Graph Structure in the Web – Aggregated by Pay-Level Domain

Oliver Lehmberg
Data and Web Science Group
University of Mannheim
Germany
oli@informatik.uni-mannheim.de

Robert Meusel
Data and Web Science Group
University of Mannheim
Germany
robert@informatik.uni-mannheim.de

Christian Bizer
Data and Web Science Group
University of Mannheim
Germany
chris@informatik.uni-mannheim.de

ABSTRACT

Previous research on the overall graph structure of the World Wide Web mostly focused on the page level, meaning that the graph that directly results from hyperlinks between individual web pages was analyzed. This paper aims to provide additional insights about the macroscopic structure of the World Web Web by analyzing an aggregated version of a recent web graph. The graph covers over 3.5 billion web pages and 128 billion hyperlinks between pages. It was crawled in the first half of 2012. We aggregate this graph by pay-level domain (PLD), meaning that all pages that belong to the same pay-level domain are represented by a single node and that an arc exists between two nodes if there is at least one hyperlink between pages of the corresponding pay-level domains. The resulting PLD graph covers 43 million PLDs and contains 623 million arcs between PLDs. Analyzing this aggregated graph allows us to present findings about linkage patterns between complete websites and not only individual HTML pages. In this paper, we present basic statistics about the PLD graph, such as degree distributions, top-ranked PLDs, distances and diameter. We analyze whether the bow-tie structure introduced by Broder et al. can also be identified in our PLD graph and reveal a backbone of highly interlinked websites within the graph. We group the websites by top-level domain and report findings about the overall linkage within and between different top-level domains. In a last experiment, we use data from the Open Directory Project (DMOZ) to categorize websites by topic and report findings about linkage patterns between websites belonging to different topical categories.

Categories and Subject Descriptors

H.3.4 [**INFORMATION STORAGE AND RETRIEVAL**]: System and Software - World Wide Web

Keywords

World Wide Web, Web Graph, Network Analysis, Graph Analysis, Web Mining, Web Science

WebSci'14, June 23–26, 2014, Bloomington, IN, USA.

ACM 978-1-4503-2622-3/14/06 ...$15.00.
http://dx.doi.org/10.1145/2615569.2615674.

1. INTRODUCTION

With the growth of the World Wide Web, the corresponding web graph has evolved in size and complexity. Knowledge about the macroscopic structure of this graph is useful within various application domains. It forms the basis for designing the ranking methods of web search engines. In turn, search engine optimization (SEO) efforts exploit the link structure of the Web in order to fool search engines and increase the ranking of target websites. An extreme appearance of such manipulations are spam networks, which consist of large numbers of websites that are created for the sole purpose of influencing rankings. Knowledge about the graph structure of the Web is also important for the seed selection of general web crawlers. Knowledge about linkage patterns within specific topical domains can help focused crawlers to adapt their crawling strategy. In addition, such patterns might also revile interesting findings about the social mechanisms that govern a specific domain.

We have extracted a large web graph from the 2012 version of the Common Crawl.[1] The graph covers over 3.5 billion web pages which are connected by over 128 billion hyperlinks. It is the largest web graph that is currently (May 2014) available to the public. In [13], we analyze the structure of this graph on page level. The paper updates the findings of Broder *et al.* [6], who did a similar page-level analysis over a decade ago, concerning in- and outdegree distributions, the distributions of the sizes of weakly and strongly connected components, distances within the graph, as well as the size of the components of the bow tie. In contrast to earlier studies and using the power law fitting methodology proposed Clauset *et al.* in [7], we find that the in- and outdegree distributions on page-level do not follow power laws.

In this paper, we analyze an aggregated version of the same graph. This aggregation is based on the pay-level domain of each HTML page. In addition to the analysis on page level, the PLD aggregation gives us the possibility to draw conclusions about the connectivity amongst complete websites[2] and not only individual pages. The paper makes the following contributions: (1) We report the results of a structural analysis of the PLD graph including degree distributions, distances, and diameter (Section 3). We present several rankings of pay-level domains which are particularly interesting for search engine optimization and spam detection. We analyze whether the bow-tie structure introduced by Broder *et al.* can also be identified in our PLD graph and report the sizes of the components of this bow tie (Section 3.4). Additional experiments (2) reveal a backbone of highly interlinked websites (Section 3.6) and enable us to (3) infer a two-layer model explaining the largest

[1] `http://commoncrawl.org`
[2] We will use the term website synonymously to the term pay-level domain.

strongly connected component and the low average shortest path length within the graph (Section 3.7). (4) We group the websites by top-level domain and report findings about the linkage within and between different top-level domains[3] (Section 4). In addition, (5) we use data from the Open Directory Project[4] to categorize the websites by topic and describe the structural properties of the websites assigned to different topical categories (Section 5). (6) The page- and PLD graph as well as the code used for the analysis are made publicly available via the *WebDataCommons.org* website[5] as we want to encourage other researchers to validate our findings and to be able to further analyze the graphs.

2. CRAWLING STRATEGY AND PLD COVERAGE

The web crawl from which we extracted our web graphs was collected in the first half of 2012 by the Common Crawl Foundation. Their crawler used a breath-first crawling strategy and was seeded with over 71 million different URLs from previous crawls and from Wikpedia pages. Unfortunately, the Common Crawl Foundation does not provide detailed information about banned pages, crawling limitations, as well as the stopping conditions of the crawler. Additional statistics about the crawl are provided in [18]. We aggregate the web graph by merging all pages from the same pay-level domain into a single node and remove internal and duplicate arcs. This results in a PLD graph containing 43 million nodes and 623 million arcs. We used the WebGraph library [4] to shrink the page graph into the PLD graph.

While we do not know the overall number of HTML pages on the Web, we know how many PLDs were registered at the time of crawling. This allows us to estimate the percentage of all registered PLDs that are covered by our graph. The number of registered domains is frequently reported by Verisign. In their report from October 2012[6] about the second quarter of the same year, they state a total of 240 million registered domain names. With our graph covering 43 million domains, this means we have (at least partial[7]) data about 18% of all domains that were registered at that time. The report further states that only 66% of all ".com" and ".net" domains contain real websites, meaning that one third of all registered domains forward to other domains or do not contain any web pages. We hence assume that our graph effectively covers more than the 18% that we can state with certainty: Assuming that the 66% hold for all domains, only 158 million domains have to be considered, resulting in a coverage of 27%.

