A Layered Approach for Investigating the Topological Structure of Communities in the Web

Mike Thelwall[1]
School of Computing and Information Technology, University of Wolverhampton, 35/49 Lichfield Street, Wolverhampton, WV1 1EQ, UK.
Email: m.thelwall@wlv.ac.uk

Abstract

A layered approach for identifying communities in the Web is presented and explored by applying the Flake Exact Community Identification Algorithm to the UK academic Web. Although community or topic identification is a common task in information retrieval, a new perspective is developed by: (a) the application of Alternative Document Models, shifting the focus from individual pages to aggregated collections based upon Web directories, domains and entire sites; (b) the removal of internal site links; and (c) the adaptation of a new fast algorithm to allow fully automated community identification using all possible single starting points. The overall topology of the graphs in the three least aggregated layers was first investigated and found to include a large number of isolated points but, surprisingly, with most of the remainder being in one huge connected component, exact proportions varying by layer. The community identification process then found that the number of communities far exceeded the number of topological components, indicating that community identification is a potentially useful technique, even with random starting points. Both the number and size of communities identified was dependant on the parameter of the algorithm, with very different results being obtained in each case. In conclusion, the UK academic Web is embedded with layers of non-trivial communities and, if it is not unique in this, then there is the promise of (a) improved results for information retrieval algorithms that can exploit this additional structure, and (b) the application of the technique directly to partially automate Web metrics tasks such as that of finding all pages related to a given subject hosted by a single country's universities.

Introduction

The task of identifying connected communities of topic based pages on the Web is one that is important in information retrieval to improve the standard text matching vector space model (Salton & McGill, 1983) in order to give improved precision and recall in search engine searches (Kleinberg, 1999). But the individual Web page is not necessarily the correct or only useful unitary entity for the purpose of analysing the Web. For example, inter-site links have been singled out as more important that intra-site links (Kleinberg, 1999; Flake *et al*., 2000, 2002), showing the need for alternative perspectives. Moreover, search engines also implicitly recognize that the Web is not a collection of unrelated pages by returning results organized by site. From a Web metrics perspective, aggregating pages into clusters using alternative document models (ADMs) based upon directories, domains and multi-domain sites has been previously found to be a fruitful technique (Thelwall, 2002a, 2003; Thelwall & Harries, 2003; Thelwall & Wilkinson, 2003a). In the light of all these indicators, it is logical to explore the potential for clustering information on the

[1] Thelwall, M. (2003). A layered approach for investigating the topological structure of communities in the Web. *Journal of Documentation*, 59(4), 410-429.

Web based upon different levels of aggregation than that of the page. This will be termed a *layered approach*. A successful outcome would be unlikely to result in the surpassing of the page as the basic unit of information for search engines; rather it would add an extra tool to the armoury of Web information retrieval (IR) practitioners. For example, a page could be clustered into new communities based upon its directory, domain and site and this information could be incorporated into existing classification structures in a probabilistic model in order to improve overall results (e.g. the approach of Gao *et al*., 2001 or Xi & Fox, 2001). Multiple overlapping approaches are typical of search engine algorithms (Arasu *et al*., 2001; Page, 2001). The first aim of this paper is not the construction of such an improved Web IR system but only to investigate the community structure of a section of the Web in a systematic way in order to ascertain whether this approach is viable in principle.

As a second motivation, the topological structure of the Web is also of interest when link structures are modelled or visualised. This occurs in many disciplines including communication networks (Park *et al*., 2002; Garrido & Halavais, 2003), computer science (Broder *et al*., 2000; Pennock *et al*., 2002), geography (Brunn & Dodge, 2001), information science (Rousseau, 1997; Ingwersen, 1998; Smith, 1999; Thomas & Willett, 2000; Björneborn, 2001; Thelwall, 2002c), physics (Albert *et al*., 1999; Jung *et al*., 2002) and sociology (Boudourides & Antypas, 2002; Rogers, 2002). The development of tools to identify community structures is vital for tasks such as these that involve exhaustively identifying pages that relate to a given topic or theme. The most immediate need for comprehensiveness probably comes from Web metrics studies of departments within a national university system (Chen *et al*., 1998; Thomas & Willett, 2000; Chu *et al*., 2002; Tang & Thelwall, 2002; Li *et al*., 2003) because the Web pages associated with a department are typically spread across several different domains. It can be a very labour intensive process to identify all relevant pages, and one that is almost certainly very error-prone. In the remainder of the paper this will be termed the *Web metrics motivation*, in contrast to the *Web IR motivation* discussed in the first paragraph.

Literature Review

The Topological Structure of the Web Graph

The Web can be conceived as a mathematical graph where the pages are called nodes and are connected by the hyperlinks between pages, sometimes called arrows or directed edges in graph theory. This kind of graph is known as a directed graph. If the direction of the links is ignored, then they merely connect nodes and will be termed arcs. The resulting collection of nodes and arcs forms an undirected graph. A component in a graph is a collection of nodes such that it is possible to start at any one of them and eventually reach any other by following arrows or arcs. Directed graph components are typically smaller than those of similar undirected graphs because the direction of the arrows is not ignored. Studying the Web as a directed or undirected graph is a useful technique in order to be able to gain information about its overall link structure (Rousseau, 1997; Albert *et al*, 1999; Broder *et al*., 2000; Baeza-Yates & Castillo, 2001; Huberman, 2001; Jung *et al*., 2002; Pennock *et al*., 2002; Thelwall & Wilkinson, 2003b). This is described by Borgman and Furner (2002) as bibliometric analysis of the Web. It is of interest here because community identification algorithms act on the topological structure of the Web only, ignoring page contents.

