

# Small-World Linkage and Co-Linkage

*Lennart Björneborn*

Royal School of Library and Information Science, Denmark

E-mail: lb@db.dk

## ABSTRACT

The paper presents ideas from a current research project concerned with link structures and small-world phenomena on the WWW, with possible implications for knowledge discovery or 'web mining'. The project includes case studies of so-called co-linkage chains consisting of co-linking and co-linked web nodes (analogous to bibliographic couplings and co-citations) in a context of researchers' homepages and published bookmark lists. Key concepts are so-called transversal links and transversal co-linkages (on co-linkage chains) functioning as short cuts or 'weak ties' between heterogeneous subject domains and interest communities on the Web. According to a hypothesis in the project, transversal links make the Web more strongly connected and 'crumpled up' by creating small-world phenomena in the shape of short distances between nodes in the Web graph.

**KEYWORDS:** WWW, link structure analysis, citation analysis, transversal links, co-linkage chains, small-world phenomena, knowledge discovery, web mining

## INTRODUCTION

The Web may be conceived as an ecological system [e.g., 7], that is self-organised and multi-agent constructed by millions of laymen, researchers, institutions, companies, etc., that dynamically create, adapt and remove web pages and links. The local 'anarchistic' behaviour of these diverse web 'weavers' is usually considered to have negative consequences on the global performance of the Web as a hypertext system. But so-called transversal links [2] functioning as short cuts or 'weak ties' between heterogeneous subject domains and interest communities may be a usable feature of this 'imperfect' behaviour, with possible implications for knowledge discovery or 'web mining' [5]. This and other ideas presented in the paper stem from a current research project concerned with link structures and small-world phenomena [11] on the Web, drawing on graph theory and bibliometric citation analysis [2].

## TRANSVERSAL LINKS

By connecting heterogeneous subject domains on the Web, transversal links may affect possibilities for human serendipity and computer-supported knowledge discovery, when unexpected but potentially useful information is encountered and extracted. The idea of transversal links is inspired by Bush's [4] vision of 'Memex' with associative 'trails' that interlink text paragraphs, e.g., transversely to classificational hierarchies with implications for scientific innovativeness. A human or digital agent traversing the Web by following links from web page to web page has the possibility to move from one subject domain (e.g., in information science) to another 'distant' domain (e.g., in creativity research) using a single transversal link (e.g., on a researcher's published bookmark list) as a short cut. According to a hypothesis in the project, transversal links make the Web more strongly connected and 'crumpled up' by creating small-world phenomena in the shape of short distances between nodes in the Web graph. Transversal links thus give a new significance to the social network analytic notion of 'the strength of weak ties' [6].

## SMALL-WORLD PHENOMENA

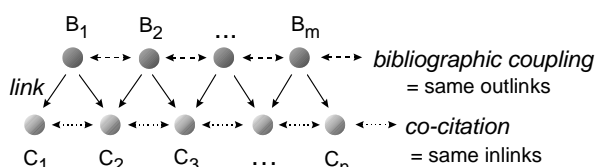
In graph theoretic terms so-called small-world networks [11] have highly clustered nodes as in regular graphs, yet characteristic path lengths between pairs of nodes are short as in random graphs. In a small-world network it is sufficient with a very small percentage of edges functioning as short cuts (i.e. 'transversal links') connecting 'distant' nodes of the network. Small-world phenomena occur in a wide variety of biological, technological and social networks [11]. There is still a lack of research on small-world phenomena and their possible usabilities regarding different types of nodes and edges in informational networks such as the Web [1,2], bibliographic [8] and citation databases, semantic networks, thesauri, etc. In this context, the research project is concerned with developing methods to identify and locate two different types of transversal edges on the Web: (1) on directed link paths, (2) on so-called co-linkage chains consisting of co-linking and co-linked nodes (analogous to bibliographic couplings and co-citations, cf. fig. 1).

## LINK PATHS AND CO-LINKAGE CHAINS

One way of locating transversal links on link paths between web pages would be to use large-scale link structure data (as in the Connectivity Server [3]) and select

start and end web pages from two heterogeneous scientific domains, and then identify transversal links on the shortest link path (if any [3]) connecting the web pages. Large-scale data is essential since the exclusion of relatively few transversal links may affect the size of strongly connected components in the Web graph. In the research project efforts are made to establish access to such large-scale link structure data.

