The Journal of Web Science, 2015, 1: 33–47

The Graph Structure in the Web – Analyzed on Different Aggregation Levels

Robert Meusel[1], Sebastiano Vigna[2], Oliver Lehmberg[1] and Christian Bizer[1]

[1]*Data and Web Science Group, University of Mannheim, Germany*
{robert,oliver,chris}@informatik.uni-mannheim.de

[2]*Laboratory for Web Algorithmics, Università degli Studi di Milano, Italy*
vigna@acm.org

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 analyze a large web graph. The graph was extracted from a large publicly accessible web crawl that was gathered by the Common Crawl Foundation in 2012. The graph covers over 3.5 billion web pages and 128.7 billion hyperlinks. We analyze and compare, among other features, degree distributions, connectivity, average distances, and the structure of weakly/strongly connected components. We conduct our analysis on three different levels of aggregation: page, host, and pay-level domain (PLD) (one "dot level" above public suffixes).

Our analysis shows that, as evidenced by previous research (Serrano *et al.*, 2007), some of the features previously observed by Broder *et al.*, 2000 are very dependent on artifacts 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 (Donato *et al.*, 2005; Boldi *et al.*, 2002; Baeza-Yates and Poblete, 2003), 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 by Broder *et al.* 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 of the page and host graph are not power laws, contrarily to what was previously reported for much smaller crawls, although they might be heavy tailed. If we aggregate at pay-level domain, however, a power law emerges. 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 (Boldi and Vigna, 2013).

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

ISSN 2332-4031; DOI 10.1561/106.00000003

1 Introduction

The evolution of the World Wide Web (WWW) 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" (Hall and Tiropanis, 2012). With the evolution of the 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 gather evidence for the social phenomena governing the growth of the Web (Hall and Tiropanis, 2012). 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. In addition, 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 this paper we present a study of the structure of the Web using one of the largest—or the largest—publicly accessible web graph dataset. The analyzed graph was extracted from a web crawl, which contains pages gathered in the first half of 2012. The graph covers 3.5 billion crawled pages and 128 billion unique links between the crawled pages. We will briefly discuss the basic statistics about the graph as degree distributions, connectivity of pages and distance distribution. In addition we calculate the bow-tie structure of the graph, and draw comparisons with the latest comparable analysis of this structure presented by Broder *et al.*, 2000.

To get a deeper understanding of the structure of the graph, we also analyze it on two coarser aggregation levels: host and pay-level domain. A definition of those two aggregations levels is given in the next section.

The article is structured as follows: After introducing the necessary terms and definitions we give a brief overview of publications analyzing different aspects of web graphs as well as available datasets. In Section 4 the dataset as well as the methodology which was used to extract the data from a public web crawl is described. In addition this section gives an overview of the used methods to analyze the graph. The following three sections (Section 5, 6, and 7) report our analysis at the three different aggregation levels. In Section 8 we discuss in more detail the outcome of the analysis of the different levels of aggregation. Section 9 summarizes our findings and presents open challenges.

The article presents findings for the page and the PLD graph that

we already reported in two papers (Meusel *et al.*, 2014; Lehmberg *et al.*, 2014). In addition, we complete the findings with an analysis of the host graph. We then discuss the similarities and differences of the findings of the different aggregation levels.

2 Definitions

In this article we make use of different terms which might have slight different meanings depending on the background of the reader. To avoid confusions, we state brief definitions in the following.

Page: A page is defined as a single page within the web crawl, uniquely identified by its URL (Berners-Lee *et al.*, 1994), for example http://graph.webdatacommons.org/index.html.

Host: The host of a URL is defined in RFC 1738 (Berners-Lee *et al.*, 1994): informally, everything after the protocol double slash and the following slash, but excluding port and authentication information. The host of the previous URL is graph.webdatacommons.org.

Pay-Level Domain: The pay-level domain of a URL is determined from its host using the *Public Suffix List* published by the Mozilla Foundation.[1] The PLD of a host is defined as one dot level above that the public suffix of the host: for example, a.com for b.a.com (as .com is on the public suffix list) and c.co.uk for a.b.c.co.uk (as .co.uk is on the public suffix list). The PLD of our example is webdatacommons.org.

The different aggregation levels are visualized in Figure 1. Figure 1a shows an example of a page-level graph with hosts and pay-level domains indicated by dashed shapes. The host-level and pay-level aggregations, where hyperlinks between pages are merged per host respectively pay-level domains are shown in Figure 1b and 1c.

3 Related Work

In this section we discuss related work analyzing the structure of the Web. We first focus on publications analyzing web graphs on page level. We then state additional research using another aggregation level of the graph for their analysis. The section closes with an overview of publicly accessible web graph datasets and discusses their characteristics.

3.1 Page Graph

In spite of the importance of knowledge about the structure of the Web, 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.*, 2000 using an AltaVista crawl of 200 million pages and 1.5 billion links.[2] 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 (LSCC) containing 28% of the nodes. Nodes and paths leading to the LSCC are assigned to the IN component of the bow tie. Nodes and paths leading away from the LSCC are assigned to the OUT component.[3]

In addition, Broder *et al.* show that the indegree distribution, the outdegree distribution and the distribution of the sizes 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.*, 2007 analyzing 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 (Donato *et al.*, 2005; Boldi *et al.*, 2002; Baeza-Yates and Poblete, 2003; Zhu *et al.*, 2008) using different (possibly regional) crawls gathered by different crawlers provided quite different pictures of the web graph (e.g., that "daisy" of Donato *et al.*, 2005 or the "teapot" of Zhu *et al.*, 2008).

In particular, recent strong and surprising results by Achlioptas *et al.*, 2009 have shown that, in principle, most heavy-tailed (and even power-law) distributions observed in web crawls may be just an artifact 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 found previously, and heavy-tailed (often, power-law) strongly connected component 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.

3.2 Aggregated Graphs

Hirate *et al.*, 2008 analyze at the host level the components of a large (3.2 billion pages) crawl gathered between 2004 and 2005. 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.

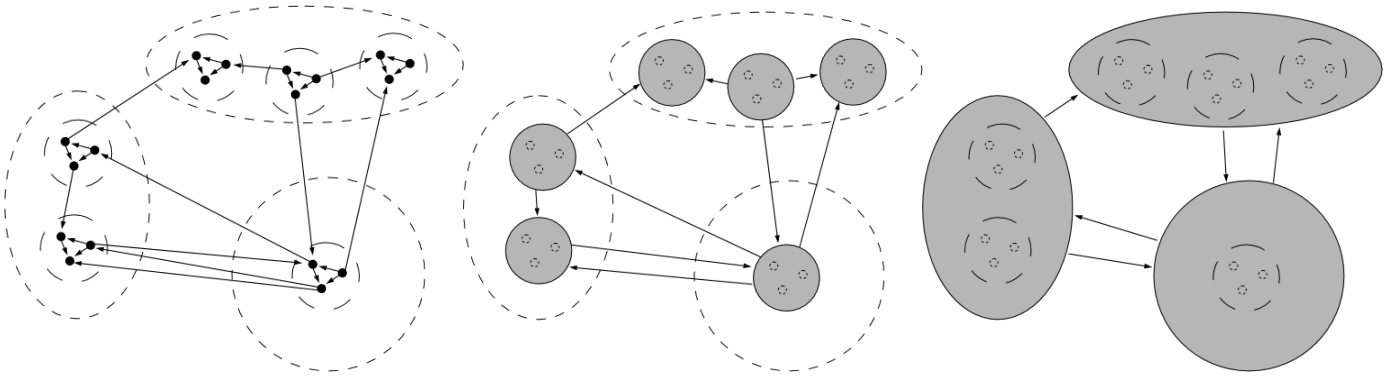
Zhu *et al.*, 2008 analyze the structure of the Chinese Web. They compared their results on three different aggregation levels: the page level, the host level and the domain level. On the page level, they found a large IN component, which disappears on the host and the domain level. Analogously, the LSCC and OUT components of the host and domain level are larger than on the page level.