The Web is incredibly well interconnected from a topological perspective: for example if two pages from an AltaVista crawl were chosen at random then it is probable that they would be connected via a chain of (forwards and backwards) links between Web pages (Broder *et al*., 2000). As a result of this, the undirected and undirected components in the AltaVista crawl were much too big to be useful for community identification purposes: the Web is simply too well interconnected. It seems likely that this is still true today and is generally true for the databases of large commercial search engines. An explanation given for this phenomenon is that it is likely that a cluster of highly interlinked pages on the Web will be connected to other clusters by isolated links, known as shortcuts or small world links (Watts & Strogatz, 1998; Björneborn, 2001). This is the phenomenon that simultaneously allows networks to have clusters of highly interlinked nodes despite a relatively low number of links being required to traverse between most pairs of nodes. The highly interconnected nature of the Web is the reason for the need for special algorithms to detach coherent communities of pages from the rest of the Web. Similar problems predate the Web (Botafogo & Shneiderman, 1991).

Communities and Topics in the Web

The term *topic* will be used to refer to sets of pages on the Web that share a common theme. These are typically identified by a combination of text analysis and link structures, whereas *communities* are identified by link structures alone. Chakrabarti *et al*. (2002) have investigated topics on the Web and found that that they do interlink in a way that should make identification through link structures possible but despite this topics can drift quickly when following links randomly on the Web, a small world type of phenomenon.

The concept of community is actually a complex one when put into practice. In the theory of communication networks, a *clique* in a graph is a connected collection of nodes such that each node has at least as many connections inside the clique as to other nodes outside (different definitions are sometimes used for this term). Even with this straightforward definition in a non-trivial connected graph each node will be in at least two cliques; itself on its own; and the entire graph. In fact one node could be in a number of valid cliques, as the diagram below illustrates.

Fig 1. A simple connected graph with many non-trivial communities.

The node 2 is in the following cliques: {2}, {1,2}, {1,2,3}, {2,3,4}, {2,3}, {1,2,3,4}. Clearly the task of identifying a non-trivial maximal set of disjoint cliques in a graph will typically have multiple solutions. Ultimately, identifying communities in a graph is not a well-defined problem and all algorithms will need to incorporate heuristics in their working definition of community.

Community and Topic Identification Algorithms

The use of links in Web IR became mainstream with Google (Brin & Page, 1998), but its algorithm did not use clustering. Kleinberg's (1999) HITS algorithm used a different approach by grouping pages into topics, using a combination of page text and link structures. Other techniques include that of Haveliwala *et al*. (2000),

which is an attempt to produce a scalable algorithm for topic based clustering although it is much less oriented on link topologies than HITS. Haveliwala (2002) has also produced a topic-based variant of PageRank without a clustering algorithm. Flake *et al*. (2000; 2002) see the reliance on semantics of HITS and similar algorithms as a weakness and developed the Community Identification Algorithm (CIA) to identify communities of Web pages through link structures alone. The major advance for this algorithm over previous similar approaches was that it was able to process a huge graph in a computationally tractable way. The communities produced are guaranteed to be connected if the seed sets were, and as a result of being produced by the algorithm are cliques. The set of communities identified by applications of the algorithm will be a subset of the communities in the Web, but there is no known concise description of this subset, other than the tautological: the set of sets which are produced by the algorithm.

The computational viability of the CIA for large graphs is gained from a clever repurposing of the maximal flow algorithm (Ford & Fulkerson, 1956), which makes this type of topological community identification possible for the first time for huge graphs such as the Web. There are two CIA algorithms: exact CIA and approximate CIA. The exact algorithm calculates a community in one step whereas the approximate version proceeds iteratively, starting with a seed and gradually extending the network by finding pages that are linked to, or themselves link to the community and then using the exact version to select only those that connect most to the community. The approximate version is the one used in practice by its inventors because it they deal with the situation where the theoretical graph covered, the whole Web, is too huge to be stored in memory and in any case is not immediately accessible but must be generated on demand by downloading Web pages.

The exact CIA is described in detail in (Flake *et al.*, 2000, 2002) but its salient points and simple consequences of its organization will be described here. Essentially, with a directed graph G and a seed set S of nodes (i.e. Web pages), which would normally be the nodes of a connected subgraph of G, the task is to find a (possibly) larger, subgraph E' with $E \subset E' \subset G$ such that E' is in some sense well connected to the seed set relative to the rest of the graph. This is where the modified maximal flow algorithm is used. Firstly, all graph links are complemented by the addition of links in the opposite direction, if one is not already present. The links in the resulting graph are all assigned the same number *k*, called the flow capacity, where *k* is a variable parameter of the algorithm. Secondly, artificial source (S) and sink (T) nodes are added to the graph. Thirdly, links are added from the source to each community member in E and given an infinite capacity. Lastly, links are added from each node in the original G to the sink and given a capacity of 1. An example of this is shown in Fig. 2 for a simple network of two nodes connected by a single arrow, where the seed set is {A}.

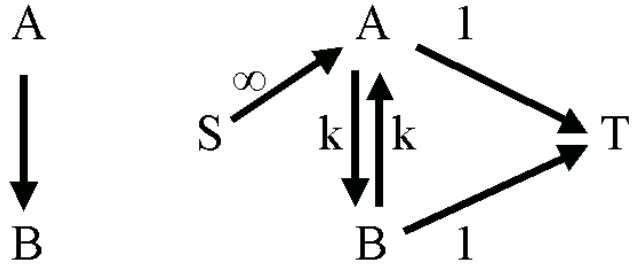


Fig 2. A simple directed graph on the left and its weighted modified version on the right with seed set {A} for processing to find the maximal flow from S to T.

