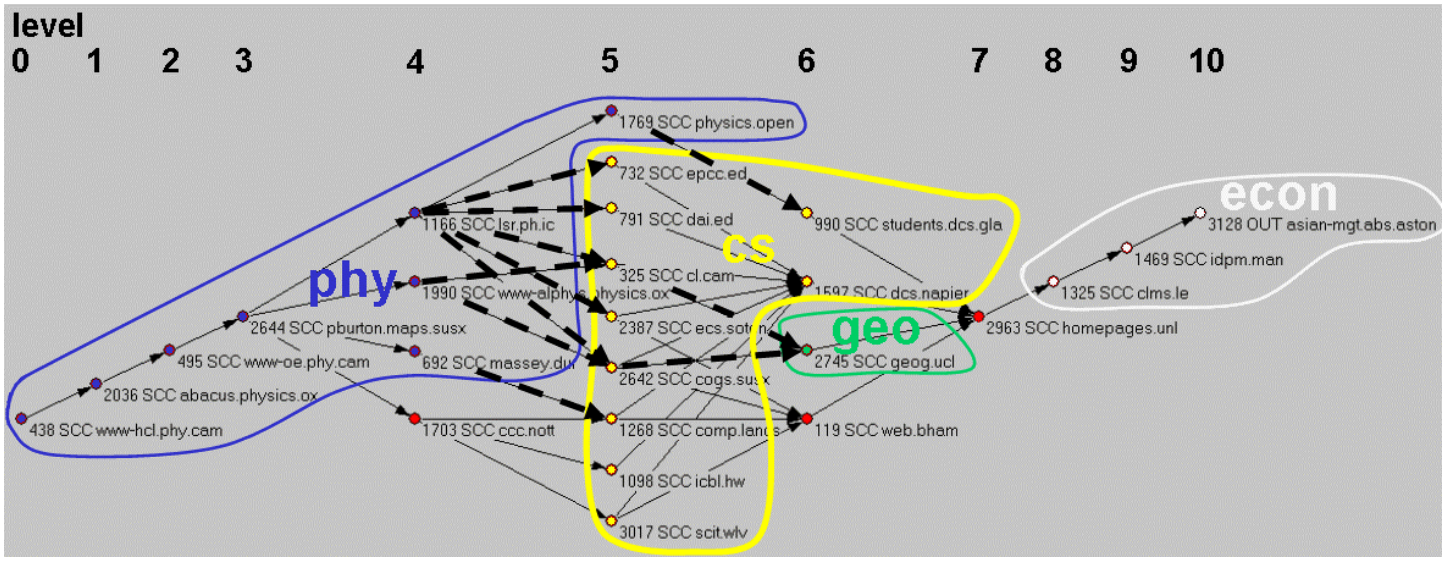


Figure 7.3. Path net containing all shortest link paths of length 10 between node 438 (www-hcl.phy.cam.ac.uk) and node 3128 (asian-mgt.abs.aston.ac.uk) with enclosed topical areas physics (phy), computer science (cs), geography (geo) and economics/management (econ). Non-enclosed nodes are generic-type. Transversal links are marked with dashed bold links. Levels show link distances from start node. Due to limited space, initial and final nodes in the path net are drawn together. See Appendix 6 for affiliations of nodes in the path net.

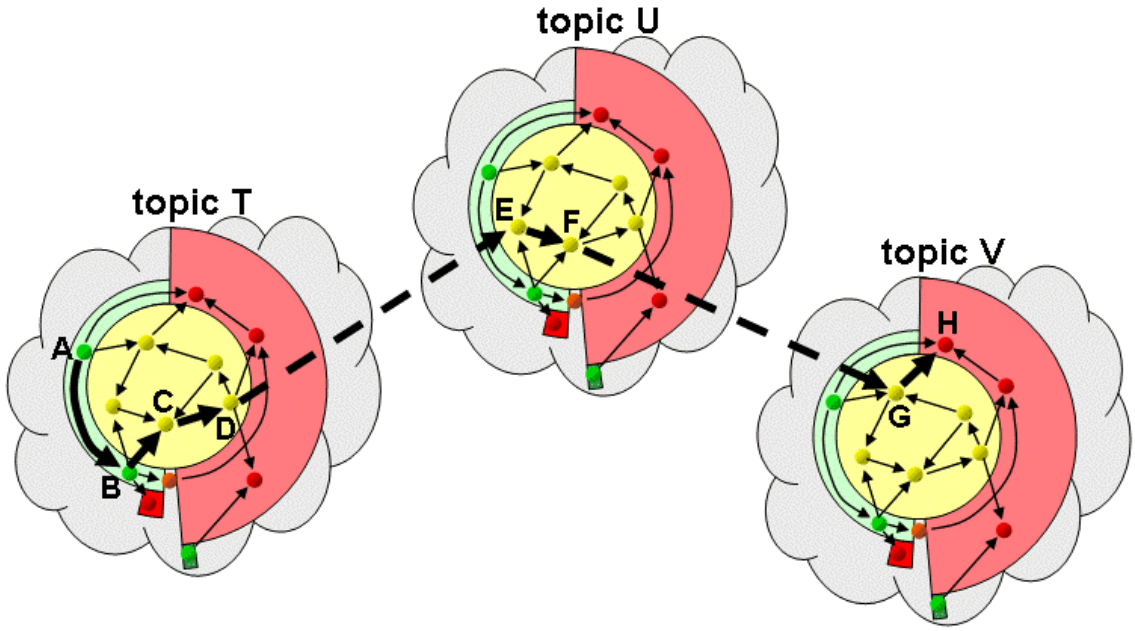


Figure 7.4. Example of shortest link path (bold links) between nodes A and H crossing three topic clusters, each of which with 'corona'-like graph components. Transversal (inter-topic) links are marked with dashed bold links.

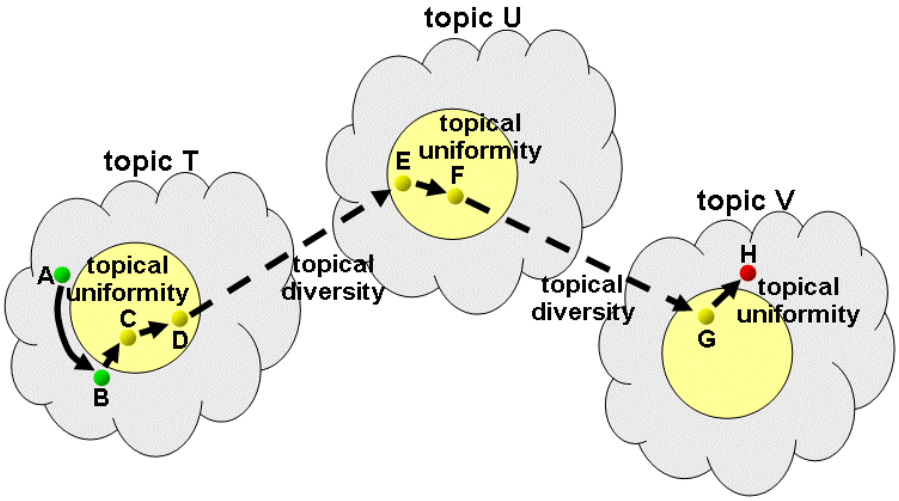


Figure 7.5. Simplified version of Fig. 7.4 with a shortest link path comprising steps of topical uniformity and diversity to reach from node A to H crossing three topic clusters, each of which with a strongly connected component (SCC) denoted by an inner circle. Transversal (inter-topic) links are marked with dashed links.

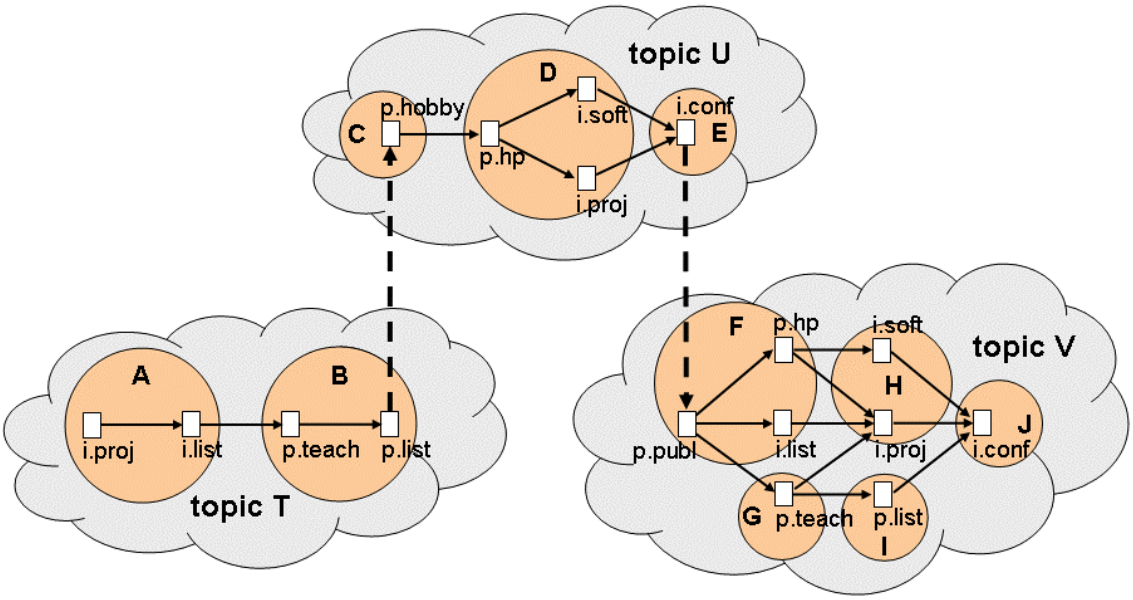


Figure 7.8. Intra-cluster genre drift and inter-cluster topic drift along shortest link paths from web site A in topic cluster T to site J in topic cluster V. Transversal inter-topic links are denoted with dashed bold links.

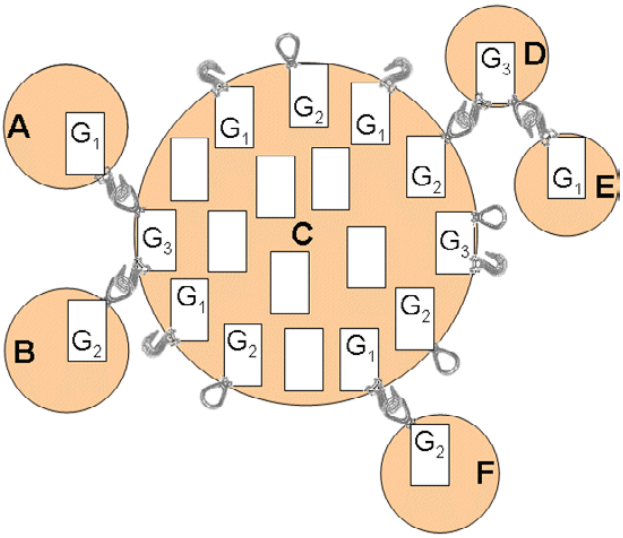


Figure 7.12. Some web page genres may function as outlink-prone hook genres (G1), inlink-prone lug genres (G2), or combined hook&lug genres (G3), here pulling web sites A-F close together.

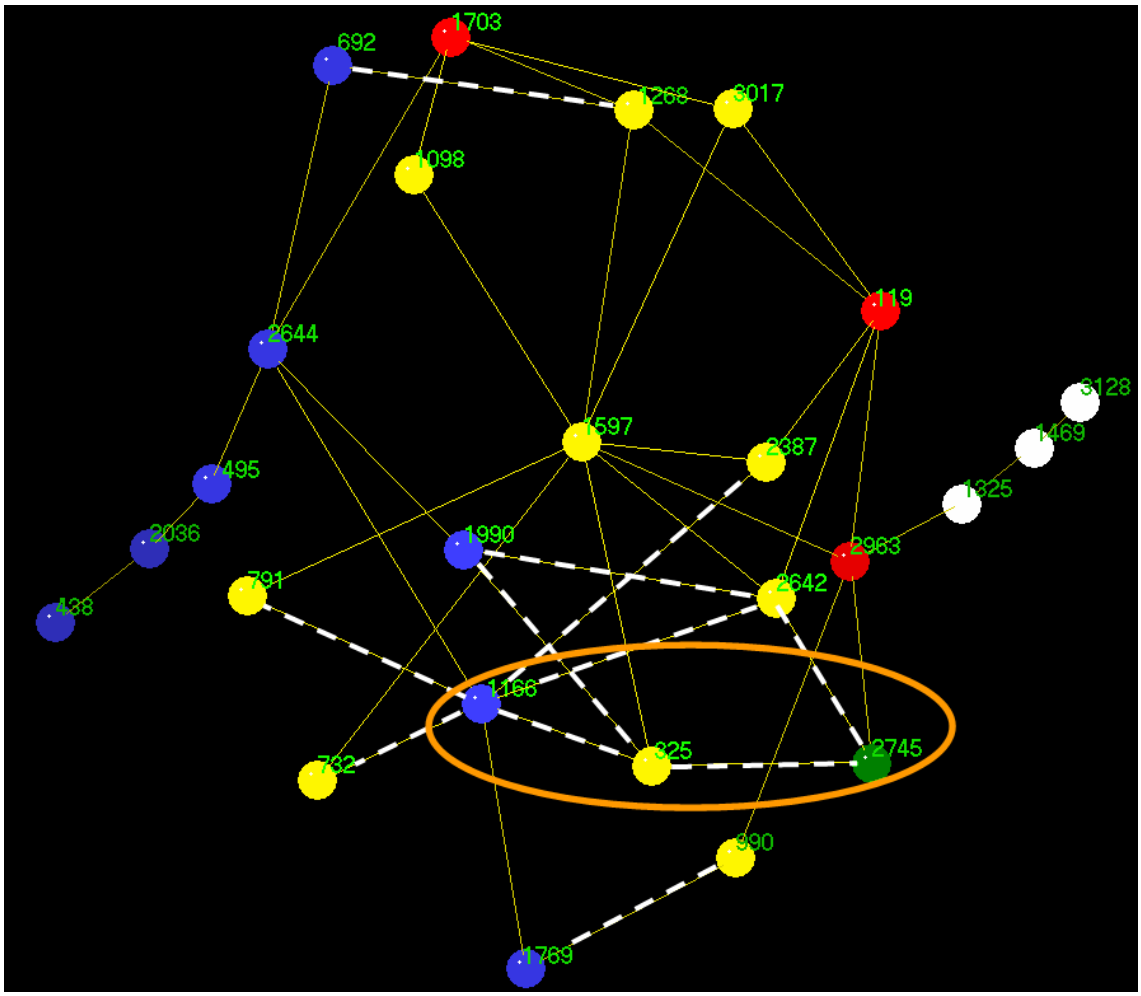


Figure 7.14. 3D visualization made in network software tool *Kinemage* of the same path net as in Fig. 7.3 containing all shortest link paths (length 10) between node 438 (*www-hcl.phy.cam.ac.uk*) and node 3128 (*asian-mgt.abs.aston.ac.uk*). Blue nodes denote physics subsites, yellow computer science, green geography, white economics/management, and red generic-type subsites. Encircled nodes 1168, 325 and 2745 exemplify topical domains pulled together by transversal links. Transversal links are marked with dashed bold lines. See Appendix 6 for affiliations of nodes.

Figure 7.15. Crumpled-up web space with three crumpled-up topic clusters.

Appendices

Appendix 1. UK Higher Education Map

Source: Burden (2003).[91] 'UK Sensitive Map : Universities : Version 5'. Available: *http://www.scit.wlv.ac.uk/ukinfo/uk.map.html* [visited 27.9.2003]

"This map shows all recognised Universities, University Colleges and Higher Education Colleges in the United Kingdom except for residential colleges within universities.

- Red means a University
- Orange means a University sector college
- Green means the UK campus of a foreign institution
- Blue marks the actual location of a town or city
- Purple means a professional, postgraduate or other institution that doesn't fit into the above" categories

The stars indicate town or city names. In many cases, there are several universities or colleges in a town or city. In such cases the star links to the university or college bearing the name of the town or city. Please Note: Due to limited space, universities and sites in London and the South East of England are not shown geographically but appear as a separate list on the right hand side of the map." (ibid.)

[91] "This map was created and is maintained by Peter Burden of the School of Computing and Information Technology of the University of Wolverhampton as a spare time activity" (Burden, 2003: http://www.scit.wlv.ac.uk/ukinfo/uk.map.html).

Appendix 2. UK Universities and Colleges

The list below (Burden, 2001) contains all universities and colleges in the UK Higher Education Map (cf. Appendix 1) as of July 2001. The list was downloaded from the Internet Archive (*www.archive.org*) in order to get an overview over the universities and colleges at the time of the original 2001 web crawl: *http://web.archive.org/web/20010707114102/http://www.scit.wlv.ac.uk/ukinfo/alpha.html* [visited 27.9.2003]

The list is kept intact in its original form. The 109 universities and colleges (cf. Appendix 3) included in the 2001 UK data set are marked in bold and highlighted:

"Universities.

- **University of Aberdeen**
- **University of Abertay , Dundee**
- **Anglia Polytechnic University, Chelmsford**
- **Anglia Polytechnic University, Cambridge**
- **Aston University, Birmingham**
- **The University of Bath**
- **Birmingham University**
- **Bournemouth University**
- **Bradford University**
- **University of Brighton**
- **Bristol University**
- **Brunel University, Uxbridge, West London**
- **University of Buckingham**
- **University of Cambridge**
- **University of Central England , Birmingham**
- **University of Central Lancashire , Preston**
- **City University, Central London**
- **Coventry University**
- Cranfield University

- **Derby University**
- **University of Dundee**
- **Durham University**
- **University of East Anglia, Norwich**
- **University of East London**
- **Edinburgh University**
- **University of Essex, Colchester**
- **Exeter University**
- **Glasgow Caledonian University**
- **University of Glamoragn, Pontypridd**
- **Glasgow University**
- **University of Greenwich, London**
- **London Guildhall University**
- **Heriot Watt University, Edinburgh**
- **University of Hertfordshire, Hatfield**
- **Huddersfield University**
- **University of Hull**
- The University of the Highlands and Islands Project, Inverness
- **Keele University, Staffordshire**
- **University of Kent, Canterbury**
- **Kingston University, South West London**
- **Lancaster University**
- **University of Leeds**
- **Leeds Metropolitan University**
- **University of Leicester**
- **De Montfort University, Leicester**
- **De Montfort University, Bedford**

- **De Montfort University , Lincoln**
- **De Montfort University , Milton-Keynes**
- **University of Lincolnshire and Humberside , Lincoln**
- **University of Lincolnshire and Humberside , Hull**
- **Liverpool University**
- **Liverpool John Moores University**
- University of London Colleges, Schools, Institutes and Teaching Hospitals
    - University of London
    - University of London St.Bartholomew's and the Royal London School of Medicine and Dentistry
    - University of London Birkbeck College
    - **University of London Goldsmiths College**
    - University of London Charing Cross & Westminster Medical School
    - University of London Heythrop College
    - **University of London Imperial College of Science, Technology and Medicine**
    - **University of London King's College**
    - Courtauld Institute of Art, London
    - **University of London Queen Mary and Westfield College**
    - **University of London Royal Holloway**
    - **University of London University College**
    - University of London Imperial College at Wye , [London Link]
    - University of London Imperial College at Wye , [Kent Link]
    - University of London Eastman Dental Institute
    - Institute of Child Health , London
    - University of London Institute of Cancer Research
    - Institute of Neurology , London
    - University of London Institute of Education

  - University of London Institute of Psychiatry
  - University of London Royal College of Physicians
  - The Royal College of Music , London
  - University of London School of Advanced Studies
  - University of London St. George's Hospital Medical School , London SW17
  - University of London, London Business School
  - **University of London School of Economics and Political Science**
  - School of Pharmacy , London
  - University of London Royal Free Hospital School of Medicine
  - University of London Royal Postgraduate Medical School
  - University of London School of Hygiene and Tropical Medicine
  - **University of London School of Oriental and African Studies**
  - University of London School of Slavonic and East European Studies
  - University of London United Medical and Dental Schools
- **Loughborough University**
- **University of Luton**
- **Manchester Metropolitan University**
- **Manchester Metropolitan University, Crewe**
- University of Manchester
  - **University of Manchester**
  - Manchester Business School
  - **University of Manchester Institute of Science and Technology , (UMIST)**
- **Middlesex University, West London**
- **Napier University, Edinburgh**
- **Newcastle University**
- **University of Northumbria , Newcastle**

- **University of Northumbria , Carlisle**
- **University of North London**
- **Nottingham University**
- **Nottingham Trent University**
- **The Open University , Milton Keynes**
- **Oxford University**
- **Oxford Brookes University**
- **Paisley University**
- **Plymouth University**
- **The University of Portsmouth**
- Queen's University of Belfast
    - **Queen's University Belfast**
    - St.Mary's University College, Belfast
    - Stranmillis University College, Belfast
- **Reading University**
- **Robert Gordon University, Aberdeen**
- **St.Andrews University**
- **University of Salford**
- **The University of Sheffield**
- **Sheffield Hallam University**
- **University of Southampton**
- **South Bank University, London**
- **Staffordshire University**
- **Stirling University**
- **The University of Strathclyde , Glasgow**
- **Sunderland University**
- **University of Surrey , Guildford**

- **Sussex University, Brighton**
- **University of Teesside**
- **Thames Valley University, Slough**
- **University of Ulster**
- University of Wales
    - Univerity of Wales
    - **University of Wales Aberystwyth**
    - **University of Wales Bangor**
    - **University of Wales Cardiff**
    - University of Wales College of Medicine Cardiff
    - **University of Wales Lampeter**
    - **University of Wales College Newport**
    - **University of Wales Swansea**
    - **University of Wales Institute, Cardiff**
- **University of Warwick**
- **University of the West of England, Bristol**
- **University of Westminster, London**
- **University of Wolverhampton**
- **University of York**

University sector colleges.

I.e. Institutions which are not universities but run a substantial number of degree courses usually in conjunction with a local university.