3. ANALYSIS OF THE PLD GRAPH

In the following, we first have a closer look at the distribution of indegree, outdegree (Section 3.1), and PageRank (Section 3.2), and identify the top ranked PLDs according to several ranking methods (Section 3.3). Next, we present our findings about the connectivity of the PLD graph and examine whether the bow-tie structure introduced by Broder *et al.* [6] can also be found in our graph. In Section 3.5, we examine paths and reachability of websites in the bow-tie components. Afterwards, we investigate the robustness of the PLD graph and have a closer look at groups of strongly linked websites which connect a large part of the PLD graph (Section 3.6). Finally, we combine these findings into a two-layer model which explains the largest strongly connected component and the low average shortest path length within the graph (Section 3.7).

3.1 In- and Outdegree Distributions

Figure 1 displays the in- and outdegree distributions of the PLD graph, showing the number of PLDs (y-axis) for a certain degree (x-axis) using a log-log scale. The grey dots show the actual distribution of the degree values. We also include the *Fibonacci Binning* [19] of the degree values into the figure to give a better impression of the distribution. In order to test the hypothesis that the distribution follows a power law, we employ the methodology proposed by Clauset *et al.* [7]. As concrete implementation, we use the `plfit`[8] tool to estimate the power law that fits our data with the maximum-likelihood. We also perform a goodness-of-fit test. The test produces a p-value, which tells us to reject the hypothesis that the distribution follows a power law if the value is smaller than 0.1. In the diagrams, the black line represents the best-fitting power law for the given distribution.

For the indegree distribution, the best-fitting pow law has an exponent of 2.40 and starts at a degree of 3 062. The p-value of the best fit for the indegree distribution is $0.43 \pm (0.01)$, meaning that the distribution follows a power law. The best-fitting power law for the outdegree distribution starts at 496 and has an exponent of 2.39. The p-value of this fit is $0 \pm (0.01)$. It is thus very likely that the outdegree distribution does not follow a power law. In [13], we found the largest outdegree value within the page graph to be three orders of magnitude smaller than the largest indegree value. Within the PLD graph the largest outdegree and indegree values are quite similar.

Both the indegree and the outdegree distribution (Figure 1) show several outliers above the rest of the distribution. In addition, both degree distributions display spikes at an indegree of roughly 3 000 and an outdegree of roughly 8 500. Examining a sample of those data points, we find that the corresponding websites can be classified as spam sites or domain seller sites. This has also been observed by Fetterly *et al.* [10] for the degree distributions on page level. Beside obvious spam sites, some companies register a separate PLD for every city that matters to their business. An example is a group of job-search websites following the pattern "*-jobs.co.uk", while each website links to all the other websites.

3.2 PageRank Distribution

In addition to the in- and outdegree, we also examined the distribution of PageRank values. PageRank is a popular measure for the prestige of a website, as it cannot be tricked by spammers as easily as indegree [14]. The PageRank distribution has been shown by Pandurangan *et al.* [15] to have approximately the same power-law exponent as the indegree distribution.

The right-most diagram in Figure 1 shows the PageRank distribution for the PLD graph. We can report a best-fit power law exponent of 2.27, which differs by 0.13 from the exponent of the indegree distribution. Generally, we can say that the PageRank distribution is much cleaner than the distribution of the indegree and does not contain any extreme outliers (like spikes within the distribution).

[3] More precisely *public suffixes* which are domain endings under which a domain name can be registered. Examples are ".com" or ".co.uk".

[4] `http://dmoz.org`

[5] `http://webdatacommons.org/hyperlinkgraph/`

[6] `http://www.verisigninc.com/assets/domain-name-brief-oct2012.pdf`

[7] We can say for sure that we have at least one page from each of these domains. Again, it is not possible to determine whether our data contains all pages from a specific domain.

[8] `https://github.com/ntamas/plfit`

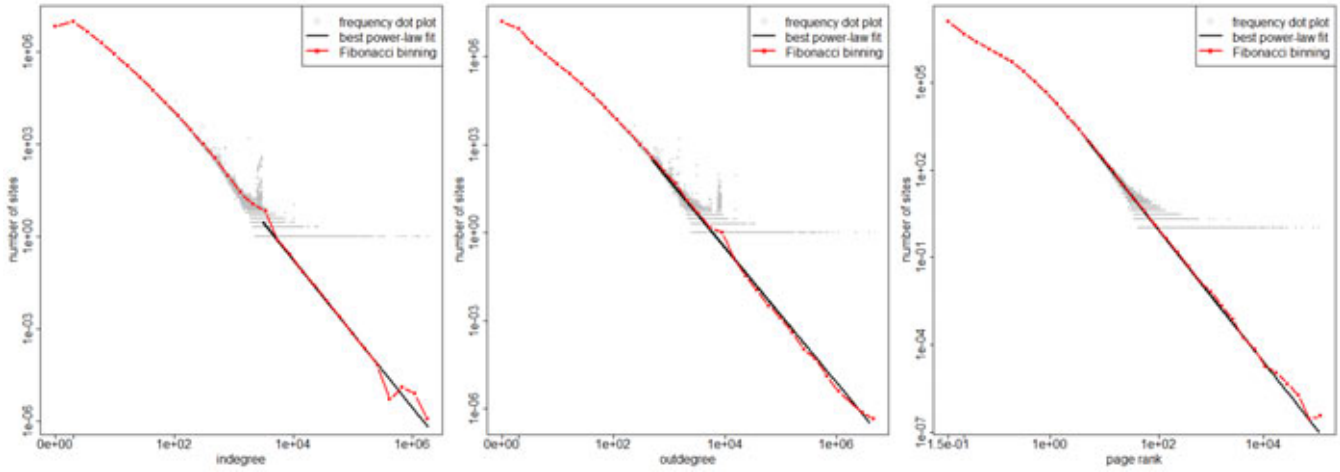

Figure 1: Distributions of indegree, outdegree and PageRank of the PLD graph

3.3 Top Ranked PLDs

We now have a look at the top ranked websites with respect to the in- and outdegree as well as the PageRank. This gives us an impression about the most important websites in our graph. Table 1 lists the top 20 websites in our PLD graph, ordered by their out- respectively indegree.

Regarding the outdegree, the highest ranked website is *blogspot.com*. We assume the reason for this high rank to be our aggregation. On this website, every user can create a personal blog and is provided with a sub domain under *blogspot.com*. Hence, this high outdegree can be assumed to be the sum of the outdegrees of all personal blogs hosted by *blogspot*.

Ordered by indegree, the most prominent website is *wordpress.org*. We assume that this is due to a large number of blogs that use *wordpress*' blog software and hence all of them also set links to the central *wordpress.org* website. The second ranked website is *youtube.com* followed by the online encyclopedia *wikipedia.org*.

Table 2 shows the top 20 websites according to their PageRank values. In this table, we find many websites that are also in the top list for the indegree. The highest ranked websites are *wordpress.com*, *gmpg.org* and *youtube.com*. The website *gmpg.org* provides a vocabulary for describing relationships. From the high PageRank we can assume this vocabulary is used frequently by rather popular websites.