The network can be thought of as a flow system for water, where the source of water is at S and its destination is T. The arrows represent pipes with the number indicating the maximum quantity of water that can flow through them and the head indicating the permissible direction of flow. The question can then be asked as to how much water can flow from S to T. Applying the maximal flow algorithm to the network can solve this problem. A by-product of the algorithm is something called a minimal cut. This is a line separating the network through which no flow is possible and ideally it will break away the highly connected portion E' of G around E from the rest of G. The size of E' is potentially dependant on the parameter k, with larger values giving bigger communities.

When a value of *k* is chosen for the network in Fig. 2 and the maximal flow from S to T has been calculated then the nodes still reachable from S along arrows that are not saturated are in the community. If $k = 1$ then the maximal flow is two: 1 flows along the path S-A-T and one flows along the path S-A-B-T. Now both arrows out of A are full and so no more flow is possible from S to T. The community is just A because A can be reached from S but there are no unsaturated arrows leading from A. If $k = 2$ then there is a different result. The maximal flow is still two with the same paths. But now the arrow from A to B is not saturated, it has a spare capacity of one. The blockages are now A-T and B-T. The nodes still connected to S at the end are A and B, giving a different community. This example illustrates how the choice of *k* affects the size of community. If *k* is greater than the number of arrows in the network then all nodes connected to any node in the initial community will find themselves in the community calculated by the CIA, but if $k = 1$ then the algorithm will return only the original community.

The Alternative Document Models and the Layered Approach

The ADMs are essentially methods for grouping together Web pages for the purpose of counting links. They were introduced in order to develop new metrics through reducing anomalies created by inappropriate implicit conceptualisations of documents on the Web (Thelwall, 2002a; Thelwall & Wilkinson, 2003a). All were then used to investigate the impact of source page type on Web linking in a scholarly context (Thelwall & Harries, 2002). The descriptions below are taken from Thelwall (2002a).

- *Individual Web page* Each separate HTML file is treated as a document for the purposes of extracting links. Each unique link URL is treated as pointing to a separate document for the purposes of finding link targets. URLs are truncated before any internal target marker '#' character found to avoid multiple references to different parts of the same page.
- *Directory* All HTML files in the same directory are treated as a document. All target URLs are automatically shortened to the position of the last slash, and links from multiple pages in the same directory are combined and duplicates eliminated.
- *Domain name* As above except all HTML files with the same domain name are treated as a single document for both link sources and link targets. In particular, this clusters together all pages hosted by a single subdomain of a university site.
- *University* As above except that all pages belonging to a university are treated as a single document for both link sources and link targets.

In the context of identifying communities, the ADMs will not be treated as competing conceptualisations, as originally developed for metrics, but instead as *layers of structure* in the Web. Each layer has the potential to give different and complementary information about how its contents relate to the rest of the Web.

The Research Questions

The issue to be addressed is whether there are significant community structures in the Web in any or all layers of the document model. In the light of the use of inter-site links only in community and topic identification algorithms (Kleinberg, 1999, Flake *et al*., 2002) attention will be restricted to the Web after the removal of all internal site links. Unfortunately, it is impractical to crawl the whole Web for a single research question and also infeasible to extract the topological structure of the Web from the freely available large crawl data sets at archive.org. Taking a random sample of the Web (e.g. Thelwall, 2002b) is also not a sensible approach since links between sites are likely to be relatively rare unless the group of sites forms a coherent collection in some way. As a result the focus will be upon national academic domains. This is a viable choice since several of these have been previously investigated and found to be highly interlinked (Smith & Thelwall, 2002; Thelwall, 2002a; Thelwall & Tang, 2002). It also directly addresses the Web metrics motivation for community identification. The limitations that this imposes will be discussed at the end.

For a national academic Web stripped of links within individual university Web sites and each of the four document models, the specific questions to be investigated are the following.

- Does a topological decomposition of the graph yield non-trivial results: i.e. a significant number of communities that have more than one node but do not tend to form one huge component?
- Are there values of k in the single seed exact CIA that give non-trivial results: i.e. a significant number of communities that have more than one node but do not form a complete connected component?
- When significant community formation is identified, are the communities produced predominantly robust in the sense of being frequently rediscovered for different seed values?

Methodology

The approach to be used is that of applying the algorithm to various ADM versions of a large crawl: that of the UK academic Web. The UK was chosen as (a) the largest academic Web that it was practical to crawl and process using the software and hardware available, and (b) a Web that had been previously crawled and extensively analysed which gives the considerable advantage of allowing most of the mirror sites to be bypassed in the crawl (Thelwall, 2001c; Thelwall, 2002a).

The experiment to be conducted will be to use the CIA with each node in each data set as a single seed set, using k = 2, 4, 8, 16 and 32. The communities identified will be summarised by size and compared to the topological structure of this modified Web space in order to ascertain whether the results do indeed show real communities that are different from the connected components obtained from the overall topological structure. Investigations into whether different k values give similar results and how robust the communities are to the choice of seed will also be reported.

A database of the link structure of the UK academic Web as of July, 2002 was created by a specialist information science Web crawler (Thelwall, 2001ab) and has been placed in a publicly available database for free access by researchers (http://cybermetrics.wlv.ac.uk/database). It consists of the set of pages that was obtained by starting at the home page or an alternative starting point for each of 111 university institutions (listed on the site) and then following links recursively. It is not in any sense a definitive crawl because it will exclude pages that are not linked to. The database was synthesized by removing all links to pages within the same university and all links to domains outside the set under consideration. Finally, all completely isolated pages that were neither the source nor target of any remaining links were then removed. The remainder formed the raw data for the experiment. The set was then processed to form three new ones for the directory, domain and university layer, making a total of four alternative representations of the interlinking structure of the UK academic Web. These were then converted into a numerical code form for efficient and compact main memory storage purposes. The conversion process included heuristics to merge URLs for the same pages, principally alternative home pages and domain names.