- **Bath Spa University College**
- Belfast Royal Hospitals
- Bolton Institute
- Bretton Hall College, Wakefield
- Buckinghamshire Chilterns University College, High Wycombe

- Camborne School of Mines , Cornwall
- Welsh College of Music and Drama , Cardiff
- Trinity College Carmarthen
- Cumbria College of Art and Design, Carlisle
- **Canterbury Christ Church University College**
- Cheltenham & Gloucester College of Higher Education
- Chester College of HE
- **University College Chichester**
- Royal Agricultural College, Cirencester
- Northern School of Contemporary Dance , Leeds
- Central School of Speech & Drama , London
- Dartington College of Arts
- Edinburgh College of Art
- Edge Hill University College, Ormskirk, Lancashire
- Falmouth College of Arts
- Glasgow School of Art
- Guildford College of Further and Higher Education
- **Harper Adams University College, Newport, Shropshire**
- College of Guidance Studies , Hextable, Kent
- Homerton College, Cambridge
- Kent Institute of Art and Design
- King Alfred's College, Winchester
- University College of St.Martin , Lancaster, Ambleside and Carlisle
- Bishop Grosseteste College, Lincoln
- Liverpool Hope University College
- London College of Fashion
- The London College of Printing

- The London Institute , [Chelsea College, London College of Fashion, Camberwell College, Central St.Martins College]
- Moray House Institute of Education, Edinburgh
- Northern College , Aberdeen
- **University College Northampton**
- North East Wales Institute of Higher Education, Wrexham
- Newman College of Higher Education, Birmingham
- Norwich School of Art and Design
- University College of St.Mark & St.John , Plymouth
- **Queen Margaret University College, Edinburgh**
- Ravensbourne College of Design and Communication , London
- Royal College of Nursing
- University of Surrey Roehampton , London SW15
- Royal Academy of Music , London
- The Royal College of Art
- Royal Northern College of Music , Manchester
- Royal Scottish Academy of Music and Drama
- The Royal Veterinary College
- University College Scarborough , [North Riding College]
- The Scottish Agricultural College , Edinburgh, Aberdeen and Ayr
- Heriot Watt University, Borders Campus, Galashiels [formerly Scottish College of Textiles]
- Rose Bruford College, Sidcup, Kent
- Southampton Institue of HE
- Saint Andrew's College, Glasgow
- University of Durham Stockton Campus
- **Surrey Institute of Art and Design**
- Swansea Institute

- Trinity and All Saints University College, Leeds
- Trinity College of Music , London
- St.Mary's College, Twickenham, West London
- College of Ripon and York St.John [University of Leeds]
- University College, Warrington
- The University of Birmingham Westhill , Birmingham
- Westminster College , Oxford
- Wimbledon School of Art
- **University College, Worcester**
- University College Writtle , Chelmsford

International colleges and universities

Foreign institutions with campuses in the United Kingdom

- Huron University USA in London
- The American International University in London , Richmond

Professional and Postgraduate Institutions

- The College of Law
- Henley Management College
- Inns of Court School of Law
- Institute for System Level Integration , Livingston"

(Burden, 2001)

Appendix 3. Included 109 UK universities

domain name	name used on university homepage
abdn.ac.uk	University of Aberdeen
aber.ac.uk	University of Wales, Aberystwyth
anglia.ac.uk	Anglia Polytechnic University, Cambridge & Chelmsford
aston.ac.uk	Aston University, Birmingham
bangor.ac.uk	University of Wales, Bangor
bath.ac.uk	University of Bath
bathspa.ac.uk	Bath Spa University College
bham.ac.uk	University of Birmingham
bournemouth.ac.uk	Bournemouth University
brad.ac.uk	Bradford University
bris.ac.uk	University of Bristol
brookes.ac.uk	Oxford Brookes University
brunel.ac.uk	Brunel University (West London)
bton.ac.uk	University of Brighton
buckingham.ac.uk	University of Buckingham
cam.ac.uk	University of Cambridge
cant.ac.uk	Canterbury Christ Church University College
cf.ac.uk	Cardiff University (University of Wales)
chichester.ac.uk	University College Chichester
city.ac.uk	City University, London
coventry.ac.uk	Coventry University
derby.ac.uk	University of Derby
dmu.ac.uk	De Montfort University, Leicester, Bedford, Milton-Keynes
dundee.ac.uk	University of Dundee
dur.ac.uk	University of Durham
ed.ac.uk	University of Edinburgh
essex.ac.uk	University of Essex
ex.ac.uk	University of Exeter
gcal.ac.uk	Glasgow Caledonian University
gla.ac.uk	University of Glasgow
glam.ac.uk	University of Glamorgan
goldsmiths.ac.uk	Goldsmiths College, University of London
gre.ac.uk	University of Greenwich
harper-adams.ac.uk	Harper Adams University College, Newport, Shropshire
herts.ac.uk	University of Hertfordshire
hud.ac.uk	University of Huddersfield
hull.ac.uk	University of Hull
hw.ac.uk	Heriot-Watt University (Edinburgh)
ic.ac.uk	Imperial College of Science, Technology and Medicine, University of London
kcl.ac.uk	King's College London, University of London
keele.ac.uk	Keele University
king.ac.uk	Kingston University (South West London)
lamp.ac.uk	University of Wales, Lampeter
lancs.ac.uk	Lancaster University
lboro.ac.uk	Loughborough University
le.ac.uk	University of Leicester
leeds.ac.uk	University of Leeds
lgu.ac.uk	London Guildhall University
liv.ac.uk	University of Liverpool
livjm.ac.uk	Liverpool John Moores University
lmu.ac.uk	Leeds Metropolitan University
lse.ac.uk	London School of Economics and Political Science, University of London
luton.ac.uk	University of Luton
man.ac.uk	University of Manchester
mdx.ac.uk	Middlesex University, London
mmu.ac.uk	Manchester Metropolitan University
napier.ac.uk	Napier University, Edinburgh
ncl.ac.uk	University of Newcastle upon Tyne
newport.ac.uk	University of Wales College, Newport
northampton.ac.uk	University College Northampton
nott.ac.uk	University of Nottingham

ntu.ac.uk	Nottingham Trent University
open.ac.uk	Open University, Milton Keynes
ox.ac.uk	University of Oxford
paisley.ac.uk	University of Paisley
plym.ac.uk	University of Plymouth
port.ac.uk	University of Portsmouth
qmced.ac.uk	Queen Margaret University College, Edinburgh
qmw.ac.uk	Queen Mary, University of London
qub.ac.uk	Queen's University Belfast
rdg.ac.uk	University of Reading
rgu.ac.uk	Robert Gordon University (Aberdeen)
rhul.ac.uk	Royal Holloway, University of London
salford.ac.uk	University of Salford
sbu.ac.uk	South Bank University, London
shef.ac.uk	University of Sheffield
shu.ac.uk	Sheffield Hallam University
soas.ac.uk	School of Oriental and African Studies, University of London
soton.ac.uk	University of Southampton
staffs.ac.uk	Staffordshire University
st-and.ac.uk	University of St Andrews
stir.ac.uk	University of Stirling
strath.ac.uk	University of Strathclyde, Glasgow
sunderland.ac.uk	University of Sunderland
surrart.ac.uk	Surrey Institute of Art and Design
surrey.ac.uk	University of Surrey, Guildford
susx.ac.uk	University of Sussex, Brighton
swan.ac.uk	University of Wales, Swansea
tay.ac.uk	University of Abertay, Dundee
tees.ac.uk	University of Teesside, Middlesbrough
tvu.ac.uk	Thames Valley University, Slough
uce.ac.uk	University of Central England, Birmingham
ucl.ac.uk	University College London
uclan.ac.uk	University of Central Lancashire, Preston
uea.ac.uk	University of East Anglia, Norwich
uel.ac.uk	University of East London
ukc.ac.uk	University of Kent at Canterbury
ulh.ac.uk	University of Lincolnshire and Humberside (today: University of Lincoln)
ulst.ac.uk	University of Ulster
umist.ac.uk	University of Manchester Institute of Science and Technology
unl.ac.uk	University of North London (today: London Metropolitan University)
unn.ac.uk	Northumbria University, Newcastle
uwe.ac.uk	University of the West of England, Bristol
uwic.ac.uk	University of Wales Institute at Cardiff
warwick.ac.uk	University of Warwick
wlv.ac.uk	University of Wolverhampton
wmin.ac.uk	University of Westminster, London
worc.ac.uk	University College Worcester
york.ac.uk	University of York

Appendix 4. Variant domain names of 109 UK universities

canonical domain name	variant name 1	variant name 2
abdn.ac.uk	aberdeen.ac.uk	
aber.ac.uk	aberystwyth.ac.uk	
anglia.ac.uk		
aston.ac.uk		
bangor.ac.uk		
bath.ac.uk		
bathspa.ac.uk		
bham.ac.uk	birmingham.ac.uk	
bournemouth.ac.uk		
brad.ac.uk	bradford.ac.uk	
bris.ac.uk	bristol.ac.uk	
brookes.ac.uk	oxford-brookes.ac.uk	
brunel.ac.uk		
bton.ac.uk	brighton.ac.uk	
buckingham.ac.uk		
cam.ac.uk	cambridge.ac.uk	
cant.ac.uk		
cf.ac.uk	cardiff.ac.uk	
chichester.ac.uk		
city.ac.uk		
coventry.ac.uk	cov.ac.uk	
derby.ac.uk		
dmu.ac.uk		
dundee.ac.uk		
dur.ac.uk	durham.ac.uk	
ed.ac.uk	edinburgh.ac.uk	
essex.ac.uk	sx.ac.uk	
ex.ac.uk	exeter.ac.uk	
gcal.ac.uk		
gla.ac.uk	glasgow.ac.uk	
glam.ac.uk	glamorgan.ac.uk	
goldsmiths.ac.uk	gold.ac.uk	
gre.ac.uk	greenwich.ac.uk	
harper-adams.ac.uk	haac.ac.uk	
herts.ac.uk	hertfordshire.ac.uk	
hud.ac.uk	huddersfield.ac.uk	
hull.ac.uk		
hw.ac.uk	heriot-watt.ac.uk	
ic.ac.uk		
kcl.ac.uk		
keele.ac.uk		
king.ac.uk	kingston.ac.uk	
lamp.ac.uk	lampeter.ac.uk	
lancs.ac.uk	lancaster.ac.uk	
lboro.ac.uk	loughborough.ac.uk	
le.ac.uk	leicester.ac.uk	
leeds.ac.uk		
lgu.ac.uk		
liv.ac.uk	liverpool.ac.uk	
livjm.ac.uk		
lmu.ac.uk		
lse.ac.uk		
luton.ac.uk		
man.ac.uk	manchester.ac.uk	mcc.ac.uk
mdx.ac.uk		
mmu.ac.uk		
napier.ac.uk		
ncl.ac.uk	newcastle.ac.uk	
newport.ac.uk		
northampton.ac.uk		
nott.ac.uk	nottingham.ac.uk	

ntu.ac.uk		
open.ac.uk		
ox.ac.uk	oxford.ac.uk	
paisley.ac.uk		
plym.ac.uk	plymouth.ac.uk	
port.ac.uk	portsmouth.ac.uk	
qmced.ac.uk	qmuc.ac.uk	
qmw.ac.uk		
qub.ac.uk		
rdg.ac.uk	reading.ac.uk	
rgu.ac.uk		
rhul.ac.uk	rhbnc.ac.uk	
salford.ac.uk		
sbu.ac.uk		
shef.ac.uk	sheffield.ac.uk	
shu.ac.uk		
soas.ac.uk		
soton.ac.uk	southampton.ac.uk	
staffs.ac.uk	staffordshire.ac.uk	
st-and.ac.uk	st-andrews.ac.uk	
stir.ac.uk	stirling.ac.uk	
strath.ac.uk	strathclyde.ac.uk	
sunderland.ac.uk	sund.ac.uk	
surrart.ac.uk		
surrey.ac.uk		
susx.ac.uk	sussex.ac.uk	
swan.ac.uk	swansea.ac.uk	
tay.ac.uk	abertay.ac.uk	
tees.ac.uk	teeside.ac.uk	
tvu.ac.uk		
uce.ac.uk		
ucl.ac.uk		
uclan.ac.uk		
uea.ac.uk		
uel.ac.uk		
ukc.ac.uk		
ulh.ac.uk	humber.ac.uk	lincoln.ac.uk
ulst.ac.uk	ulster.ac.uk	
umist.ac.uk		
unl.ac.uk		
unn.ac.uk	northumbria.ac.uk	
uwe.ac.uk		
uwic.ac.uk		
warwick.ac.uk		
wlv.ac.uk	wolverhampton.ac.uk	
wmin.ac.uk	westminster.ac.uk	
worc.ac.uk		
york.ac.uk		

Appendix 5. Subsites per university

rank	university domain name	university	# sub-sites	%
1	cam	University of Cambridge	582	7,59
2	ox	University of Oxford	515	6,72
3	ed	University of Edinburgh	321	4,19
4	gla	University of Glasgow	297	3,87
5	man	University of Manchester	266	3,47
6	ic	Imperial College	251	3,27
7	soton	University of Southampton	249	3,25
8	ucl	University College London	227	2,96
9	open	Open University, Milton Keynes	174	2,27
10	strath	University of Strathclyde, Glasgow	171	2,23
11	bris	University of Bristol	167	2,18
12	bham	University of Birmingham	164	2,14
13	leeds	University of Leeds	162	2,11
14	ncl	University of Newcastle upon Tyne	149	1,94
15	umist	University of Manchester Institute of Science and Technology	125	1,63
16	nott	University of Nottingham	119	1,55
17	ulst	University of Ulster	93	1,21
18	salford	University of Salford	90	1,17
19	qmw	Queen Mary, University of London	89	1,16
20	lboro	Loughborough University	87	1,13
21	lancs	Lancaster University	82	1,07
22	qub	Queen's University Belfast	81	1,06
23	rdg	University of Reading	78	1,02
24	mmu	Manchester Metropolitan University	76	0,99
25	st-and	University of St Andrews	76	0,99
26	dmu	De Montfort University	75	0,98
27	hw	Heriot-Watt University	73	0,95
28	stir	University of Stirling	67	0,87
29	essex	University of Essex	65	0,85
30	swan	University of Wales, Swansea	65	0,85
31	plym	University of Plymouth	63	0,82
32	port	University of Portsmouth	63	0,82
33	shef	University of Sheffield	63	0,82
34	warwick	University of Warwick	62	0,81
35	bangor	University of Wales, Bangor	61	0,80
36	dur	University of Durham	61	0,80
37	gcal	Glasgow Caledonian University	61	0,80
38	surrey	University of Surrey, Guildford	61	0,80
39	cf	Cardiff University	60	0,78
40	kcl	King's College London	60	0,78
41	york	University of York	58	0,76
42	mdx	Middlesex University, London	57	0,74
43	anglia	Anglia Polytechnic University	55	0,72
44	brad	Bradford University	55	0,72
45	dundee	University of Dundee	52	0,68
46	liv	University of Liverpool	52	0,68
47	ntu	Nottingham Trent University	52	0,68
48	unn	Northumbria University, Newcastle	52	0,68
49	sunderland	University of Sunderland	50	0,65
50	le	University of Leicester	49	0,64
51	susx	University of Sussex	49	0,64
52	uea	University of East Anglia	48	0,63
53	aston	Aston University, Birmingham	47	0,61
54	rhul	Royal Holloway, University of London	47	0,61
55	bton	University of Brighton	46	0,60
56	napier	Napier University, Edinburgh	46	0,60
57	sbu	South Bank University, London	46	0,60

58	city	City University, London	45	0,59
59	ukc	University of Kent at Canterbury	45	0,59
60	lse	London School of Economics and Political Science	44	0,57
61	ex	University of Exeter	43	0,56
62	paisley	University of Paisley	41	0,53
63	shu	Sheffield Hallam University	41	0,53
64	herts	University of Hertfordshire	40	0,52
65	abdn	University of Aberdeen	38	0,50
66	hull	University of Hull	37	0,48
67	goldsmiths	Goldsmiths College, University of London	36	0,47
68	newport	University of Wales College, Newport	35	0,46
69	livjm	Liverpool John Moores University	34	0,44
70	wlv	University of Wolverhampton	34	0,44
71	coventry	Coventry University	33	0,43
72	uwe	University of the West of England, Bristol	33	0,43
73	wmin	University of Westminster	33	0,43
74	derby	University of Derby	32	0,42
75	tees	University of Teesside, Middlesbrough	31	0,40
76	gre	University of Greenwich	29	0,38
77	aber	University of Wales, Aberystwyth	28	0,37
78	uce	University of Central England, Birmingham	27	0,35
79	bournemouth	Bournemouth University	26	0,34
80	king	Kingston University (South West London)	25	0,33
81	staffs	Staffordshire University	25	0,33
82	tay	University of Abertay, Dundee	24	0,31
83	hud	University of Huddersfield	22	0,29
84	uwic	University of Wales Institute at Cardiff	22	0,29
85	keele	Keele University	21	0,27
86	brookes	Oxford Brookes University	20	0,26
87	glam	University of Glamorgan	20	0,26
88	unl	University of North London	20	0,26
89	bath	University of Bath	18	0,23
90	tvu	Thames Valley University, Slough	16	0,21
91	uel	University of East London	16	0,21
92	ulh	University of Lincolnshire and Humberside	16	0,21
93	brunel	Brunel University (West London)	15	0,20
94	soas	School of Oriental and African Studies, University of London	11	0,14
95	cant	Canterbury Christ Church University College	10	0,13
96	rgu	Robert Gordon University (Aberdeen)	10	0,13
97	luton	University of Luton	9	0,12
98	qmced	Queen Margaret University College, Edinburgh	9	0,12
99	lmu	Leeds Metropolitan University	7	0,09
100	uclan	University of Central Lancashire	7	0,09
101	lamp	University of Wales, Lampeter	6	0,08
102	lgu	London Guildhall University	6	0,08
103	worc	University College Worcester	6	0,08
104	northampton	University College Northampton	4	0,05
105	bathspa	Bath Spa University College	3	0,04
106	buckingham	University of Buckingham	2	0,03
107	chichester	University College Chichester	1	0,01
108	harper-adams	Harper Adams University College, Newport, Shropshire	1	0,01
109	surrart	Surrey Institute of Art and Design	0	0,00
	total		**7669**	**100,00**

Appendix 6. Affiliations in path net with longest path length 10

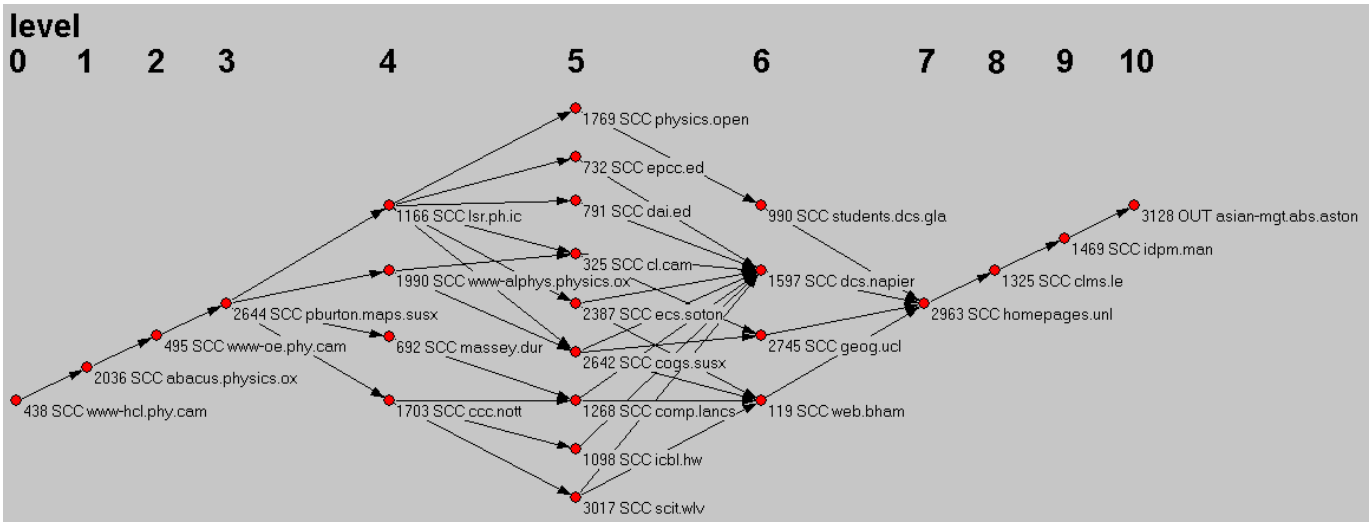


Figure 5.12. Section 5.3.1. All shortest link paths (path length 10) between node 438 (*www-hcl.phy.cam.ac.uk*) and node 3128 (*asian-mgt.abs.aston.ac.uk*).

path net level	id	com-pon-ent	short domain name	affiliation
0	438	SCC	www-hcl.phy.cam	Hitachi Cambridge Laboratory, University of Cambridge
1	2036	SCC	abacus.physics.ox	Quantum Optoelectronics Group, Clarendon Laboratory, Oxford
2	495	SCC	www-oe.phy.cam	Optoelectronics Cavendish Laboratory, University of Cambridge
3	2644	SCC	pburton.maps.susx	Sussex Centre for Optical and Atomic Physics
4	692	SCC	massey.dur	Atomic & Molecular Physics Group, University of Durham
4	1166	SCC	lsr.ph.ic	Laser optics & spectroscopy group, The Blackett Laboratory, Imperial College
4	1703	SCC	ccc.nott	Official webpages of University of Nottingham
4	1990	SCC	www-alphys.physics.ox	Atomic & Laser Physics ["sub-department"], Oxford
5	325	SCC	cl.cam	Computer Lab (Computer Science department), University of Cambridge
5	732	SCC	epcc.ed	Edinburgh Parallel Computing Centre, University of Edinburgh
5	791	SCC	dai.ed	Department of Artificial Intelligence, University of Edinburgh
5	1098	SCC	icbl.hw	Institute for Computer Based Learning, Heriot-Watt University
5	1268	SCC	comp.lancs	Computing Department, Lancaster University
5	1769	SCC	physics.open	Department of Physics and Astronomy, Open University
5	2387	SCC	ecs.soton	Department of Electronics and Computer Science, University of Southampton
5	2642	SCC	cogs.susx	School of Cognitive and Computing Sciences (COGS), University of Sussex
5	3017	SCC	scit.wlv	School of Computing and Information Technology, Univ. of Wolverhampton
6	119	SCC	web.bham	Personal webpages of University of Birmingham
6	990	SCC	students.dcs.gla	Student pages, Dept. of Computing Science, University of Glasgow
6	1597	SCC	dcs.napier	School of Computing, Napier University
6	2745	SCC	geog.ucl	Department of Geography, University College of London
7	2963	SCC	homepages.unl	Personal homepages, University of North London
8	1325	SCC	clms.le	Centre for Labour Market Studies, University of Leicester
9	1469	SCC	idpm.man	Institute for Development Policy and Management, University of Manchester
10	3128	OUT	asian-mgt.abs.aston	The Aston Centre for Asian Business and Management, Aston Business School

Appendix 7. Sample of 189 SCC subsites by topic

Table sorted by topic groups: 'hum/soc' (A-E) and 'nat/tech' (F-L), cf. Section 6.2.

id	short domain name	meta topic	topic group	subsite topic	subsite genre	subsite affiliation
755	edina.ed	Generic	-	Collaborative National uni service	National uni service	Edinburgh Data and Information Access, Univ. of Edinburgh
400	members.emma.cam	Generic	-	College homepage	College homepage	Emmanuel College, Univ. of Cambridge
637	mk.dmu	Generic	-	College homepage	College homepage	Milton Keynes Campus, De Montfort Univ.
1819	wadham.ox	Generic	-	College homepage	College homepage	Wadham College, Univ. of Oxford
2002	trinity.ox	Generic	-	College homepage	College homepage	Trinity College, Univ. of Oxford
2008	home.jesus.ox	Generic	-	College homepage	College homepage	Jesus College, Univ. of Oxford
279	edec.brookes	Generic	-	Learning technology	National uni service	Electronic Design Education Consortium, Oxford Brookes Univ.
897	wblt.gcal	Generic	-	Learning technology	Uni service	Web-Based Learning and Teaching, Department of Learning and Educational Development, Glasgow Caledonian Univ.
958	iteu.gla	Generic	-	Learning technology	Uni service	Information Technology Education Unit, Univ. of Glasgow
824	omni.bus.ed	Generic	-	Learning technology	Uni service	Omni.bus facility (supports teaching applications and WWW services), Univ. of Edinburgh
1545	ilrs.mdx	Generic	-	Library & Learning Service	Library & Learning Service	Information and Learning Resource Services, Middlesex Univ.
2692	iserv.tay	Generic	-	Library & Learning Service	Library & Learning Service	Information Services, Univ. of Abertay Dundee
2693	learn5.tay	Generic	-	Library & Learning Service	Library & Learning Service	Information Services, Univ. of Abertay Dundee
33	libweb.anglia	Generic	-	Library service	Library service	Univ. Library Service, Anglia Polytechnic Univ.
66	library.bangor	Generic	-	Library service	Library service	Library & Information Service, Anglia Polytechnic Univ.
294	silver.bton	Generic	-	Library service	Library service	Information Services, Univ. of Brighton
607	library.coventry	Generic	-	Library service	Library service	Lanchester Library, Coventry Univ.
857	libwww.essex	Generic	-	Library service	Library service	Albert Sloman Library, Univ. of Essex
924	lib.gla	Generic	-	Library service	Library service	Glasgow Univ. Library
3012	library.warwick	Generic	-	Library service	Library service	Library, Univ. of Warwick
1895	libsun1.jr2.ox	Generic	-	Library service	Library service	Eprint server, Univ. of Oxford
31	union.aber	Generic	-	Students' union	Students' union homepage	Guild of Students, Univ. of Wales, Aberystwyth
756	eusa.ed	Generic	-	Students' union	Students' union homepage	Edinburgh Univ. Students' Association
1051	uhsu.herts	Generic	-	Students' union	Students' union homepage	Students' Union, Univ. of Hertfordshire
1308	www-lsu.lboro	Generic	-	Students' union	Students' union homepage	Students' Union, Univ. of Loughborough
2144	quis.qub	Generic	-	Students' union	Students' union homepage	Clubs and Societies, Queen's Univ. of Belfast
1703	ccc.nott	Generic	-	Uni homepage	Uni homepage	Univ. of Nottingham
1780	www3.open	Generic	-	Uni homepage	Uni homepage	Open Univ.
2220	www2.rhul	Generic	-	Uni homepage	Uni homepage	Royal Holloway, Univ. of London
625	staff.dmu	Generic	-	Uni service	Uni service	De Montfort Univ.
825	visres.ed	Generic	-	Uni service	Uni service	Visual Resources, Univ. of Edinburgh
1489	net.man	Generic	-	Uni service	Uni service	Manchester Computing - Communications, Operations & Systems, Univ. of Manchester
2348	careers.soton	Generic	-	Uni service: careers advisory	Uni service	Careers Advisory Service, Univ. of Southampton

