Graph Structure in the Web — Revisited

or A Trick of the Heavy Tail

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

Sebastiano Vigna
Laboratory for Web Algorithmics
Università degli Studi di Milano
Italy
vigna@acm.org

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

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

ABSTRACT

Knowledge about the general graph structure of the World Wide Web is important for understanding the social mechanisms that govern its growth, for designing ranking methods, for devising better crawling algorithms, and for creating accurate models of its structure. In this paper, we describe and analyse a large, publicly accessible crawl of the web that was gathered by the Common Crawl Foundation in 2012 and that contains over 3.5 billion web pages and 128.7 billion links. This crawl makes it possible to observe the evolution of the underlying structure of the World Wide Web within the last 10 years: we analyse and compare, among other features, degree distributions, connectivity, average distances, and the structure of weakly/strongly connected components.

Our analysis shows that, as evidenced by previous research [17], some of the features previously observed by Broder *et al.* [10] are very dependent on artefacts of the crawling process, whereas other appear to be more structural. We confirm the existence of a giant strongly connected component; we however find, as observed by other researchers [12, 5, 3], very different proportions of nodes that can reach or that can be reached from the giant component, suggesting that the "bow-tie structure" as described in [10] is strongly dependent on the crawling process, and to the best of our current knowledge is not a structural property of the web.

More importantly, statistical testing and visual inspection of size-rank plots show that the distributions of indegree, outdegree and sizes of strongly connected components are not power laws, contrarily to what was previously reported for much smaller crawls, although they might be heavy tailed. We also provide for the first time accurate measurement of distance-based features, using recently introduced algorithms that scale to the size of our crawl [8].

WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2576928.

Categories and Subject Descriptors

H.3.4 [**Information storage and retrieval**]: Systems and software—World Wide Web (WWW)

Keywords

World Wide Web; Web Graph; Network Analysis; Graph Analysis; Web Mining

1. INTRODUCTION

The evolution of the World Wide Web is summarized by Hall and Tiropanis as the development from "the web of documents" in the very beginning, to "the web of people" in the early 2000's, to the present "web of data and social networks" [13]. With the evolution of the World Wide Web (WWW), the corresponding web graph has grown and evolved as well.

Knowledge about the general graph structure of the web graph is important for a number of purposes. From the structure of the web graph, we can provide evidence for the social phenomena governing the growth of the web [13]. Moreover, the design of *exogenous* ranking mechanisms (i.e., based on the links between pages) can benefit from deeper knowledge of the web graph, and the very process of crawling the web can be made more efficient using information about its structure. Finally, studying the web can help to detect rank manipulations such as spam networks, which publish large numbers of "fake" links in order to increase the ranking of a target page.

In spite of the importance of knowledge about the structure of the web graph, the latest publicly accessible analysis of a large global crawl is nearly a decade old. The first, classic work about the structure of the web as a whole was published by Broder *et al.* [10] in 2000 using an AltaVista crawl of 200 million pages and 1.5 billion links.[1] A second similar crawl was used to validate the results.

One of their main findings was a *bow-tie* structure within the web graph: a giant strongly connected component containing 28% of the nodes. In addition, Broder *et al.* show that the indegree distribution, the outdegree distribution and the distribution of the sizes

[1]Throughout the paper, we avoid redundant use of the $\approx$ symbol: all reported figures are rounded.

of strongly connected components are heavy tailed. The paper actually claims the distributions to follow power laws, but provides no evidence in this sense except for the fact that the data points in the left part of the plots are gathered around a line. The authors comment also on the fact that the initial part of the distributions displays some concavity on a log-log plot, which requires further analysis.

An important observation that has been made by Serrano *et al.* [17] analysing four crawls gathered between 2001 and 2004 by different crawlers with different parameters is that *several properties of web crawls are dependent on the crawling process*. Maybe a bit optimistically, Broder *et al.* claimed in 2000 that "These results are remarkably consistent across two different, large AltaVista crawls. This suggests that our results are relatively insensitive to the particular crawl we use, provided it is large enough". We now know that this is not true: several studies [12, 5, 3, 21] using different (possibly regional) crawls gathered by different crawlers provided quite different pictures of the web graph (e.g., that "daisy" of [12] or the "teapot" of [21]).