The Graph Structure Analysis Environment program (Thelwall & Wilkinson, 2003b) was used to find the basic topological structure of each layer. A new program was then written to apply the CIA to the data sets. It is identical in basic algorithm to the original version but was designed with a new data structure to keep all links in main memory permanently. This allowed it to use a much faster non-dynamic data structure to store the large graphs, which sped up the algorithm to the extent that it was feasible to run it repeatedly with a large range of initial seed sets. The data was processed in a lab of 20 PCs with each one processing a single document model and k value but all possible single seed starting sets. The longest time to complete for any program was seven days, for the domain model with $k = 32$. On each computer most seeds were typically processed very quickly but highly connected ones required a much longer time to handle.

The data obtained will not be tested for goodness of fit with Lotka's law (Nicholls, 1986) or other mathematical models, the primary interest here being the overall distribution of communities rather than any formal mathematical statement of properties.

Results

The topological component sizes in the different layers will be reported first in order to form a basis for comparison with the results of the CIA algorithm.

Topological Structures

University Layer Components

There was only one undirected component, which contained the whole set of 111 universities. There was one directed component containing 108 of the universities. The other three were in components of size one. No universities had been removed from the data set through being isolated.

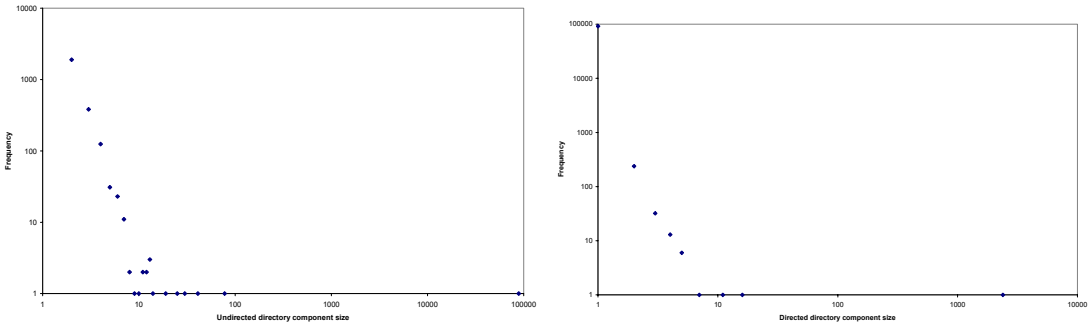
Domain Layer Components

There were 6,754 separate domains. Of these, 6,750 formed one large component and there were two other components of size 2. There was one large

directed component of 2,297 domains, one component of two and the rest were all of size 1. Almost half of the domains had been removed for being isolated, 5,630 in all.

Directory Layer Components

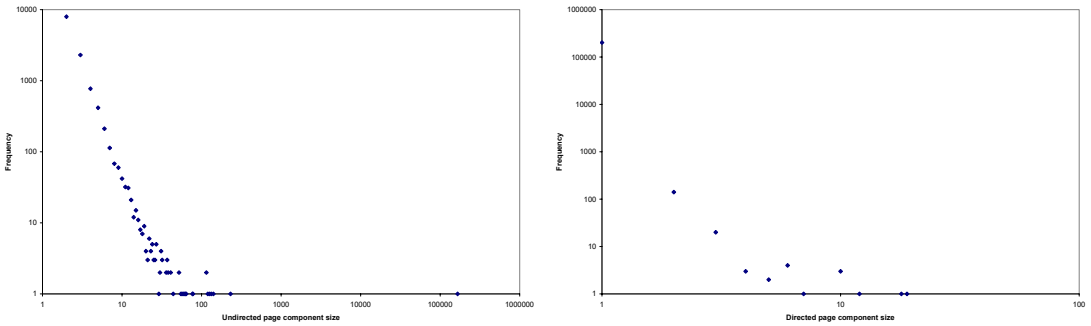
There were many more separate undirected components at the directory level. Figure 3 illustrates the results using logarithmic scales to highlight the power law like distribution, although an anomaly is also present. There were 94,983 separate directories in the data set, 88,849 of which were in one huge component. There were many more separate directed components, with the largest being of size 2,388 and 91,901 being components of size 1. Figure 4 illustrates the results. Most of the directories (615,842 or 87%) had been removed for being isolated.



Figs. 3-4. The size of directory-based undirected and directed components on the Web.

Page Layer Components

There were 202,338 separate pages in the data set, 165,710 of which were in one huge undirected component. Almost all pages were in directed components of size 1, 201,864 in total. The largest directed component was only of size 19. Most of the files (6,049,036 or 97%) had been removed for being isolated.



Figs. 5-6. The size of page-based undirected and directed components on the Web.

General statistics about the size of the data set are summarised in Table 1 for comparison purposes with the communities found.

Table 1. The size of each data set, after the elimination of isolated nodes.

	Nodes in data set	Largest undirected component size
File	202,338	165,710
Directory	94,983	88,849
Domain	6,754	6,750
University	111	111

Community Structures

University Layer Communities

The results for the university model were straightforward. When k = 2, 106 starting universities lead to a community containing all 111, the rest leading to no communities. This indicates a high degree of interconnectivity. When k = 4 and 8, only one university did not seed a community of all 111. For larger values (k=16) all starting universities returned a community of 111. At the university level the pattern of interconnectivity is clearly sufficiently great to negate the possibility of interesting community formation. If the data set had been much larger and international then this may well have been different, although some sites do attempt to maintain a list of all university home pages and the inclusion of one of these would go a long way towards making all universities interconnected.