2916	eltc.umist	Generic	-	Uni service: language training	Uni service	English Language Teaching Centre, Univ. of Manchester Institute of Science and Technology
736	caad.ed	Hum/Soc	A	Architecture	Dept homepage	Department of Architecture, Univ. of Edinburgh
1043	lih.gre	Hum/Soc	A	Architecture: Landscape architecture	Collaborative project homepage	LIH Landscape Information Hub (Landscape Architecture, Design and Planning), Univ. of Greenwich
2981	media.uwe	Hum/Soc	A	Art & Media	Research group homepage	Digital Media Research Centre, Faculty of Art, Media and Design, Univ. of the West of England
1547	cea.mdx	Hum/Soc	A	Arts	Centre homepage	Lansdown Centre for Electronic Arts, School of Arts, Middlesex Univ.
110	artsweb.bham	Hum/Soc	A	Arts & Humanities	Faculty homepage	Arts and Humanities dept.s, Univ. of Birmingham
2099	hum.port	Hum/Soc	A	Hum./Soc.	Faculty homepage	Faculty of Humanities and Social Sciences, Univ. of Portsmouth
1444	fmg.lse	Hum/Soc	B	Business	Research group homepage	Financial Markets Group, London School of Economics and Political Science
1279	lums.lancs	Hum/Soc	B	Business	School homepage	Lancaster Univ. Management School
1546	mubs.mdx	Hum/Soc	B	Business	School homepage	Middlesex Univ. Business School
1735	nbs.ntu	Hum/Soc	B	Business	School homepage	Nottingham Business School, Univ. of Nottingham
2128	econ.qmw	Hum/Soc	B	Economics	Dept homepage	Department of Economics, Queen Mary, Univ. of London
2394	economics.soton	Hum/Soc	B	Economics	Dept homepage	Department of Economics, Univ. of Southampton
1445	is.lse	Hum/Soc	B	Economics: Information systems	Dept homepage	Department of Information Systems, London School of Economics and Political Science
217	law.bris	Hum/Soc	B	Law	Faculty homepage	Faculty of Law, Univ. of Bristol
120	iel.bham	Hum/Soc	B	Law	Institute homepage	The Institute of European Law, Univ. of Birmingham
1849	oiprc.ox	Hum/Soc	B	Law/Economics/Info.Policy	Centre homepage	Oxford Intellectual Property Research Centre, Univ. of Oxford
1952	politics.ox	Hum/Soc	B	Political science	Dept homepage	Department of Politics and International Relations, Univ. of Oxford
2067	politics.plym	Hum/Soc	B	Political science	Dept homepage	Department of Politics, Univ. of Plymouth
74	weblife.bangor	Hum/Soc	C	Education: Learning technology research	Research project homepage	School of Education, Univ. of Wales, Bangor
2543	hilt.cdlr.strath	Hum/Soc	C	Library & Information Science	Research project homepage	High-Level Thesaurus project, Centre for Digital Library Research, Andersonian Library, Univ. of Strathclyde
2581	business.dis.strath	Hum/Soc	C	Library & Information Science	Research project homepage	Business Information and the Internet, Department of Information Science, Univ. of Strathclyde
109	www-clg.bham	Hum/Soc	C	Linguistics	Centre homepage	Centre for Corpus Linguistics, Univ. of Birmingham
871	speech.essex	Hum/Soc	C	Linguistics	Research group homepage	Speech Group, Department of Language and Linguistics, Univ. of Essex
2370	arch.soton	Hum/Soc	D	Archaeology	Dept homepage	Department of Archaeology, Univ. of Southampton
364	www-mcdonald.arch.cam	Hum/Soc	D	Archaeology	Institute homepage	The McDonald Institute for Archaeological Research, Univ. of Cambridge
2110	lime.qmced	Hum/Soc	D	Ethnography	Research project homepage	Living Memory Project (LiMe), Queen Margaret Univ. College, Edinburgh
2068	geog.plym	Hum/Soc	D	Geography	Dept homepage	Department of Geography, School of Geography and Geology, Univ. of Plymouth
2227	gg.rhul	Hum/Soc	D	Geography	Dept homepage	Department of Geography, Royal Holloway, Univ. of London
691	www-ibru.dur	Hum/Soc	D	Geography: International law	Research group homepage	International Boundaries Research Unit, Department of Geography, Univ. of Durham
199	epi.bris	Hum/Soc	E	Medicine: Social Medicine	Dept homepage	Department of Social Medicine, Univ. of Bristol

998	medusa.psy.gla	Hum/Soc	E	Psychology	Dept homepage	Department of Psychology, Univ. of Glasgow
1494	psy.man	Hum/Soc	E	Psychology	Dept homepage	Department of Psychology, Univ. of Manchester
2777	rizzo.psychol.ucl	Hum/Soc	E	Psychology: Statistics	Teaching resource pages	UCL psychology statistics demonstrations, Univ. College London
1448	booth.lse	Hum/Soc	E	Sociology: history	Online archive homepage	Charles Booth Online Archive, London School of Economics and Political Science
2361	chst.soton	Hum/Soc	E	Sociology: human service technology	Centre homepage	Centre for Human Service Technology, Dept.of Social Work Studies, Univ. of Southampton
2673	medsocbsa.swan	Hum/Soc	E	Sociology: Medical Sociology	Research group homepage	Medical Sociology Group, Department of Sociology, Univ. of Warwick
23	irs.aber	Nat/Tech	F	Agriculture	Institute homepage	Institute of Rural Studies, Univ. of Wales, Aberystwyth
2781	slamdunk.geol.ucl	Nat/Tech	F	Earth sciences	Personal homepages	Slamdunk, Department of Geological Sciences, Univ. College London
245	palaeo.gly.bris	Nat/Tech	F	Earth sciences	Research group homepage	Palaeontology Research Group, Department of Earth Sciences, Univ. of Bristol
1966	teachserv.earth.ox	Nat/Tech	F	Earth sciences: learning technology	Teaching resource pages	Earth Sciences Teaching Network, Univ. of Oxford
1617	cluwrr.ncl	Nat/Tech	F	Environmental studies	Centre homepage	Centre for Land Use and Water Resources Research (CLUWRR), Univ. of Newcastle upon Tyne
1195	iceo.ic	Nat/Tech	F	Environmental studies	Centre homepage: identical with id 1187	Environment Office, Imperial College London
1187	gse.ic	Nat/Tech	F	Environmental studies	Centre homepage: identical with id 1195	Environment Office, Imperial College London
2271	ties.salford	Nat/Tech	F	Environmental studies	Institute homepage	Telford Institute of Environmental Systems, Univ. of Salford
2468	ruwpa.st-and	Nat/Tech	F	Environmental studies	Research group homepage	RUWPA, Centre for Research into Ecological and Environmental Modelling (CREEM), Univ. of St Andrews
1375	medalus.leeds	Nat/Tech	F	Environmental studies	Research project homepage	MEDALUS (Mediterranean Desertification and Land Use), Univ. of Leeds
1358	env.leeds	Nat/Tech	F	Environmental studies	School homepage	School of the Environment, Univ. of Leeds
2852	envam1.env.uea	Nat/Tech	F	Environmental studies	Server standard page	School of Environmental Sciences & School of Mathematics, Univ. of East Anglia
10	oceanlab.abdn	Nat/Tech	F	Zoology: Marine ecology	Centre homepage	Ocean Research Lab, Department of Zoology, Univ. of Aberdeen
2458	smub.st-and	Nat/Tech	F	Zoology: Marine Mammal Biology	Research group homepage	Sea Mammal Research Unit, Gatty Marine Laboratory, Univ. of St Andrews
1883	bioch.ox	Nat/Tech	G	Biochemistry	Dept homepage	Department of Biochemistry, Univ. of Oxford
970	biochem.gla	Nat/Tech	G	Biochemistry	Division homepage	Division of Biochemistry & Molecular Biology, Univ. of Glasgow
3059	ysbl.york	Nat/Tech	G	Biochemistry: structural biology	Lab homepage	York Structural Biology Laboratory, Dept. of Chemistry, Univ. of York
3008	globin.bio.warwick	Nat/Tech	G	Bioscience	Research group resource pages	Brunel Bioinformatics Group, Institute for Cancer Genetics and Pharmagenomics, Department of Biological Sciences, Brunel Univ.
129	biosciences.bham	Nat/Tech	G	Bioscience	School homepage	School of Biosciences, Univ. of Birmingham
200	bio.bris	Nat/Tech	G	Bioscience	School homepage	School of Biological Sciences, Univ. of Bristol
316	bio.cam	Nat/Tech	G	Bioscience	School homepage	School of the Biological Sciences, Univ. of Cambridge
1530	teaching-biomed.man	Nat/Tech	G	Bioscience: learning technology	Teaching resource pages	School of Biological Sciences and the Faculty of Medicine, Dentistry & Nursing, Univ. of Manchester
1844	medicine.ox	Nat/Tech	G	Medicine	Division homepage	Medical Sciences Division, Univ. of Oxford

1710	holmes.cancres.nott	Nat/Tech	G	Medicine: Cancer research	Personal resource pages	Cancer Research Laboratories, School of Pharmaceutical Science, Univ. of Nottingham
349	wbic.cam	Nat/Tech	G	Medicine: Clinical Medicine	Centre homepage	Wolfson Brain Imaging Centre, School of Clinical Medicine, Univ. of Cambridge
2771	ilo.ucl	Nat/Tech	G	Medicine: Laryngology & Otology	Institute homepage	Institute of Laryngology & Otology, Univ. College London
1548	hebes.mdx	Nat/Tech	G	Medicine: Nursing, Environmental...	School homepage	School of Health, Biological and Environmental Sciences at Middlesex Univ.
1885	eye.ox	Nat/Tech	G	Medicine: Ophthalmology	Dept homepage	Department of Ophthalmology, Oxford Univ.
763	orthopaedic.ed	Nat/Tech	G	Medicine: Orthopaedics	Dept homepage	Department of Orthopaedic Surgery, Univ. of Edinburgh
2795	physiol.ucl	Nat/Tech	G	Medicine: Physiology	Dept homepage	Physiology Department, Univ. College London
2828	madeira.physiol.ucl	Nat/Tech	G	Medicine: Physiology	Lab resource pages	Neurophysiology Research lab, Department of Physiology, Univ. College London
344	mrc-cpe.cam	Nat/Tech	G	Medicine: Proteins	Centre homepage	Centre for Protein Engineering, Univ. of Cambridge (+ MRC, Medical Research Council)
345	scop.mrc-lmb.cam	Nat/Tech	G	Medicine: Proteins	Research group homepage	SCOP (Structural Classification of Proteins), Univ. of Cambridge
2165	tess.pt.qub	Nat/Tech	G	Medicine: Quantitative Pathology	Lab resource pages	Quantitative Pathology Web Server, Queen's Univ. Belfast
649	mediate.dmu	Nat/Tech	G	Medicine: Radiology	Research project homepage	MEDIATE (Medical Image Description And Training Environment), De Montfort Univ.
409	perch.cimr.cam	Nat/Tech	G	Medicine: Structural Medicine	Centre homepage	Structural Medicine, Department of Haematology in the Univ. of Cambridge + Cambridge Institute for Medical Research (CIMR)
945	vir.gla	Nat/Tech	G	Medicine: Virology	Division homepage	Division of Virology, Univ. of Glasgow
1701	vsms.nott	Nat/Tech	G	Pharmacology: molecular science	School homepage	The Virtual School of Molecular Sciences, The Department of Pharmaceutical Sciences, Univ. of Nottingham
2515	cn.stir	Nat/Tech	G	Psychology: Computational neuroscience	Research group homepage	Computational Neuroscience, Department of Psychology, Univ. of Stirling
1120	ps.ic	Nat/Tech	H	Chemical engineering	Centre homepage	Centre for Process Systems Engineering, Imperial College London
2734	chemeng.ucl	Nat/Tech	H	Chemical engineering	Dept homepage	Department of Chemical Engineering, Univ. College London
917	chem.gla	Nat/Tech	H	Chemistry	Dept homepage	Department of Chemistry, Univ. of Glasgow
1235	ch.kcl	Nat/Tech	H	Chemistry	Dept homepage	Chemistry Department, King's College London
208	chm.bris	Nat/Tech	H	Chemistry	Research group resource pages	School of Chemistry, Univ. of Bristol
227	dougal.chm.bris	Nat/Tech	H	Chemistry	Research group resource pages	School of Chemistry, Univ. of Bristol
422	ccdc.cam	Nat/Tech	H	Chemistry: Crystallography	Centre homepage	Cambridge Crystallographic Data Centre, Department of Chemistry, Univ. of Cambridge
1206	sc.ic	Nat/Tech	H	Materials science	Centre homepage	Centre for Electronic Materials & Devices, The Blackett Laboratory [Physics dept], Imperial College of Science, Technology and Medicine
365	msm.cam	Nat/Tech	H	Materials science	Dept homepage	Department of Materials Science and Metallurgy, Univ. of Cambridge
1879	materials.ox	Nat/Tech	H	Materials science	Dept homepage	Department of Materials, Univ. of Oxford
2095	sun1.sms.port	Nat/Tech	I	Astronomy	N/A in Internet Archive	Institute of Cosmology and Gravitation, Univ. of Plymouth
979	astro.gla	Nat/Tech	I	Astronomy	Research group homepage	Astronomy & Astrophysics Group, Department of Physics & Astronomy, Univ. of Glasgow
2382	integral.soton	Nat/Tech	I	Astronomy	Research project homepage	Southampton Integral Web Pages (Integral = ESA satellite), Univ. of

						Southampton
209	phy.bris	Nat/Tech	I	Physics	Dept homepage	Department of Physics, Univ. of Bristol
414	phy.cam	Nat/Tech	I	Physics	Dept homepage	Department of Physics, Univ. of Cambridge
1112	phywww.phy.hw	Nat/Tech	I	Physics	Intranet	Internal network page, Department of Physics, Heriot-Watt Univ.
2031	www-teaching.physics.ox	Nat/Tech	I	Physics	Teaching resource pages	Physics Practical Course, Physics Department, Univ. of Oxford
1794	yan.open	Nat/Tech	I	Physics & Astronomy	Dept homepage	Department of Physics and Astronomy, Open Univ.
2465	star-www.st-and	Nat/Tech	I	Physics & Astronomy	School homepage	School of Physics and Astronomy, Univ. of St Andrews
1904	atm.ox	Nat/Tech	I	Physics: Atmospheric, Oceanic and Planetary Physics	Centre homepage	Atmospheric, Oceanic and Planetary Physics, Department of Physics, Univ. of Oxford
2553	cnqo.phys.strath	Nat/Tech	I	Physics: Computational Nonlinear & Quantum Optics	Research group homepage	Computational Nonlinear & Quantum Optics Group, Department of Physics and Applied Physics, Univ. of Strathclyde
1529	reynolds.ph.man	Nat/Tech	I	Physics: Liquid crystal	Research group homepage	Nonlinear and Liquid Crystal Physics Group, Manchester Centre for Nonlinear Dynamics, Univ. of Manchester
290	nuclphys.eng.bton	Nat/Tech	I	Physics: Nuclear physics	Research group homepage	Nuclear Physics Research Group, Univ. of Brighton
1493	arthur.ph.man	Nat/Tech	I	Physics: Nuclear physics	Research group homepage	Manchester Nuclear Physics Group, Univ. of Manchester
950	ppewww.ph.gla	Nat/Tech	I	Physics: Particle physics	Research group homepage	Particle Physics Experimental (PPE) Group, Department of Physics and Astronomy, Univ. of Glasgow
439	poco.phy.cam	Nat/Tech	I	Physics: Polymers & Colloids	Research group homepage	Polymers & Colloids Group, Department of Physics, Univ. of Cambridge
2036	abacus.physics.ox	Nat/Tech	I	Physics: Quantum Optoelectronics	Research group homepage	Quantum Optoelectronics Group, Department of Physics, Oxford
1159	sp.ph.ic	Nat/Tech	I	Physics: Space & atmospheric physics	Research group homepage	Space & Atmospheric Physics Group, Physics Department, Imperial College London
2799	tampa.phys.ucl	Nat/Tech	I	Physics: Theoretical Atomic and Molecular Physics	Research group homepage	Theoretical Atomic and Molecular Physics and Astrophysics Group, Department of Physics and Astronomy, Univ. College London
893	maths.gcal	Nat/Tech	J	Mathematics	Dept homepage	Mathematics Department, Glasgow Caledonian Univ.
1122	ma.ic	Nat/Tech	J	Mathematics	Dept homepage	Department of Mathematics, Imperial College London
1412	maths.liv	Nat/Tech	J	Mathematics	Dept homepage	Department of Mathematical Sciences, Univ. of Liverpool
2218	ma.rhul	Nat/Tech	J	Mathematics	Dept homepage	Department of Mathematics, Royal Holloway, Univ. of London
2113	maths.qmw	Nat/Tech	J	Mathematics	School homepage	School of Mathematical Sciences, Queen Mary, Univ. of London
475	can.dpmms.cam	Nat/Tech	J	Mathematics/Statistics	Dept homepage	Department of Pure Mathematics & Mathematical Statistics, Univ. of Cambridge
815	pgprospectus.maths.ed	Nat/Tech	J	Mathematics/Statistics	Postgraduate prospectus homepage	Postgraduate Prospectus, Department of Mathematics and Statistics, Univ. of Edinburgh
2492	mcs.st-and	Nat/Tech	J	Mathematics/Statistics	School homepage	School of Mathematics and Statistics, Univ. of St Andrews
2467	www-vortex.mcs.st-and	Nat/Tech	J	Mathematics: Applied math.	Research group homepage	Waves, Vortices and Turbulence Research Group, Applied Mathematics, Institute of Mathematics, Univ. of St Andrews
1130	metric.ma.ic	Nat/Tech	J	Mathematics: Learning technology	Teaching resource pages	Mathematics Education Technology Research, Imperial College
933	rss2001.gla	Nat/Tech	J	Statistics	Conference homepage	RSS2001, International Conference of the Royal Statistical Society, Univ. of Glasgow
255	aer.bris	Nat/Tech	K	Engineering: Aerospace	Dept homepage	Department of Aerospace Engineering, Univ. of Bristol
2286	pse.sbu	Nat/Tech	K	Engineering: Building technology	Research group homepage	Construction Management Research Group, Faculty of the Built Environment, South Bank Univ.

2277	surveying.salford	Nat/Tech	K	Engineering: Building technology	School homepage	School of Construction and Property Management, Univ. of Salford
2542	cad.strath	Nat/Tech	K	Engineering: Computer aided design	Centre homepage	CAD Centre, Design Manufacture & Engineering Management, Univ. of Strathclyde
168	xanadu.bournemouth	Nat/Tech	K	Engineering: Design + Computing	School resource pages	Applied Psychology and Computing Support Server, School of Design, Engineering & Computing, Univ. of Bournemouth
1188	washer.ee.ic	Nat/Tech	K	Engineering: Electrical and Electronic Engineering	Startpage without content: "Under construction"	Department of Electrial and Electronic Engineering, Imperial College, Univ. of London
179	vader.brad	Nat/Tech	K	Engineering: Electronic and Electrical Engineering: Cryptography	Research group homepage	Cryptography & Computer Communications Security Group, Department of Electronic and Electrical Engineering, Bradford Univ.
2044	icte.paisley	Nat/Tech	K	Engineering: Electronic Engineering & Physics: Information & Communications Technologies in Education	Research group homepage	ICTE Research Group (Information & Communications Technologies in Education), Dept of Electronic Engineering & Physics, Univ. of Paisley
2050	eep.paisley	Nat/Tech	K	Engineering: Electronical engineering + Physics	Dept homepage	Dept of Electronic Engineering & Physics, School of Information and Communication Technologies, Univ. of Paisley
189	manuel.brad	Nat/Tech	K	Engineering: Electronics & Telecommunciation	Dept homepage	Department of Electronics & Telecommunications, Univ. of Bradford
241	lara.enm.bris	Nat/Tech	K	Engineering: Engineering mathematics: Artificial Intelligence	Personal resource pages	Computational Intelligence Group, Department of Engineering Mathematics, Univ. of Bristol
1040	fseg.gre	Nat/Tech	K	Engineering: Fire safety	Research group homepage	Fire Safety Engineering Group (FSEG), Univ. of Greenwich
330	www-sigproc.eng.cam	Nat/Tech	K	Engineering: Signal processing	Research group homepage	Signal Processing Laboratory, Department of Engineering, Univ. of Cambridge
388	www-control.eng.cam	Nat/Tech	K	Engineering: System Engineering	Research group homepage	Control Group, Department of Engineering, Univ. of Cambridge
2396	trg.soton	Nat/Tech	K	Engineering: Transportation Research	Research group homepage	Transportation Research Group, Department of Civil and Environmental Engineering, Univ. of Southampton
1762	telematics.open	Nat/Tech	K	Telematics	Dept homepage	Department of Telematics, Faculty of Technology, Open Univ.
3034	cpc.wmin	Nat/Tech	L	Computer science	Centre homepage	Centre for Parallel Computing, Cavendish School of Computing Science, Univ. of Westminster
325	cl.cam	Nat/Tech	L	Computer science	Dept homepage	Univ. of Cambridge Computer Lab (Computer Science Teaching and Research department)
1612	cs.ncl	Nat/Tech	L	Computer science	Dept homepage	Department of Computing Science, Univ. of Newcastle
1792	linux.open	Nat/Tech	L	Computer science	Journal homepage	Linux Gazette, hosted at Open Univ.
2155	strule.cs.qub	Nat/Tech	L	Computer science	Personal resource pages	the Strule server, School of Computer Science, Queen's Univ. of Belfast
1550	cs.mdx	Nat/Tech	L	Computer science	School homepage	School of Computing Science, Middlesex Univ.
2448	soc.staffs	Nat/Tech	L	Computer science	School homepage	School of Computing, Staffordshire Univ.
3020	scitsc.wlv	Nat/Tech	L	Computer science	School homepage	School of Computing and Information Technology, Univ. of Wolverhampton
2163	main.cs.qub	Nat/Tech	L	Computer science	Server standard page	School of Computer Science, Queen's Univ. of Belfast
2078	area51.upsu.plym	Nat/Tech	L	Computer science	Students' union society homepage	TermiSoc, Student Union Computing Society, Univ. of Plymouth
2387	ecs.soton	Nat/Tech	L	Computer science + Electronics	Dept homepage	Department of Electronics and Computer Science, Univ. of Southampton

644	cms.dmu	Nat/Tech	L	Computer science + Engineering	Faculty homepage	Faculty of Computing Sciences and Engineering, De Montfort Univ.
2342	kingfisher.cms.shu	Nat/Tech	L	Computer science + Management science	School resource pages	School of Computing and Management Sciences, Sheffield Hallam Univ.
276	wwwcms.brookes	Nat/Tech	L	Computer science + Mathematics	School homepage	School of Computing and Mathematical Sciences, Oxford Brookes Univ.
2240	nicve.salford	Nat/Tech	L	Computer science: Virtual environments	Centre homepage	Centre for Virtual Environments, Univ. of Salford
776	icsa.informatics.ed.ac	Nat/Tech	L	Informatics	Institute homepage	Institute for Computing Systems Architecture, Division of Informatics, Univ. of Edinburgh
2899	starform.infj.ulst	Nat/Tech	L	Informatics	Personal homepage	Bill McMillan, Senior Lecturer, Faculty of Informatics, Univ. of Ulster
592	web.soi.city	Nat/Tech	L	Informatics	School homepage	School of Informatics, City Univ., London
957	jetai.mcg.gla	Nat/Tech	L	Informatics + Psychology	Conference homepage	JETAI Conference ["Journées Européennes des Techniques Avancées de L'Informatique"], Multimedia Communications Group, Department of Psychology, Univ. of Glasgow
774	aiai.ed	Nat/Tech	L	Informatics: Artificial intelligence	Institute homepage	Artificial Intelligence Applications Institute, Centre for Intelligent Systems, School of Informatics, Univ. of Edinburgh
775	cogsci.ed	Nat/Tech	L	Informatics: Cognitive science	Centre homepage	Cognitive Science, Division of Informatics, Univ. of Edinburgh
790	ltg.ed	Nat/Tech	L	Informatics: Language technology	Research group homepage	Language Technology Group, Institute for Communicating and Collaborative Systems, Division of Informatics, Univ. of Edinburgh
2902	crossconnect.infm.ulst	Nat/Tech	L	Informatics: Learning technology	Research project homepage	Faculty of Informatics, Univ. of Ulster
2322	ir.shef	Nat/Tech	L	Library & Information Science	Research group homepage	Information Retrieval Group, Department of Information Studies, Univ. of Sheffield
id	**short domain name**	**meta topic**	**topic group**	**subsite topic**	**subsite genre**	**subsite affiliation**

Appendix 8. Glänzel & Schubert (2003): classification scheme

Glänzel & Schubert (2003). 'A new classification scheme of science fields and subfields designed for scientometric evaluation purposes'. *Scientometrics*, 56(3): 357-367.
Table 1. Fields and subfields of sciences, social sciences and arts & humanities. pp. 359-360.