Dill *et al.*, 2002 compared several induced subgraphs identified by common technical features (location, content, etc.) as well as the host graph. The alleged indegree power law exponent for their graph is 2.34, albeit no indication is reported on how to the value was computed, and on its statistical reliability. 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.

[1] http://publicsuffix.org/list/

[2] Throughout the paper, we avoid redundant use of the ≈ symbol: all reported figures are rounded.

[3] The components of the bow tie are described in more detail in Section 4.4.3. Figure 9 shows an exemplary visualization of a bow tie.



(a) page graph (b) host graph (c) PLD graph

Figure 1: Different aggregation levels of the graph

3.3 Web Graph Datasets

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 than the crawls which has been analyzed so far. Nonetheless, billion-scale representative crawls have not been publicly 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 web crawls.

A few exceptions exist, but they have significant problems. The AltaVista webpage connectivity dataset, distributed by Yahoo! as part of the WebScope program,[4] 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[5], 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[6], 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,[7] and it is focused mostly on English web pages.

[4] http://webscope.sandbox.yahoo.com/catalog.php?datatype=g

[5] http://www.lemurproject.org/clueweb09.php/

[6] http://www.lemurproject.org/clueweb12.php/

[7] 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 numbers reported within this paper refer to the actually crawled pages.

4 Datasets and Methodology

This section first describes the web crawl from which the analyzed graph was extracted. Then we explain the extraction methodology that was used to generate the graph from the crawl and state some basic statistics about the graph at different aggregation levels. The section is completed by an overview of the methodology that will be used for the analysis of the graph in the following.

4.1 Common Crawl Datasets

The web crawl which was used to extract the hyperlink graph for our analysis is provided by the *Common Crawl Foundation*.[8] The foundation was founded by Gil Elbaz and has the mission to gather and maintain web crawls which are publicly accessible and can be used by everyone. So far they released over 15 web crawl corpora and until mid of 2014 they release a new crawl almost each month. The crawl corpus, which was used for our analysis was released in August 2012 and was gathered in the first half of 2012 and is so far the largest crawl published by the Common Crawl Foundation. The crawl contains 3.83 billion web documents, of which over 3.53 billion (92%) are of mime-type `text/html`. This dataset was gathered using a web crawler which employed a breadth-first-search crawling strategy, together with heuristics to detect spam pages. 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 publicly 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 chosen amount of seeds in combination with the crawling strategy 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 (Spiegler, 2013). In addition, it is likely that the large number of seeds used caused the large number of pages with indegree zero (20% of the graph) found in the graph.

Figure 2 shows in log-log scale the distribution of hosts and pay-level domains with respect to the number of crawled pages of each

[8] http://commoncrawl.org

host/PLD. The x-axis shows the number of crawled pages in the original corpus whereas the y-axis shows the number of hosts/PLDs including a certain number of crawled pages. The shape of the two distributions appears to be heavy tailed, with a large number of hosts/PLDs having a small number of crawled pages and a small number of hosts/-PLDs with a large number of crawled pages.

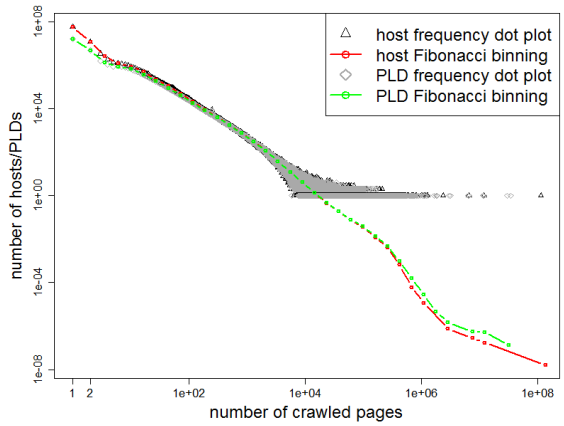


Figure 2: Frequency plot of the pages per host and PLD distribution

Table 1 reports the five hosts with the largest number of crawled pages in the dataset (i.e., the five rightmost in Figure 2). It is remarkable that within those five hosts the number of crawled pages differs by two orders of magnitude.

Table 1: Top 5 hosts by number of crawled pages

host	# crawled pages
youtube.com	113 453 983
amazon.com	11 933 190
flickr.com	6 512 766
en.wikipedia.org	2 353 610
amazon.co.jp	1 275 624

4.2 Extraction Methodology

Associated with the web 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 (Bizer *et al.*, 2013). [9] We first collected for each crawled page its URL, mime-type, links to other pages, link 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.[10] 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, index all URLs and remove duplicate links.

4.3 Graph Datasets

Applying the extraction methodology as described above we extracted the hyperlink graph, where each node represents a page, and each arc represent a link from a page to another. Beside this graph on page level, we aggregated two further graph datasets, namely on host and PLD level. Nodes in such graphs represent sets of pages with the 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 2 provides basic statistics about the size of the extracted and calculated graphs. All graphs are available for download from the WebDataCommons website.[11]

Table 2: Sizes of the graph for different granularities

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

4.3.1 Relevance and Coverage

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[12] about the second quarter of the same year, they state a total of 240 million registered domain names.[13] With our graph covering 43 million domains, this means we have (at least partial)[14] 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.

4.4 Analysis Methodology

In the following sections we will analyze the graph using the three different aggregation levels: page, host and pay-level domain. For all graphs we analyze the degree distribution and connected components. Following the findings of Broder *et al.*, 2000 we calculate the

[9] A detailed description of the used extraction framework can be found at http://webdatacommons.org/framework

[10] 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 organizations is, unfortunately, quite difficult, as link and page selection strategies could differ.

[11] http://webdatacommons.org/hyperlinkgraph

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

[13] The report is talking about registered domains, which is the intended meaning of a PLD. As you pay for a pay-level domain, this is a domain that can be registered.

[14] 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.

bow-tie structure of each graph. For all three graphs we conclude with the calculation of the diameter and distances. Most of those analyzes have been performed using the "big" version of the WebGraph framework described by Boldi and Vigna, 2004, 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[15] (under a suitable permutation of the node identifiers it is common to obtain slightly more than one bit per link).

The whole graph on page level occupied in compressed form just 57.5 GB, which made it possible to run resource-intensive computations such as that of strongly connected components.

4.4.1 Distributions

In our analysis we try to fit a power-law statistical distribution function for various distributions, like degree and components. Most of the 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 (Willinger *et al.*, 2009). 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.