In particular, recent strong and surprising results [1] have shown that, in principle, most heavy-tailed (and even power-law) distributions observed in web crawls may be just an artefact of the crawling process itself. It is very difficult to predict when and how we will be able to understand fully whether this is true or not.

Subsequent studies confirmed the existence of a large strongly connected component, usually significantly larger than previously found, and heavy-tailed (often, power-law) distributions. However, such studies used even smaller web crawls while the size of the web was approaching the tera scale, and provided the same, weak visual evidence about distribution fitting. While no crawl can claim to represent the web as a whole (even large search engines crawl only a small portion of the web, geographically, socially and economically selected) the increase in scale of the web requires the analysis of crawls an order of magnitude larger. Nonetheless, billion-scale representative crawls have not been available to the scientific community until very recently. Thus, only large companies such as Google, Yahoo!, Yandex, and Microsoft had updated knowledge about the structure of large crawls of the WWW.

A few exceptions exist, but they have significant problems. The AltaVista webpage connectivity dataset, distributed by Yahoo! as part of the WebScope program, has in theory 1.4 billion nodes, but it is extremely disconnected: half of the nodes are isolated (no links incoming or outgoing) and the largest strongly connected component is less than 4% of the whole graph, which makes it entirely unrepresentative. We have no knowledge of the crawling process, and URLs have been anonymised, so no investigation of the causes of these problems is possible.

The ClueWeb09 graph, gathered in 2009 within the U.S. National Science Foundation's Cluster Exploratory (CluE), has a similar problem due to known mistakes in the link construction, with a largest strongly connected component that is less the 3% of the whole graph. As such, these two crawls cannot be used to infer knowledge about the structure of the web.

The ClueWeb12 crawl, released concurrently with the writing of this paper, has instead an accurate link structure, and contains a largest strongly connected component covering 76% of the graph. The crawl, however, is significantly smaller than the graph used in this paper, as it contains 1.2 billion pages,[2] and it is focused mostly on English web pages.

[2]Note that the web graph distributed with ClueWeb09 and ClueWeb12 appears to be much larger because all *frontier* nodes have been included in the graph. The number we report are those of the actually crawled pages.

In this paper, we try to update the original studies on the structure of the web and its current state. We revisit and update the findings of previous research to give an up-to-date view of the web graph today, using a crawl that is significantly larger (3.5 billion pages) than the ones used in previous work.

We repeat previous measurement, observing interesting differences, and provide new, previously unknown data, such as the distance distribution. The crawl[3] as well as the hyperlink graph[4] are publicly available, so to encourage other researchers and analysts to replicate our results and investigate in further interesting topics.

2. DATASET AND METHODOLOGY

The object of study of this paper is a large web crawl gathered by the Common Crawl Foundation[5] in the first half of 2012. The crawl contains 3.83 billion web documents, of which over 3.53 billion (92%) are of mime-type `text/html`. The crawler used by the Common Crawl (CC) Foundation for the crawl is based on a breath-first visiting strategy, together with heuristics to detect spam pages. In addition heuristics were used to reduce the number of crawled pages with duplicate or no content. Such heuristics, in principle, may cut some of the visiting paths and make the link structure sparser. The crawl was seeded with the list of pay-level-domain names from a previous crawl and a set of URLs from Wikipedia. The list of seeds was ordered by the number of external references. Unfortunately this list is not public accessible, but we estimated that at least 71 million different seeds were used, based on our observations on the ratio between pages and domains. The selected amount of seeds in combination with the methodology are likely to affect the distribution of host sizes, as popular websites were crawled more intensely: for example, `youtube.com` is represented by 93.1 million pages within the crawl [18]. In addition, it is likely that the large number of seeds used in the multiple phases of the crawl caused the large number of pages of indegree zero (20% of the graph) found in the graph.