1. AGRICULTURE & ENVIRONMENT
- A1 Agricultural Science & Technology
- A2 Plant & Soil Science & Technology
- A3 Environmental Science & Technology
- A4 Food & Animal Science & Technology

2. BIOLOGY (ORGANISMIC & SUPRAORGANISMIC LEVEL)
- Z1 Animal Sciences
- Z2 Aquatic Sciences
- Z3 Microbiology
- Z4 Plant Sciences
- Z5 Pure & Applied Ecology
- Z6 Veterinary Sciences

3. BIOSCIENCES (GENERAL, CELLULAR & SUBCELLULAR BIOLOGY; GENETICS)
- B0 Multidisciplinary Biology
- B1 Biochemistry/Biophysics/Molecular Biology
- B2 Cell Biology
- B3 Genetics & Developmental Biology

4. BIOMEDICAL RESEARCH
- R1 Anatomy & Pathology
- R2 Biomaterials & Bioengineering
- R3 Experimental/Laboratory Medicine
- R4 Pharmacology & Toxicology
- R5 Physiology

5. CLINICAL AND EXPERIMENTAL MEDICINE I (GENERAL & INTERNAL MEDICINE)
- I1 Cardiovascular & Respiratory Medicine
- I2 Endocrinology & Metabolism
- I3 General & Internal Medicine
- I4 Hematology & Oncology
- I5 Immunology

6. CLINICAL AND EXPERIMENTAL MEDICINE II (NON-INTERNAL MEDICINE SPECIALTIES)
- M1 Age & Gender Related Medicine
- M2 Dentistry
- M3 Dermatology/Urogenital System
- M4 Ophthalmology/Otolaryngology
- M5 Paramedicine
- M6 Psychiatry & Neurology
- M7 Radiology & Nuclear Medicine
- M8 Rheumatology/Orthopedics
- M9 Surgery

7. NEUROSCIENCE & BEHAVIOR
- N1 Neurosciences & Psychopharmacology
- N2 Psychology & Behavioral Sciences

8. CHEMISTRY
- C0 Multidisciplinary Chemistry
- C1 Analytical, Inorganic & Nuclear Chemistry
- C2 Applied Chemistry & Chemical Engineering
- C3 Organic & Medicinal Chemistry
- C4 Physical Chemistry
- C5 Polymer Science
- C6 Materials Science

9. PHYSICS
- P0 Multidisciplinary Physics
- P1 Applied Physics
- P2 Atomic, Molecular & Chemical Physics
- P3 Classical Physics
- P4 Mathematical & Theoretical Physics
- P5 Particle & Nuclear Physics
- P6 Physics of Solids, Fluids And Plasmas

10. GEOSCIENCES & SPACE SCIENCES
- G1 Astronomy & Astrophysics
- G2 Geosciences & Technology
- G3 Hydrology/Oceanography
- G4 Meteorology/Atmospheric & Aerospace Science & Technology
- G5 Mineralogy & Petrology

11. ENGINEERING
- E1 Computer Science/Information Technology
- E2 Electrical & Electronic Engineering
- E3 Energy & Fuels
- E4 General & Traditional Engineering

12. MATHEMATICS
- H1 Applied Mathematics
- H2 Pure Mathematics

13. SOCIAL SCIENCES I (GENERAL, REGIONAL & COMMUNITY ISSUES)
- S1 Education & Information
- S2 General, Regional & Community Issues

14. SOCIAL SCIENCES II (ECONOMICAL & POLITICAL ISSUES)
- O1 Economics, Business & Management
- O2 History, Politics & Law

15. ARTS & HUMANITIES
- U1 Arts & Literature
- U2 Language & Culture
- U3 Philosophy & Religion

Appendix 9. HERO (2001). RAE: Units of Assessment

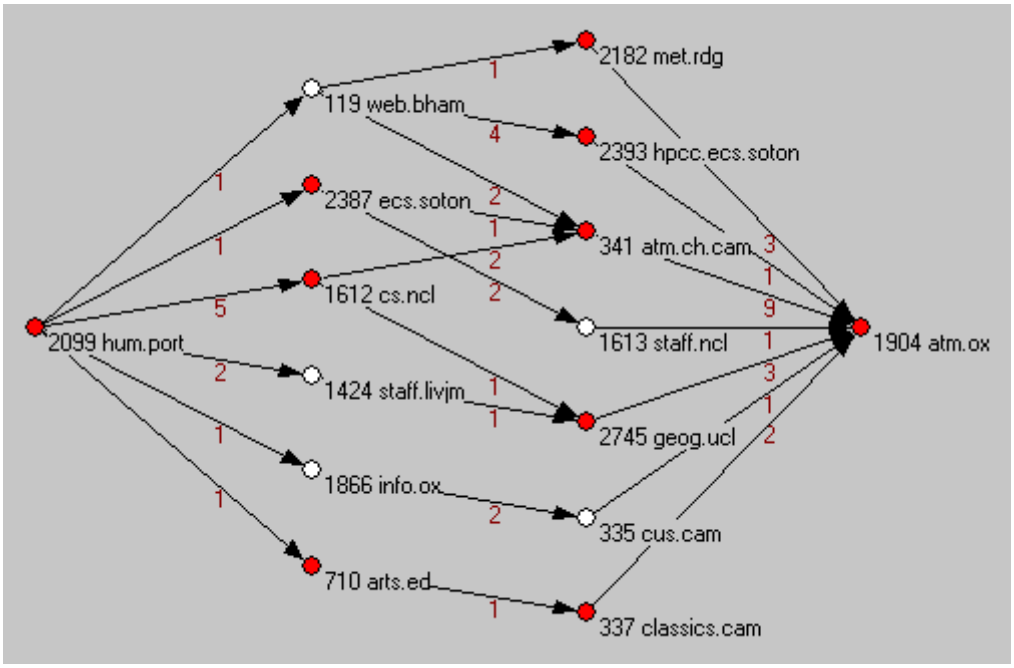
Higher Education & Research Opportunities in the UK. 2001 Research Assessment Exercise: The Outcome : Units of Assessment.
Available: *http://www.hero.ac.uk/rae/Pubs/4_01/section4.htm*

	I Medical and Biological Sciences
1	Clinical Laboratory Sciences
2	Community-based Clinical Subjects
3	Hospital-based Clinical Subjects
4	Clinical Dentistry
5	Pre-clinical Studies
6	Anatomy
7	Physiology
8	Pharmacology
9	Pharmacy
10	Nursing
11	Other Studies and Professions Allied to Medicine
12	Discontinued for 2001
13	Psychology
14	Biological Sciences
15	Agriculture
16	Food Science and Technology
17	Veterinary Science
	II Physical Sciences and Engineering
18	Chemistry
19	Physics
20	Earth Sciences
21	Environmental Science
22	Pure Mathematics
23	Applied Mathematics
24	Statistics and Operational Research
25	Computer Science
26	General Engineering
27	Chemical Engineering
28	Civil Engineering
29	Electrical and Electronic Engineering
30	Mechanical, Aeronautical and Manufacturing Engineering
31	Mineral and Mining Engineering
32	Metallurgy and Materials

	III Social Sciences
33	Built Environment
34	Town and Country Planning
35	Geography
36	Law
37	Anthropology
38	Economics and Econometrics
39	Politics and International Studies
40	Social Policy and Administration
41	Social Work
42	Sociology
43	Business and Management Studies
44	Accounting and Finance
68	Education
69	Sports-related Subjects
	IV Area Studies and Languages
45	American Studies
46	Middle Eastern and African Studies
47	Asian Studies
48	European Studies
49	Celtic Studies
50	English Language and Literature
51	French
52	German, Dutch and Scandinavian Languages
53	Italian
54	Russian, Slavonic and East European Languages
55	Iberian and Latin American Languages
56	Linguistics
	V Arts and Humanities
57	Classics, Ancient History, Byzantine and Modern Greek
58	Archaeology
59	History
60	History of Art, Architecture and Design
61	Library and Information Management
62	Philosophy
63	Theology, Divinity and Religious Studies
64	Art and Design
65	Communication, Cultural and Media Studies
66	Drama, Dance and Performing Arts
67	Music

Appendix 10. 10 path nets including subsite affiliations

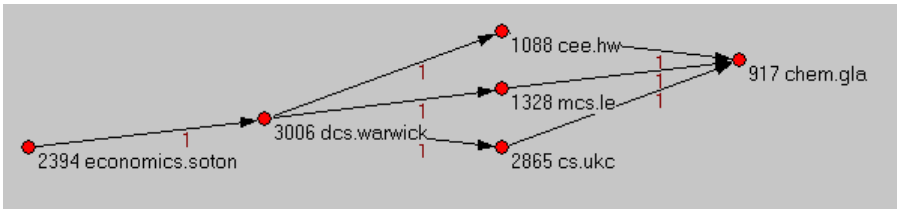
See Appendix 11 for more extensive node data from the 10 path nets.



Path net HN01. White nodes denote generic-type subsites. Counts of *page level* links are shown.

path level	id	short domain name	affiliation
0	2099	hum.port	Faculty of Humanities and Social Sciences, Univ. of Portsmouth
1	119	web.bham	Personal web pages at Univ. of Birmingham
1	710	arts.ed	Faculty of Arts, Univ. of Edinburgh
1	1424	staff.livjm	Staff web pages, Liverpool John Moores University
1	1612	cs.ncl	School of Computing Science, Univ. of Newcastle upon Tyne
1	1866	info.ox	official Oxford University web pages
1	2387	ecs.soton	Dept of Electronics and Computer Science, Univ. of Southampton
2	335	cus.cam	The Central Unix Service (CUS), Univ. of Cambridge
2	337	classics.cam	Faculty of Classics and Museum of Classical Archaeology, Univ. of Cambridge
2	341	atm.ch.cam	Centre for Atmospheric Science, Univ. of Cambridge
2	1613	staff.ncl	Staff web pages, Univ. of Newcastle upon Tyne
2	2182	met.rdg	Dept of Meteorology, Univ. of Reading
2	2393	hpcc.ecs.soton	High Performance Computing, Dept of Electronics and Computer Science, Univ. of Southampton
2	2745	geog.ucl	Dept of Geography, University College of London
3	1904	atm.ox	Atmospheric, Oceanic and Planetary Physics, Dept of Physics, Univ. of Oxford

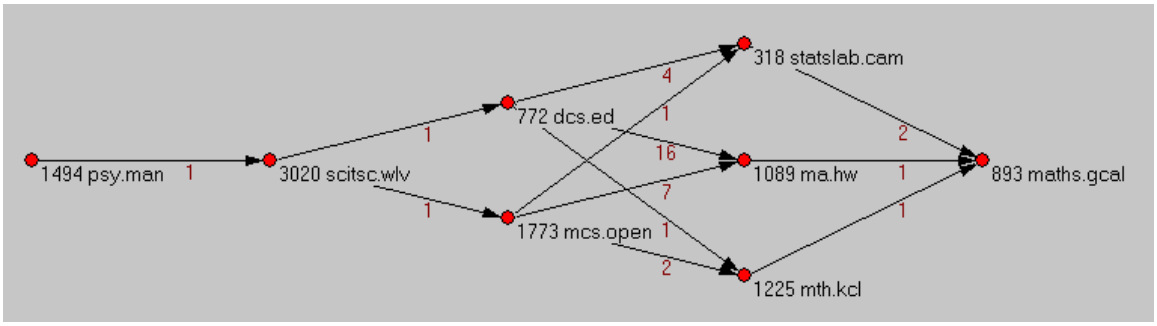
Path net HN01. Affiliations. All subsites were visited in this path net.



Path net HN02. Counts of *page level* links are shown.

path level	id	short domain name	affiliation
0	2394	economics.soton	Dept of Economics, Univ. of Southampton
1	3006	dcs.warwick	Dept of Computer Science, Univ. of Warwick
2	1088	cee.hw	Dept of Computing and Electrical Engineering, Heriot-Watt University
2	1328	mcs.le	Dept of Mathematics and Computer Science, Univ. of Leicester
2	2865	cs.ukc	Computing Laboratory (Dept.), Univ. of Kent, Canterbury
3	917	chem.gla	Dept of Chemistry, Univ. of Glasgow

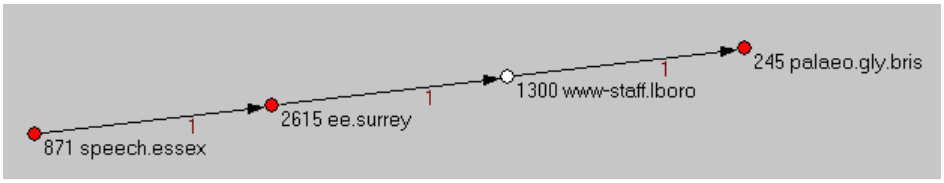
Path net HN02. Affiliations. All subsites were visited in this path net.



Path net HN03. Counts of *page level* links are shown.

path level	id	short domain name	affiliation
0	1494	psy.man	Dept of Psychology, Univ. of Manchester
1	3020	scitsc.wlv	School of Computing and Information Technology, Univ. of Wolverhampton
2	772	dcs.ed	Dept of Computer Science, Univ. of Edinburgh
2	1773	mcs.open	Faculty of Mathematics and Computing, Open University
3	318	statslab.cam	Statistical Laboratory, Faculty of Mathematics, Univ. of Cambridge.
3	1089	ma.hw	School of Mathematical and Computer Sciences, Heriot-Watt University
3	1225	mth.kcl	Mathematics Dept, King's College, Univ. of London
4	893	maths.gcal	Mathematics Dept, Glasgow Caledonian University

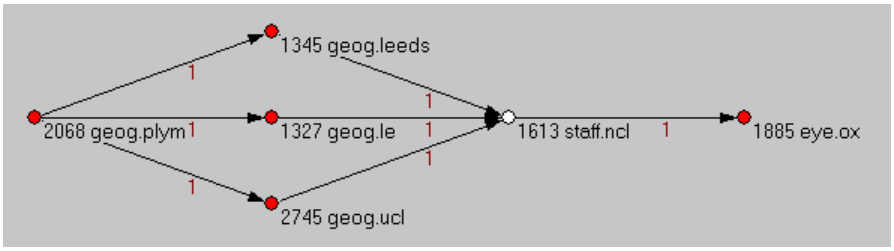
Path net HN03. Affiliations. All subsites were visited in this path net.



Path net HN04. White node denotes generic-type subsite. Counts of *page level* links are shown.

path level	id	short domain name	affiliation
0	871	speech.essex	Speech Group, Dept of Language and Linguistics, Univ. of Essex
1	2615	ee.surrey	Dept of Electrical Engineering, Univ. of Surrey
2	1300	www-staff.lboro	Staff web pages, Loughborough University
3	245	palaeo.gly.bris	Palaeontology Research Group, Dept of Earth Sciences, Univ. of Bristol

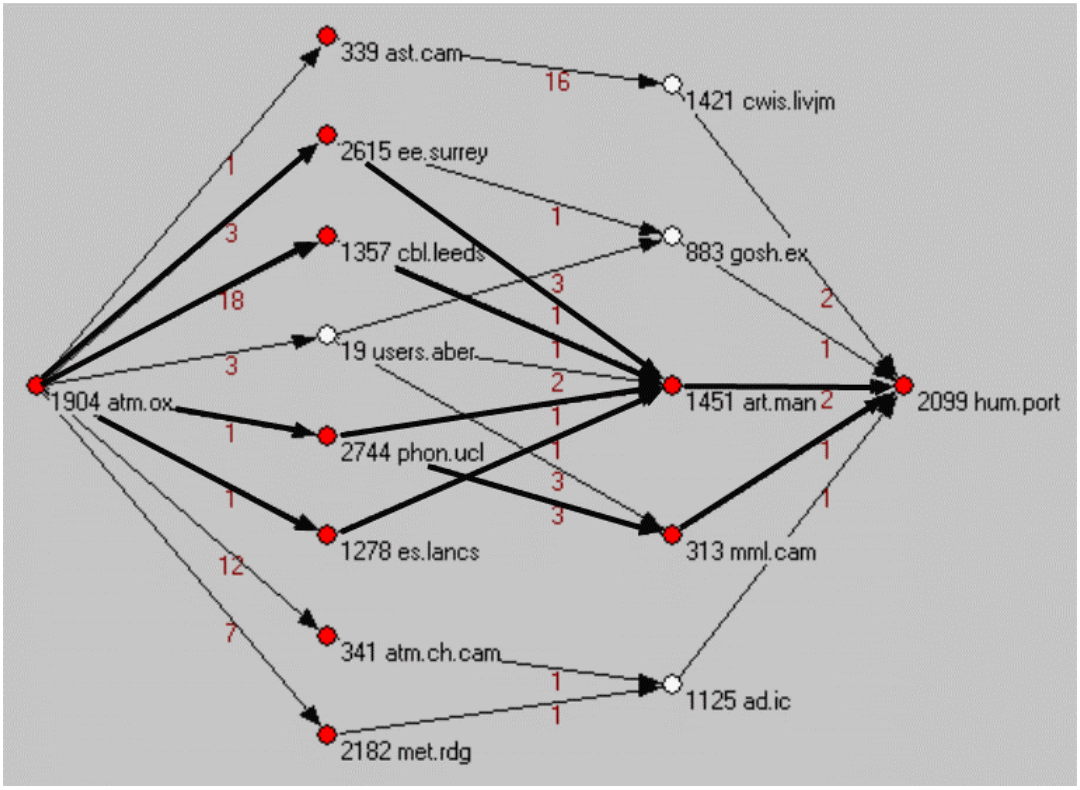
Path net HN04. Affiliations. All subsites were visited in this path net.



Path net HN05. White node denotes generic-type subsite. Counts of *page level* links are shown.

path level	id	short domain name	affiliation
0	2068	geog.plym	Dept of Geographical Sciences, Univ. of Plymouth
1	1327	geog.le	Dept of Geography, Univ. of Leicester
1	1345	geog.leeds	School of Geography, Univ. of Leeds
1	2745	geog.ucl	Dept of Geography, University College of London
2	1613	staff.ncl	Staff web pages, Univ. of Newcastle upon Tyne
3	1885	eye.ox	Dept of Ophthalmology, Oxford University

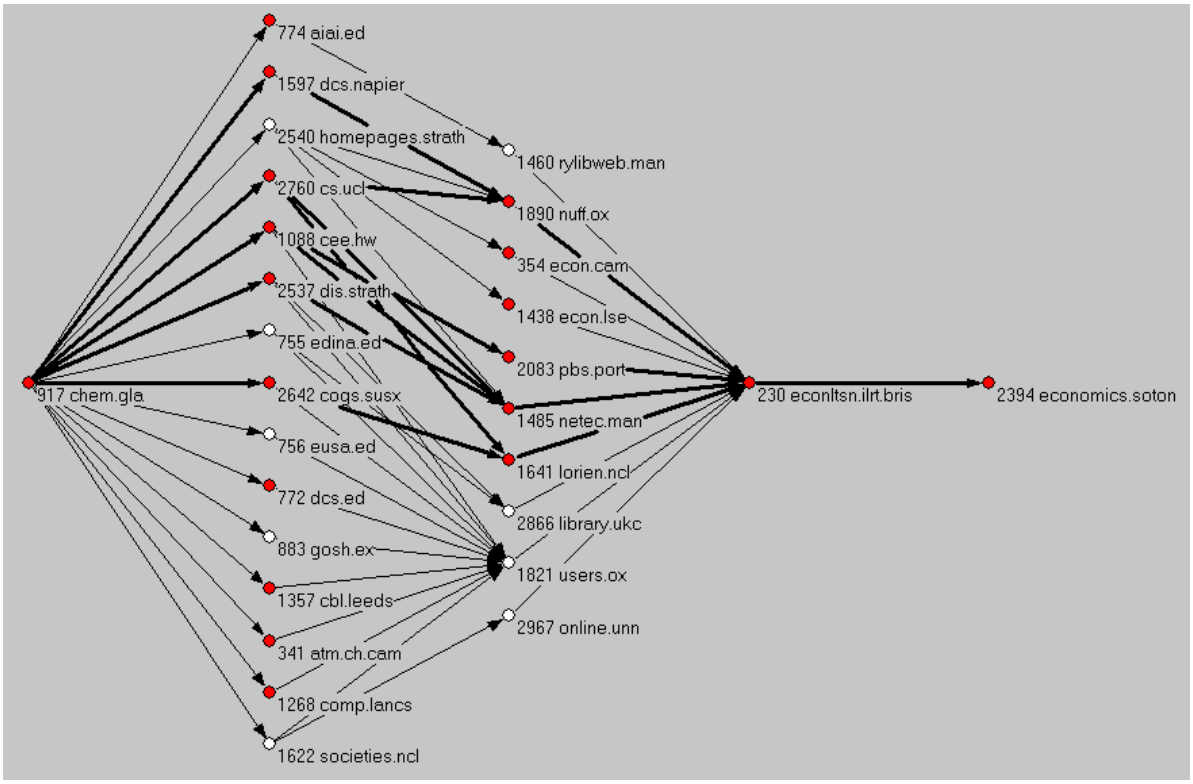
Path net HN05. Affiliations. All subsites were visited in this path net.