	Website	Outdegree	Website	Indegree
1	blogspot.com	3 898 561	wordpress.org	1 822 440
2	wordpress.com	2 249 553	youtube.com	1 319 548
3	youtube.com	1 078 938	wikipedia.org	1 243 291
4	wikipedia.org	862 705	gmpg.org	1 156 727
5	serebella.com	699 609	blogspot.com	1 034 450
6	refertus.info	668 271	google.com	782 660
7	top20directory.com	650 884	wordpress.com	710 590
8	typepad.com	551 360	twitter.com	646 239
9	botw.org	496 645	yahoo.com	554 251
10	tumblr.com	496 045	flickr.com	339 231
11	dmoz.org	476 890	facebook.com	314 051
12	vindhetviahier.nl	424 646	apple.com	312 396
13	jcsearch.com	423 918	miibeian.gov.cn	289 605
14	startpagina.nl	392 543	vimeo.com	269 003
15	yahoo.com	371 087	tumblr.com	226 596
16	tatu.us	370 918	joomla.org	201 863
17	freeseek.org	362 310	amazon.com	196 690
18	lap.hu	352 668	w3.org	196 507
19	blau-webkatalog .com	312 924	nytimes.com	193 907
20	allepaginas.nl	276 578	sourceforge.net	189 663

Table 1: Top 20 websites ordered by in- and outdegree

3.4 Bow-Tie Structure

Broder *et al.* [6] have set up the hypothesis that the macroscopic structure of the Web has the form of a bow tie. The bow-tie structure has a large strongly connected component (LSCC) as its core. The sets containing the remaining nodes that can reach the LSCC or that can be reached from there are called IN and OUT respectively. The nodes that do not belong to any of these three sets are either TENDRILS, if they can either be reached from IN or can reach OUT, TUBES if they are located on a connected path from IN to OUT without passing the LSCC, or DISCONNECTED otherwise.

In this section we test whether we also find a bow-tie structure in our PLD graph and determine the sizes of the components. Before determining this structure, we examine the overall connectedness of the PLD graph, using a weakly connected component (WCC) analysis. We discover a giant WCC, covering 39 374 588 (91.8%) websites. The largest strongly connected component (SCC) in the PLD graph contains 22 274 865 (51.9%) PLDs. The next largest components are much smaller with maximum sizes of around 1 000 PLDs for SCCs and less than 100 PLDs for WCCs.

Having identified the largest WCC and SCC, we now examine the bow-tie structure of our PLD graph. The results are shown in Figure 2. We find that the LSCC is referenced by a small IN component, consisting of 7.65% of all websites. The OUT component contains 30.98% of all nodes. TUBES and TENDRILS together amount to 1.24%.

Table 3 shows the sizes of the different bow-tie components of our PLD graph in comparison to the sizes in the page graph from which we constructed the PLD graph. The LSCC has a similar relative size. In contrast, the relative sizes of the IN and OUT component are exchanged. While the page graph has a large IN component (cf. [13]), we can only find a rather small one in the PLD graph. For the OUT component, the opposite effect appears. In order to understand this effect, one has to keep the applied aggregation in mind. Pages from the IN component are now counted to the LSCC whenever at least one page of the same PLD receives one link from a page in the LSSC component. The same holds for

	Website	PR		Website	PR
1	wordpress.org	113 388	11	apple.com	23 929
2	gmpg.org	111 173	12	phpbb.com	22 329
3	youtube.com	88 206	13	miibeian.gov.cn	22 165
4	twitter.com	54 644	14	hugedomains.com	20 793
5	wikipedia.org	54 081	15	facebook.com	20 254
6	blogspot.com	40 901	16	joomla.org	18 146
7	google.com	40 799	17	flickr.com	17 966
8	wordpress.com	28 018	18	adobe.com	17 903
9	yahoo.com	27 594	19	linkedin.com	16 083
10	networkadvertising.org	27 395	20	w3.org	15 539

Table 2: Top 20 websites ordered by PageRank (PR)

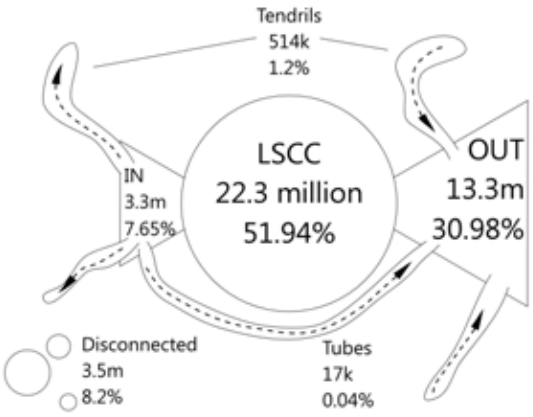

Figure 2: Bow-tie structure of the PLD graph

the OUT component, for PLDs including at least one page pointing back to the LSCC. In addition, we have to keep in mind that the crawl is biased towards some PLDs. For example *youtube.com* contains more than one million pages where other PLDs contain less than ten crawled pages. This skewed distribution in the number of pages per PLD can substantially change the picture, as we measure the sizes of the components relative to the total number of nodes in the respective graph. Beside these general factors, we find that a very large number of pages from the IN component belong to the LSCC after aggregating the page graph to the PLD graph. This means that pages from the IN component often belong to a website that has at least one page in the LSCC. Most pages from the OUT component, however, do not change the component after the aggregation. Hence, these pages belong to websites that are located completely in the OUT component.

	LSCC	IN	OUT	TENDR. TUBES	DISC
PLD graph	51.94	7.65	30.98	1.24	8.20
page graph [13]	51.28	31.96	6.05	4.87	5.84

Table 3: Comparison of the sizes of the different bow-tie components between PLD and page graph in percentages.

3.5 Distances and Diameter

As we now have an impression about the macroscopic structure of the graph and have determined which sets of websites could potentially reach others, we are interested in the distances between pairs of websites. We use the Hyperball method by [5] to calculate an approximation of the distance distribution for our graph (technically, we computed five runs, which give us an approximation with the relative standard deviation 7.66%).

In the PLD graph, $42.42 \pm 3.59\%$ of the pairs of nodes are connected by a directed path between them. Moreover the length of the average shortest path is 4.27 ± 0.085. This means that from large proportion of the websites it is possible to reach almost all other websites by crossing only three others. Figure 3 shows the distance distribution, concentrated around the average. Using Hyperball, we can further estimate the diameter, i.e. the longest shortest path, to be at least 48.