We thus follow the methodological suggestions of Clauset *et al.*, 2009. We use the `plfit`[16] tool to attempt a maximum-likelihood fitting of a power-law tail starting from each possible value, 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. A p-value greater than 0.1 is then considered a reasonable statistical evidence for a power law. We report in all cases the starting point and exponent provided by maximum-likelihood fitting, even if the p-value does not give sufficient statistical evidence.

Another methodology which can be applied to determine the best possible fit for a given distribution is proposed by Malevergne *et al.*, 2009 and Malevergne *et al.*, 2005. In comparison to estimate a p-value for one single computed best fit, they suggest the comparison of two best fitted functions (e.g. power-law and lognormal) which gives an evidence of the closest fit. Alstott J, 2014 implement this methodology in their work, which is unfortunately inapplicable for distributions with over 3.5 billion data points.[17]

In addition, we aggregate the data points using *Fibonacci binning* (Vigna, 2013), to show the approximate shape of the distribution. Fibonacci binning is analogous to common data-visualization practices such as base-2 exponential binning (sometimes called *logarithmic* binning), but it uses approximately the golden ratio as a base. More precisely, the bounds of the bins are defined by Fibonacci numbers, which are approximately spaced as the golden ratio, and a data point is plotted in the center of the bin using the average of the values in the bin.

Finally, we display data using *size-rank* plots, as suggested by Li *et al.*, 2005 to find visual evidence of power laws. 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 from the "cloud of points" problem.

4.4.2 Components

We calculate for each graph the weakly and strongly connected components.

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 artifacts.

The strongly connected components (SCC) are easier to interpret. We use WebGraph (Boldi and Vigna, 2004) which uses *lazy* techniques to generate successor lists (i.e., successors lists are never actually stored in memory in uncompressed form). This technique made it even possible to compute the strongly connected components of the 3.5 billion node graph (page level), which needed one terabyte of main memory.

4.4.3 Bow-Tie Structure

From the giant strongly connected component, one 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:

- 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 *directed 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.

4.4.4 Distances and Diameters

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

[15] The information-theoretical lower bound is $\log \binom{n^2}{m}$ for a graph with n nodes and m arcs. While most graphs needs approximately this number of bits to be represented, compression uses the statistical skewness of a specific subclass (i.e., web graphs) to represent a graph using much less bits—in our case, about one eighth.

[16] https://github.com/ntamas/plfit

[17] We initiated the comparison of best power-law fit and best lognormal fit on a one terabyte memory machine with 40 cores but the calculations did not terminate within seven days.

graphs that are not strongly connected. The data we report has been computed using HyperBall (Boldi and Vigna, 2013), a diffusion-based algorithm that computes an approximation of the distance distribution. We report, for each datum, the empirical standard error computed by the jackknife resampling method.

5 Analysis of the Page Graph

In this section we report our findings for the page graph and compare them (if possible) to the findings of Broder *et al.*, 2000, who did a comprehensive analysis of the structure of the Web in 2000 using a comparable web graph dataset.

5.1 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.[18]

Broder *et al.* report an average degree of 7.5 links per page. Similarly low values can be found in crawls from the same years – for instance, in the crawls made by the Stanford WebBase project.[19] 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 which has an average degree of 45.1.[20] 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 3a and 4a 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 that have degree d. Note that *we included the data for degree zero*, which is omitted in most of the literature.

We then try to fit a power law to a tail of the data. We always indicate the best fit for the distributions by the black line, where the most left points represents the minimal node from which the best fit was found. 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 in Figure 7a 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.

In all cases, 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 behavior, thus, explain very little of the process of link formation in the Web.

The values we report are slightly different from those of Broder *et al.*, who found 2.09 (2.72, respectively) as power-law exponent for the indegree (outdegree, respectively). But in fact they are incomparable, as our fitting process likely 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 7a.

5.2 High Indegree Pages

The three web pages with the highest indegree are the root pages of YouTube, WordPress and Google.[21] 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 artifact of the large number of pages crawled from YouTube.[22]

5.3 Components

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

Figure 5a shows the distribution of the sizes of the weakly connected components of the 2012 crawl using a visualization similar to the previous figures. The largest component (rightmost gray point) contains about around 94% of the whole graph, and it is slightly larger than the 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 0 ± 0.01, and the law covers only to 1% of the distribution.

Figure 6a shows the distribution of the sizes of the strongly connected components. The largest component (rightmost gray 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 ± 0.01.

In Figure 8a we show the size-rank plots of both distributions, which confirm again that the apparent fitting in the previous figures is an artifact of the frequency plots (the rightmost gray points are again the giant components).

5.4 The Bow Tie

From the strongly connected components it is easy to compute the size of the parts of the bow tie. The bow tie structure of the page graph is shown in Figure 9.

Table 3 compares the sizes of the different components of the bow-tie structure between the web graph discussed in this paper (columns

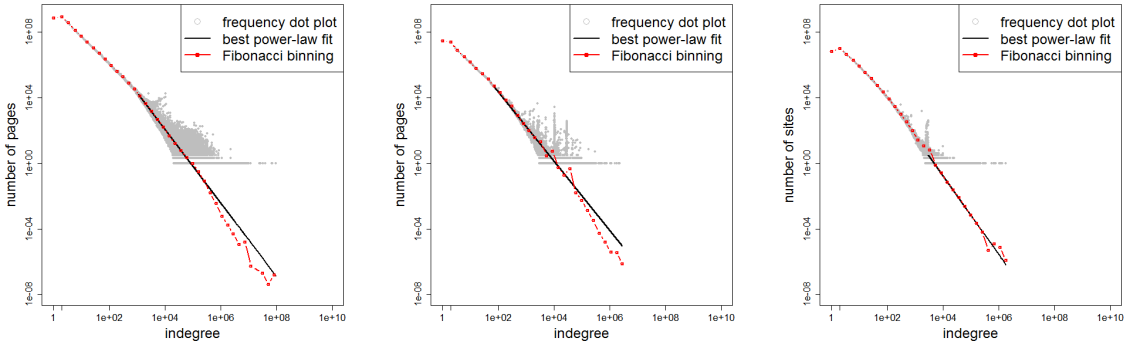
[18] Technically speaking, the *density* of a graph is the ratio between the number of arcs and the *square* of the number of nodes, but for very sparse graphs one obtains abysmally small numbers that are difficult to interpret.

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

[20] 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.

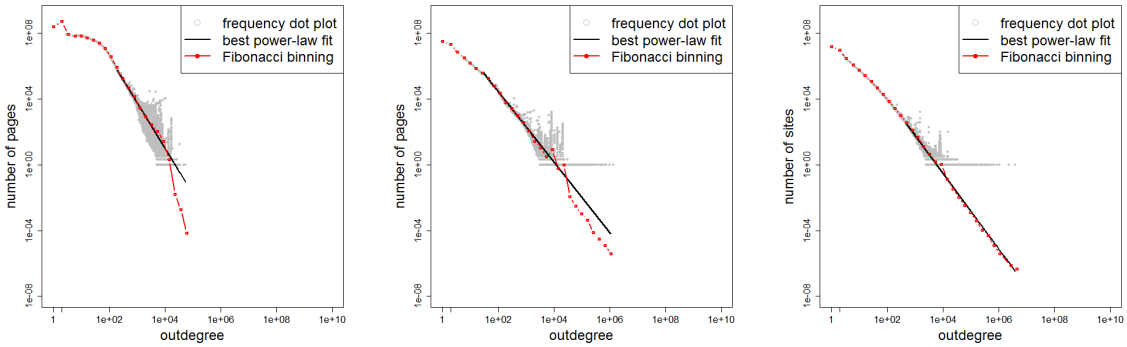
[21] The root page of a website refers to the page which is directed to by the pure website URL.