Path net NH01. Generic-type subsites are marked with white nodes. Followed link paths in bold contain non-generic subsites only. Counts of *page level* links are shown.

path level	visited node	id	short domain name	affiliation
0	x	1904	atm.ox	Atmospheric, Oceanic and Planetary Physics, Dept of Physics, Univ. of Oxford
1	.	19	users.aber	Personal web pages, Univ. of Aberdeen
1	.	339	ast.cam	Institute of Astronomy, School of Physical Sciences, Univ. of Cambridge
1	.	341	atm.ch.cam	Centre for Atmospheric Science, Univ. of Cambridge
1	x	1278	es.lancs	Dept of Environmental Science, Lancaster University
1	x	1357	cbl.leeds	Computer Based Learning Unit, Univ. of Leeds
1	.	2182	met.rdg	Dept of Meteorology, Univ. of Reading
1	x	2615	ee.surrey	Dept of Electrical Engineering, Univ. of Surrey
1	x	2744	phon.ucl	Dept of Phonetics and Linguistics, University College London
2	x	313	mml.cam	Faculty of Modern and Medieval Languages, Univ. of Cambridge
2	.	883	gosh.ex	GOSH (Guild of Students Home Page), Univ. of Exeter
2	.	1125	ad.ic	Academic and Administrative Services, Imperial College, London
2	.	1421	cwis.livjm	Campus Wide Information Service, Liverpool John Moores University
2	x	1451	art.man	Faculty of Arts, Univ. of Manchester
3	x	2099	hum.port	Faculty of Humanities and Social Sciences, Univ. of Portsmouth

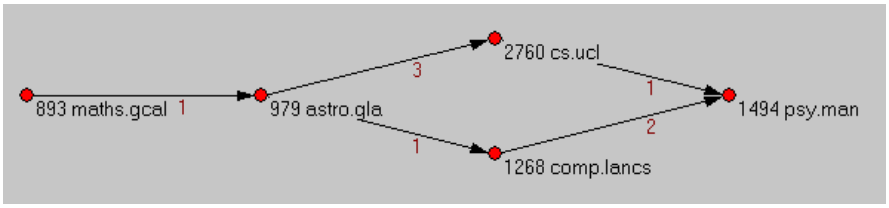
Path net NH01. Affiliations including visited subsites on followed link paths.



Path net NH02. Generic-type subsites marked with white nodes. Followed link paths in bold. See Appendix 11 for counts of *page level* links.

path level	visited node	id	short domain name	affiliation
0	x	917	chem.gla	Dept of Chemistry, Univ. of Glasgow
1		341	atm.ch.cam	Centre for Atmospheric Science, Univ. of Cambridge
1		755	edina.ed	Edinburgh Data and Information Access, Univ. of Edinburgh, (National Data Centre)
1		756	eusa.ed	Edinburgh University Students' Association
1		772	dcs.ed	Dept of Computer Science, Univ. of Edinburgh
1		774	aiai.ed	Artificial Intelligence Applications Institute, School of Informatics, Univ. of Edinburgh
1		883	gosh.ex	GOSH (Guild of Students Home Page), Univ. of Exeter
1	x	1088	cee.hw	Dept of Computing and Electrical Engineering, Heriot-Watt University
1		1268	comp.lancs	Computing Dept, Lancaster University
1		1357	cbl.leeds	Computer Based Learning Unit, Univ. of Leeds
1	x	1597	dcs.napier	School of Computing, Napier University
1		1622	societies.ncl	Club & Society Home Pages, Univ. of Newcastle upon Tyne
1	x	2537	dis.strath	Dept of Computer and Information Sciences, Univ. of Strathclyde
1		2540	homepages.strath	Personal web pages, Univ. of Strathclyde
1	x	2642	cogs.susx	School of Cognitive and Computing Sciences, Univ. of Sussex
1	x	2760	cs.ucl	Dept of Computer Science, University College London
2		354	econ.cam	Faculty of Economics and Politics, Univ. of Cambridge
2		1438	econ.lse	Dept of Economics, London School of Economics
2		1460	rylibweb.man	University Library, Univ. of Manchester
2	x	1485	netec.man	NetEc, economic subject gateway mirrored at Univ. of Manchester
2	x	1641	lorien.ncl	Dept of Chemical and Process Engineering, Newcastle University
2		1821	users.ox	Personal web pages, Univ. of Oxford
2	x	1890	nuff.ox	Nuffield College, Univ. of Oxford (Social sciences: Economics, Politics, and Sociology)
2	x	2083	pbs.port	Portsmouth Business School, Univ. of Portsmouth
2		2866	library.ukc	Library, Univ. of Kent, Canterbury
2		2967	online.unn	official homepage of Northumbria University
3	x	230	econltsn.ilrt.bris	Learning and Teaching Support Network Centre for Economics, Univ. of Bristol
4	x	2394	economics.soton	Dept of Economics, Univ. of Southampton

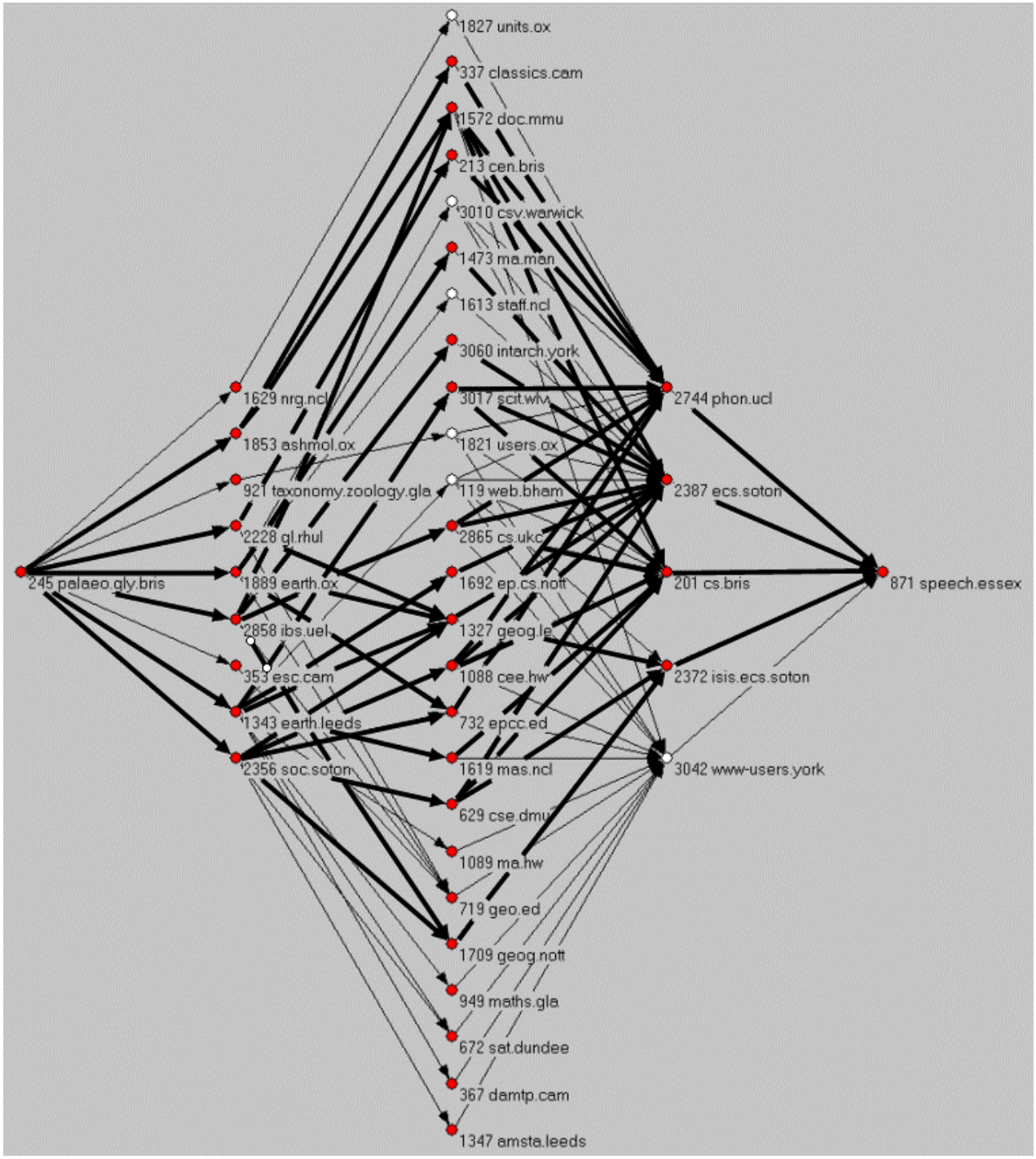
Path net NH02. Affiliations including visited subsites on followed link paths.



Path net NH03. Counts of *page level* links are shown.

path level	id	short domain name	affiliation
0	893	maths.gcal	Mathematics Dept, Glasgow Caledonian University
1	979	astro.gla	Astronomy & Astrophysics Group, Dept of Physics & Astronomy, Glasgow
2	1268	comp.lancs	Computing Dept, Lancaster University
2	2760	cs.ucl	Dept of Computer Science, University College London
3	1494	psy.man	Dept of Psychology, Univ. of Manchester

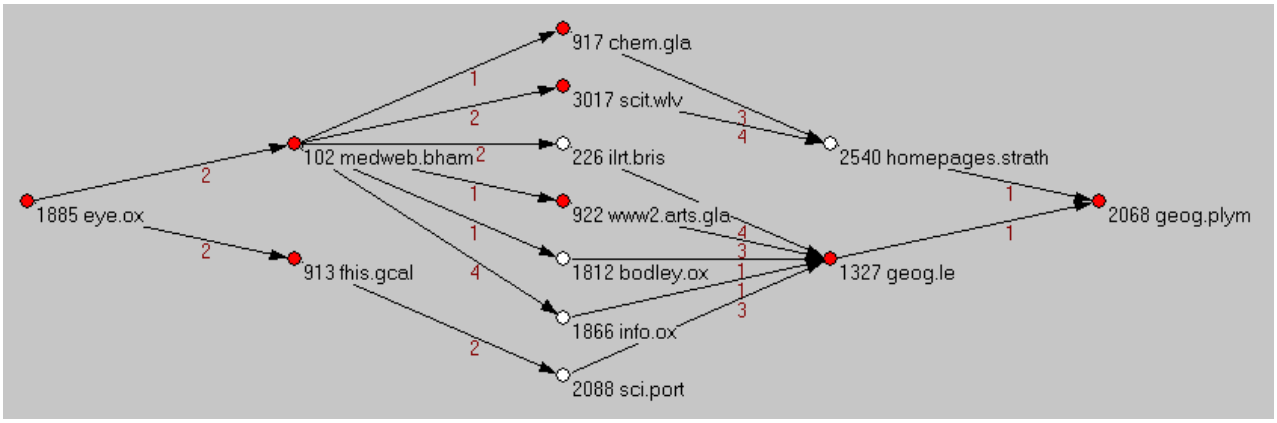
Path net NH03. Affiliations. All subsites were visited in this path net.



Path net NH04. Generic-type subsites marked with white nodes. Followed link paths in bold. Affiliations listed below. See Appendix 11 for counts of *page level* links.

path level	visited node	id	short domain name	affiliation
0	x	245	palaeo.gly.bris	Palaeontology Research Group, Dept of Earth Sciences, Univ. of Bristol
1		353	esc.cam	Dept of Earth Sciences, Univ. of Cambridge
1		921	taxonomy.zoology.gla	Taxonomy, Systematics, and Bioinformatics, Univ. of Glasgow
1	x	1343	earth.leeds	School of Earth Sciences, Univ. of Leeds
1		1629	nrg.ncl	Newcastle Research Group [Fossil Fuels and Environmental Geochemistry Institute], Univ. of Newcastle upon Tyne
1	x	1853	ashmol.ox	The Ashmolean Museum, Museum of Art & Archaeology, Univ. of Oxford
1	x	1889	earth.ox	Dept of Earth Sciences, Univ. of Oxford
1	x	2228	gl.rhul	Dept of Geology, Royal Holloway Univ. of London
1	x	2356	soc.soton	Southampton Oceanography Centre, Univ. of Southampton
1	x	2858	ibs.uel	Internet Biodiversity Service, Univ. of East London
2		119	web.bham	Personal web pages at Univ. of Birmingham
2	x	213	cen.bris	Dept of Civil Engineering, Univ. of Bristol
2	x	337	classics.cam	Faculty of Classics and Museum of Classical Archaeology, Univ. of Cambridge
2		367	damtp.cam	Dept of Applied Mathematics & Theoretical Physics, Univ. of Cambridge
2	x	629	cse.dmu	Faculty of Computing Sciences and Engineering, De Montfort University
2		672	sat.dundee	Dundee Satellite Receiving Station, Univ. of Dundee
2		719	geo.ed	Geography; School of Earth, Environmental and Geographical Sciences, Edinburgh
2	x	732	epcc.ed	Edinburgh Parallel Computing Centre, Univ. of Edinburgh
2		949	maths.gla	Dept of Mathematics, Univ. of Glasgow
2	x	1088	cee.hw	Dept of Computing and Electrical Engineering, Heriot-Watt University
2		1089	ma.hw	School of Mathematical and Computer Sciences, Heriot-Watt University
2	x	1327	geog.le	Dept of Geography, Univ. of Leicester
2		1347	amsta.leeds	School of Mathematics, Univ. of Leeds
2	x	1473	ma.man	Dept of Mathematics, Univ. of Manchester
2	x	1572	doc.mmu	Dept of Computing and Mathematics, Manchester Metropolitan University
2		1613	staff.ncl	Staff web pages, Univ. of Newcastle upon Tyne
2	x	1619	mas.ncl	School of Mathematics and Statistics, Univ. of Newcastle upon Tyne
2	x	1692	ep.cs.nott	Electronic Publishing Research Group, School of Computer Science and Information Technology, Univ. of Nottingham
2	x	1709	geog.nott	School of Geography, Faculty of Law and Social Sciences, Univ. of Nottingham
2		1821	users.ox	Personal web pages, Univ. of Oxford
2		1827	units.ox	Personal web pages, Univ. of Oxford
2	x	2865	cs.ukc	Computing Laboratory (Dept.), Univ. of Kent, Canterbury
2		3010	csv.warwick	Official web pages of Univ. of Warwick
2	x	3017	scit.wlv	School of Computing and Information Technology, Univ. of Wolverhampton
2	x	3060	intarch.york	Internet Archaeology, electronic journal at Dept of Archaeology, Univ. of York
3	x	201	cs.bris	Dept of Computer Science, Univ. of Bristol
3	x	2372	isis.ecs.soton	The Image, Speech and Intelligent Systems research group, Dept of Electronics and Computer Science, Univ. of Southampton
3	x	2387	ecs.soton	Dept of Electronics and Computer Science, Univ. of Southampton
3	x	2744	phon.ucl	Dept of Phonetics and Linguistics, University College London
3		3042	www-users.york	Personal web pages, Univ. of York
4	x	871	speech.essex	Speech Group, Dept of Language and Linguistics, Univ. of Essex

Path net NH04. Affiliations including visited subsites on followed link paths.



Path net NH05. Generic-type subsites marked with white nodes. Counts of *page level* links are shown.

path level	id	short domain name	affiliation
0	1885	eye.ox	Dept of Ophthalmology, Oxford University
1	102	medweb.bham	School of Medicine, Univ. of Birmingham
1	913	fhis.gcal	Faculty of Health, Glasgow Caledonian University
2	226	ilrt.bris	Institute for Learning and Research Technology, Univ. of Bristol
2	917	chem.gla	Dept of Chemistry, Univ. of Glasgow
2	922	www2.arts.gla	Faculty of Arts, Univ. of Glasgow
2	1812	bodley.ox	Bodleian Library, Univ. of Oxford
2	1866	info.ox	official Oxford University web pages
2	2088	sci.port	Faculty of Science, Univ. of Portsmouth
2	3017	scit.wlv	School of Computing and Information Technology, Univ. of Wolverhampton
3	1327	geog.le	Dept of Geography, Univ. of Leicester
3	2540	homepages.strath	Personal web pages, Univ. of Strathclyde
4	2068	geog.plym	Dept of Geographical Sciences, Univ. of Plymouth

Path net NH05. Affiliations. All subsites were visited in this path net.

Appendix 11. Summary node data from 10 path nets

Sorted by (1) path net, (2) path net level, and (3) subsite id.

path net	path net level	path length	path start	path end	visited node	id	short domain name	subsite topic (cf. legend Append.16)	first time indexed in Internet Archive	core	betweenness centrality in UK subweb	bc rank among 7669 subsites	in-distance	out-distance	in-neighbors in UK subweb	out-neighbors in UK subweb	inlinks in UK subweb	outlinks in UK subweb	in-neighbors in path net	out-neighbors in path net	inlinks in path net	outlinks in path net
HN01	0	3	hum	atm	x	2099	hum.port	hum/soc	1997-07-28	23	0.0001461	655	3.27	3.09	17	16	25	22	0	6	0	11
HN01	1	3	hum	atm	x	119	web.bham	generic	1998-01-15	53	0.0041168	31	2.53	2.42	68	210	114	846	1	3	1	7
HN01	1	3	hum	atm	x	710	arts.ed	hum	1997-01-01	21	0.0002204	503	2.96	3.49	18	13	38	22	1	1	1	1
HN01	1	3	hum	atm	x	1424	staff.livjm	generic	1999-10-11	29	0.0004992	291	3.14	2.96	9	49	11	63	1	1	2	1
HN01	1	3	hum	atm	x	1612	cs.ncl	cs	1998-12-12	53	0.0019715	78	2.62	2.56	75	145	292	679	1	2	5	3
HN01	1	3	hum	atm	x	1866	info.ox	generic	1997-01-05	53	0.0083196	11	2.25	2.73	259	120	939	293	1	1	1	2
HN01	1	3	hum	atm	x	2387	ecs.soton	cs+ee	1997-06-25	53	0.0079356	13	2.45	2.33	117	327	377	1110	1	2	1	3
HN01	2	3	hum	atm	x	335	cus.cam	generic	1997-12-24	46	0.0015281	109	2.82	2.59	23	136	57	301	1	1	2	1
HN01	2	3	hum	atm	x	337	classics.cam	hum	1996-12-27	35	0.0006469	225	3.00	2.86	21	52	32	104	1	1	1	2
HN01	2	3	hum	atm	x	341	atm.ch.cam	atm	1997-06-28	40	0.0011598	144	2.54	2.92	74	24	274	134	3	1	5	9
HN01	2	3	hum	atm	x	1613	staff.ncl	generic	1998-12-12	47	0.0021993	66	2.75	2.55	41	162	92	255	1	1	2	1
HN01	2	3	hum	atm	x	2182	met.rdg	met	1997-01-09	45	0.0016762	98	2.67	2.74	62	66	227	235	1	1	1	3
HN01	2	3	hum	atm	x	2393	hpcc.ecs.soton	cs+ee	1997-01-30	26	0.0000478	1033	3.05	3.10	15	29	55	73	1	1	4	1
HN01	2	3	hum	atm	x	2745	geog.ucl	geo	1997-06-12	39	0.0018515	86	2.72	2.77	56	68	148	148	2	1	2	3
HN01	3	3	hum	atm	x	1904	atm.ox	atm	1997-04-04	35	0.0005643	256	3.17	2.93	24	48	110	102	7	0	20	0
HN02	0	3	econ	chem	x	2394	economics.soton	econ	2000-03-02	3	0.0000001	1901	4.03	3.57	1	2	3	2	0	1	0	1
HN02	1	3	econ	chem	x	3006	dcs.warwick	cs	1998-12-03	53	0.0026981	48	2.41	2.61	134	98	565	248	1	3	1	3
HN02	2	3	econ	chem	x	1088	cee.hw	cs+ee	1997-12-10	53	0.0103513	8	2.35	2.24	148	518	608	3380	1	1	1	1
HN02	2	3	econ	chem	x	1328	mcs.le	cs+ma	1997-02-01	53	0.0005872	249	2.57	2.63	71	93	250	393	1	1	1	1
HN02	2	3	econ	chem	x	2865	cs.ukc	cs	1998-12-12	53	0.0024727	55	2.50	2.50	91	125	446	569	1	1	1	1
HN02	3	3	econ	chem	x	917	chem.gla	chem	1997-01-26	36	0.0014124	119	2.83	2.82	39	46	96	83	3	0	3	0
HN03	0	4	psy	math	x	1494	psy.man	psych	1996-12-21	9	0.0000014	1570	3.04	4.28	9	1	20	1	0	1	0	1
HN03	1	4	psy	math	x	3020	scitsc.wlv	cs	1996-12-26	53	0.001728	95	2.21	3.28	249	8	459	8	1	2	1	2
HN03	2	4	psy	math	x	772	dcs.ed	cs	1996-11-02	53	0.0068109	16	2.28	2.49	191	185	1309	835	1	3	1	21

HN03	2	4	psy	math	x	1773	mcs.open	cs+ma	1997-01-25	53	0.0007711	194	2.62	2.72	51	73	303	199	1	3	1	10
HN03	3	4	psy	math	x	318	statslab.cam	math	1997-06-14	53	0.0025677	53	2.39	2.77	127	72	2209	375	2	1	5	2
HN03	3	4	psy	math	x	1089	ma.hw	cs+ma	1997-12-11	53	0.0064728	18	2.35	2.60	191	186	884	636	2	1	23	1
HN03	3	4	psy	math	x	1225	mth.kcl	math	1997-01-12	53	0.0013741	123	2.61	2.75	72	105	224	530	2	1	3	1
HN03	4	4	psy	math	x	893	maths.gcal	math	1998-12-12	8	0.0000009	1627	3.14	3.75	7	1	14	1	3	0	4	0
HN04	0	3	speech	palaeo	x	871	speech.essex	ling	1997-12-21	15	0.0000706	912	3.24	3.19	5	13	9	19	0	1	0	1
HN04	1	3	speech	palaeo	x	2615	ee.surrey	el.engineer	1997-05-11	53	0.0080852	12	2.32	2.43	187	213	619	566	1	1	1	1
HN04	2	3	speech	palaeo	x	1300	www-staff.lboro	generic	1997-03-28	45	0.0015585	107	2.95	2.62	18	115	21	186	1	1	1	2
HN04	3	3	speech	palaeo	x	245	palaeo.gly.bris	geo	1997-12-24	20	0.0002119	512	3.24	3.52	17	15	30	17	1	0	2	0
HN05	0	3	geogr	eye	x	2068	geog.plym	geo	2001-01-24	23	0.0002129	511	3.49	3.02	4	35	5	46	0	3	0	6
HN05	1	3	geogr	eye	x	1327	geog.le	geo	1997-07-26	46	0.0039168	33	2.62	2.54	91	175	216	565	1	1	3	1
HN05	1	3	geogr	eye	x	1345	geog.leeds	geo	1996-12-23	40	0.002737	46	2.84	2.65	41	143	86	463	1	1	1	2
HN05	1	3	geogr	eye	x	2745	geog.ucl	geo	1997-06-12	39	0.0018515	86	2.72	2.77	56	68	148	148	1	1	2	1
HN05	2	3	geogr	eye	x	1613	staff.ncl	generic	1998-12-12	47	0.0021993	66	2.75	2.55	41	162	92	255	3	1	4	2
HN05	3	3	geogr	eye	x	1885	eye.ox	med	1998-12-06	4	0.0000019	1530	3.75	3.80	2	2	4	4	1	0	2	0
NH01	0	3	atm	hum	x	1904	atm.ox	met	1997-04-04	35	0.0005643	256	3.17	2.93	24	48	110	102	0	8	0	46
NH01	1	3	atm	hum	.	19	users.aber	generic	2000-09-01	53	0.00366	36	2.77	2.39	34	250	63	512	1	3	3	8
NH01	1	3	atm	hum	.	339	ast.cam	astro	1997-12-11	53	0.0041995	28	2.44	2.68	139	101	931	1848	1	1	1	16
NH01	1	3	atm	hum	.	341	atm.ch.cam	atm	1997-06-28	40	0.0011598	144	2.54	2.92	74	24	274	134	1	1	12	1
NH01	1	3	atm	hum	x	1278	es.lancs	environ	1997-05-05	32	0.0006984	212	3.17	2.78	20	51	32	95	1	1	1	1
NH01	1	3	atm	hum	x	1357	cbl.leeds	learn/gen	1998-12-02	53	0.0129657	6	2.12	2.64	387	91	4475	245	1	1	18	1
NH01	1	3	atm	hum	.	2182	met.rdg	met	1997-01-09	45	0.0016762	98	2.67	2.74	62	66	227	235	1	1	7	1
NH01	1	3	atm	hum	x	2615	ee.surrey	el.engineer	1997-05-11	53	0.0080852	12	2.32	2.43	187	213	619	566	1	2	3	2
NH01	1	3	atm	hum	x	2744	phon.ucl	ling	1997-01-28	48	0.0012989	129	2.56	2.78	70	64	208	171	1	2	1	4
NH01	2	3	atm	hum	x	313	mml.cam	ling	1998-04-15	30	0.0006109	237	2.88	3.09	19	30	44	73	2	1	6	1
NH01	2	3	atm	hum	.	883	gosh.ex	generic	1997-07-15	45	0.0013909	122	2.86	2.57	37	133	68	217	2	1	4	1
NH01	2	3	atm	hum	.	1125	ad.ic	generic	1996-12-28	26	0.000605	241	2.93	3.16	27	24	38	33	2	1	2	1
NH01	2	3	atm	hum	.	1421	cwis.livjm	generic	1998-12-06	45	0.0041473	30	2.72	2.57	63	154	115	362	1	1	16	2
NH01	2	3	atm	hum	x	1451	art.man	hum	1997-06-25	40	0.0035525	38	2.61	2.76	68	87	134	227	5	1	6	2
NH01	3	3	atm	hum	x	2099	hum.port	hum/soc	1997-07-28	23	0.0001461	655	3.27	3.09	17	16	25	22	5	0	7	0
NH02	0	4	chem	econ	x	917	chem.gla	chem	1997-01-26	36	0.0014124	119	2.83	2.82	39	46	96	83	0	15	0	21
NH02	1	4	chem	econ	.	341	atm.ch.cam	atm	1997-06-28	40	0.0011598	144	2.54	2.92	74	24	274	134	1	1	1	1
NH02	1	4	chem	econ	.	755	edina.ed	generic	1996-12-22	41	0.0038221	35	2.42	2.96	158	36	885	60	1	2	2	4
NH02	1	4	chem	econ	.	756	eusa.ed	generic	1998-12-07	36	0.0006022	244	2.80	2.73	27	54	45	214	1	1	1	12
NH02	1	4	chem	econ	.	772	dcs.ed	cs	1996-11-02	53	0.0068109	16	2.28	2.49	191	185	1309	835	1	1	1	10
NH02	1	4	chem	econ	.	774	aiai.ed	cs	1997-07-19	53	0.0009277	170	2.56	2.85	74	46	266	86	1	1	1	1
NH02	1	4	chem	econ	.	883	gosh.ex	generic	1997-07-15	45	0.0013909	122	2.86	2.57	37	133	68	217	1	1	1	6
NH02	1	4	chem	econ	x	1088	cee.hw	cs+ee	1997-12-10	53	0.0103513	8	2.35	2.24	148	518	608	3380	1	3	3	3
NH02	1	4	chem	econ	.	1268	comp.lancs	cs	1998-01-16	53	0.0090899	10	2.31	2.39	183	265	709	1497	1	1	1	36
NH02	1	4	chem	econ	.	1357	cbl.leeds	learn	1998-12-02	53	0.0129657	6	2.12	2.64	387	91	4475	245	1	1	1	1
NH02	1	4	chem	econ	x	1597	dcs.napier	cs	1997-01-04	53	0.0061159	19	2.47	2.46	127	226	576	1404	1	1	2	1
NH02	1	4	chem	econ	.	1622	societies.ncl	generic	1998-12-06	32	0.0005224	277	2.84	2.77	27	39	52	73	1	2	1	15
NH02	1	4	chem	econ	x	2537	dis.strath	cs/is	1996-11-14	39	0.0024583	57	2.53	2.88	101	24	256	239	1	3	1	15