Associated with the crawl is a *web graph*, in which each node represents a page and each arc between two nodes represents the existence of one or more hypertextual links between the associated pages. We extracted the web graph from the crawl with a 3-step process, using an infrastructure similar to the framework used by Bizer *et al.* to parse the Common Crawl corpus and extract structured data embedded in HTML pages [4]. We first collected for each crawled page its URL, mime-type, links to other pages, type, and, if available, the redirect URL, using 100 parallel `c1.xlarge` Amazon Elastic Compute Cloud (EC2) machine instances. We then filtered the extracted URLs by mime-type `text/html` and kept only links within HTML elements of type `a` and `link`, as we want to focus on HTML pages linking to other HTML pages.[6] Also redirects contained in HTTP header have been treated as links. Finally, we used a 40-node Amazon Elastic MapReduce cluster to compress the graph, indexing all URLs and remove duplicate links.

Additionally, we built the host graph and the pay-level-domain (PLD) graph. Nodes in such graphs represent sets of pages with the

[3]`https://commoncrawl.atlassian.net/wiki/display/CRWL/About+the+Data+Set`

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

[5]`http://commoncrawl.org/`

[6]We remark that this choice might have introduced some sparsity, as in principle the crawling process might have followed further links, such as `src` attributes of `iframe` elements. Keeping perfectly aligned the online (during the crawl) and offline (in a separate pass after the crawl) link extraction process when they are performed by different organisations is, unfortunately, quite difficult, as link and page selection strategies could differ.

same host/pay-level-domain, and there is an arc between nodes x and y if there is at least one arc from a page in the set associated with x to a page in the set associated with y. Table 1 provides basic data about the size of the graphs.

Granularity	# Nodes in millions	# Arcs in millions
Page Graph	3 563	128 736
Host Graph	101	2 043
PLD Graph	43	623

Table 1: Sizes of the graphs

3. ANALYSIS OF THE WEB GRAPH

Most of the analyses presented in the following section have been performed using the "big" version of the WebGraph framework [6], which can handle more than 2^{31} nodes. The BV compression scheme was able to compress the graph *in crawl order* at 3.52 bits per link, which is just 12.6% of the information-theoretical lower bound (under a suitable permutation of the node identifiers it is common to obtain slightly more than one bit per link). The whole graph occupied in compressed form just 57.5 GB, which made it possible to run resource intensive computations such as the computation of the strongly connected components.

3.1 Indegree & Outdegree Distribution

The simplest indicator of density of web graphs is the average degree, that is, the ratio between the number of arcs and the number of nodes in the graph.[7]

Broder *et al.* report an average degree of 7.5 links per page. Similar low values can be found in crawls of the same years—for instance, in the crawls made by the Stanford WebBase project.[8] In contrast our graph has average degree of 36.8, meaning that the average degree is factor 4.9 larger than in the earlier crawls. Similar values can be found in 2007 `.uk` crawls performed by the Laboratory for Web Algorithmics, and the ClueWeb12 crawl has average degree 45.1.[9] A possible explanation for the increase of the average degree is the wide adoption of *content management systems*, which tend to create dense websites.

Figures 1 and 2 show frequency plots of indegrees and outdegrees in log-log scale. For each d, we plot a point with an ordinate equal to the number of pages with that have degree d. Note that *we included the data for degree zero*, which is omitted in most of the literature. We then aggregate the values using *Fibonacci binning* [19] to show the approximate shape of the distribution.

Finally, we try to fit a power law to a tail of the data. This part is somewhat delicate: previous work in the late 90's has often claimed to find power laws just by noting an approximate linear shape in log-log plots: unfortunately, almost all distributions (even, sometime, non-monotone ones) look like a line on a log-log plot [20]. Tails exhibiting high variability, in particular, are very noisy (see the typical "clouds of points" in the right part of degree plots) and difficult to interpret.