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



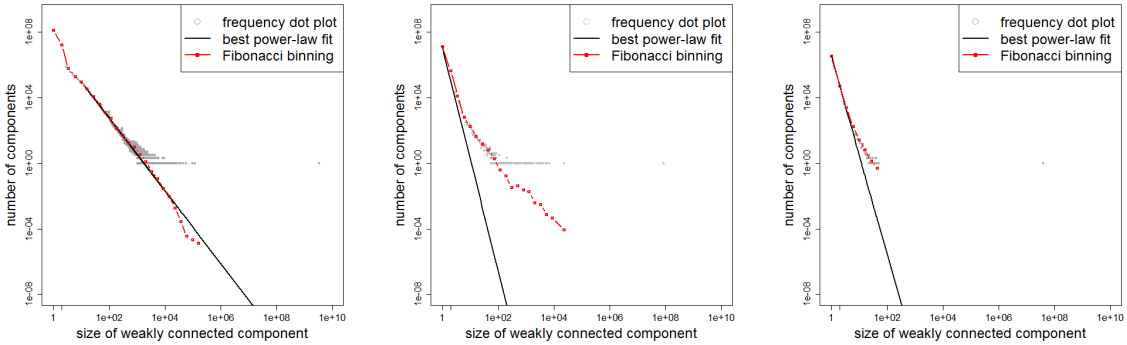
(a) page graph (b) host graph (c) PLD graph

Figure 3: Indegree distributions



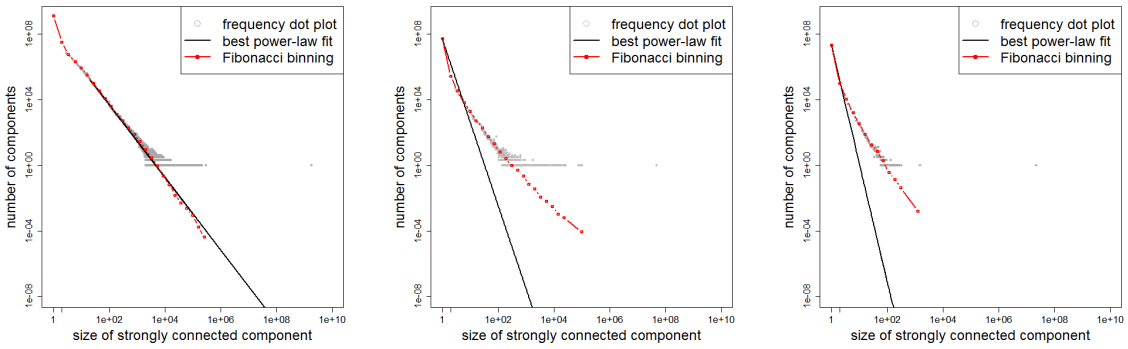
(a) page graph (b) host graph (c) PLD graph

Figure 4: Outdegree distributions



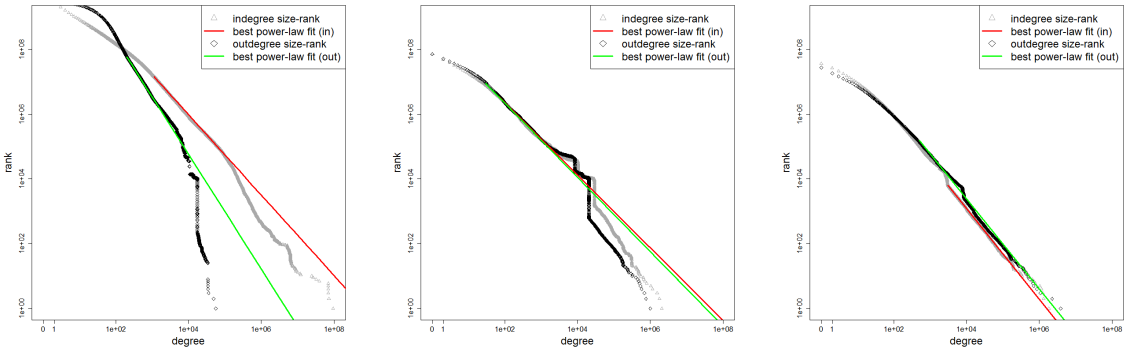
(a) page graph (b) host graph (c) PLD graph

Figure 5: WCC distributions



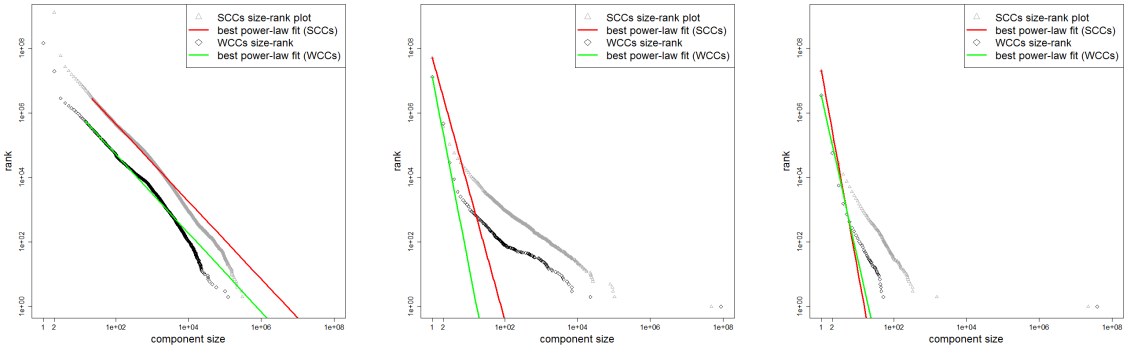
(a) page graph (b) host graph (c) PLD graph

Figure 6: SCC distributions



(a) page graph (b) host graph (c) PLD graph

Figure 7: Size-rank plot of degree distributions



(a) page graph (b) host graph (c) PLD graph

Figure 8: Size-rank plot of components

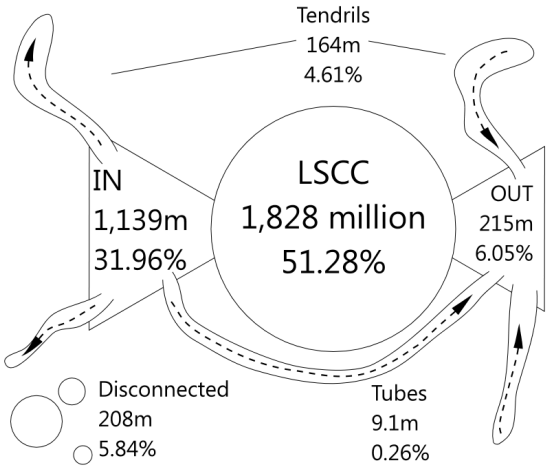


Figure 9: Bow-tie structure of the page graph

two and three) and the web graph analyzed by Broder *et al.* in 2000 (columns four and five).[23]

The main constant is the existence of an 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.