NH02	1	4	chem	econ	.	2540	homepages.strath	generic	1998-05-09	53	0.002394	61	2.71	2.57	31	164	55	616	1	4	3	8
NH02	1	4	chem	econ	x	2642	cogs.susx	cs/cog	1997-07-11	53	0.0103444	9	2.26	2.40	231	268	1319	1264	1	2	1	5
NH02	1	4	chem	econ	x	2760	cs.ucl	cs	1997-02-24	53	0.01396	3	2.16	2.39	300	265	1802	1155	1	3	1	10
NH02	2	4	chem	econ	.	354	econ.cam	econ	1997-07-11	28	0.0003497	378	2.84	3.32	32	31	84	96	1	1	1	1
NH02	2	4	chem	econ	.	1438	econ.lse	econ	1997-12-22	25	0.000327	395	3.00	3.04	18	22	81	69	1	1	1	1
NH02	2	4	chem	econ	.	1460	rylibweb.man	generic	1996-11-22	41	0.0023549	63	2.72	2.62	42	113	74	187	1	1	1	2
NH02	2	4	chem	econ	x	1485	netec.man	econ	1997-12-11	39	0.0021197	69	2.61	2.84	82	56	364	633	4	1	17	3
NH02	2	4	chem	econ	x	1641	lorien.ncl	chem	1996-12-19	39	0.0010449	161	2.86	2.69	20	68	42	95	2	1	3	1
NH02	2	4	chem	econ	.	1821	users.ox	generic	1996-12-20	53	0.0368644	1	2.17	2.18	330	507	1106	1405	11	1	80	1
NH02	2	4	chem	econ	x	1890	nuff.ox	soc	1998-12-03	39	0.0008106	185	2.70	3.03	57	45	180	133	3	1	3	4
NH02	2	4	chem	econ	x	2083	pbs.port	econ	1997-07-03	28	0.0002804	440	2.92	3.20	15	26	48	86	1	1	1	16
NH02	2	4	chem	econ	.	2866	library.ukc	generic	2000-01-25	40	0.0018051	89	2.82	2.69	43	81	79	135	2	1	12	1
NH02	2	4	chem	econ	.	2967	online.unn	generic	2000-08-15	25	0.0004443	320	3.03	2.99	11	30	20	44	1	1	9	1
NH02	3	4	chem	econ	x	230	econltsn.ilrt.bris	econ/learn	2000-05-11	29	0.0008956	178	3.03	2.85	14	64	42	287	10	1	31	3
NH02	4	4	chem	econ	x	2394	economics.soton	econ	2000-03-02	3	0.0000001	1901	4.03	3.57	1	2	3	2	1	0	3	0
NH03	0	3	math	psy	x	893	maths.gcal	math	1998-12-12	8	0.0000009	1627	3.14	3.75	7	1	14	1	0	1	0	1
NH03	1	3	math	psy	x	979	astro.gla	astro	1997-05-08	41	0.0004628	310	2.84	2.75	31	54	88	823	1	2	1	4
NH03	2	3	math	psy	x	1268	comp.lancs	cs	1998-01-16	53	0.0090899	10	2.31	2.39	183	265	709	1497	1	1	1	2
NH03	2	3	math	psy	x	2760	cs.ucl	cs	1997-02-24	53	0.01396	3	2.16	2.39	300	265	1802	1155	1	1	3	1
NH03	3	3	math	psy	x	1494	psy.man	psych	1996-12-21	9	0.0000014	1570	3.04	4.28	9	1	20	1	2	0	3	0
NH04	0	4	palaeo	speech	x	245	palaeo.gly.bris	earth	1997-12-24	20	0.0002119	512	3.24	3.52	17	15	30	17	0	9	0	11
NH04	1	4	palaeo	speech	.	353	esc.cam	earth	1997-04-13	34	0.0007641	196	2.85	3.20	49	34	98	66	1	1	1	6
NH04	1	4	palaeo	speech	.	921	taxonomy.zoology.gla	zoo	1996-11-19	23	0.0005224	276	2.93	3.05	22	22	38	49	1	1	2	1
NH04	1	4	palaeo	speech	x	1343	earth.leeds	earth	1996-12-19	34	0.0003914	351	3.04	2.97	32	37	131	65	1	9	1	15
NH04	1	4	palaeo	speech	.	1629	nrg.ncl	earth	1997-03-31	20	0.0001301	711	3.59	3.32	9	20	16	33	1	1	1	1
NH04	1	4	palaeo	speech	x	1853	ashmol.ox	hum/archaeo	1997-07-16	30	0.0001494	644	2.77	3.20	43	9	130	161	1	2	1	2
NH04	1	4	palaeo	speech	x	1889	earth.ox	earth	1997-02-05	40	0.0007379	201	2.66	2.94	54	50	101	178	1	6	1	17
NH04	1	4	palaeo	speech	x	2228	gl.rhul	earth	2001-02-20	27	0.0002529	475	3.23	3.11	15	33	18	55	1	3	1	4
NH04	1	4	palaeo	speech	x	2356	soc.soton	earth	1997-12-10	40	0.0021678	67	2.61	2.90	71	70	194	251	1	7	1	14
NH04	1	4	palaeo	speech	x	2858	ibs.uel	environ	1998-02-01	26	0.0007153	206	3.09	3.01	26	42	59	69	1	5	2	8
NH04	2	4	palaeo	speech	.	119	web.bham	generic	1998-01-15	53	0.0041168	31	2.53	2.42	68	210	114	846	1	5	1	28
NH04	2	4	palaeo	speech	x	213	cen.bris	engineer	1999-01-25	17	0.0002789	444	3.45	3.06	6	24	8	32	1	1	1	1
NH04	2	4	palaeo	speech	x	337	classics.cam	hum	1996-12-27	35	0.0006469	225	3.00	2.86	21	52	32	104	1	1	1	2
NH04	2	4	palaeo	speech	.	367	damtp.cam	ma+physic	1997-12-12	53	0.0048087	24	2.50	2.57	137	147	646	641	1	1	1	6
NH04	2	4	palaeo	speech	x	629	cse.dmu	cs+ee	1999-04-27	53	0.0009752	167	2.79	2.51	20	144	39	467	1	3	1	7
NH04	2	4	palaeo	speech	.	672	sat.dundee	earth	1997-04-27	46	0.0012281	135	2.45	2.91	103	30	192	63	2	1	10	1
NH04	2	4	palaeo	speech	.	719	geo.ed	geo	1997-06-05	52	0.0042815	27	2.41	2.84	182	65	473	110	4	1	15	1
NH04	2	4	palaeo	speech	x	732	epcc.ed	cs	1996-12-22	53	0.0039239	32	2.52	2.69	113	102	318	369	2	2	3	3
NH04	2	4	palaeo	speech	.	949	maths.gla	math	1997-05-03	53	0.0002542	474	2.64	2.84	46	55	132	220	1	1	3	1
NH04	2	4	palaeo	speech	x	1088	cee.hw	cs+ee	1997-12-10	53	0.0103513	8	2.35	2.24	148	518	608	3380	1	4	2	17
NH04	2	4	palaeo	speech	.	1089	ma.hw	ma+cs	1997-12-11	53	0.0064728	18	2.35	2.60	191	186	884	636	1	1	1	1
NH04	2	4	palaeo	speech	x	1327	geog.le	geo	1997-07-26	46	0.0039168	33	2.62	2.54	91	175	216	565	4	2	6	7
NH04	2	4	palaeo	speech	.	1347	amsta.leeds	math	1996-12-26	53	0.002656	49	2.41	2.59	117	134	415	951	1	1	2	8

NH04	2	4	palaeo	speech	x	1473	ma.man	math	1997-12-11	53	0.0003272	394	2.67	2.81	54	42	159	121	1	1	1	2
NH04	2	4	palaeo	speech	x	1572	doc.mmu	cs+ma	1998-02-04	53	0.0130101	5	2.38	2.27	127	514	441	1056	2	4	4	8
NH04	2	4	palaeo	speech	.	1613	staff.ncl	generic	1998-12-12	47	0.0021993	66	2.75	2.55	41	162	92	255	1	1	3	1
NH04	2	4	palaeo	speech	x	1619	mas.ncl	math	1998-01-12	53	0.0007442	200	2.96	2.67	17	127	27	432	1	2	1	3
NH04	2	4	palaeo	speech	x	1692	ep.cs.nott	cs/is	1997-01-17	27	0.0001001	805	2.72	3.05	26	7	75	18	1	1	1	2
NH04	2	4	palaeo	speech	x	1709	geog.nott	geo	1997-01-29	38	0.0009572	168	2.74	2.86	42	51	72	84	2	1	2	1
NH04	2	4	palaeo	speech	.	1821	users.ox	generic	1996-12-20	53	0.0368644	1	2.17	2.18	330	507	1106	1405	1	4	1	15
NH04	2	4	palaeo	speech	.	1827	units.ox	generic	1997-01-01	40	0.0010801	154	2.76	2.95	69	47	150	67	1	1	1	2
NH04	2	4	palaeo	speech	x	2865	cs.ukc	cs	1998-12-12	53	0.0024727	55	2.50	2.50	91	125	446	569	1	4	1	19
NH04	2	4	palaeo	speech	.	3010	csv.warwick	generic	1997-12-11	53	0.0114955	7	2.26	2.38	202	354	402	498	1	3	4	3
NH04	2	4	palaeo	speech	x	3017	scit.wlv	cs	1998-01-31	53	0.0264755	2	2.14	2.32	349	434	895	1482	1	3	1	4
NH04	2	4	palaeo	speech	x	3060	intarch.york	archaeo	1997-12-10	33	0.0007499	199	2.87	2.87	27	47	44	107	1	1	1	1
NH04	3	4	palaeo	speech	x	201	cs.bris	cs	1997-04-01	53	0.0030483	39	2.51	2.48	93	168	389	1016	9	1	33	2
NH04	3	4	palaeo	speech	x	2372	isis.ecs.soton	cs	1996-12-28	38	0.0001355	698	2.78	3.13	37	15	105	20	4	1	21	1
NH04	3	4	palaeo	speech	x	2387	ecs.soton	cs	1997-06-25	53	0.0079356	13	2.45	2.33	117	327	377	1110	13	1	42	3
NH04	3	4	palaeo	speech	x	2744	phon.ucl	ling	1997-01-28	48	0.0012989	129	2.56	2.78	70	64	208	171	11	1	15	2
NH04	3	4	palaeo	speech	.	3042	www-users.york	generic	1999-05-08	53	0.0056106	20	2.53	2.43	94	226	238	526	13	1	33	1
NH04	4	4	palaeo	speech	x	871	speech.essex	ling	1997-12-21	15	0.0000706	912	3.24	3.19	5	13	9	19	5	0	9	0
NH05	0	4	eye	geogr	x	1885	eye.ox	med	1998-12-06	4	0.0000019	1530	3.75	3.80	2	2	4	4	0	2	0	4
NH05	1	4	eye	geogr	x	102	medweb.bham	med	1997-04-22	40	0.0024722	56	2.48	2.89	94	65	166	137	1	6	2	11
NH05	1	4	eye	geogr	x	913	fhis.gcal	med	1996-12-21	24	0.0003823	359	3.20	3.04	9	59	29	68	1	1	2	2
NH05	2	4	eye	geogr	x	226	ilrt.bris	learn/gen	1998-01-10	41	0.0027692	45	2.49	2.83	113	54	317	170	1	1	2	4
NH05	2	4	eye	geogr	x	917	chem.gla	chem	1997-01-26	36	0.0014124	119	2.83	2.82	39	46	96	83	1	1	1	3
NH05	2	4	eye	geogr	x	922	www2.arts.gla	hum	1998-01-14	40	0.0017401	92	2.78	2.67	35	103	55	291	1	1	1	3
NH05	2	4	eye	geogr	x	1812	bodley.ox	generic	1997-12-10	40	0.0028413	44	2.55	2.96	127	44	343	80	1	1	1	1
NH05	2	4	eye	geogr	x	1866	info.ox	generic	1997-01-05	53	0.0083196	11	2.25	2.73	259	120	939	293	1	1	4	1
NH05	2	4	eye	geogr	x	2088	sci.port	gen/multi-sci	1997-02-10	36	0.0009013	176	3.00	2.83	30	67	42	94	1	1	2	3
NH05	2	4	eye	geogr	x	3017	scit.wlv	cs	1998-01-31	53	0.0264755	2	2.14	2.32	349	434	895	1482	1	1	2	4
NH05	3	4	eye	geogr	x	1327	geog.le	geo	1997-07-26	46	0.0039168	33	2.62	2.54	91	175	216	565	5	1	12	1
NH05	3	4	eye	geogr	x	2540	homepages.strath	generic	1998-05-09	53	0.002394	61	2.71	2.57	31	164	55	616	2	1	7	1
NH05	4	4	eye	geogr	x	2068	geog.plym	geo	2001-01-24	23	0.0002129	511	3.49	3.02	4	35	5	46	2	0	2	0
path net	**path net level**	**path length**	**path start**	**path end**	**visited node**	**id**	**short domain name**	**subsite topic**	**first time indexed in Internet Archive**	**core**	**betweenness centrality in UK subweb**	**bc rank among 7669 subsites**	**in-distance**	**out-distance**	**in-neighbors in UK subweb**	**out-neighbors in UK subweb**	**inlinks in UK subweb**	**outlinks in UK subweb**	**in-neighbors in path net**	**out-neighbors in path net**	**inlinks in path net**	**outlinks in path net**

Appendix 12. Multi-occurring subsites in 10 path nets

22 subsites occurring in more than one path net. Sorted by betweenness centrality rank, cf. Section 6.3.2.

id	short domain name	affiliation	# path nets	path nets	betw. centrality	bc rank
1821	users.ox	University of Oxford	2	NH02 NH04	0,03686	1
3017	scit.wlv	School of Computing and Information Technology, University of Wolverhampton	2	NH04 NH05	0,02648	2
2760	cs.ucl	Department of Computer Science, University College London	2	NH02 NH03	0,01396	3
1357	cbl.leeds	Computer Based Learning Unit, University of Leeds	2	NH01 NH02	0,01297	6
1088	cee.hw	Department of Computing and Electrical Engineering, Heriot-Watt University	**3**	HN02 NH02 NH04	0,01035	8
1268	comp.lancs	Computing Department, Lancaster University	2	NH02 NH03	0,00909	10
1866	info.ox	University of Oxford	2	HN01 NH05	0,00832	11
2615	ee.surrey	Department of Electrical Engineering, University of Surrey	2	HN04 NH01	0,00809	12
2387	ecs.soton	Department of Electronics and Computer Science, University of Southampton	2	NH01 NH04	0,00794	13
772	dcs.ed	Department of Computer Science [today: Division of Informatics], University of Edinburgh	2	HN03 NH02	0,00681	16
1089	ma.hw	School of Mathematical and Computer Sciences, Heriot-Watt University	2	HN03 NH04	0,00647	18
119	web.bham	University of Birmingham	2	HN01 NH04	0,00412	31
1327	geog.le	Department of Geography, University of Leicester	**3**	HN05 NH04 NH05	0,00392	33
2865	cs.ukc	Computing Laboratory (Dept.), University of Kent, Canterbury	2	HN02 NH04	0,00247	55
2540	homepages.strath	University of Strathclyde	2	NH02 NH05	0,00239	61
1613	staff.ncl	University of Newcastle upon Tyne	**3**	HN01 HN05 NH04	0,00220	66
2745	geog.ucl	Department of Geography, University College of London	2	HN01 HN05	0,00185	86
2182	met.rdg	Department of Meteorology, University of Reading	2	HN01 NH01	0,00168	98
883	gosh.ex	GOSH (Guild of Students Home Page), University of Exeter	2	NH01 NH02	0,00139	122
2744	phon.ucl	Department of Phonetics and Linguistics, University College London	2	NH01 NH04	0,00130	129
341	atm.ch.cam	Centre for Atmospheric Science, University of Cambridge	**3**	HN01 NH01 NH02	0,00116	144
337	classics.cam	Faculty of Classics and Museum of Classical Archaeology, University of Cambridge	2	HN01 NH04	0,00065	225

Appendix 13. Genres of visited source pages

The table shows subgenres of visited source pages sorted by institutional and personal meta genres, cf. Table 6-19, Section 6.4.5. Legend: The three numbers on each row are counts of page level links, pages, and subsites. For example, *34/26/17* at *i.conf* means that 34 followed links originate from 26 different source pages at 17 different subsites. If a subsite is present in two path nets, the subsite counts as two different path net subsites. Simple sums of subgenres may exceed the count on the step above since, e.g., a subsite may contain more than one page genre.

followed links	visited unique source pages	visited unique source subsites							
352	**281**	**93**	**SOURCE PAGES**						
i.conf			**34**	**26**	**17**	**CONFERENCE/WORKSHOP/SEMINAR PAGE**			
						12	12	7	Conference page
							1	1	1	Conference homepage
							7	7	5	Conference programme
										incl. alumni conference, list of keynote speeches, presentation abstracts,
										2 poster sessions, special session
							4	4	3	Other conference webpage
										incl. delegates, delegates' comments
										programme committee, technical committee
						13	7	4	Workshop page
							11	5	3	Workshop homepage
							2	2	2	Workshop programme

								incl. presentation abstracts
				9	7	6	Seminar page	
						3	3	3 Seminar programme: presentation abstract
						4	2	2 Seminar series overview
						2	2	1 Other seminar webpages
								incl. research project news and events, previous seminars
i.generic	**6**	**6**	**3**	**GENERIC/UNI SERVICE**				
				2	2	2	Uni service: studying abroad	
				4	4	1	Web server statistics	
i.list	**87**	**73**	**34**	**INSTITUTIONAL LINK LIST**				
				79	67	33	Link list (institutional)	
						46	37	23 Link list (institutional): research-related
						2	2	2 Link list (institutional): research-related + library-related
						1	1	1 Link list (institutional): generic + research-related
						5	3	2 Link list (institutional): generic: academic sites (p.list overtaken by institution?)
						2	2	2 Link list (institutional): library-related
						2	2	1 Link list (institutional) / resource guide: teaching-related
						11	10	7 Link list (institutional): teaching-related
						2	2	1 Link list (institutional): teaching-related bibliography
						7	7	1 Link list (institutional): studying abroad
						1	1	1 Link list + FAQ (institutional): museum-related
				3	3	2	Publication list (institutional)	
								incl. 2/2/1 publication lists (institutional): joint research project annual reports
				5	3	2	Staff/alumni list	
						3	1	1 Staff list: former PhD graduates and postdocs
						2	2	2 Alumni list: incl. former MSc graduates
i.proj	**20**	**16**	**14**	**RESEARCH PROJECT (INST.)**				
				11	9	7	Joint research group page	
						4	3	2 Joint research group homepage
						6	5	5 Joint research project homepage
						1	1	1 Joint research project webpage
				6	4	4	Research group page	
						3	2	2 Research group homepage
						3	2	2 Research group webpage: projects overview

				3	3	3	Research project page
					1	1	1 Research project homepage
					1	1	1 Research projects overview (dept.)
					1	1	1 Index page: TOC research project
i.publ	**8**	**8**	**4**	**PUBLICATION/GUIDE (INST.)**			
				1	1	1	Report/documentation: research project documentation
				1	1	1	Resource guide (text-rich incl. link list): teaching-related: "A brief guide to computers in Archaeology"
				1	1	1	Resource guide (institutional): entry (resource description)
				5	5	1	Research newsletter
i.soft	**6**	**6**	**4**	**SOFTWARE PROGRAM (INST.)**			
				4	4	2	Manual/documentation: software (joint research project web page)
				1	1	1	Manual/documentation: software demo (joint research project web page)
				1	1	1	Manual/tutorial (institutional)(section):software (joint research project)
i.teach	**4**	**4**	**2**	**TEACHING (INST.)**			
				4	4	2	Institutional course home page
p.hobby	**3**	**2**	**2**	**PERSONAL HOBBY PAGE (2 different persons)**			
				2	1	1	Personal hobby page (researcher): private travel
				1	1	1	Personal hobby page (researcher): Saxon shore forts
p.hp	**22**	**19**	**14**	**PERSONAL HOMEPAGE (19 different persons)**			
				3	3	3	Personal homepage (PhD student): incl. long link list; papers & presentations
				17	14	11	Personal homepage (researcher): incl. link list; CV (3); educational background, publication list
				1	1	1	Personal homepage (student)
				1	1	1	Personal homepage (technical staff)
p.list	**112**	**83**	**37**	**PERSONAL LINK LIST (53 different persons incl. 2 multi-occurring in two path nets)**			
				13	5	2	Bibliography (researcher) (3 different persons, one of which also has link list below)
				2	1	1	Link list (admin staff): bookmarks
				78	59	31	Link list (researcher) (37 different persons incl. 2 multi-occurring)
					20	15	8 Link list (researcher): research-related
					1	1	1 Link list (researcher): research-related + misc
					1	1	1 Link list (researcher): personal academic interest (grammar)
					4	4	3 Link list (researcher): teaching-related