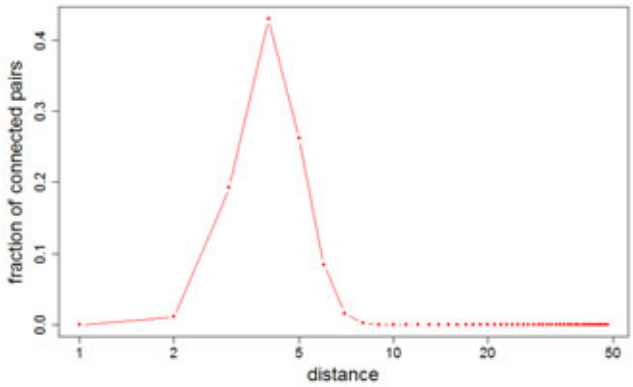

Figure 3: Distance distribution

Having calculated the basic distance measures within the whole graph, we are interested in the diameters of the different bow-tie components. Donato *et al.* [9] showed that both the IN and OUT component are very shallow. Our experiments, running breath-first search from the different components (we used a sample of 500 randomly chosen websites) in both directions lead to the following results: (a) the IN and OUT components are shallow as we found a diameter of two within both components and (b) the overall diameter of the graph is embedded in the LSCC, as the maximal diameter from the IN component is shorter than the one starting in the LSCC. This leads to the assumption that the LSCC contains chains of websites, which are very sparsely connected to other websites. We manually pick such chains from our results and have a look at the pay-level domains. As an example, one chain of length 43 has only domain names containing the keywords "sports" and "bet"/"betting". Another chain of length 26 has a comparably suspicious naming scheme. These domain names contain the keyword "ringtones" in combination with a singer's or a band's name. This indicates that very long paths are likely to be spam and arise due to the artificial or automated generation of websites.

3.6 Backbone of Highly Interlinked Websites

We further want to determine, if there are some particular websites which are responsible for the wide connectivity of the graph and the short average distances. In particular we want to find out if there exists a *hub and spoke* system, or if there is a finely woven structure which connects all websites. For this reason, we attempt to remove links from the PLD graph and measure to which degree the connectedness changes, as proposed by Broder *et al.* [6].

3.6.1 Removing Strongly Connected Websites

To determine the role of hubs, i.e. sites with very large indegree, in connecting the PLD graph, we start with removing all links to websites with an indegree of at least k. We then measure the

sizes of the largest remaining weakly and strongly connected components.

Table 4 shows the experimental results for different values of k. We observe that for a k equal to or smaller than ten large strongly connected components are practically absent. This means that, if there were no hubs, i. e. websites with high indegree, most websites could never be reached by following hyperlinks from other websites. In the case of weakly connected components, for a k of five, only half of all nodes in the graph are still weakly connected. This shows that the graph is not composed of many completely isolated components that are connected by hubs. There are rather resilient connections between the low indegree websites, but for wide navigability, hubs are needed.

	all	10 000	1 000	100	10	5	4
WCC	91.8	90.3	89.4	87.6	73.7	56.8	48.0
SCC	51.9	45.1	39.7	27.8	0.1	< 0.1	< 0.1

Table 4: Percentage of nodes in the largest components for websites with indegree smaller than k

3.6.2 Removing Weakly Connected Websites

We now remove all links to nodes with indegree less than k. Table 5 shows the sizes of the largest remaining weakly and strongly connected components. From the indegree distribution we know that the majority of nodes has low values. As we remove the links to these nodes they can no longer be part of an SCC. For a k of ten, for example, we remove all links to a total of 33 957 836 nodes, which is about 79% of our whole PLD graph. These numbers are reported in the third row of Table 5 for every k.

	10	100	1 000	10 000	100 000
WCC	62.6	48.7	38.6	28.4	18.1
SCC	16.1	1.7	0.1	< 0.1	< 0.1
% removed	79.2	97.9	99.9	> 99.9	> 99.9

Table 5: Percentage of nodes in the largest components for websites with indegree at least k

For a k of ten, we see that the size of the largest WCC decreases from 91% to 62%. This means that links to nodes with low indegree are very important for the overall connectivity of the graph. The remaining 62% of the nodes are still connected, are nodes that link to other nodes with an indegree of ten or higher. As we increase k, we see how the size of the largest WCC decreases. Concerning the largest SCC, the picture now becomes clearer. From the previous experiment we know that the nodes with indegree smaller than ten do not form a large SCC. We now see that the nodes with indegree at least ten form an SCC spanning 16% of the whole graph. Recall that we removed incoming links to 79% of all nodes, which cannot participate in an SCC any more. So this SCC contains 76% of the remaining nodes. For a k of 100, the SCC contains 1.7% of all nodes and 97.9% of all nodes cannot participate in any SCC. This means, almost 81% of the remaining nodes are included in the largest SCC. Summing up, we can say that for our PLD graph the nodes with indegree at least ten form the core of the largest strongly connected component. The nodes with indegree less than ten do not form a large strongly connected component, but a large fraction is weakly connected.

3.6.3 Using a Weighted Graph

In order to find websites that are massively interlinked, we remove all arcs from the PLD graph that represent less than k links in the page graph. This means we practically assign weights to the arcs of the PLD graph. Table 6 shows the results for k values of 500 000, 100 000, 1 000 and 100. As the resulting numbers decrease rapidly, we report absolute numbers instead of percentages. For comparison, the largest weakly connected component for a k of 100 contains 3.9 million websites, which is around 9% of the PLD graph.

	500 000	100 000	10 000	1 000	100
WCC	106	2 331	45 396	642 276	3 908 604
SCC	10	65	1 900	33 300	381 000

Table 6: Number of nodes in the largest components for websites with at least k links between them

We now focus our attention to $k = 500\,000$. In this case, the largest remaining weakly connected component contains 106 websites. These are the pay-level domains that are most frequently being linked to in our PLD graph, as each of them receives at least half a million incoming links. Figure 4 shows this sub graph, where we visualize the topical areas of the included websites, using the best fitting category from the Open Directory Project.

Looking at this sub graph, we can make some detailed observations. The WCC can be split into two parts that are only connected via *imdb.com*. The smaller part, around the domain *amazon.com*, belongs mostly to the shopping category. The other part contains various other well-known PLDs like *google.com*, *facebook.com* and *blogspot.com*. In the centre of this second part, we find an SCC of size seven containing *youtube.com*, *google.com*, *sapo.pt*, *wordpress.com*, *typepad.com*, *blogspot.com* and *blogalaxia .com*. From this centre, several links to groups of other PLDs can be observed. *youtube.com* links to several PLDs that belong to various categories. *blogspot.com* links to many blogs, blog-hosting sites and news sites. Further analysis shows that the 106 PLDs from this WCC are connected, either by inlinks or by outlinks, to a total of 10 456 257 PLDs, which is almost a quarter of all PLDs in our PLD graph.