The second type of transversal edges, in co-linkage chains, is inspired by research on literature-based knowledge discovery using so-called ‘indirect multi-step co-citations’ [9] (named ‘co-citation chains’ in the project) between different scientific fields to handle ‘undiscovered public knowledge’ [10]. Using fig. 1 on this approach: If literature in scientific domain  $C_1$  is never co-cited directly with literature  $C_n$ , then literatures  $C_2, \dots$  transitively connecting  $C_1$  to  $C_n$  may reveal implicit relationships or patterns between ideas and concepts not considered before.



**Figure 1: Co-linkage chain**

In the project case studies comprise co-linkage chains consisting of co-linked (co-cited) researchers’ homepages and co-linking (bibliographically coupled) bookmark lists (and similar types of ‘hotlists’), the Cs and Bs respectively in fig. 1. Published bookmark lists are of special interest, since their diverse contents may provide transversal relations on link paths and co-linkage chains between heterogeneous subject domains. Bookmark lists reflect trails of varied interests, preferences and actions on the Web, and thus constitute an obvious area for scientometric and webometric investigation [2]. Some of the links on such lists may reflect emerging cross-disciplinary ‘research fronts’ or ‘invisible colleges’ in the evolving interconnectedness of science.

The case studies include, e.g., a co-linkage chain of 5 co-linked and 4 co-linking nodes with research interests ranging from small-world networks to distributed knowledge systems, interdisciplinary studies, philosophy of mind, education research, and linguistics. Co-linkage chains are constructed by alternate steps ( $C_1, B_1, C_2, B_2, C_3, \dots$ ) of selecting a researcher’s bookmark list with outlinks to other researchers’ homepages, and selecting a researcher’s homepage with inlinks from other researchers’ bookmark lists. Inlink analysis is conducted using AltaVista’s advanced search features, with inherent bias of search engine coverage, performance, etc. [2]. Assessing heterogeneity between researchers’ scientific domains

is necessary in order to identify transversal co-linkages (as ‘weak ties’ contrary to the strong co-citations in [9]) between co-linking or co-linked nodes. This is not uncomplicated. Endeavours are made to establish criteria of definition, e.g., by using low co-linkage frequency compared with low co-citation frequency in citation databases.

## CONCLUDING REMARKS

The presented ideas from the research project indicate complementarities between ‘convergent’ and ‘divergent’ link structures on the Web, with subject-specific domains and interest communities (‘web clusters’) corresponding to the former type and transversal links and co-linkages to the latter. These different link structures may support exploration of the Web conducted in convergent (i.e. rational, goal-directed) ways complemented by divergent (i.e. intuitive, serendipitous) behaviour. Investigating such complementarities might give a better understanding of the complex topologies, functionalities and potentials of the Web, which might be utilised in web mining, harvesting schemes of web robots, ranking algorithms of search engines, visualisation/navigation features of browsers, etc.

## REFERENCES

1. Adamic, L. The Small World Web. Lecture Notes in Computer Science, 1696, (1999), 443-452.
2. Björneborn, L. & Ingwersen, P. Perspectives of Webometrics. *Scientometrics*, 50, 1 (2001), 65-82.
3. Broder, A. et al. Graph Structure in the Web. *Computer Networks*, 33, (2000), 309-320
4. Bush, V. As We May Think. *The Atlantic Monthly*, (July, 1945), 101-108.
5. Chakrabarti, S. et al. Mining the Web’s Link Structure. *IEEE Computer*, 32, 8 (1999), 60-67.
6. Granovetter, M.S. The Strength of Weak Ties. *American Journal of Sociology*, 78, 6 (1973), 1360-1380.
7. Huberman, B.A. & Adamic, L. Growth Dynamics of the World-Wide Web. *Nature*, 401 (Sep 9, 1999), 131.
8. Newman, M.E.J. The Structure of Scientific Collaboration Networks. *PNAS*, 98, 2 (2001), 404-409.
9. Small, H. A Passage Through Science: Crossing Disciplinary Boundaries. *Library Trends*, 48, 1 (1999), 72-108.
10. Swanson, D.R. Undiscovered Public Knowledge. *Library Quarterly*, 56, 2 (1986), 103-118.
11. Watts, D. J. & Strogatz, S.H. Collective Dynamics of ‘Small-World’ Networks. *Nature*, 393 (June 4, 1998), 440-442.