				5	4	3	Link list (researcher): teaching-related: course page	
				7	4	2	Link list (researcher): UK academic sites	
				1	1	1	Link list (researcher): friends+scientists	
				1	1	1	Link list (researcher): friends	
				1	1	1	Link list (researcher): leisure	
				2	2	1	Link list (researcher): sci-fi	
				10	8	4	Link list (researcher): sports	
				25	17	10	Link list (researcher): bookmarks (13 different persons)	
						6	5	4	Link list (researcher): bookmarks: research-related
						7	4	4	Link list (researcher): bookmarks: research-related + misc
						6	3	2	Link list (researcher): bookmarks: teaching-related
						5	4	2	Link list (researcher): bookmarks: personal academic interest
						1	1	1	Link list (researcher): bookmarks: hobby (mountain climbing)
			12	12	8	Link list (PhD student) (10 different persons: no overlap between next level subgenres)		
				4	4	3	Link list (PhD student): research-related	
				7	7	5	Link list (PhD student): bookmarks	
						2	2	2	Link list (PhD student): bookmarks: research-related
						5	5	4	Link list (PhD student): bookmarks: research-related + misc
				1	1	1	Link list (PhD student): travel	
			7	6	6	Link list (student) (5 different persons + 1 group (another 5 different persons)		
				1	1	1	Link list (student): research-related	
				1	1	1	Link list (students' group): for student's assignment	
				4	3	3	Link list (student): bookmarks	
						3	2	2	Link list (student): bookmarks: research-related + misc
						1	1	1	N/A Link list (student): bookmarks [deduced from parent page]
				1	1	1	Link list (student): sports	
						38	28	18	**Bookmarks [extracted]**
							2	1	1	Link list (admin staff): bookmarks
							25	17	10	Link list (researcher): bookmarks
							7	7	5	Link list (PhD student): bookmarks
							4	3	3	Link list (student): bookmarks
p.publ		**7**	**7**	**7**	**PUBLICATION/GUIDE (PERS.) (7 different persons)**				
				1	1	1	Book		
				1	1	1	Book (chapter)		
				1	1	1	Paper		

				1	1	1		Paper (copy version)
				1	1	1		Paper (section)
				1	1	1		Paper/resource guide: "Archaeology on the World Wide Web: a user's field-guide"
				1	1	1		Resource guide: biographical guide: "Karl Pearson: A Reader's Guide"
p.soft	**21**	**14**	**4**	**SOFTWARE PROGRAM (PERS.) (4 different persons)**				
				16	10	3		Manual/documentation (copy version)(section): software
				3	3	1		Manual/documentation (section): software
				2	1	1		Manual/tutorial (researcher): software
p.teach	**22**	**17**	**11**	**TEACHING (PERS.) (12 different persons)**				
				13	10	7		Lecturer's teaching pages (7 different persons)
						3	3	3 Course homepage (researcher)
						1	1	1 Course page (researcher): practical task
						8	5	3 Tutorial (researcher): course page
						1	1	1 Tutorial (researcher): academic writing
				9	7	5		Student's assignments (5 different persons + 1 group)
						1	1	1 Student's assignment: paper
						4	4	2 Student's assignment: text-rich resource guide (3 different persons)
						3	1	1 Student's assignment: reference list
						1	1	1 Students' group assignment

Appendix 14. Genres of visited target pages

The table shows subgenres of visited target pages sorted by institutional and personal meta genres, cf. Table 6-19, Section 6.4.5. Legend: The three numbers on each row are counts of page level links, pages, and subsites. For example, *20/12/8* at *i.conf* means that 20 followed links were received by 12 different target pages at 8 different subsites. If a subsite is present in two path nets, the subsite counts as two different path net subsites. Simple sums of subgenres may exceed the count on the step above since, e.g., a subsite may contain more than one page genre.

followed links	visited unique target pages	visited unique target subsites				
352	**249**	**93**	**TARGET PAGES**			
i.archive		**5**	**4**	**4**	**ARCHIVE/DATABASE (INST.)**	
			1	1	1	Archive homepage: British National Corpus
			2	1	1	Archive/database homepage: Dinosaurs
			1	1	1	Archive: Global climate summaries
			1	1	1	Search page (database)
i.conf		**20**	**12**	**8**	**CONFERENCE/WORKSHOP/SEMINAR PAGE**	
			13	6	4	Conference pages
				8	3	2 Conference homepage
				3	1	1 Conference session
				1	1	1 Conference webpage [conference summary]
				1	1	1 Conference webpage: local transport information

				5	5	4	Workshop/Summer school
					1	1	1 Workshop homepage
					1	1	1 Workshop programme
					1	1	1 Workshop webpage: CFP
					1	1	1 Workshop webpage: online registration
					1	1	1 Summer school homepage
				2	1	1	"Informal one-day meeting"
i.generic	**17**	**10**	**7**	**GENERIC/UNI SERVICE**			
				11	4	4	Uni service
					1	1	1 Uni service homepage
					8	2	2 Uni service: Studying abroad (guide)
					2	1	1 Uni service: local transport information
				4	4	2	Sport pages
					1	1	1 Sport Organization homepage
					3	3	1 N/A: deduced by URL: institutional sports pages at atm.cam.ac.uk
				1	1	1	Student Hall homepage
				1	1	1	N/A: deduced by URL and anchor text: institutional tourist info: photo page
i.hp	**80**	**42**	**38**	**INSTITUTIONAL HOMEPAGE**			
				8	4	3	International/national research institution
					1	1	1 Homepage: International scientific society
					1	1	1 Homepage: International research dissemination project
					5	1	**1** Homepage: International research centre
					1	1	1 Homepage: Scientific society homepage
				70	36	34	University research institution
					5	4	4 Homepage: School
					31	14	13 Homepage: Dept
					31	16	16 Homepage: Centre
					3	2	2 Homepage: Lab
				2	2	2	Other university institution
					1	1	1 Homepage: Library
					1	1	1 Homepage: Museum
i.list	**34**	**23**	**17**	**INSTITUTIONAL LINK LIST**			
				30	20	14	Link list

					13	7	6	Link list (institutional): research-related
					7	5	2	Link list (institutional): teaching-related
					4	3	2	Link list (Institutional): image map [same UK map at HN03 3020, NH04 3017, NH05 3017]
					6	5	4	Index page: TOC link list (institutional) [incl. multi-occurring HN051327, NH04 1327, NH05 1327]
				3	2	2	Publication list	
					1	1	1	Publication list (institutional): essays
					2	1	1	Publication list (institutional): joint research project web page
				1	1	1	Staff list (dept)	
i.proj		**24**	**24**	**17**	**RESEARCH PROJECT (INST.)**			
				8	8	6	Joint research project	
					3	3	3	Joint research project homepage
					3	3	2	Joint research project programme
					1	1	1	Joint research project webpage: photo page: meeting
					1	1	1	Joint research project webpage: preliminary results
				5	5	5	Research group	
					4	4	4	Research group homepage
					1	1	1	Research group webpage: research project description
				5	5	4	Research project	
					2	2	2	Research project homepage
					2	2	2	Research project webpage + resource guide
					1	1	1	Research projects overview [placed in personal directory]
				4	4	2	Research group/project webpage deduced by URLs	
				2	2	2	Grants	
					1	1	1	Scientific society: fellowship page
					1	1	1	Research grants
i.publ		**11**	**7**	**7**	**PUBLICATION/GUIDE (INST.)**			
				6	3	3	Journal	
					5	2	2	Homepage: journal
					1	1	1	N/A Journal [deduced from anchor + URL]
				2	2	2	Paper/guide	
					1	1	1	Guide: museum guidebook
					1	1	1	Political statement [copy page]
				2	1	1	N/A Research project report [deduced from anchor + URL]	
				1	1	1	FAQ: satellite imagery	

i.soft	**10**	**9**	**5**	**SOFTWARE PROGRAM (INST.)**				
				1	1	1	Software programme homepage	
				2	1	1	Download page: software	
				1	1	1	Manual/documentation: software	
				2	2	1	Manual/documentation (section): software	
				3	3	2	Manual/tutorial: software	
				1	1	1	FAQ: software	
i.teach	**25**	**18**	**13**	**TEACHING (INST.)**				
				21	15	10	Institutional course pages	
					14	10	8	Course homepage (institutional)
					3	2	2	Course page (institutional): lecture notes
					1	1	1	Tutorial (institutional)
					3	2	1	Tutorial: advice sheet (institutional)
				4	3	3	Other institutional teaching pages	
					3	2	2	National teaching project homepage
					1	1	1	Institutional teaching project homepage
p.archive	**3**	**2**	**2**	**ARCHIVE/DATABASE (PERS.)**				
				1	1	1	Discussion group	
				2	1	1	Mailing list archive	
p.hobby	**15**	**12**	**7**	**PERSONAL HOBBY PAGE (2 different persons)**				
				10	7	5	Researcher's hobby page	
					4	1	1	Personal hobby page (researcher): Greek warship
					3	3	2	Personal hobby page (researcher): sports
					1	1	1	Personal hobby page (researcher): sports (for group of runners)
					2	2	1	Personal hobby page (researchers): sci-fi
				5	5	3	Student's hobby page	
					5	5	3	Personal hobby page (student): sports
p.hp	**47**	**34**	**22**	**PERSONAL HOMEPAGES (33 different persons)**				
				46	33	21	Researcher's homepage	
					2	2	1	Personal homepage (PhD student) (1 person)
					43	30	19	Personal homepage (researcher) (30 different persons)
					1	1	1	Personal webpage (researcher): photos from attended course
				1	1	1	Student's homepage	

				1	1	1	N/A Personal homepage (student) [deduced from anchor text]
p.list	**5**	**5**	**5**	**PERSONAL LINK LIST (5 different persons)**			
				1	1	1	Link list (researcher): teaching-related
				1	1	1	Link list (researcher): info policy
				1	1	1	Link list (researcher): cultural+tourist info (China)
				1	1	1	Link list (researcher): tourist info (Glasgow)
				1	1	1	Link list (staff: technician): charity
p.proj	**5**	**4**	**4**	**RESEARCH PROJECT (PERS.) (3 different persons)**			
				1	1	1	Personal webpage (researcher): research project
				3	2	2	Personal webpage (researcher): research project homepage (cybergeography) [NH01 2745 + NH05 2745]
				1	1	1	Personal webpage (researcher): research projects overview
p.publ	**24**	**22**	**9**	**PUBLICATION/GUIDE (PERS.) (8 different persons + 4 inst. journals/archives)**			
				1	1	1	Book review: personal webpage (researcher)
				4	4	3	Paper: full text (html)
				3	3	2	Paper: full text (postscript)
				2	1	1	Report: postscript
				7	7	1	Paper: abstract (with link to full text)
				1	1	1	Paper/documentation: overview of directory containing mathematical proofs
				2	1	1	Conference poster: Index (root) directory page
				1	1	1	Workshop presentation: summary
				1	1	1	Workshop presentation [N/A: deduced from anchor text]
				1	1	1	Resource guide/tutorial
				1	1	1	Resource guide/tutorial (PhD student): Quechua (Andes language) [text-rich guide]
p.soft	**17**	**14**	**8**	**SOFTWARE PROGRAM (PERS.) (7 different persons + 1 journal + 1 research group)**			
				10	7	5	Manual/documentation: software
				5	5	2	Manual/documentation (section): software
				1	1	1	Manual/tutorial (PhD student): software: wave file format
				1	1	1	Manual/tutorial (postgraduate student): software: cross-network sms messaging
p.teach	**10**	**7**	**6**	**TEACHING (PERS.) (6 different persons)**			
				8	6	5	Lecturer's teaching pages
					5	3	3 Course homepage (researcher)
					1	1	1 Teaching page (researcher): lecturer's teaching resources

		2	2	1 Tutorial (researcher)
2	1	1	Student's assignments	
		2	1	1 Student's assignment: tutorial section: software (Multiview)

Appendix 15. Genre matrix

The adjacency matrix below is based on identified pairs of page genres interconnected by the 352 followed links in the 10 path nets, cf. Section 6.4.5.4 and the table on next page.

	i.archive	i.conf	i.generic	i.hp	i.list	i.proj	i.publ	i.soft	i.teach	p.archive	p.hobby	p.hp	p.list	p.proj	p.publ	p.soft	p.teach	
i.archive	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**
i.conf	0	9	2	4	2	4	0	0	0	0	0	10	0	0	3	0	0	**34**
i.generic	0	0	2	0	4	0	0	0	0	0	0	0	0	0	0	0	0	**6**
i.hp	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**
i.list	7	0	0	33	14	5	5	0	10	0	2	5	0	2	1	1	2	**87**
i.proj	0	2	0	9	1	3	1	2	0	0	0	2	0	0	0	0	0	**20**
i.publ	0	0	0	7	0	1	0	0	0	0	0	0	0	0	0	0	0	**8**
i.soft	0	2	0	0	0	0	0	1	0	0	0	1	0	0	0	2	0	**6**
i.teach	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	**4**
p.archive	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**
p.hobby	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	**3**
p.hp	0	0	0	8	0	2	3	1	2	0	0	4	0	0	0	2	0	**22**
p.list	3	7	7	15	8	8	2	4	7	1	11	10	4	2	17	3	3	**112**
p.proj	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**
p.publ	1	0	0	0	1	0	0	0	0	0	1	2	0	0	1	0	1	**7**
p.soft	0	0	0	2	0	0	0	2	0	2	0	9	0	0	1	5	0	**21**
p.teach	0	0	0	2	4	1	0	0	2	0	1	1	1	1	1	4	4	**22**
	11	**20**	**11**	**80**	**34**	**24**	**11**	**10**	**25**	**3**	**15**	**47**	**5**	**5**	**24**	**17**	**10**	**352**

Pairs of page meta genres connected by the 352 followed page level links. Ranked by most frequently interlinked genre pairs, cf. Table 6-31, Section 6.4.5.4.

source meta genre	target meta genre	count
i.list	i.hp	33
p.list	p.publ	17
p.list	i.hp	15
i.list	i.list	14
p.list	p.hobby	11
i.conf	p.hp	10
i.list	i.teach	10
p.list	p.hp	10
i.conf	i.conf	9
i.proj	i.hp	9
p.soft	p.hp	9
p.hp	i.hp	8
p.list	i.list	8
p.list	i.proj	8
i.list	i.archive	7
i.publ	i.hp	7
p.list	i.conf	7
p.list	i.generic	7
p.list	i.teach	7
i.list	i.proj	5
i.list	i.publ	5
i.list	p.hp	5
p.soft	p.soft	5
i.conf	i.hp	4
i.conf	i.proj	4
i.generic	i.list	4
i.teach	i.teach	4
p.hp	p.hp	4
p.list	i.soft	4
p.list	p.list	4
p.teach	i.list	4
p.teach	p.soft	4
p.teach	p.teach	4
i.conf	p.publ	3
i.proj	i.proj	3
p.hobby	p.hp	3
p.hp	i.publ	3
p.list	i.archive	3
p.list	p.soft	3
p.list	p.teach	3

source meta genre	target meta genre	count
i.conf	i.generic	2
i.conf	i.list	2
i.generic	i.generic	2
i.list	p.hobby	2
i.list	p.proj	2
i.list	p.teach	2
i.proj	i.conf	2
i.proj	i.soft	2
i.proj	p.hp	2
i.soft	i.conf	2
i.soft	p.soft	2
p.hp	i.proj	2
p.hp	i.teach	2
p.hp	p.soft	2
p.list	i.publ	2
p.list	p.proj	2
p.publ	p.hp	2
p.soft	i.hp	2
p.soft	i.soft	2
p.soft	p.archive	2
p.teach	i.hp	2
p.teach	i.teach	2
i.list	p.publ	1
i.list	p.soft	1
i.proj	i.list	1
i.proj	i.publ	1
i.publ	i.proj	1
i.soft	i.soft	1
i.soft	p.hp	1
p.hp	i.soft	1
p.list	p.archive	1
p.publ	i.archive	1
p.publ	i.list	1
p.publ	p.hobby	1
p.publ	p.publ	1
p.publ	p.teach	1
p.soft	p.publ	1
p.teach	i.proj	1
p.teach	p.hobby	1
p.teach	p.hp	1
p.teach	p.list	1
p.teach	p.proj	1
p.teach	p.publ	1

Appendix 16. Transversal link paths

In the table below, all 81 followed link paths in the 10 path nets are listed. Transversal links are denoted with bold right angle brackets (**>**) if they are research-related; with hash signs (#) if they reflect a personal non-academic link. A non-bold right angle bracket (>) denotes other transversal links. Generic (univ-service type) subsites are marked with a red (dark) 'gen'. 'Non-generic' paths do not pass any generic subsites (univ-service type). 'CS' marks a link path that passes a computer-science-related subsite (marked with yellow/grey). Abbreviations for subsite topics are listed further below.

path net	link path	path net level 0		level 1		level 2		level 3		level 4	non-gen.	cs
HN01	HN01-01	2099hum	-	119gen	-	341atm	-	1904atm				
	HN01-02	2099hum	-	119gen	-	2182met	-	1904atm				
	HN01-03	2099hum	-	119gen	-	2393cs	**>**	1904atm				cs
	HN01-04	2099hum	-	710hum	-	337hum	**>**	1904atm			**x**	**./.cs**
	HN01-05	2099hum	-	1424gen	-	2745geo	>	1904atm				
	HN01-06	2099hum	>	1612cs	#	341atm	-	1904atm			**x**	cs
	HN01-07	2099hum	>	1612cs	>	2745geo	>	1904atm			**x**	cs
	HN01-08	2099hum	-	1866gen	-	335gen	-	1904atm				
	HN01-09	2099hum	#	2387cs/ee	#	341atm	-	1904atm			**x**	cs
	HN01-10	2099hum	#	2387cs/ee	-	1613gen	-	1904atm				cs
HN02	HN02-01	2394econ	**>**	3006cs	-	1088cs/ee	#	917chem			**x**	cs
	HN02-02	2394econ	**>**	3006cs	-	1328cs/ma	>	917chem			**x**	cs
	HN02-03	2394econ	**>**	3006cs	-	2865cs	#	917chem			**x**	cs
HN03	HN03-01	1494psy	>	3020cs	-	772cs	**>#**	318ma	-	893ma	**x**	cs
	HN03-02	1494psy	>	3020cs	-	772cs	-	1089cs/ma	-	893ma	**x**	cs
	HN03-03	1494psy	>	3020cs	-	772cs	>	1225ma	-	893ma	**x**	cs
	HN03-04	1494psy	>	3020cs	-	1773cs/ma	-	318ma	-	893ma	**x**	cs
	HN03-05	1494psy	>	3020cs	-	1773cs/ma	-	1089cs/ma	-	893ma	**x**	cs
	HN03-06	1494psy	>	3020cs	-	1773cs/ma	-	1225ma	-	893ma	**x**	cs
HN04	HN04-01	871ling	**>**	2615ee	-	1300gen	-	245earth				
HN05	HN05-01	2068geo	-	1327geo	-	1613gen	-	1885med				
	HN05-02	2068geo	-	1345geo	-	1613gen	-	1885med				
	HN05-03	2068geo	-	2745geo	-	1613gen	-	1885med				
NH01	NH01-01	1904atm	-	1278environ	-	1451hum(incl. geo)	-	2099hum				**./.cs**
	NH01-02	1904atm	-	1357learn/gen	-	1451hum	-	2099hum				
	NH01-03	1904atm	**>#**	2615ee	**>**	1451hum	-	2099hum			**x**	**./.cs**
	NH01-04	1904atm	#	2744ling	-	1451hum	-	2099hum			**x**	**./.cs**
	NH01-05	1904atm	#	2744ling	-	313ling	-	2099hum			**x**	**./.cs**
NH02	NH02-01	917chem	#	1088cs/ee	>	1485econ	-	230econ/learn	-	2394econ	**x**	cs
	NH02-02	917chem	#	1088cs/ee	>	2083econ	-	230econ/learn	-	2394econ	**x**	cs
	NH02-03	917chem	#	1597cs	>	1890soc(incl. econ)	-	230econ/learn	-	2394econ	**x**	cs
	NH02-04	917chem	#	2537cs/is	>	1485econ	-	230econ/learn	-	2394econ	**x**	cs
	NH02-05	917chem	**>**	2642cs/cog	>	1641chem	**>**	230econ/learn	-	2394econ	**x**	cs
	NH02-06	917chem	#	2760cs	**>**	1485econ	-	230econ/learn	-	2394econ	**x**	cs
	NH02-07	917chem	#	2760cs	**>**	1641chem	**>**	230econ/learn	-	2394econ	**x**	cs
	NH02-08	917chem	#	2760cs	>	1890soc(incl. econ)	-	230econ/learn	-	2394econ	**x**	cs
NH03	NH03-01	893ma	**>**	979astro	>	1268cs	**>**	1494psy			**x**	cs
	NH03-02	893ma	**>**	979astro	>	2760cs	**>**	1494psy			**x**	cs

NH04	NH04-01	245earth	-	1343earth	-	1327geo	**>**	2372cs/ee	**>**	871ling	**x**	cs
	NH04-02	245earth	-	1343earth	-	1327geo	**>**	2387cs/ee	**>**	871ling	**x**	cs
	NH04-03	245earth	-	1343earth	**>**	1619ma	>	201cs	**>**	871ling	**x**	cs
	NH04-04	245earth	-	1343earth	**>**	1692cs/is	-	2387cs/ee	**>**	871ling	**x**	cs
	NH04-05	245earth	-	1343earth	>	3017cs	-	201cs	**>**	871ling	**x**	cs
	NH04-06	245earth	-	1343earth	>	3017cs	-	2387cs/ee	**>**	871ling	**x**	cs
	NH04-07	245earth	-	1343earth	>	3017cs	>	2744ling	-	871ling	**x**	cs
	NH04-08	245earth	-	1853hum/archaeo	-	337hum	-	2744ling	-	871ling	**x**	**./.cs**
	NH04-09	245earth	-	1853hum/archaeo	**>**	1572cs/ma	-	201cs	**>**	871ling	**x**	cs
	NH04-10	245earth	-	1853hum/archaeo	**>**	1572cs/ma	-	2387cs/ee	**>**	871ling	**x**	cs
	NH04-11	245earth	-	1853hum/archaeo	**>**	1572cs/ma	>	2744ling	-	871ling	**x**	cs
	NH04-12	245earth	-	1889earth	**>**	732cs/ma	>	2744ling	-	871ling	**x**	cs
	NH04-13	245earth	-	1889earth	-	1327geo	**>**	2372cs/ee	**>**	871ling	**x**	cs
	NH04-14	245earth	-	1889earth	-	1327geo	**>**	2387cs/ee	**>**	871ling	**x**	cs
	NH04-15	245earth	-	1889earth	**>**	1473ma	>	2387cs/ee	**>**	871ling	**x**	cs
	NH04-16	245earth	-	2228earth	**>**	213ee	#	2744ling	-	871ling	**x**	**./.cs**
	NH04-17	245earth	-	2228earth	-	1327geo	**>**	2372cs/ee	**>**	871ling	**x**	cs
	NH04-18	245earth	-	2228earth	-	1327geo	**>**	2387cs/ee	**>**	871ling	**x**	cs
	NH04-19	245earth	-	2356earth	**>**	629cs/ee	-	201cs	**>**	871ling	**x**	cs
	NH04-20	245earth	-	2356earth	**>**	629cs/ee	-	2372cs/ee	**>**	871ling	**x**	cs
	NH04-21	245earth	-	2356earth	**>**	629cs/ee	-	2387cs/ee	**>**	871ling	**x**	cs
	NH04-22	245earth	-	2356earth	**>**	732cs	>	2744ling	-	871ling	**x**	cs
	NH04-23	245earth	-	2356earth	**>**	1088cs/ee	-	201cs	**>**	871ling	**x**	cs
	NH04-24	245earth	-	2356earth	**>**	1088cs/ee	-	2387cs/ee	**>**	871ling	**x**	cs
	NH04-25	245earth	-	2356earth	**>**	1088cs/ee	>	2744ling	-	871ling	**x**	cs
	NH04-26	245earth	-	2356earth	-	1327geo	**>**	2372cs/ee	**>**	871ling	**x**	cs
	NH04-27	245earth	-	2356earth	-	1327geo	**>**	2387cs/ee	**>**	871ling	**x**	cs
	NH04-28	245earth	-	2356earth	-	1709geo	**>**	2372cs/ee	**>**	871ling	**x**	cs
	NH04-29	245earth	-	2858environ	**>**	1572cs/ma	-	201cs	**>**	871ling	**x**	cs
	NH04-30	245earth	-	2858environ	**>**	1572cs/ma	-	2387cs/ee	**>**	871ling	**x**	cs
	NH04-31	245earth	-	2858environ	**>**	1572cs/ma	>	2744ling	-	871ling	**x**	cs
	NH04-32	245earth	-	2858environ	-	1709geo	**>**	2372cs/ee	**>**	871ling	**x**	cs
	NH04-33	245earth	-	2858environ	**>**	2865cs	-	201cs	**>**	871ling	**x**	cs
	NH04-34	245earth	-	2858environ	**>**	2865cs	-	2387cs/ee	**>**	871ling	**x**	cs
	NH04-35	245earth	-	2858environ	**>**	2865cs	>	2744ling	-	871ling	**x**	cs
	NH04-36	245earth	-	2858environ	-	3060archaeo	**>**	2387cs/ee	**>**	871ling	**x**	cs
NH05	NH05-01	1885med	-	102med	-	226gen	-	1327geo	-	2068geo		
	NH05-02	1885med	-	102med	**>**	917chem	-	2540gen	-	2068geo		
	NH05-03	1885med	-	102med	**>**	922hum	>	1327geo	-	2068geo	**x**	**./.cs**
	NH05-04	1885med	-	102med	-	1812gen	-	1327geo	-	2068geo		
	NH05-05	1885med	-	102med	-	1866gen	-	1327geo	-	2068geo		
	NH05-06	1885med	-	102med	>	3017cs	-	2540gen	-	2068geo		
	NH05-07	1885med	-	913med	-	2088gen/multi-sci	-	1327geo	-	2068geo		

abbrev.	**subsite topic**
astro	astronomy
atm	atmospheric sciences
chem	chemistry
cs	computer science
cs/cog	computer science & cognitive sciences
cs/ee	computer science & electrical engineering/electronics
cs/is	computer science & info. science/technology
cs/ma	computer science & mathematics
earth	earth sciences
econ	economics
econ/learn	economics: learning tech.
ee	electrical engineering
environ	environmental studies
gen	generic
geo	geography
hum	arts & humanities
hum/archae	archaeology
learn/gen	generic: learning technology
ling	linguistics
ma	mathematics
med	medicine
met	meteorology
psy	psychology
soc	social sciences