The largest strongly connected component comprises ten websites with adult content. Further inspection of the next largest SCCs for a k of 500 000 reveals the previously mentioned one SCC with seven PLDs. The domain *universehotels*, represented with five different top-level domains, forms another strongly connected component. Another eight SCCs with a total of 48 PLDs contain adult content.

3.7 Two-Layer Model of the PDL Graph

Combining all our observations, we hypothesize that the structure of the PLD graph can be explained using a *Two-Layer Model*, as depicted in Figure 5. The *Low Degree Layer* (LDL) contains the majority of websites that are sparsely connected, forming the giant weakly connected component. Within the *High Degree Layer* (HDL) we find websites with high indegree and large amounts of links between the websites.

From Table 4, we know that a large strongly connected component does not exist for nodes with indegree ten or lower. In case we include nodes with indegree up to 100, we find such a component.[9] Thus, we infer that the LDL consists of nodes with a maximum

[9]From our experiments, this is the smallest value of k that resulted in a large SCC. A more exact definition of this border is left for future research.

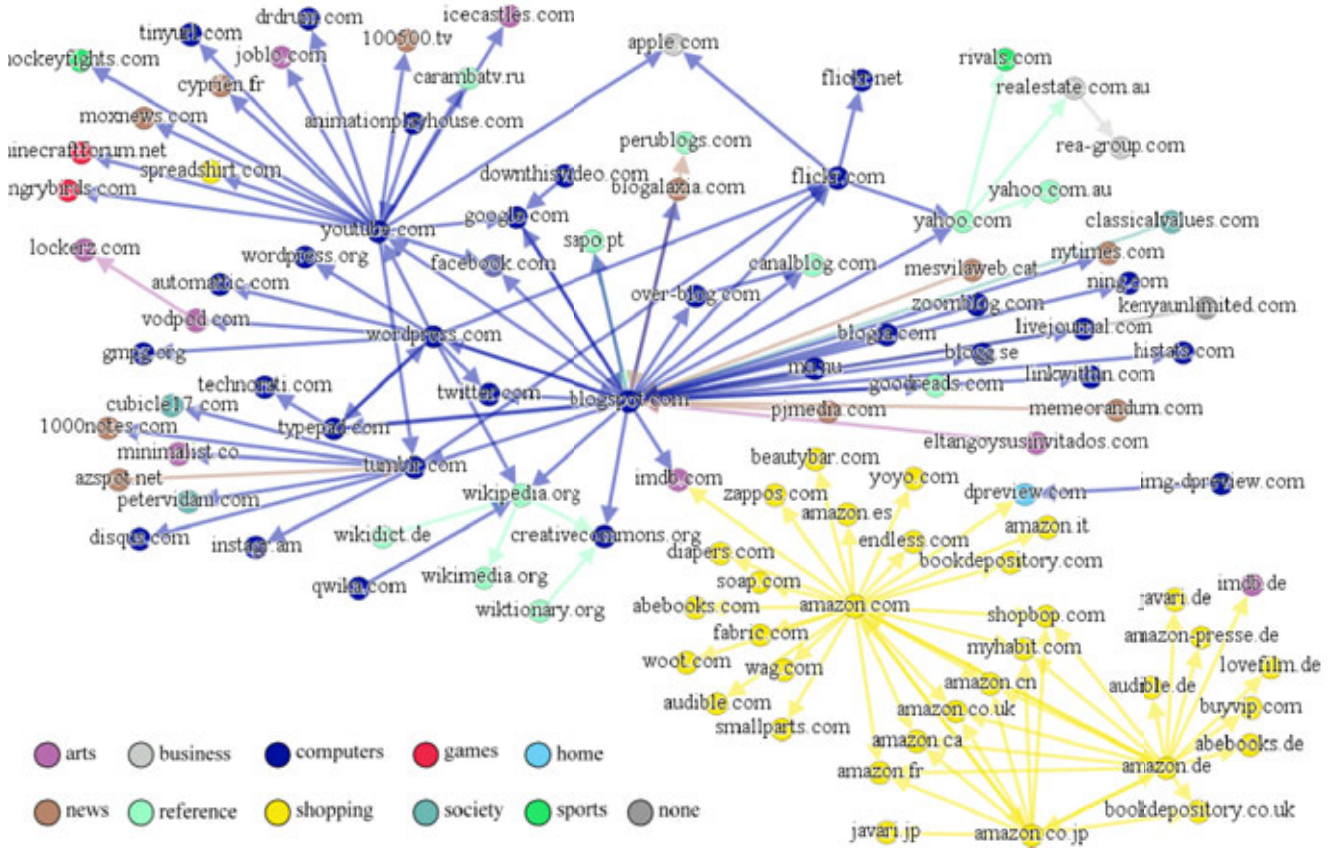

Figure 4: Largest WCC with more than 500 000 links

indegree ranging between ten and 100. As we know from the distribution of the indegree, this is the majority of all nodes. There are 34.6 million nodes with an indegree up to ten and additional 7.3 million nodes with indegree between 11 and 100. This leaves 881 thousand nodes with an indegree higher than 100. Hence, the LDL contains between 80% and 97% of all nodes of our PLD graph. Assuming the LDL includes all nodes with indegree up to 100, we find that 30 million nodes of this layer (71% of the PLD graph) being directly connected to the HDL formed by the remaining nodes.

We now compare the *Two-Layer Model* to the bow-tie structure and outline how the bow tie can, partially, be explained by our hypothesis. A large part of the giant WCC, containing 91.8% of all nodes, is explained by the LDL, which has a WCC containing between 73.7% to 87.6% of all nodes. The LSCC of the bow-tie structure comprises 51.9% of all nodes and can only be formed by the HDL. We find that 781 722 (89%) nodes from the HDL belong to the LSCC. The links between these nodes and the LDL create the LSCC with 22.3 million nodes. Hence, the nodes from the HDL are also likely to have a rather high outdegree. The remaining nodes from the HDL mostly belong to the OUT component. Only a few can also be found in the other components of the largest WCC.

We further suppose that the HDL provides short-cuts through the graph. The nodes with indegree higher than 100, which are definitely located in the HDL, are source or target of 421 million links, which is 68% of all links in the graph. 124 million of these links have both their source and target inside the HDL, leaving 297 million inter-layer links. This linkage pattern can be used to explain the small distances between a large proportion of pairs of nodes, as shown in Section 3.5.