[7]Technically speaking, the *density* of a graph is the ratio between the *square* of the number of nodes and the number of arcs, but for very sparse graphs one obtains abysmally small numbers that are difficult to interpret.

[8]`http://dbpubs.stanford.edu:8091/~testbed/doc2/WebBase/`

[9]We remark that all these values are actually an underestimation, as they represent the average number of outgoing arcs *in the web graph built from the crawl*. The average number of links per page can be higher, as several links will point outside the graph.

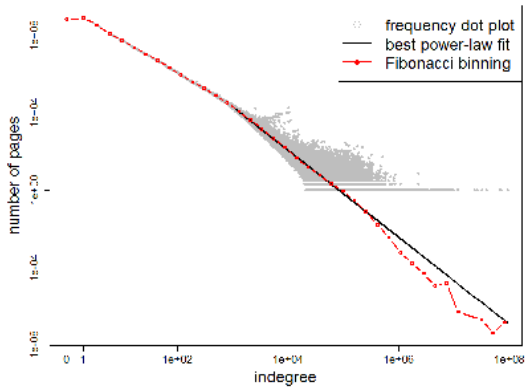

Figure 1: Frequency plot of the indegree distribution

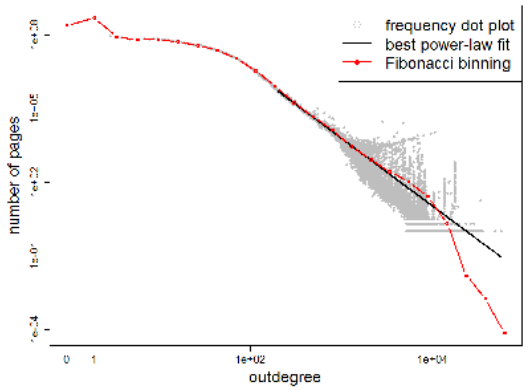

Figure 2: Frequency plot of the outdegree distribution

We thus follow the methodological suggestions of Clauset *et al.* [11]. We use the `plfit`[10] tool to attempt a maximum-likelihood fitting of a power law starting from each possible degree, keeping the starting point and the exponent providing the best likelihood. After that we perform a goodness-of-fit test and estimate a p-value.

The first important fact we report is that *the p-value of the best fits is 0* (± 0.01). In other words, from a statistical viewpoint, in spite of some nice graphical overlap the tail of the distribution is *not* a power law. We remark that this paper applies for the first time a sound methodology to a large dataset: it is not surprising that the conclusions diverge significantly from previous literature.

To have some intuition about the possibility of a heavy tail (i.e., that the tail of the distribution is not exponentially bounded) we draw the *size-rank* plot, as suggested in [14]. The size-rank plot is the discrete version of the complementary cumulative distribution function in probability: if the data fits a power law it should display as a line on a log-log scale. Concavity indicates a superpolynomial decay. Size-rank plots are monotonically decreasing functions, and do not suffer the "cloud of points" problem.

Figure 3 shows the size-rank plot of the degree distributions of our graph and the best power-law fit: from what we can ascertain visually, there is a clear concavity, indicating once again that the tail of the distribution is not a power law. The concavity leaves open the possibility of a non-fat heavy tail, such as that of a lognormal distribution.

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

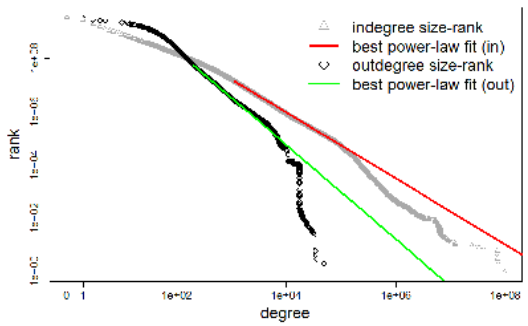

Figure 3: Size-rank plot of degree distributions

In any case, the tails providing the best fit characterize a very small fraction of the probability distribution: for indegrees, we obtain an exponent 2.24 starting at degree 1 129, whereas for outdegrees we obtain an exponent 2.77 starting at 199, corresponding, respectively, to 0.4% and less than 2% of the probability mass (or, equivalently, fraction of nodes). Models replicating this behaviour, thus, explain very little of the process of link formation in the web.