Table 3: Comparison of sizes of bow-tie components of the page graph

	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

5.5 Diameter and Distances

In the page graph, $48.15 \pm 2.14\%$ of the pairs of pages have a connecting directed path. Moreover, the average distance between connected pairs is 12.84 ± 0.09 and the *harmonic diameter* (the harmonic mean of all distances, see Marchiori and Latora, 2000 and Boldi and Vigna, 2012 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* (Backstrom *et al.*, 2012).

We can also estimate that the graph has a diameter of at least 5 282 (the maximum number of iterations of a HyperBall run). Figure 10 shows the distance distribution, sharply concentrated around the average.

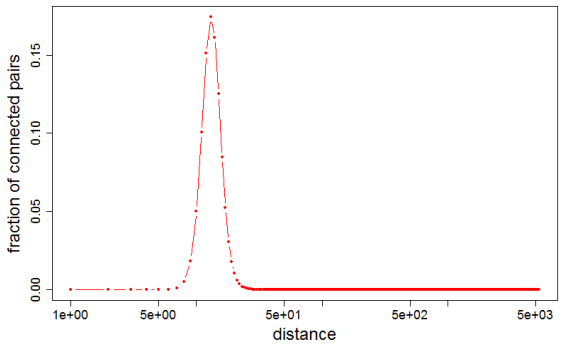


Figure 10: Distance distribution within the page graph

6 Analysis of the Host Graph

In this section we analyze the graph aggregated on host level, following the same steps of the analysis of the page graph.

6.1 Indegree & Outdegree Distribution

The average degree is 20.2, which is 0.55 times the average degree of the page graph. Figures 3b and 4b show frequency plots of indegrees and outdegrees in log-log scale. Our fitting procedure results in an exponent of 2.12 starting at 69 for indegrees and an exponent of 2.14 starting at 29 for outdegrees. As for the page graph, the p-value of the best fits is 0 (± 0.01) for both distributions, and the same comments apply. However, compared to the page graph, the x_{min} values are much smaller, indicating that these distributions have in principle more explanatory value on the host level than on the page level, as they cover a larger fraction of the data points. Further, the exponents for the indegrees of both graphs are quite similar (difference of 0.12) while those for the outdegrees vary by 0.63.

Figure 7b shows the size-rank plot of the degree distributions of our host graph and the best power-law fit. The sharp drop at degree 10 000 of both curves results from the number of high spikes that can be observed for the degree distributions (see Figure 3b and 4b). We will discuss this phenomenon in Section 7.1.

6.2 Top Ranked Hosts

In Table 4 we show the top 20 hosts by indegree, PageRank (Page *et al.*, 1998) (setting the damping factor α to 0.85) and harmonic centrality (Boldi and Vigna, 2014).[24] While most of the sites are the

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

[24] For the computation of PageRank and harmonic centrality we again used HyperBall (Boldi and Vigna, 2013) from the WebGraph library and the PageRank

Table 4: The 20 top web hosts by PageRank, indegree and harmonic centrality (boldfaced entries are unique to the list they belong to)

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
twitter.com	tumblr.com	wordpress.org
en.wikipedia.org	twitter.com	flickr.com
tumblr.com	google.com	facebook.com
promodj.com	flickr.com	**apple.com**
google.com	**rtalabel.org**	vimeo.com
networkadvertising.org	**wordpress.com**	creativecommons.org
phpbb.com	**mp3shake.com**	**amazon.com**
ytmnd.com	**w3schools.com**	**adobe.com**
miibeian.gov.cn	**domains.lycos.com**	**myspace.com**
flickr.com	**staff.tumblr.com**	**w3.org**
blog.fc2.com	**club.tripod.com**	**bbc.co.uk**
tw.yahoo.com	creativecommons.org	**nytimes.com**
facebook.com	vimeo.com	**yahoo.com**
addthis.com	miibeian.gov.cn	**microsoft.com**
parallels.com	facebook.com	**guardian.co.uk**
creativecommons.org	phpbb.com	**imdb.com**

same, some noise appears because some sites are highly linked for technical or political reasons (for instance, `gmpg.org` is the reference for a vocabulary that describes relationships). 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 6 146.

6.3 Components

Following our scheme of analysis, we now analyze the weakly connected components of our host graph. Figure 5b shows the distribution of their sizes. The largest component (rightmost gray point) contains about around 87% of the whole graph. Again, we show the max-likelihood power-law fit starting at 1 with exponent 6.82. Note that these values are completely different from all our other distributions: in this case, the best fit is with the *very first points* of the distribution; the p-value is again 0 ± 0.01.

Regarding the distribution of strongly connected components, displayed in Figure 6b we find the largest component (rightmost gray point) contains 47% of the nodes. We show a fitted power law also starting at 1 with a comparably high exponent 5.07, which again has a p-value of 0 ± 0.01.

In Figure 8b we show the size-rank plots of both distributions, which confirm again that the apparent fitting in the previous figures is an artifact of the frequency plots.

6.4 The Bow Tie

As we have done it for the page graph, we now extract the bow-tie components for the host graph. Table 6 shows the sizes of the bow-tie components.

Compared to the page graph we find a slightly smaller LSCC and a larger DISCONNECTED component. Also, the IN component shrinks while the OUT component seems to grow. The reason for these changes lies in the aggregation process itself: hosts are of wildly different sizes (see Figure 2), which implies that shrinking them to a single node varies significantly the size of the bow-tie components. Also, pages from different components of the bow tie are merged into single nodes. Finally, because of the breadth-first nature of the visit, a large number of hosts visited late has a very small number of pages, which explains the growth of the OUT component.

6.5 Diameter and Distances

In the host graph, only $34.59 \pm 0.79\%$ of the pairs are connected by a directed path, which is a lower percentage than for the page graph. Figure 11 shows the distance distribution in the PLD graph. Also, the average distance of 5.3 ± 0.001 as well as the harmonic diameter of 14.34 ± 0.32 are much smaller than for the page graph. The lower bound for the diameter is 261.

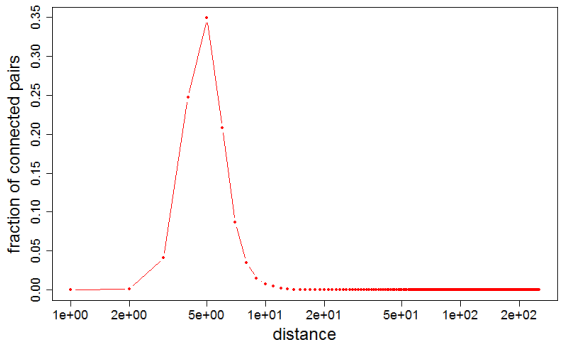


Figure 11: Distance distribution of the host graph

7 Analysis of the PLD Graph

In this section we focus on the analysis of the pay-level-domain graph.