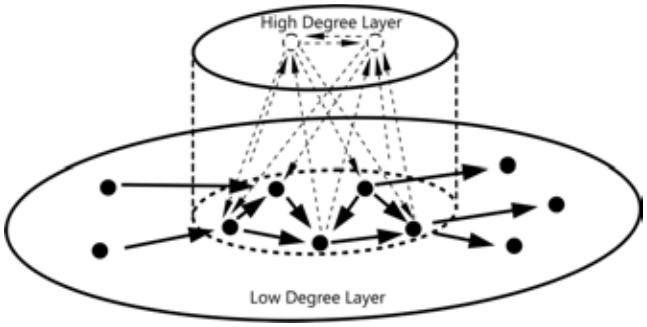

Figure 5: Two-Layer Model of the PLD Graph

An interesting question that we did not investigate and which is left for future research is how our hypothesized *Two-Layer Model* is in agreement with the Core-Periphery Structure as described by Rombach *et al.* [16].

4. LINKAGE BETWEEN TOP-LEVEL DOMAINS

Besides the general connectivity and linkage on PLD level, we are also interested in linkage patterns that exist between websites that belong to different public suffixes. For this analysis we select the top ten public suffixes (PS) based on the number of websites included in our PLD graph. In addition, we introduce a group "others" that covers the remaining websites. Altogether, websites from

the top ten PS amount for a total of 78.98% of all nodes of the PLD graph. Nodes belonging to the group of the ".com" websites cover almost half of all nodes of our graph. The distribution of the different PS in our dataset generally agrees with the Verisign Domain Industry Brief from October 2012.

Figure 6 shows the links between these 11 groups. In this diagram, the outermost circle labelled with percentages represents the total number of links for each group. Directly adjacent to this circle are two smaller bars for each group. The outer bar represents the number of inlinks while the inner bar represents the number of outlinks. Further to the centre there is another circle labelled with absolute numbers. This circle, again, represents the total number of links. From this circle ribbons spanning through the middle represent interconnections between the different groups. Incoming links have a white gap between the ribbon spanning in the middle and the circle part labelled with absolute numbers. Also, the ribbon has the same colour as the group that it originates from.

The two largest public suffixes, ".com" and ".de", exhibit an outstandingly high number of intra-group links. On the opposite, a rather low number of such links exists within the ".org", ".info" and ".net" suffixes. Besides, we can observe a general trend. Public suffixes with country code top-level domains (TLDs) tend to have a larger fraction of intra-group links than those with general TLDs, such as ".org" or ".net". The ".com" TLD does not follow this trend. An explanation for this exception is that the ".com" TLD is used accross all countries. This also manifests in the fact that the largest fraction of inter-group links for every public suffix is from and to the ".com" TLD.

Further inspection of the ".de" TLD links reveals an interesting fact. A total of 11 171 PLDs are either linking to *verleihcenter.de* or *verleihcenter.eu*. All those websites are hosted on nine different IP-addresses, meaning they mainly share the same servers. The summed outdegree of all PLDs hosted on these nine IP-addresses amounts to 21 801 852 which is almost half of all intra-".de"-links. Also, these spam PLDs are responsible for the spikes in the in- and outdegree distributions of the PLD graph (see section 3.1).

5. LINKAGE OF CATEGORIZED WEBSITES

Adding category information using data obtained from DMOZ to our websites, we add a content-based perspective to our analysis. Combined with our large amount of linkage information, we can analyse the websites of each topical category for their structural properties. Besides a visualization of the linkage between the topical categories, we detail statistics about incoming and outgoing links as well as the average clustering coefficient for each category.

5.1 Category Statistics

We matched the websites for all 15 relevant DMOZ top-level categories,[10] excluding the categories "regional" and "world" as they provide a geographical and not a content based categorisation. Within DMOZ categories can be maintained for different granularities, meaning that a large number of categorizations are done on page level. As we apply the categorisation to our PLD graph, we obtain multiple categories for some pay-level domains. For example, the video portal *youtube.com* is found in almost all categories. In order to obtain a distinct categorisation we only use those websites that are assigned only a single category. This results in 743 686 distinctly categorised PLDs out of the 796 251 PLDs we obtained from the matching of our PLD graph with the DMOZ data.

[10]The data set is available as public RDF data: `http://www.dmoz.org/docs/en/rdf.html`

Figure 6: Linkage of websites grouped by TLDs

Table 7 shows the number of matched websites for each category. The numbers reported refer to pay-level domains that were categorised distinctly. The second column gives the absolute number and the third column contains the percentage of this number relative to the total number of categorized websites.

The largest categories we obtain are "business" and "society". The categories with the least number of websites are "home" and "news".

Category	Number of PLDs	Percentage
business	176 890	26
society	99 801	14
arts	75 978	11
shopping	67 477	10
recreation	54 721	8
computers	52 995	8
sports	43 334	6
science	28 717	4
health	28 571	4
adult	12 475	2
reference	14 329	2
kids and teens	11 742	2
games	10 885	2
home	7 577	1
news	4 118	1

Table 7: Websites distinctly categorized by DMOZ

5.2 Linkage between Categories

Combining the DMOZ data with our PLD graph gives us the opportunity to analyse the linkage between the categories. For the visualisation of our results, we use the same diagram layout as in Section 4. In Figure 7, the outer circle represents the categories, sized by their respective number of links. In the centre of the cir-

cle, the interconnections of the categories are displayed by ribbons spanning in between. Although we hoped to find groups of categories that show a clearly distinguishable linkage pattern, our results show that there are no clear preferences concerning which categories link to each other.

Looking at the figure, we see that the "computers" category plays a dominant role. It has the largest number of links, although it is only the sixth largest category by the number of websites. We observe that for connections with other categories, it has more outgoing than incoming links. The opposite applies for the "business" category, which receives more incoming links than outgoing links.

Overall, we can say that most categories have a large fraction of internal links. However, this does not endure for some categories. For example, the "shopping" and "reference" categories only have few internal links. In the case of the "shopping" category, this may be explained by the fact that different shopping websites are competitors and hence do not link to each other. This was also hypothesized by Broder *et al.* [6], assuming that those are mainly located in the OUT component.

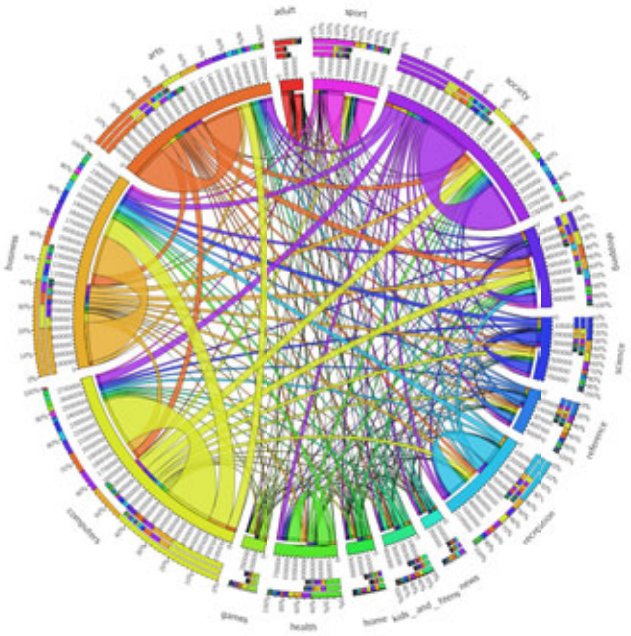

Figure 7: Linkage of websites grouped by categories

5.3 In- and Outdegree

In a next step, we have a closer look at the distribution of links among the categories. To this end, we determine the total number of links as well as the average in- and outdegree. Note that, contrary to the previous section, we now take all links, including those to and from uncategorised websites, into account.