The values we report are slightly different than those of Broder *et al.*, who found 2.09 respectively 2.72 as power-law exponent for the indegree respectively outdegree. But in fact they are incomparable, as our fitting process used different statistical methods.

Finally, the largest outdegree is three magnitudes smaller than the largest indegree. This suggests that the decay of the indegree distribution is significantly slower than that of the outdegree distribution, a fact confirmed by Figure 3.

3.2 High Indegree Pages and Hosts

The three web pages with highest indegree are the starting pages of YouTube, WordPress and Google. Other six pages from YouTube from the privacy, press and copyright sections of this website appear within the top 10 of pages ranked by their indegree. This is an artefact of the large number of pages crawled from YouTube.[11]

The list of *hosts* with the highest indegree (in the host graph) is more interesting: in Table 2 we show the top 20 hosts by indegree, PageRank [16] and harmonic centrality [9]. While most of the sites are the same, some noise appears because some sites are highly linked for technical or political reasons. In particular, the site `miibeian.gov.cn` must be linked by every Chinese site, hence the very high ranking. PageRank is as usual very correlated to degree, and cannot avoid ranking highly this site, whereas harmonic centrality understands its minor importance and ranks it at position 6146.

3.3 Components

Following the steps of Broder *et al.*, we now analyse the weakly connected components (WCC) of our web graph.

Weakly connected components are difficult to interpret—in theory, unless one has two seed URLs reaching completely disjoint regions of the web (unlikely), one should always find a single weakly connected component. The only other sources of disconnection are crawling and/or parsing artefacts.

Figure 4 shows the distribution of the sizes of the weakly connected components using a visualization similar to the previous figures. The largest component (rightmost grey point) contains about around 94% of the whole graph, and it is slightly larger than the

[11]The highest ranked pages are listed at `http://webdatacommons.org/hyperlinkgraph/top_degree_pages.html`.

PageRank	Indegree	Harmonic Centrality
gmpg.org	wordpress.org	youtube.com
wordpress.org	youtube.com	en.wikipedia.org
youtube.com	gmpg.org	twitter.com
livejournal.com	en.wikipedia.org	google.com
tumblr.com	tumblr.com	wordpress.org
en.wikipedia.org	twitter.com	flickr.com
twitter.com	google.com	facebook.com
networkadvertising.org	flickr.com	**apple.com**
promodj.com	**rtalabel.org**	vimeo.com
skriptmail.de	wordpress.com	creativecommons.org
parallels.com	**mp3shake.com**	**amazon.com**
tistory.com	w3schools.com	**adobe.com**
google.com	domains.lycos.com	**myspace.com**
miibeian.gov.cn	**staff.tumblr.com**	**w3.org**
phpbb.com	**club.tripod.com**	**bbc.co.uk**
blog.fc2.com	creativecommons.org	**nytimes.com**
tw.yahoo.com	vimeo.com	**yahoo.com**
w3schools.com	miibeian.gov.cn	**microsoft.com**
wordpress.com	facebook.com	**guardian.co.uk**
domains.lycos.com	phpbb.com	**imdb.com**

Table 2: The 20 top web hosts by PageRank, indegree and harmonic centrality (boldfaced entries are unique to the list they belong to)

one reported by Broder *et al.* (91.8%). Again, we show the max-likelihood power-law fit starting at 14 with exponent 2.22, which however excludes the largest component. The p-value is again 0, and the law covers only to 1% of the distribution.