7.1 In- and Outdegree Distributions

Figures 3c and 4c show frequency plots of in- and outdegrees in log-log scale, using the same techniques of the previous sections, whereas Figure 7c shows the corresponding size-rank plot; in all cases, we again display as usual the best power-law tail fitted by maximum likelihood.

For the indegree distribution, the best-fit power 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 ± 0.01, meaning that the tail of the distribution follows a power law. This is indeed the *first* distribution in this paper for which we find significant statistical evidence of a tail fitting a power law. The fitted tail contains however just the 0.0148% of the whole distribution.

The best-fitting power law for the outdegree distribution starts at 496, has an exponent of 2.39 and covers 0.32% of the distribution; he p-value is again 0 ± 0.01. We remark that the largest outdegree value within the page graph is three orders of magnitude smaller than the largest indegree value: however, within the PLD graph the largest outdegree and indegree values are comparable.

Both the indegree and the outdegree distribution (Figure 3c and 4c) show several outliers above the rest of the distribution. In addition, both degree distributions spike, at an indegree of roughly 3 000 and an

Gauss-Seidel parallel implementation from the LAW library (http://law.di.unimi.it/software/). The latter was run until the ℓ_1 norm of the error was smaller than 5×10^{-15}.

outdegree of roughly 8 500, respectively. We find similar outliers for the page graph. Examining a sample of those data points, we found that the corresponding websites can be classified as spam sites or domain reseller sites. This has also been observed by Fetterly *et al.*, 2004 for the degree distributions at the 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.

7.2 Top Ranked PLDs

Similar to Section 6.2, Table 5 shows the top 20 PLDs by indegree, PageRank and harmonic centrality. The main difference to Table 4 is a significant increase in uniformity of the rankings. The elements unique to a specific ranking are 9 instead of 26. The first ten entries are largely the same. Still, we see `miibeian.gov.cn` highly ranked by indegree and PageRank (whereas its rank by harmonic centrality is 3 243)

Table 5: The 20 top PLDs by PageRank, indegree and harmonic centrality (boldfaced entries are unique to the list they belong to)

PageRank	Indegree	Harmonic Centrality
wordpress.org	wordpress.org	youtube.com
gmpg.org	youtube.com	wikipedia.org
youtube.com	wikipedia.org	wordpress.org
twitter.com	gmpg.org	blogspot.com
wikipedia.org	blogspot.com	google.com
blogspot.com	google.com	twitter.com
google.com	wordpress.com	wordpress.com
wordpress.com	twitter.com	yahoo.com
yahoo.com	yahoo.com	gmpg.org
networkadvertising.org	flickr.com	apple.com
apple.com	facebook.com	facebook.com
phpbb.com	apple.com	flickr.com
miibeian.gov.cn	miibeian.gov.cn	**microsoft.com**
hugedomains.com	vimeo.com	w3.org
facebook.com	**tumblr.com**	adobe.com
joomla.org	joomla.org	vimeo.com
flickr.com	**amazon.com**	sourceforge.net
adobe.com	w3.org	**typepad.com**
linkedin.com	nytimes.com	nytimes.com
w3.org	sourceforge.net	**bbc.co.uk**

The PageRank distribution for graphs with power-law indegree distributions has been suggested by Pandurangan *et al.*, 2002 to have approximately the same power-law exponent as the indegree distribution, so it was natural to test this property after our findings.

The Figure 12 shows the PageRank distribution for the PLD graph. We can report a best-fit power law exponent of 2.27 (starting at a rank of 418), which differs by 0.13 from the exponent of the indegree distribution. The p-value is 0.10 ± 0.01, meaning that we find statistical evidence.

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).

7.3 Components

The largest weakly connected component of the PLD graph covers 91.8% of all websites, and the largest strongly connected component contains 51.9% of all PLDs. Figure 5c and 6c show the distributions of all WCCs and SCCs in the PLD graph. In Figure 8c we also show the size-rank plots of both distributions. The WCC and SCC power-law fitting results in a p-value of 0 ± 0.01, where the best fit in both

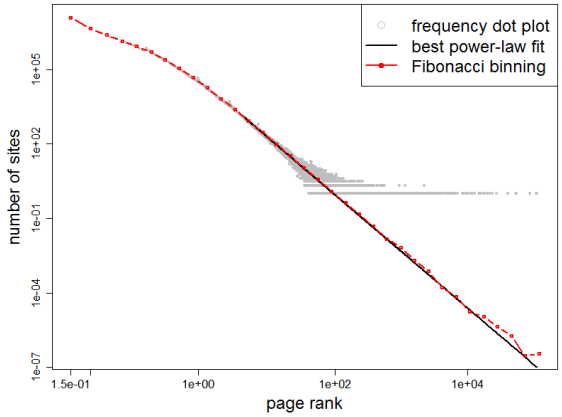


Figure 12: PageRank distribution of the PLD graph

cases start at 1 and we find a exponent of 6.82 for the WCC and 7.17 for the SCC distribution.

7.4 Bow-Tie Structure

As in the two previous sections, we calculate the bow-tie structure within the PLD graph and determine the sizes of the components. Table 6 shows the sizes of the bow-tie components. The LSCC has a similar relative size as in the page graph, and by this is again slightly larger as the LSCC of the host graph. The relative size of disconnected nodes as well as the TUBES and TENDRILS components has decreased again. In addition, we observe a further reduced size of the IN component and an increased size of the OUT component in comparison to the host graph.

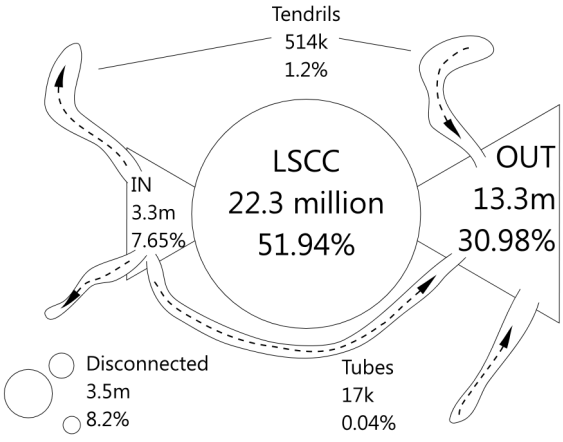


Figure 13: Bow-tie structure of the PLD graph

7.5 Distances and Diameter

In the PLD graph, $42.42 \pm 3.59\%$ of all pairs of nodes are connected by a directed path. Figure 14 shows the distance distribution in the PLD graph. The length of the average shortest path is 4.27 ± 0.085 and the harmonic diameter is 9.55 ± 0.34. This means that a large fractions

of pairs of PLDs which are mutually reachable are actually connected through at most three PLDs. The lower bound on the diameter is 48.