Appendix 17. Distribution of source genres with transversal outlinks

Distribution of transversal outlinks from source genres (divided in institutional and personal) to target genres. The list is sorted by frequency for each source genre.

source genre	target genre	# transversal outlinks	sub total
i.conf	i.hp	2	8
i.conf	i.list	2	
i.conf	p.hp	2	
i.conf	i.proj	1	
i.conf	p.publ	1	
i.generic	i.list	4	4
i.list	i.hp	12	27
i.list	i.generic	5	
i.list	i.proj	3	
i.list	i.list	2	
i.list	i.publ	2	
i.list	p.hobby	2	
i.list	i.teach	1	
i.proj	i.soft	2	5
i.proj	i.conf	1	
i.proj	i.hp	1	
i.proj	p.hp	1	
i.publ	i.proj	1	1
i.soft	i.soft	1	3
i.soft	p.hp	1	
i.soft	p.soft	1	
p.hobby	p.hp	1	1
p.hp	p.hp	2	5
p.hp	i.hp	1	
p.hp	i.soft	1	
p.hp	p.soft	1	
p.list	p.publ	9	45
p.list	p.hobby	8	
p.list	i.generic	4	
p.list	i.hp	4	
p.list	i.list	4	
p.list	p.hp	3	
p.list	p.list	3	
p.list	i.proj	2	
p.list	i.publ	2	
p.list	i.soft	2	
p.list	p.proj	2	
p.list	p.archive	1	
p.list	p.teach	1	
p.publ	i.list	1	2
p.publ	p.publ	1	
p.soft	i.soft	2	4
p.soft	i.hp	1	
p.soft	p.soft	1	
p.teach	p.soft	3	7
p.teach	i.list	1	
p.teach	p.hobby	1	
p.teach	p.list	1	
p.teach	p.teach	1	
		112	**112**

Appendix 18. Distribution of target genres with transversal inlinks

Distribution of transversal inlinks to target genres (divided in institutional and personal) from source genres. The list is sorted by frequency for each target genre.

source genre	target genre	# transversal inlinks	sub total
i.proj	i.conf	1	1
i.list	i.generic	5	9
p.list	i.generic	4	
i.list	i.hp	12	21
p.list	i.hp	4	
i.conf	i.hp	2	
i.proj	i.hp	1	
p.hp	i.hp	1	
p.soft	i.hp	1	
i.generic	i.list	4	14
p.list	i.list	4	
i.conf	i.list	2	
i.list	i.list	2	
p.publ	i.list	1	
p.teach	i.list	1	
i.list	i.proj	3	7
p.list	i.proj	2	
i.conf	i.proj	1	
i.publ	i.proj	1	
i.list	i.publ	2	4
p.list	i.publ	2	
i.proj	i.soft	2	8
p.list	i.soft	2	
p.soft	i.soft	2	
i.soft	i.soft	1	
p.hp	i.soft	1	
i.list	i.teach	1	1
p.list	p.archive	1	1
p.list	p.hobby	8	11
i.list	p.hobby	2	
p.teach	p.hobby	1	
p.list	p.hp	3	10
i.conf	p.hp	2	
p.hp	p.hp	2	
i.proj	p.hp	1	
i.soft	p.hp	1	
p.hobby	p.hp	1	
p.list	p.list	3	4
p.teach	p.list	1	
p.list	p.proj	2	2
p.list	p.publ	9	11
i.conf	p.publ	1	
p.publ	p.publ	1	
p.teach	p.soft	3	6
i.soft	p.soft	1	
p.hp	p.soft	1	
p.soft	p.soft	1	
p.list	p.teach	1	2
p.teach	p.teach	1	
		112	**112**

Appendix 19. Source genres with transversal outlinks

The table shows subgenres of visited transversal source pages sorted by institutional and personal meta genres, cf. Table 6-19, Section 6.4.5. Legend: The three numbers on each row are counts of transversal outlinks, pages, and subsites. For example, *8/6/6* at *i.conf* means that 8 followed transversal outlinks originate from 6 different source pages at 6 different subsites. If a subsite is present in two path nets, the subsite counts as two different path net subsites. Simple sums of subgenres may exceed the count on the step above since, e.g., a subsite may contain more than one page genre.

followed links	visited unique source pages	visited unique source subsites				
112	**95**	**48**	**TRANSVERSAL SOURCE PAGES**			
i.conf			**8**	**6**	**6**	**CONFERENCE/WORKSHOP/SEMINAR PAGE**
			2	2	2	Workshop page
				1	1	1 Workshop homepage
				1	1	1 Workshop programme
			6	4	4	Seminar page
				2	2	2 Seminar programme: presentation abstract
				4	2	2 Seminar series overview
i.generic			**4**	**4**	**1**	**GENERIC/UNI SERVICE**
			4	4	1	Web server statistics
i.list			**27**	**27**	**15**	**INSTITUTIONAL LINK LIST**
			24	24	14	Link list (institutional)
				15	15	10 Link list (institutional): research-related
				3	3	2 Link list (institutional): generic: academic sites

					1	1	1 Link list (institutional): teaching-related
					5	5	1 Link list (institutional): studying abroad
				3	3	2	Publication list (institutional)
							incl. 2/2/1 publication lists (institutional): joint research project annual reports
i.proj	**5**	**4**	**4**	**RESEARCH PROJECT (INST.)**			
				2	1	1	Joint research group homepage
				1	1	1	Joint research project homepage
				1	1	1	Research group homepage
				1	1	1	Research projects overview (dept.)
i.publ	**1**	**1**	**1**	**PUBLICATION/GUIDE (INST.)**			
				1	1	1	Report/documentation: research project documentation
i.soft	**3**	**3**	**2**	**SOFTWARE PROGRAM (INST.)**			
				1	1	1	Manual/documentation: software (joint research project web page)
				1	1	1	Manual/documentation: software demo (joint research project web page)
				1	1	1	Manual/tutorial (institutional)(section):software (joint research project)
p.hobby	**1**	**1**	**1**	**PERSONAL HOBBY PAGE**			
				1	1	1	Personal hobby page (researcher): Saxon shore forts
p.hp	**5**	**5**	**4**	**PERSONAL HOMEPAGE (19 different persons)**			
				3	3	3	Personal homepage (PhD student): incl. long link list; papers & presentations
				17	14	11	Personal homepage (researcher): incl. link list; CV (3); educational background, publication list
				1	1	1	Personal homepage (student)
				1	1	1	Personal homepage (technical staff)
p.list	**45**	**34**	**18**	**PERSONAL LINK LIST (23 different persons incl.1 multi-occurring)**			
				7	1	1	Bibliography (researcher) (also has link list below)
				2	1	1	Link list (admin staff): bookmarks
				30	26	15	Link list (researcher) (16 different persons incl. 1 multi-occurring)
					4	4	3 Link list (researcher): research-related
					1	1	1 Link list (researcher): personal academic interest (grammar)
					3	2	2 Link list (researcher): UK academic sites
					1	1	1 Link list (researcher): friends+scientists
					1	1	1 Link list (researcher): leisure
					2	2	1 Link list (researcher): sci-fi
					8	7	3 Link list (researcher): sports
					10	8	6 Link list (researcher): bookmarks (6 different persons)
						1	1 1 Link list (researcher): bookmarks: research-related
						4	3 3 Link list (researcher): bookmarks: research-related + misc

						5	4	2	Link list (researcher): bookmarks: personal academic interest
			3	3	3	Link list (PhD student) (3 different persons)		
					1	1	1	Link list (PhD student): research-related
					2	2	2	Link list (PhD student): bookmarks
						1	1	1	Link list (PhD student): bookmarks: research-related
						1	1	1	Link list (PhD student): bookmarks: research-related + misc
			3	3	3	Link list (student) (3 different persons)		
					2	2	2	Link list (student): bookmarks
						1	1	1	Link list (student): bookmarks: research-related + misc
						1	1	1	N/A Link list (student): bookmarks [deduced from parent page]
					1	1	1	Link list (student): sports
						16	13	9	**Bookmarks [extracted]**
						2	1	1	Link list (admin staff): bookmarks
						10	8	6	Link list (researcher): bookmarks
						2	2	2	Link list (PhD student): bookmarks
						2	2	2	Link list (student): bookmarks
p.publ		**2**	**2**	**2**	**PUBLICATION/GUIDE (PERS.) (2 different persons)**				
					1	1	1	Paper/resource guide: "Archaeology on the World Wide Web: a user's field-guide"
					1	1	1	Resource guide: biographical guide: "Karl Pearson: A Reader's Guide"
p.soft		**4**	**3**	**3**	**SOFTWARE PROGRAM (PERS.) (3 different persons)**				
					2	2	2	Manual/documentation (copy version)(section): software
					2	1	1	Manual/tutorial (researcher): software
p.teach		**7**	**5**	**5**	**TEACHING (PERS.) (5 different persons)**				
					3	3	3	Lecturer's teaching pages (3 different persons)
						1	1	1	Course homepage (researcher)
						1	1	1	Tutorial (researcher): course page
						1	1	1	Tutorial (researcher): academic writing
					4	2	2	Student's assignments (2 different persons)
						1	1	1	Student's assignment: paper
						3	1	1	Student's assignment: reference list

Appendix 20. Target genres with transversal inlinks

The table shows subgenres of visited transversal target pages sorted by institutional and personal meta genres, cf. Table 6-19, Section 6.4.5. Legend: The three numbers on each row are counts of transversal links, pages, and subsites. For example, *21/16/15* at *i.hp* means that 21 followed transversal links were received by 16 different target pages at 15 different subsites. If a subsite is present in two path nets, the subsite counts as two different path net subsites. Simple sums of subgenres may exceed the count on the step above since, e.g., a subsite may contain more than one page genre.

followed links	visited unique target pages	visited unique target subsites				
112	**94**	**47**	**TRANSVERSAL TARGET PAGES**			
i.conf			**1**	**1**	**1**	**CONFERENCE/WORKSHOP/SEMINAR PAGE**
			1	1	1	Summer school homepage
i.generic			**9**	**5**	**3**	**GENERIC/UNI SERVICE**
			5	1	1	Uni service: Studying abroad (guide)
			3	3	1	N/A: deduced by URL: institutional sports pages at atm.cam.ac.uk
			1	1	1	N/A: deduced by URL and anchor text: institutional tourist info: photo page
i.hp			**21**	**16**	**15**	**INSTITUTIONAL HOMEPAGE**
			1	1	1	Homepage: International research dissemination project
			1	1	1	Homepage: School
			9	5	4	Homepage: Dept
			8	8	8	Homepage: Centre
			2	1	1	Homepage: Lab

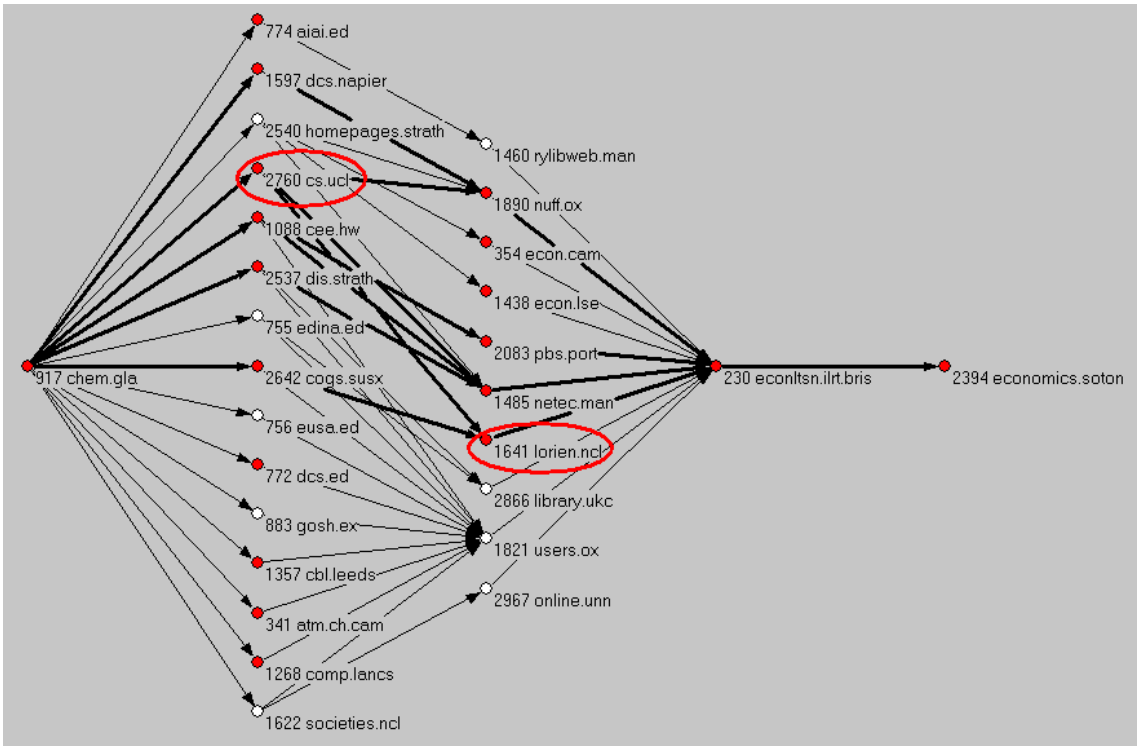
i.list	**14**	**9**	**9**	**INSTITUTIONAL LINK LIST**				
				11	7	7	Link list	
						8	4	4 Link list (institutional): research-related
						3	3	3 Link list (Institutional): same UK map at HN03 3020, NH04 3017, NH05 3017
				3	2	2	Publication list	
						1	1	1 Publication list (institutional): essays
						2	1	1 Publication list (institutional): joint research project web page
i.proj	**7**	**7**	**5**	**RESEARCH PROJECT (INST.)**				
				1	1	1	Joint research project webpage: preliminary results	
				1	1	1	Research group homepage	
				1	1	1	Research project homepage	
				2	2	2	Research project webpage + resource guide	
				2	2	1	N/A: Research group/project webpage deduced by URLs	
i.publ	**4**	**3**	**3**	**PUBLICATION/GUIDE (INST.)**				
				1	1	1	Homepage: journal	
				1	1	1	Political statement [copy page]	
				2	1	1	N/A Research project report [deduced from anchor + URL]	
i.soft	**8**	**7**	**4**	**SOFTWARE PROGRAM (INST.)**				
				2	1	1	Download page: software	
				1	1	1	Manual/documentation: software	
				2	2	1	Manual/documentation (section): software	
				2	2	1	Manual/tutorial: software	
				1	1	1	FAQ: software	
i.teach	**1**	**1**	**1**	**TEACHING (INST.)**				
				1	1	1	Institutional teaching project homepage	
p.archive	**1**	**1**	**1**	**ARCHIVE/DATABASE (PERS.)**				
				1	1	1	Discussion group	
p.hobby	**11**	**10**	**6**	**PERSONAL HOBBY PAGE**				
				8	7	5	Researcher's hobby page	
						2	1	1 Personal hobby page (researcher): Greek warship
						3	3	2 Personal hobby page (researcher): sports
								incl. one N/A: deduced from URL and source page
						1	1	1 Personal hobby page (researcher): sports (for group of runners)
						2	2	1 Personal hobby page (researchers): sci-fi
				3	3	2	Student's hobby page	
						3	3	2 Personal hobby page (student): sports

								incl. one N/A: deduced from URL and source page
p.hp	**10**	**9**	**8**	**PERSONAL HOMEPAGES**				
				9	8	7	Researcher's homepage	
					8	7	6	Personal homepage (researcher)
								incl.1 N/A: deduced from anchor text + URL
					1	1	1	Personal webpage (researcher): photos from attended course
				1	1	1	Student's homepage	
					1	1	1	N/A Personal homepage (student) [deduced from anchor text]
p.list	**4**	**4**	**4**	**PERSONAL LINK LIST**				
				1	1	1	Link list (researcher): research topics	
				1	1	1	Link list (researcher): info policy	
				1	1	1	Link list (researcher): tourist info (Glasgow)	
				1	1	1	Link list (staff: technician): charity	
p.proj	**2**	**2**	**2**	**RESEARCH PROJECT (PERS.)**				
				1	1	1	Personal webpage (researcher): research project	
				1	1	1	Personal webpage (researcher): research project homepage (cybergeography)	
p.publ	**11**	**11**	**5**	**PUBLICATION/GUIDE (PERS.)**				
				1	1	1	Book review: personal webpage (researcher)	
				2	2	2	Paper: full text (html)	
				7	7	1	Paper: abstract (with link to full text)	
				1	1	1	Workshop presentation [N/A: deduced from anchor text]	
p.soft	**6**	**6**	**4**	**SOFTWARE PROGRAM (PERS.)**				
				3	3	3	Manual/documentation: software	
				3	3	1	Manual/documentation (section): software	
p.teach	**2**	**2**	**1**	**TEACHING (PERS.)**				
				2	2	1	Tutorial (researcher)	

Appendix 21. Example of transversal source page: bookmark list

In path net NH02, cf. figure below, a computer scientist at the Department of Computer Science, University College London (node 2760 *cs.ucl.ac.uk* at path net level 1) has a bookmark list (retrieved from the Internet Archive – indexed in the Archive 22 July 2001: *web.archive.org/web/20010722225655/http://www.cs.ucl.ac.uk/staff/K.Vekaria/bookmarks.html*)

The bookmark list (entirely reproduced below) contains a transversal link to the Department of Chemical and Process Engineering, Newcastle University (node 1641 *lorien.ncl.ac.uk* at path net level 2) regarding evolutionary computation and genetic programming. The transversal link is highlighted in the bookmark list. The bookmark list also includes general and leisure-related links as shown below.



Bookmarks for Kanta VEKARIA

ANN
http://web.archive.org/web/20010722225655/http://robotics.stanford.edu/people/nilsson/mlbook.html

Soar
Latest version of Soar FAQ (TEXT)
Encoding A Task in Soar

search engines
AltaVista Search: Main Page
Infoseek

Resources
Raffaele Gaioni - GA: Resources
New Scientist Planet Science - News, jobs and more from the leading weekly magazine
BIOLOGY 151 - GENERAL BIOLOGY I
Description of NK-landscape model
choice
World Guide To Vegetarianism
UK Street Map
The Weather Channel
Sun SITE Northern Europe
Numerical Recipes in C
The Collection of Computer Science Bibliographies
Contents of ACM Journals
Xenon Labs: The Universal Currency Converter(tm)
UCL Library: UCL-restricted Electronic Journals listed by title
UCL Library Home Page
InterBook homepage
Timetables
Welcome to How Stuff Works
Hotel Reservations: Online discount travel reservations for hotels, resorts, &, Inns

Artificial Intelligence
Artificial Evolution Archive
Early Views of Intelligence
Topic: areas/fuzzy/systems/
CMU Artificial Intelligence Repository

Leisure
Veggies Unite!
Restaurants UK.
Fine Art Prints and Posters from Artprint Collection
Amazon.com: Books, Music & More!
Music Boulevard - The World's #1 Online Music Store!
The Little Prince

The Internet Movie Database. A database of more than 125,000 movies
Albemarle of London's West End Theatre Guide - what's on in London Theater
Shotokan-Ryu Kata
O'Kaigan Shotokan Karate-Do - Shotokan Karate Katas
Other Martial Arts links
The Karate Union of Great Britain's Home Page
Shotokan Karate Magazine
The London Pages - Vegetarian Restaurant Pages
Film Finder
Icons, Textures, and Art Stuff
The Backgrounds Archive
CINEASC UK Internet Film Guide
Just-So Literary Postcards
Albemarle of London's West End Theatre Guide - what's on in London Theater
Hotel Reservations: Online discount travel reservations for hotels, resorts, &, Inns
Picks' Tutorial to the Simpsons | Image Gallery

Evolutionary Computation

Justinian Rosca's Home Page
William M. Spears: Publications
The Genetic Programming Notebook
GARAGe Software
Advances in Genetic Programming: Volume 2
NSI Online Library - Pete Angeline
William Langdon
Thesis -- Introduction
http://web.archive.org/web/20010722225655/http://www-cs-faculty.stanford.edu/~koza/gp97.html
John Koza's Home Page
http://web.archive.org/web/20010722225655/http://lorien.ncl.ac.uk/sorg/
[LB note: Internet Archive's conversion of original link targeted to http://lorien.ncl.ac.uk/sorg]
Genetic Programming FAQ
Publications by C.H.M. van Kemenade
Welcome to ENCORE!
MSU GARAGe home page
Publications, A. Wu

Comp Based Learning

Stewart Wilson
Episodic Learner Model - Adaptive Remote Tutor
CALM Home Page
Computational Mathetics
Computer Based Learning Unit: Projects

Java

Java Blast

Mocha - the Java decompiler

CGI stuff

The CGI Resource Index: Programs and Scripts: Perl
WWW Security FAQ: CGI Scripts
ELECTRONIC FORUMS AND OTHER EC-RELATED ACTIVITIES
QuantumWave Interactive - Controlling Shockwave from JavaScript
OR-LIBRARY
ICEC '98 (WCCI '98) CFP --- Int'l Conference on Evolutionary Computation
IlliGAL Home Page
Send a message
GECCO 99
Welcome to GoLive CyberStudio 3
NCSTRL Home Page
Barclaycard Netlink Home Page
BBC News | Front Page | front page
: By Thread
Amazon.co.uk Thanks You
Artificial Evolution 99 (EA'99)
WSC4