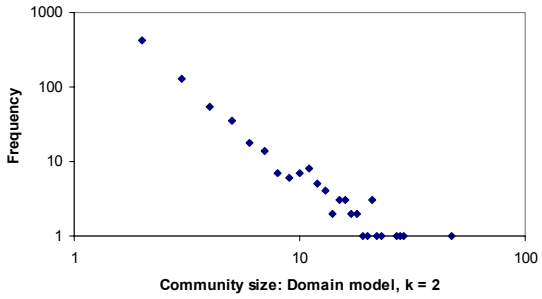
Domain Layer Communities

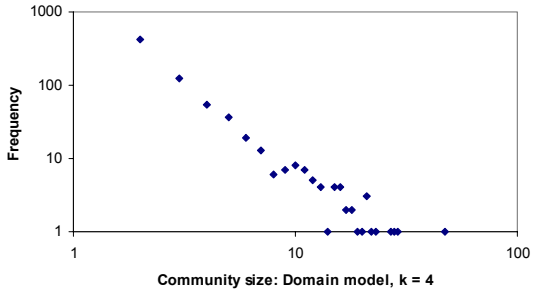
The results of the domain model are presented as graphs in Figs 7 to 11 showing the frequency of unique communities of each size. Logarithmic scales are again used to emphasize the power law type of distributions found. This kind of distribution should not be totally unexpected since for low values of k communities may be dominated by node origin inlink and outlink degrees, known to follow a power law at least at the page level (Broder *et al*., 2000; Thelwall & Wilkinson, 2003b) and for large values of k the communities may be dominated by the connectivity of the underlying graph, another power law candidate (Broder *et al*., 2000; Thelwall & Wilkinson, 2003a). The interaction between the two is non-trivial, however, and so the power laws discovered are a genuinely new finding.

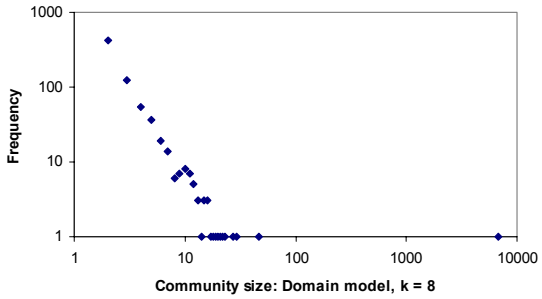
There is a clear trend in the graphs for increasing k values. When k = 8 and above a community of size 6,740 appears. This is an enormous multi-disciplinary, multi-institution very general collection, with eight seeds (abdn.ac.uk, dur.ac.uk, ex.ac.uk, leeds.ac.uk, shef.ac.uk, ucl.ac.uk, scit.wlv.ac.uk, york.ac.uk – note that domains are identified with the variant starting with www. so that www.ex.ac.uk = ex.ac.uk). Presumably these sites all host pages that link extensively to the rest of the UK academic Web. The scit.wlv.ac.uk domain, for example hosts links to all university home pages (www.scit.wlv.ac.uk/ukinfo/). A high degree of inlinking or outlinking for a domain is required to set up a large community since its maximum size is k times the total number of distinct in and outlinks.

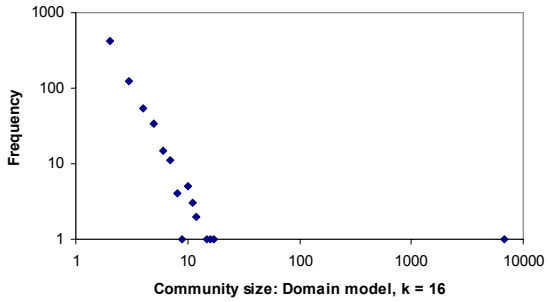
The apparent thinning out of the graphs from k = 8 to the very sparse k = 32 is caused by increasingly many starting points leading to the big group. When k = 8 it is generated by eight starting domains and when k = 32 the number has risen to 580. Essentially, at the larger k values this huge community is swallowing up many of the smaller ones. The huge community is not interesting precisely because it is too big to be topic specific or anything other than very general. The remaining communities may be more topic-specific, particularly for the higher k values.

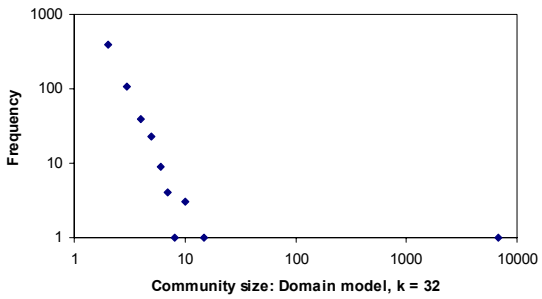
With k = 2 and k = 4 only two pairs of starting domains produced the same community, showing that the communities produced are very delicate. When k=8 two additional communities with 8 and 3 seeds appear. When k=16 the large community has 57 seeds, and when k=32 it has 580 showing a degree of robustness. However, even for k=32 there are only three other communities with more than one seed, one with three and two with two.











Figs 7-11. The size of unique communities identified from the domain model follows a power law for $k = 2$ and 4 and a power law with an anomaly for $k = 8$ to 32.