Table 8 lists the average in- and outdegree as well as the total number of outgoing links for each category. We observe the highest average indegree for the categories "computers", "kids and teens" and "news". The categories with the highest average outdegree are the same, but in a different order: "kids and teens", "news" and "computers". As mentioned before, the "business" and "shopping" categories have a low average outdegree. These sites most likely refrain from placing links to their competitors. For most categories, the average indegree is higher than the average outdegree, except for the categories "adult" and "games".

Category	Avg. indegree	Avg. outdegree	total in-links	total out-links
adult	103	116	1 297 359	1 451 297
arts	91	69	6 918 096	5 282 143
business	70	29	12 455 999	5 142 553
computers	279	133	14 807 829	7 091 805
games	101	102	1 101 279	1 113 447
health	109	48	3 129 440	1 381 696
home	177	117	1 348 016	890 686
kids and teens	257	169	3 024 563	1 988 132
news	233	155	962 152	641 919
recreation	83	52	4 559 801	2 882 297
reference	173	98	2 491 173	1 408 763
science	107	62	3 091 737	1 805 970
shopping	79	24	5 365 656	1 622 528
society	81	57	8 134 803	5 763 999
sports	65	41	2 832 955	1 813 262

Table 8: Degree statistics of websites distinctly categorized by DMOZ ordered by category

5.4 Clustering Coefficient

We now also take the vicinity of the categorised websites into account and compute the clustering coefficient. By this, we can get an impression about the structural environment the categorised websites are embedded in. Due to the computation time required to determine the clustering coefficient, we cannot use all PLDs, but used a sample of 1 000 pay-level domains per category.

Figure 8 plots the clustering coefficients for all categories. The dark blue line represents the average clustering coefficient as obtained from our sample. The light blue area around this line visualizes the positive and negative standard deviation.

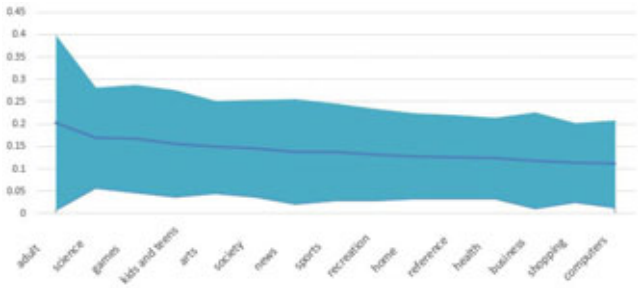

Figure 8: Average clustering coefficient by category

All values obtained for our sample fall in a range between 0.11 and 0.19 with standard deviations from 0.09 to 0.19. The highest clustering coefficients are observed for the categories "adult", "science" and "games". These categories also show the largest standard deviation. The lowest values are again observed for the "business" and "shopping" categories, which are now unexpectedly joined by the "computers" category.

5.5 Summary

Figure 9 shows a diagram summarizing our findings about the topical categories. To fit all this data into one chart, we computed several ratios to reduce the number of dimensions. The x-axis charts the ratio of in- and outdegree, which we know from Section 5.3. A value of one means both are the same, while a value higher than one means the indegree is higher than the outdegree. The y-axis represents the average number of links per website, incorporating data from Section 5.1. Finally, the size of each bubble is relative to the clustering coefficient as reported in Section 5.4.

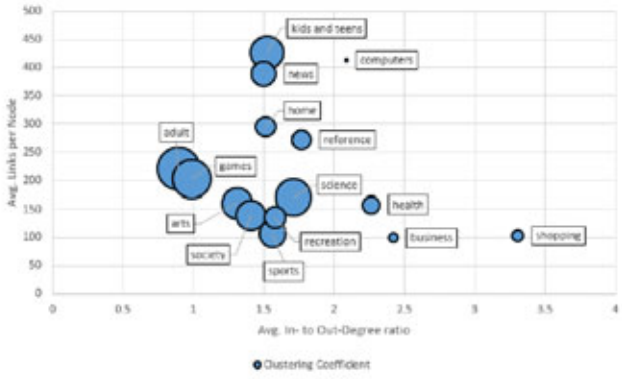

Figure 9: Structural description of the DMOZ categories

6. RELATED WORK

This section gives an overview of related research in the area of web graph analysis. First, we discuss research that analyzes the Web graph on page level. Afterwards, we describe related work analyzing web graphs on a higher aggregation level.

6.1 Page-Level Graphs

Most existing studies analyze the structure of web graphs on page level. Among these studies, one of the most influential and often cited is the work of Broder *et al.* [6] from the year 2000. Their study is based on two crawls from AltaVista with over 200 million pages and 1.5 billion links each. They introduce the bow-tie structure which gives an overview of the macroscopic structure of the Web. Further, they find that the average distance in their graph is about 16, if there is a directed path between two pages. According to them, such a path exists in only 25% of all cases. In addition, Broder *et al.* claim that in- and outdegree distributions exhibit a power law. Kumar *et al.* [12] and Barabási and Albert [2] reported similar results.

The inner structure of the bow-tie components was analyzed by Donato *et al.* [9] in their work from 2005 using four different crawls. They found weakly connected components in IN and OUT to be rather small and that most webpages are isolated, meaning that the paths to and from the LSCC are rather short. Concluding, they propose a *"daisy"* structure to replace the bow-tie structure of the Web. In this *"daisy"* structure, the IN and OUT components are represented by many small WCCs around the LSCC. Within our PLD graph, we could identify similar characteristics, where we observe a maximum length of paths inside the IN and OUT component of two.

Using different data sets, various numbers for the proportions of the bow-tie components have been reported in the past. Examples include the works of Serrano *et al.* [17], Boldi and Codenotti [3], Baeza-Yates and Poblete [1] and Zhu *et al.* [20]. As the reported numbers show strong variations, it can be assumed that the crawling strategy has a non-negligible impact on the size of the components. But, as in our case, details about the exact crawling strategy and parameters are not always known.