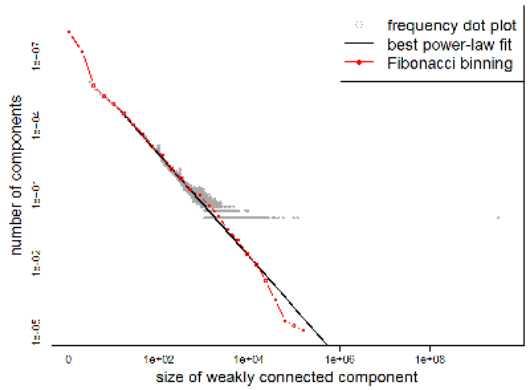

Figure 4: Frequency plot of the distribution of WCCs

More interestingly, we now analyse the strongly connected components (SCC). Computing the strongly connected components of a 3.5 billion node graph was no easy task: it required one terabyte of core memory and, in fact, the computation was only possible because WebGraph [6] uses *lazy* techniques to generate successor lists (i.e., successors lists are never actually stored in memory in uncompressed form).

Figure 5 shows the distribution of the sizes of the strongly connected components. The largest component (rightmost grey point) contains 51.3% of the nodes. Again, we show a fitted power law starting at 22 with exponent 2.20, which however excludes the largest component, and fits only to 8.9% of the distribution. The p-value is again 0.

In Figure 6 we show the size-rank plots of both distributions, which confirm again that the apparent fitting in the previous figures is an artefact of the frequency plots (the rightmost grey points are again the giant components).

3.4 The Bow Tie

Having identified the giant strongly connected component, we can determine the so-called *bow tie*, a depiction of the structure of the web suggested by Broder *et al.*. The bow tie is made of six different components:

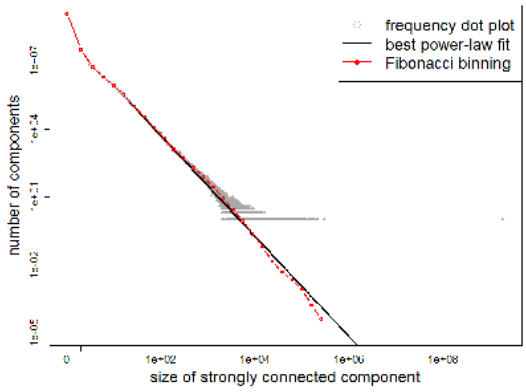

Figure 5: Frequency plot of the distribution of SCCs

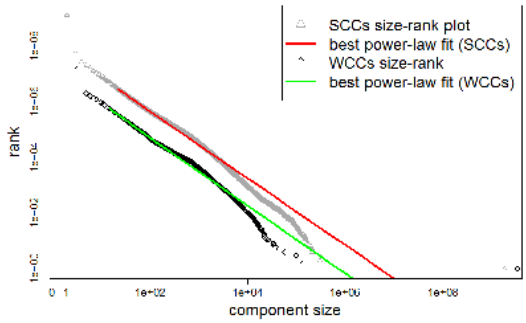

Figure 6: Size-rank plot of the distribution of components

- the core is given by the giant strongly connected component (LSCC);
- the IN component contains non-core pages that can reach the core via a directed path;
- the OUT component contains non-core pages that can be reached from the core;
- the TUBES are formed by non-core pages reachable from IN and that can reach OUT;
- pages reachable from IN, or that can reach OUT, but are not listed above, are called TENDRILS;
- the remaining pages are DISCONNECTED.

All these components are easily computed by visiting the *direct acyclic graph of strongly connected components* (SCC DAG): it is a graph having one node for each strongly connected component with an arc from x to y if some node in the component associated with x is connected with a node in the component associated with y. Such a graph can be easily generated using WebGraph's facilities. Figure 7 shows the size of bow-tie component.

Table 3 compares the sizes of the different components of the bow-tie structure between the web graph discussed in this paper (column two and three) and the web graph analysed by Broder *et al.* in 2000 (column four and five).[12]

[12]Broder *et al.* did not report the number of nodes belonging to the TUBE component separately, as they define as TUBE as a TENDRIL from the IN component hooked into the TENDRIL of a node from the OUT component.