Table 2. Largest number of seeds for a single community with the domain model.

k	2	4	8	16	32
Seeds	2	2	8	57	580

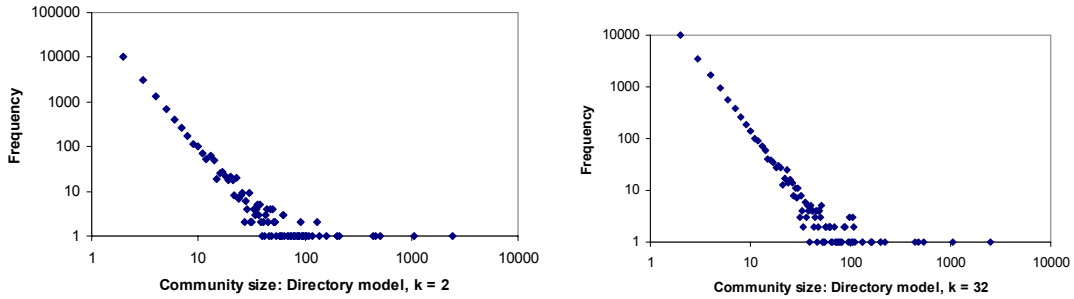
Directory Layer Communities

The directory model showed a very similar trend to the domain model, except that many more communities were identified and although around 89% of unique communities had only one seed in each case, there were more robust multiple seed communities, even for $k = 2$. Figures 12 and 13 show the distribution of unique community sizes. Again a large community is evident, but this time there are many more communities and a range of larger ones. There are quite large communities even for $k = 2$. The largest community is of size 2,504 and emerges when $k = 8$. This is a collection of pages created by many academics using Nikos Drakos' LaTeX2HTML program, each of which contains two links to the home pages of the software. The two directories from which these come are the two seeds of the community: cbl.leeds.ac.uk/nikos/tex2html/doc/latex2html and cbl.leeds.ac.uk/nikos.

There were other large communities of electronic documents. For example, a community of size 202 was seeded by the directory ast.cam.ac.uk/ioa/preprint and all the other members of the community were papers starting with virtual directory addresses xxx.soton.ac.uk/find/astro-ph/1. One other member was xxx.soton.ac.uk/find/astro-ph/1/jacco+th.+van+loon/0/1/0/all/1. Another interesting

community for k=32 is that seeded by a directory containing a page listing all UK university home pages, www.scit.wlv.ac.uk/ukinfo. This community contains 98 directories, none of which are university home pages! The other members are all directories that link to the UK university list. The reason that no university home page appears in the list is that these will all have many other pages that link to them, which act to separate them from the community. Like the Drakos directories, this community is created only by links *to* its seeds.

The graphs are all very similar and so only two are shown. There are many more communities with multiple seeds and the figure increases as k increases.



Figs 12-13. The size of unique communities identified from the directory model follows a power law for $k = 2$ and 4 and a power law with anomalies for $k = 2$ to 32.

There are an increasing number of the more robust, multiple seed communities than for the domain model and this number increases as k increases, as can be seen from tables 3 and 4.

Table 3. Distribution of multiple seeds for the directory model communities, $k = 2$.

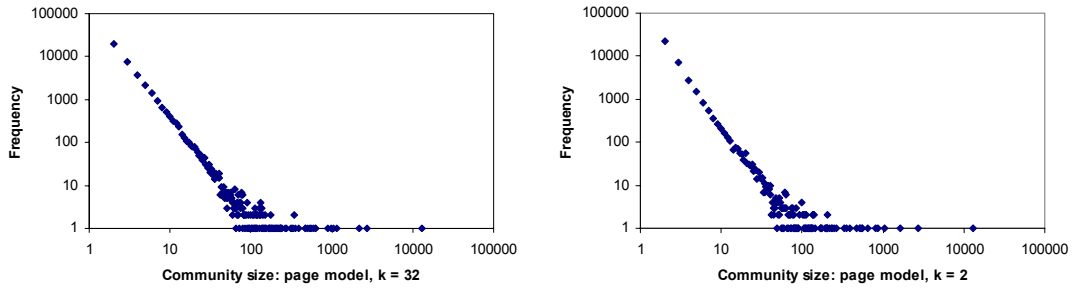
Seeds	1	2	3	4
Frequency	15225	1905	1	9

Table 4. Distribution of multiple seeds for the directory model communities, $k = 32$.

Seeds	1	2	3	4	5	6	7	8	9	10	11	12	13	14	19	25	30
Frequency	15621	1950	396	141	36	25	12	4	1	1	2	2	3	1	1	1	1

Page Layer Communities

The page model shows a similar pattern to the previous two. When $k = 2$, the outlier is a community of size 13,113 created by a mirror copy of the FOLDOC archive at the university of Brighton, each page of which contains a credit link to the original home page at Imperial College, foldoc.doc.ic.ac.uk, which is the single seed for the community. The second biggest community is created by pages generated by LaTeX2HTML that link to the creator's (old) home page at cbl.leeds.ac.uk/nikos/personal.html.



Figs 14-15. The size of unique communities identified from the page/file model follows a power law for $k = 2$ and 4 and a power law with anomalies for $k = 2$ to 32.

There are further communities with more than one seed for the page model, in fact enough to be able to plot the number of seeds for each community. As can be seen from Figure 16, these also follow a power law. Since the page model graph will be much less connected than the others, this distribution may be starting to reflect connected components of the underlying Web graph.

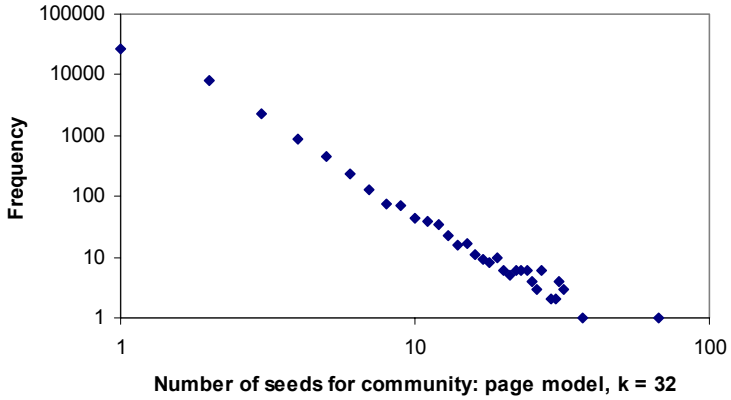


Fig 16. The seeds for communities found by the page model follows a power law for $k = 32$.