We have already published a page level analysis [13] of the same graph that was used to create the aggregated graph presented in this paper. For the in- and outdegree distributions, we report an estimate of the power law using maximum-likelihood fitting and perform a goodness-of-fit test [7]. The test shows that both distributions do not follow a power law. Further, we report that approximately 48% of all nodes are connected by a directed path and the average distance is around 12.84. The graph exhibits a bow-tie structure with large LSCC (51.28%) and IN (31.96%) components and a small OUT component (6.05%).

6.2 Aggregated Graphs

Only a few studies have been published that analyze the web graph on a higher aggregation level.

Hirate *et al.* [11] gathered a large crawl between 2004 and 2005, which they analysed using a host level reduction. Although their reduction method is the same as ours, they did not use the same aggregation level. As the host may contain sub domains, the resulting graph has several nodes per pay-level domain. They found that the bow tie shows a rather small IN component (10%) and quite large LSCC and OUT components with (41%) of all nodes each.

Zhu *et al.* [20] analysed the structure of the Chinese Web. They compared their results on three different aggregation levels: the page level, the host level and the domain level. On the page level, they found a large IN component, which disappears on the host and the domain level. Analogously, the LSCC and OUT components of the host and domain level are larger than on the page level.

Dill *et al.* [8] compared several sub graphs, one of which is the "hostgraph". The indegree power law exponent for their graph is 2.34 and hence close to the value we reported. Concerning the bow-tie structure, they found an LSCC of 82%. This result led them to the conclusion that almost every website has a page belonging to the LSCC.

Interestingly, the trend that can be seen in the results of Hirate *et al.* [11] and Zhu *et al.* [20] can also be found in our results: On the page level, the IN component is larger than the OUT component, but on an aggregated level, the opposite is true. Obviously, the aggregation from the page level to a higher level shrinks the IN component of the bow tie and simultaneously inflates the OUT component. The results of Dill *et al.* [8] do not really fit to these observations. As we do not have further information about the methodology Dill *et al.* applied to gather their data, we cannot say whether this is as phenomenon that has changed since 2002 or whether the results of Dill *et al.* are an artefact of their crawling strategy.

7. CONCLUSION

In this paper, we have analysed the aggregated version of the largest publicly available web graph, extracted from a web crawl from 2012. This PLD graph includes at least one page of between 18% and 27% of all registered domains from the time of the crawl. We have shown that using a PLD aggregation level, we can overcome the effect of website-internal links on graph measures, and hence give insights about the graph structure on website level.

First, we analysed basic graph statistics of our graph and showed that the outdegree distribution does not follow a power law. Further, we found a diameter of at least 48 in our graph and an average distance of 4.27. When examining the overall graph structure, we calculated the bow-tie structure and detected a similar shift in IN and OUT components in comparison to the page graph as previously observed by Zhu *et al.* [20].

Further, we hypothesize a two-layer model explaining the LSCC and the short distances in the graph. This model describes a large group of finely woven, weakly connected websites with a relatively low degree – the Low Degree Layer. This layer is responsible for the overall connectedness of the graph. In addition, we identify a second layer of websites with high degrees – the High Degree Layer – which we assume is responsible for the short average distance within the whole graph.

Looking at the linkage between PLDs belonging to different top-level domains, we found two patterns: generic top-level domains, for example ".org", have more external links than links pointing to websites within the same top-level domain. Public suffixes based on country-specific top-level domains have a larger proportion of internal links. The ".com" domains do not follow these patterns, which we presume to be due to its usage across all languages and countries.

As we grouped pay-level domains by the topical category obtained from DMOZ, we measured more detailed properties of the different categories. By combining these properties, we find indications for how categories can be distinguished on an aggregate level.

8. REFERENCES

[1] R. Baeza-Yates and B. Poblete. Evolution of the chilean web structure composition. In *Web Congress, 2003. Proceedings. First Latin American*, pages 11–13, 2003.

[2] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286:209–512, October 1999.

[3] P. Boldi, B. Codenotti, M. Santini, and S. Vigna. Structural properties of the african web. In *WWW '02*, volume 66, 2002.

[4] P. Boldi and S. Vigna. The webgraph framework I: compression techniques. In *WWW '04*, pages 595–602. ACM, 2004.

[5] P. Boldi and S. Vigna. In-core computation of geometric centralities with HyperBall: A hundred billion nodes and beyond. In *ICDMW 2013*. IEEE, 2013.

[6] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web. *Computer networks*, 33(1):309–320, 2000.

[7] A. Clauset, C. R. Shalizi, and M. E. Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

[8] S. Dill, R. Kumar, K. S. Mccurley, S. Rajagopalan, D. Sivakumar, and A. Tomkins. Self-similarity in the web. *ACM Trans. Internet Technol.*, 2(3):205–223, Aug. 2002.

[9] D. Donato, S. Leonardi, S. Millozzi, and P. Tsaparas. Mining the inner structure of the web graph. In *WebDB*, pages 145–150, 2005.

[10] D. Fetterly, M. Manasse, and M. Najork. Spam, damn spam, and statistics: Using statistical analysis to locate spam web pages. *International Workshop on the Web*, pages 1–6, 2004.

[11] Y. Hirate, S. Kato, and H. Yamana. Web structure in 2005. In *Algorithms and models for the web-graph*, pages 36–46. Springer, 2008.

[12] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Trawling the web for emerging cyber-communities. *Computer Networks*, 31(11 - 16):1481 – 1493, 1999.

[13] R. Meusel, S. Vigna, O. Lehmberg, and C. Bizer. Graph structure in the web — revisited: A trick of the heavy tail. In *Proc. of WWW Companion '14*, pages 427–432, 2014.

[14] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: bringing order to the web. pages 1–17, 1999.

[15] G. Pandurangan, P. Raghavan, and E. Upfal. Using Pagerank to characterize web structure. *Computing and Combinatorics*, pages 330–339, 2002.

[16] M. P. Rombach, M. A. Porter, J. H. Fowler, and P. J. Mucha. Core-periphery structure in networks. *arXiv preprint arXiv:1202.2684*, 2012.

[17] M. Serrano, A. Maguitman, M. Boguñá, S. Fortunato, and A. Vespignani. Decoding the structure of the www: A comparative analysis of web crawls. *ACM Transactions on the Web (TWEB)*, 1(2):10, 2007.

[18] S. Spiegler. Statistcs of the common crawl corpus 2012. Technical report, SwiftKey, June 2013. Document viewed on September 16th 2013 from https docs.google.comfiled1_9698uglerxB9nAglvaHkEgU-iZNm1TvVGuCW7245-WGvZq47teNpb_uL5N9.

[19] S. Vigna. Fibonacci binning. *CoRR*, abs/1312.3749, 2013.

[20] J. J. H. Zhu, T. Meng, Z. Xie, G. Li, and X. Li. A teapot graph and its hierarchical structure of the chinese web. *WWW '08*, pages 1133–1134, 2008.