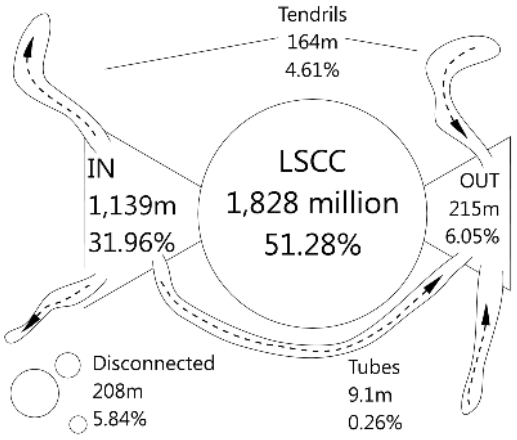

Figure 7: Bow-tie structure of the web graph

The main constant is the existence of a LSCC, which in our graph has almost doubled in relative size. We also witness a much smaller OUT component and a larger IN component. The different proportions are most likely to be attributed to different crawling strategies (in particular, to our large number of nodes with indegree zero, which cannot belong to the LSCC or OUT component). Unfortunately, basic data such as the seed size, the type of visit strategy, etc. are not available for the Broder *et al.* crawl. Certainly, however, the web has become significantly more dense and connected in the last 13 years.

	Common Crawl 2012		Broder *et al.*	
Component	# nodes (in thousands)	% nodes (in %)	# nodes (in thousands)	% nodes (in %)
LSCC	1 827 543	51.28	56 464	27.74
IN	1 138 869	31.96	43 343	21.29
OUT	215 409	6.05	43 166	21.21
TENDRILS	164 465	4.61	43 798	21.52
TUBES	9 099	0.26	-	-
DISC.	208 217	5.84	16 778	8.24

Table 3: Comparison of sizes of bow-tie components

3.5 Diameter and Distances

In this paper we report, for the first time, accurate measurements of distance-related features of a large web crawl. Previous work has tentatively used a small number of breadth-visit samples, but convergence guarantees are extremely weak (in fact, almost non-existent) for graphs that are not strongly connected. The data we report have been computed using HyperBall [8], a diffusion-based algorithm that computes an approximation of the distance distribution (technically, we computed four runs with relative standard deviation 9.25%). We report, for each datum, the empirical standard error computed by the jackknife resampling method.

In our web graph, $48.15 \pm 2.14\%$ of the pairs of pages have a connecting directed path. Moreover, the average distance is 12.84 ± 0.09 and the *harmonic diameter* (the harmonic mean of all distances, see [15] and [7] for motivation) is 24.43 ± 0.97. These figures should be compared with the 25% of connected pairs and the average distance 16.12 reported by Broder *et al.* (which however have been computed averaging the result of few hundred breadth-first samples): even if our crawl is more than 15 times larger, it is significantly more connected, in contrast to commonly accepted predictions of logarithmic growth of the diameter in terms of the

number of nodes. This is a quite general phenomenon: the average distance between Facebook users, for instance, has been steadily going down as the network became *larger* [2].

We can also estimate that the graph has a diameter of at least 5 282 (the maximum number of iteration of a HyperBall run). Figure 8 shows the distance distribution, sharply concentrated around the average.

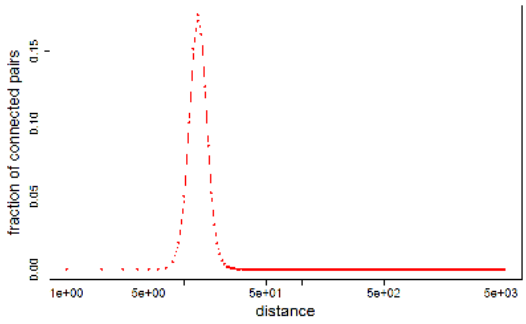

Figure 8: Distance distribution

4. CONCLUSION

We have reported a number of graph measurements on the largest web graph that is available to the public outside companies such as Google, Yahoo, Yandex, and Microsoft. Comparing our results with previous measurements performed in the last 13 years, and with previous literature on significantly smaller crawls, we reach the following conclusions:

- The average degree has significantly increased, almost by a factor of 5.
- At the same time, the connectivity of the graph (the percentage of connected pairs) has increased (almost twice) and the average distance between pages has decreased, in spite of a predicted growth that should have been logarithmic in the number of pages.
- While we can confirm the existence of a large strongly connected component of growing size, witnessing again the increase in connectivity, the structure of the rest of the web appears to be very dependent on the specific web crawl. While it is always possible to compute the components of the bow tie of Broder *et al.*, the proportion of the components is not intrinsic.
- The distribution of indegrees and outdegrees is extremely different. Previous work on a smaller scale did not detect or underplayed this fact, in part because of the little size of the concave (on a log-log plot) part of the distribution in smaller crawls. In our dataset, the two distributions have very little in common.
- The frequency plots of degree and component-size distributions are visually identical to previous work. However, using proper statistical tools, neither degree nor component-size distributions fit a power law. Moreover, visual inspection of the size-rank plots suggests that their tails are not fat (i.e., they decrease faster than a polynomial), in contrast with assumptions taken for granted in the current literature. Our data, nonetheless, leaves open the possibility of a heavy tail (e.g., lognormal).

5. ACKNOWLEDGEMENTS

The extraction of the web graph from the Common Crawl was supported by the FP7-ICT project PlanetData (GA 257641) and by an Amazon Web Services in Education Grant award. Sebastiano Vigna has been supported by the EU-FET grant NADINE (GA 288956), which provided part of the high-end hardware on which the analysis was performed.

6. REFERENCES

[1] D. Achlioptas, A. Clauset, D. Kempe, and C. Moore. On the bias of traceroute sampling: Or, power-law degree distributions in regular graphs. *Journal ACM*, 56(4):21:1–21:28, 2009.

[2] L. Backstrom, P. Boldi, M. Rosa, J. Ugander, and S. Vigna. Four degrees of separation. In *ACM Web Science 2012: Conference Proceedings*, pages 45–54. ACM Press, 2012.

[3] R. Baeza-Yates and B. Poblete. Evolution of the Chilean web structure composition. In *Proc. of Latin American Web Conference 2003*, pages 11–13, 2003.

[4] C. Bizer, K. Eckert, R. Meusel, H. Mühleisen, M. Schuhmacher, and J. Völker. Deployment of RDFa, microdata, and microformats on the web - a quantitative analysis. In *Proc. of the In-Use Track International Semantic Web Conference 2013*, Oct 2013.

[5] P. Boldi, B. Codenotti, M. Santini, and S. Vigna. Structural properties of the African web. In *Proc. WWW'02*, 2002.

[6] P. Boldi and S. Vigna. The WebGraph framework I: Compression techniques. In *Proc. WWW'04*, pages 595 – 602. ACM, 2004.

[7] P. Boldi and S. Vigna. Four degrees of separation, really. In *ASONAM 2012*, pages 1222–1227. IEEE Computer Society, 2012.

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

[9] P. Boldi and S. Vigna. Axioms for centrality. *Internet Math.*, 2014. To appear.

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

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

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

[13] W. Hall and T. Tiropanis. Web evolution and web science. *Computer Networks*, 56(18):3859 – 3865, 2012.

[14] L. Li, D. L. Alderson, J. Doyle, and W. Willinger. Towards a theory of scale-free graphs: Definition, properties, and implications. *Internet Math.*, 2(4), 2005.

[15] M. Marchiori and V. Latora. Harmony in the small-world. *Physica A: Statistical Mechanics and its Applications*, 285(3-4):539 – 546, 2000.

[16] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the web. Technical Report SIDL-WP-1999-0120, Stanford Digital Library Technologies Project, Stanford University, 1998.

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

[18] S. Spiegler. Statistics of the Common Crawl Corpus 2012. Technical report, SwiftKey, June 2013.

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

[20] W. Willinger, D. Alderson, and J. C. Doyle. Mathematics and the Internet: A source of enormous confusion and great potential. *Notices of the AMS*, 56(5):586–599, 2009.

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