Discussion and Limitations

Very interesting and surprising results were found in the overall topological structure of the Web layers, principally that the documents associated with inter-site links were very highly connected. At the domain level, almost all were in a single component. At the page level, whilst 97% of pages were neither the source nor target of any inter-site link, the vast majority of the rest formed one large connected component. These results show that a topological decomposition of the UK academic Web is not useful for community identification purposes because the components are too big to have any kind of topic-specific focus.

Despite the high degree of connectedness in the Web graph, there are many topological communities in the UK academic Web at the page, directory and domain level. The university level was too broad to be useful, at least in the context of a single country. The communities are not necessarily coherent topics but can also be organised by a common information need and can be bound together by an ‘outsider’, as evidenced in both cases by the directory community seeded by the UK universities list. Many of the communities were not very robust, in terms of only being generated by one seed, as illustrated by the general change in distribution of sizes for different

values of *k*. It can also be seen that values of *k* which are too large can result in much of the community structure being lost through being swamped in one that is too big to be useful.

There are four main limitations of this study.

- Only one national academic Web has been used and although it seems likely that the results would extend to others, this has not been proven.
- The results cover only an academic domain and from the Web IR perspective they would need to be extended to the rest of the Web in order to be useful. Specifically, it is possible that commercial Web sites are much less well interlinked and that as a result these would not tend to create significant communities. It seems more likely, however, that the approaches described here would work, although different *k* values may be needed and perhaps the page and directory models would be less successful because of their lower level of aggregation.
- No evidence has been provided for the more complex task of ascertaining whether the 'communities' identified are in fact coherent in a useful sense, from either the Web IR or the Web metrics perspectives.
- From the Web IR perspective it has not been demonstrated that the approach will scale to the whole Web. This is a particular concern for the page and directory models with their huge number of nodes.

Conclusions

From the motivation of using these results in a general Web IR algorithm, the fact that communities do emerge from the structure of the Web is promising. Probably for the general Web, there will be fewer inter-site links and so the community structure will be sparser. This evidence for additional structure is however a potential extra ingredient that could be added to existing search engine algorithms in order to improve results by identifying more accurately the topic of a page through the community structures it participates in at all layers of the Web. The different layers and different *k* values can potentially reveal different aspects of a single page, improving the chance of matching it correctly to different queries. Commercial search engine designers are therefore recommended to investigate whether this approach is capable of improving IR performance when incorporated with their existing algorithms. This may involve using several of the layers simultaneously and could even include multiple *k* values for each layer if these were found to give usefully complementary information.

From the Web metric perspective the approach is recommended as a tool for those seeking to identify coherent collections of Web pages, for example all computing related pages in UK university Web sites (Li *et al*., 2003). The CIA can be expected to be an automated tool capable of using Web link structures to suggest new pages, directories or domains that may be related to those already known.

References

Albert, R., Jeong, H. & Barabási, A. L. (1999). Diameter of the World-Wide Web, *Nature*, 401, 130-131.

Arasu, A., Cho, J., Garcia-Molina, H., Paepcke, A. & Raghavan, S. (2001). Searching the Web, *ACM Transactions on Internet Technology*, 1(1), 2-43.

Baeza-Yates, R. & Castillo, C. (2001). Relating Web characteristics with link based Web page raking, In *Proceedings of SPIRE 2001*, IEEE CS Press, Laguna San Rafael, Chile, pp. 21-32, November 2001.

Björneborn, L. (2001). Small-world linkage and co-linkage. In: *Proceedings of the 12th ACM Conference on Hypertext and Hypermedia* (pp. 133-134). New York: ACM Press.

Borgman, C & Furner, J. (2002). Scholarly communication and bibliometrics. In: Cronin, B. (ed.), *Annual Review of Information Science and Technology 36*, pp. 3-72, Medford, NJ: Information Today Inc.

Botafogo, R. A. & Shneiderman, B. (1991). Identifying aggregates in hypertext structures. *Proceedings of Hypertext 1991*, pp. 63-74, ACM: San Antonio, Texas, USA.

Boudourides, M. & Antypas, G. (2002). A Simulation of the Structure of the World-Wide Web. Available: http://www.socresonline.org.uk/7/1/boudourides.html

Brin, S. & Page, L. (1998). The Anatomy of a large scale hypertextual Web search engine, Computer Networks and ISDN Systems, 30(1-7), 107-117. Available at http://citeseer.nj.nec.com/brin98anatomy.html

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

Brunn, S. D. & Dodge, M. (2001). Mapping the "worlds" of the world wide Web: (Re)Structuring global commerce through hyperlinks, *American Behavioral Scientist,* 44(10), 1717-1739

Chakrabarti, S., Joshi, M. M., Punera, K. & Pennock, D. M. (2002). The structure of broad topics on the Web, WWW2002, Available: http://www2002.org/CDROM/refereed/338/

Chen, C., 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(4), 353-373.

Chu, H., He, S. & Thelwall, M. (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.

Flake, G.W., Lawrence, S., Giles, C.L. & Coetzee, F.M. (2000). Efficient identification of Web communities, *Proceedings of the 6th International Conference on Knowledge Discovery and Data Mining*. New York: ACM Press, pp 150-160.

Flake, G.W., Lawrence, S., Giles, C.L. & Coetzee, F.M. (2002). Self-organization and identification of Web communities, *IEEE Computer*, 35, 66-71.