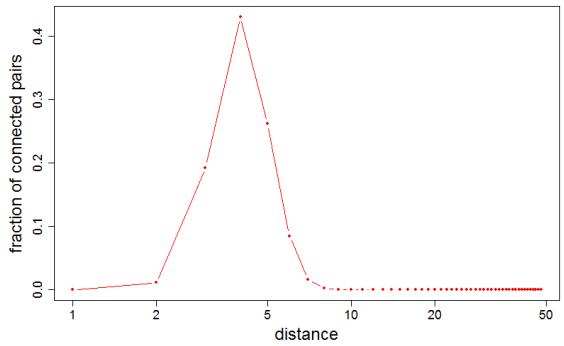


Figure 14: Distance distribution of the PLD graph

8 Discussion

In the previous three sections we have analyzed different aspects of the three graph aggregation levels—page, host and PLD. Table 6 summarizes our findings. From the basic statistics, we can see that the arcs per node ratio decreases with every aggregation level, as the effect of host- and pay-level-domain internal hyperlinks is overcome.

8.1 The Bow-Tie Structure at Different Aggregation Levels

Regarding the sizes of components, and the composition of the graphs based on the different aggregation levels, we partly can confirm the behavior observed by Zhu *et al.*, 2008. In Figure 15 we visualize the different sizes of IN, LSCC, OUT, TUBES/TENDRILS and the DISCONNECTED components at the three different aggregation levels. The rather large IN component in the page graph decreases over the aggregation levels (from 32% down to 8%), where the OUT component grows almost by the same factor (from 6% up to 31%). This behavior was also reported by Zhu *et al.*, 2008 for the Chinese Web. In contrast to their result, the LSCC in our graph at all aggregation levels has almost the same relative size and we cannot confirm a growth of this component when aggregating as reported by Zhu *et al.*. As the web corpus, from which the graph was extracted was gathered using breath-first selection strategy with a limited number of seeds, pages within the IN component most likely belong to the same host/PLD. This leads to the decreasing relative size when aggregating. Pages within the OUT component more likely belong to different hosts/PLDs, which results in an increase of the relative size.

8.2 Fitting Power-Law Distributions

In the past, a large number of works and studies have performed a visual fitting of the tail of a given distribution to a power law, meaning that a line is drawn graphically, and the line traverses the "cloud of points" generated by the high variability in the tail of the distribution. As argued in detail by Willinger *et al.*, 2009, this has led to a number of incorrect classifications. In this paper, instead, we have used the sound statistical methodology proposed by Clauset *et al.*, 2009, and size-rank plots plus Fibonacci binning (Vigna, 2013) to give a visual clue of the actual shape of the tail of the distribution.

Table 6: Overview of Results of the different graph analysis

	Page	Host	PLD
Basic Statistics			
# of Nodes (mil.)	3 563	101	43
# of Arcs (mil.)	128 736	2 043	623
Arcs per Node	36.8	20.2	14.5
Reachability			
Connected pairs	48.15%	34.59%	42.42%
	±2.14	±0.79	±3.59
Avg. distance	12.84	5.30	4.27
	±0.09	±0.001	±0.085
Harmonic diameter	24.43	14.34	9.55
	±0.97	±0.32	±0.34
Diameter(at least)	5 282	261	48
indegree			
γ	2.24	2.12	**2.4**
x_{min}	1 129	69	3 062
outdegree			
γ	2.77	2.14	2.39
x_{min}	199	29	496
WCC			
largest	0.94	0.87	0.92
γ	2.22	6.82	6.04
x_{min}	14	1	1
SCC			
largest	0.51	0.47	0.52
γ	2.20	5.07	7.17
x_{min}	22	1	1
Bow Tie			
IN	0.32	0.17	0.08
OUT	0.06	0.20	0.31
TEND+TUBE	0.05	0.02	0.01
DISC	0.06	0.13	0.08

Values for the power laws in **boldface** are statistically significant

An interesting example showing the need to accurate mathematical methods are the distributions of the weakly and strongly connected components of the host graph. Inspecting them visually, one could say that the calculated power-laws are not the best fit at all, as they only fit to the first point of the distribution (cf. Figure 5b and 6b). We have manually shifted the starting point of the distributions to 4 and 8 and calculated the new exponent for the best power-law fit. Regarding the resulting Figures 16 and 17 the drawn power-law curve seems to fit better. But from a statistical point of view the likelihood of those distributions is not good at all. In other words, trying to fit visually a power law not only might give you the false impression of having found one: it might even let you choose the *less fitting* power law.

In our example one might of course argue that the distribution is made of two power laws, or that the real fitting is the one starting at a later position, as our goal is to fit the tail. But such a judgment is entirely subjective—it should be replaced by a definite (maybe different) fitting strategy.

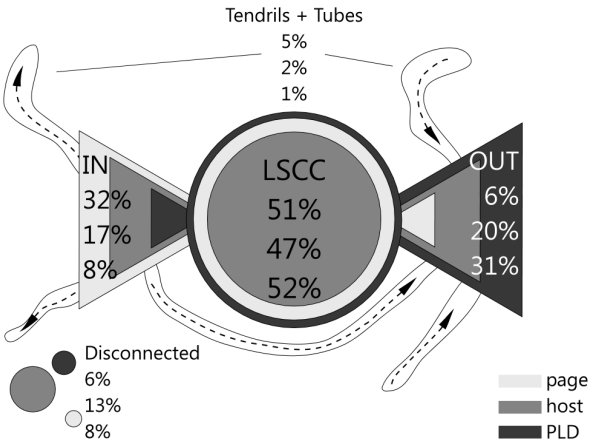


Figure 15: The bow tie on different aggregation levels

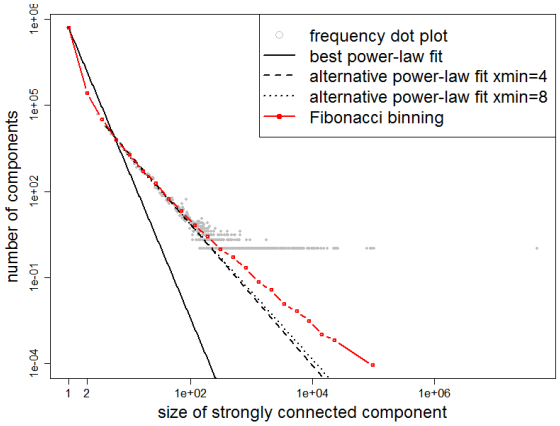


Figure 16: Frequency plot of the distribution of SCCs within the host graph with alternative power-law fits

8.3 The Role of Aggregation

A very delicate issue is the role of *aggregation* in the detection of power-law tails. Power laws (and more generally heavy-tailed distribution) should exhibit, when sampled, some rare, but not the rarest, large element (e.g., there should be nodes of high degree). As discussed in detail by Willinger *et al.*, 2009, measurements of the Internet topology performed at the end of the '90s concluded that the Internet Autonomous Systems graph had a power-law degree distribution and thus very high degree nodes that would have been the perfect target of terrorist attacks (!) as they would have destroyed the Internet's connectivity. Unfortunately, these high-degree nodes never existed: they were simply artifacts of the tool used to build the distribution, which did not understand some kind of layer-2 technology (e.g., Asynchronous Transfer Mode (ATM)), thus classifying large sets of computers connected by such technology as a single point. In other words, part of the reason of the myth of the "power law of the Internet" was aggregation artifacts. Indeed, the existence of nodes of degree so high to defy common engineering sense should have rang an alarm.