Ford, L. R. & Fulkerson, D. R. (1956). Maximal flow through a network, *Canadian Journal of Mathematics*, 8(3), 399-404.

Gao, J. Walker, S., Robertson, S., Cao, G., He, H., Zhang, M. & Nie, J-Y (2001). TREC-10 Web Track Experiments at MSRA 384-392. TREC 2001. Available: http://trec.nist.gov/pubs/trec10/t10_proceedings.html.

Garrido, M. & Halavais, A. (2003, to appear). Mapping Networks of Support for the Zapatista Movement: Applying Social Network Analysis to Study Contemporary Social Movements. *In*: M. McCaughey & M. Ayers (Eds). *Cyberactivism: Critical Practices and Theories of Online Activism*. London: Routledge.

Haveliwala, T.H., Gionis, A. & Indyk P. (2000). Scalable techniques for clustering the Web. In WebDB 2000. Available: http://www.research.att.com/conf/Webdb2000/program.html

Haveliwala, T. (2002). Topic-Sensitive PageRank, Proceedings of the Eleventh International World Wide Web Conference, May 2002. Available: http://www-db.stanford.edu/~taherh/papers/topic-sensitive-pagerank.pdf

Huberman, B. A. (2001). *The Laws of the Web: Patterns in the Ecology of Information*, Cambridge, MA: The MIT Press.

Ingwersen, P. (1998). The calculation of Web Impact Factors. *Journal of Documentation*, 54, 236-243.

Jung, S., Kim, S. & Kahng, B. (2002). Geometric fractal growth model for scale-free networks, *Physical Review E*, 65(5), No. 056101. Available: http://phya.snu.ac.kr/~kahng/PRE56101.pdf

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

Li, X., Thelwall, M., Musgrove, P. & Wilkinson, D. (2003, to appear). The relationship between the links/Web Impact Factors of Computer Science departments in the UK and their RAE (Research Assessment Exercise) ranking in 2001. *Research Evaluation*.

Nicholls, P. T. (1986). Empirical validation of Lotka's law. *Information Processing and Management*, 22, 417-419.

Page, L. (2001). Method for node ranking in a linked database, United States Patent 6,285,999.

Park, H. W., Barnett, G. A. & Nam, I. (2002). Hyperlink-affiliation network structure of top Web sites: Examining affiliates with hyperlink in Korea. *Journal of the American Society for Information Science,* 53(7), 592-601.

Pennock, D., Flake, G., Lawrence, S., Glover, E. & Giles, C. L. (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.

Rogers, R. (2002). Operating issue networks on the Web, *Science as Culture*, 11(2), 191-214.

Rousseau, R. (1997). Sitations: an exploratory study, *Cybermetrics*, 1. Available: http://www.cindoc.csic.es/cybermetrics/articles/v1i1p1.html

Salton, G. & McGill, M. J. (1983). *Introduction to Modern Information Retrieval*. New York: McGraw-Hill.

Smith, A. G. (1999). A tale of two web spaces: comparing sites using Web Impact Factors. *Journal of Documentation*, 55(5), 577-592.

Smith, A. & Thelwall, M. (2002). Web Impact Factors for Australasian universities, *Scientometrics*, 54(3), 363-380.

Tang, R. & Thelwall, M. (2002). Disciplinary differences in US academic departmental Web site interlinking. State University of New York.

Thelwall, M. & Tang, R. (2002). Disciplinary and linguistic considerations for academic Web linking: An exploratory hyperlink mediated study with Mainland China and Taiwan, University of Wolverhampton.

Thelwall, M. (2001a). A publicly accessible database of UK university Website links and a discussion of the need for human intervention in Web crawling, University of Wolverhampton. Available: http://www.scit.wlv.ac.uk/~cm1993/papers/a_publicly_accessible_database.pdf.

Thelwall, M. (2001b). A Web crawler design for data mining, *Journal of Information Science*, 27(5), 319-325.

Thelwall, M. (2001c). Extracting macroscopic information from web links, *Journal of the American Society for Information Science and Technology*, 52 (13), 1157-1168.

Thelwall, M. (2002a). Conceptualizing documentation on the Web: an evaluation of different heuristic-based models for counting links between university Web sites, *Journal of the American Society of Information Science and Technology*, 53(12), 995-1005.

Thelwall, M. (2002b). Methodologies for Crawler Based Web Surveys, *Internet Research: Electronic Networking and Applications*, 12(2), 124-138.

Thelwall, M. (2002c). An initial exploration of the link relationship between UK university Web sites, *ASLIB Proceedings*, 54(2), 118-126.

Thelwall, M. (2003, to appear). Methods for reporting on the targets of links from national systems of university Web sites. *Information Processing and Management*.

Thelwall, M. & Harries, G. (2003, to appear). 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*.

Thelwall, M. & Smith, A. (2002). A study of the interlinking between Asia-Pacific University Web sites, *Scientometrics* 55(3), 335-348.

Thelwall, M. & Wilkinson, D. (2003a, to appear). Three target document range metrics for university Web sites. *Journal of the American Society of Information Science and Technology*.

Thelwall, M. & Wilkinson, D. (2003b, to appear). Graph structure in some national academic Webs: Power laws with anomalies. *Journal of the American Society of Information Science and Technology*.

Thomas, O. & Willett, P. (2000). Webometric analysis of departments of Librarianship and information science. *Journal of Information Science*, 26(6), 421-428.

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

Xi, W. & Fox, E. A. (2001) Machine Learning Approaches for Homepage Finding Tasks at TREC-10, TREC 2001 online proceedings. Available: http://trec.nist.gov/pubs/trec10/notebook_papers/VTexplainTREC10.pdf