In this paper, while we witness distributions that appear graphically to be heavy-tailed, we cannot provide a proper power-law fit unless we aggregate pages at the PLD level and consider the indegree distribution of the associated graph. At that point, suddenly the p-

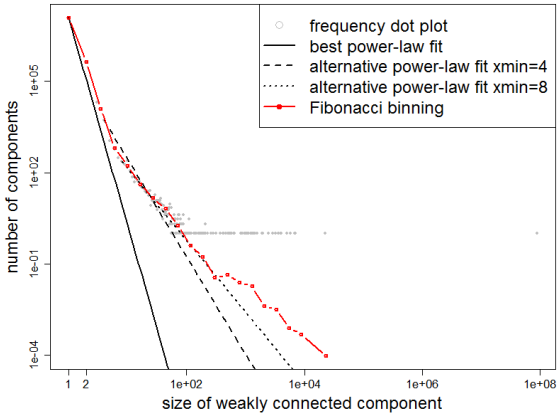


Figure 17: Frequency plot of the distribution of WCCs within the host graph with alternative power-law fits

value associated to the best fit by maximum likelihood jumps from 0 to 0.43. We found a milder evidence also for the distribution of PageRank values. It is thus a natural question whether is just the aggregation level that is "right" (and thus an artifact), or whether the page and host data are just too noisy. We do not have an answer at this time, but we find very intriguing having been able to fit correctly at least some power-law tail.

9 Conclusion

In this article we have presented the results of an analysis of the so far largest hyperlink graph that is available to the public outside companies such as Google, Yahoo!, Yandex, and Microsoft. The graph covers over 3.5 billion pages, linked by 128 billion hyperlinks. Beside the analysis of the page graph, we analyzed the host and pay-level-domain aggregation of the graph. Comparing our results with previous measurements performed in the last 15 years, and with previous literature on significantly smaller crawls, we reach the following conclusions:

- The average degree of the page graph has significantly increased, almost by a factor of 5 in comparison to the findings of Broder *et al.*, 2000.
- At the same time, the connectivity of the page 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.
- As also shown in Meusel *et al.*, 2014 and Lehmberg *et al.*, 2014 for page and PLD graph, we confirm the existence of a large strongly connected component also in the host graph.
- Although the average degree in host and PLD graph decrease due to the removal of influence of host and PLD internal link the relative sizes of the LSCC are almost stable among the different levels of aggregation.
- Even if the size of the IN and OUT component depends on the crawling strategy as discussed in Meusel *et al.*, 2014, analyzing the changes of relative size from page via host to PLD

level we can confirm an decrease of the IN component and an increase of the OUT component, already mentioned by Zhu *et al.*, 2008. In contrast the relative size of the LSCC stays almost the same among all three aggregation levels.

- Modeling the distribution of indegrees as well as outdegrees in the different aggregation levels is difficult. The clouds of points in the page graph distributions turn into outliers and spikes in the host and PLD graph. A manual inspection indicates as possible causes spam networks and domain resellers.

- Similarly to the observations of Meusel *et al.*, 2014, the frequency plots of degree and component-size distributions of host and PLD graph are visually identical to previous work. However, using proper statistical tools, we can only find a power-law tail within the indegree distribution of the PLD graph and the PageRank distribution of the same aggregation level.

Our findings form a basis for further research and analysis. One problem which still needs to be solved is the real mathematical distribution of degree and components within the Web. Although we have shown, in contrast to the assumptions of the last decade, that there is no statistical evidence of a power-law tail for most distributions related to web graph (the only exception being the indegree distribution of the PLD graph), the question of which is the correct distribution (and, more importantly, which process govern its formation) remains unanswered.

Acknowledgments

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.

References

Achlioptas, D., A. Clauset, D. Kempe, and C. Moore. 2009. “On the bias of traceroute sampling: Or, power-law degree distributions in regular graphs”. *Journal ACM*. 56(4): 21:1–21:28. URL: http://dl.acm.org/citation.cfm?id=1538905.

Alstott J Bullmore E, P. D. 2014. “powerlaw: A Python Package for Analysis of Heavy-Tailed Distributions”. *PLoS ONE*. 9. DOI: 10.1371/journal.pone.0085777. URL: http://www.plosone.org/article/info:doi/10.1371/journal.pone.0085777.

Backstrom, L., P. Boldi, M. Rosa, J. Ugander, and S. Vigna. 2012. “Four Degrees of Separation”. In: *ACM Web Science 2012: Conference Proceedings*. ACM Press. 45–54. URL: http://arxiv.org/abs/1111.4570.

Baeza-Yates, R. and B. Poblete. 2003. “Evolution of the Chilean Web structure composition”. In: *Proc. of Latin American Web Conference 2003*. 11–13. DOI: 10.1109/LAWEB.2003.1250276. URL: http://www.cwr.cl/la-web/2003/stamped/02_baeza-yates-poblete.pdf.

Berners-Lee, T., L. Masinter, and M. McCahill. 1994. “RFC 1738: Uniform Resource Locators (URL)”. URL: https://www.ietf.org/rfc/rfc1738.txt.

Bizer, C., K. Eckert, R. Meusel, H. Mühleisen, M. Schuhmacher, and J. Völker. 2013. “Deployment of RDFa, Microdata, and Microformats on the Web - A Quantitative Analysis”. In: *Proc. of the In-Use Track ISWC’13*. URL: http://link.springer.com/chapter/10.1007/978-3-642-41338-4_2.

Boldi, P., B. Codenotti, M. Santini, and S. Vigna. 2002. “Structural properties of the African web”. In: *Proc. WWW’02*. URL: http://vigna.di.unimi.it/ftp/papers/www2002b/poster.pdf.

Boldi, P. and S. Vigna. 2004. “The WebGraph Framework I: Compression Techniques”. In: *Proc. WWW’04*. ACM. 595–602. URL: http://vigna.di.unimi.it/ftp/papers/WebGraphI.pdf.

Boldi, P. and S. Vigna. 2012. “Four Degrees of Separation, Really”. In: *ASONAM 2012*. IEEE Computer Society. 1222–1227. URL: http://arxiv.org/abs/1205.5509.

Boldi, P. and S. Vigna. 2013. “In-Core Computation of Geometric Centralities with HyperBall: A Hundred Billion Nodes and Beyond”. In: *ICDMW 2013*. IEEE. URL: http://arxiv.org/abs/1308.2144.

Boldi, P. and S. Vigna. 2014. “Axioms for Centrality”. *Internet Math.* URL: http://www.tandfonline.com/doi/abs/10.1080/15427951.2013.865686#.Va-Jrvntkc8.

Broder, A., R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. 2000. “Graph structure in the Web: experiments and models”. *Computer Networks*. 33(1–6): 309–320. URL: http://www.sciencedirect.com/science/article/pii/S1389128600000839.

Clauset, A., C. R. Shalizi, and M. E. J. Newman. 2009. “Power-Law Distributions in Empirical Data”. *SIAM Rev.* 51(4): 661–703. ISSN: 0036-1445. DOI: 10.1137/070710111. URL: http://dx.doi.org/10.1137/070710111.

Dill, S., R. Kumar, K. S. Mccurley, S. Rajagopalan, D. Sivakumar, and A. Tomkins. 2002. “Self-similarity in the web”. *ACM Trans. Internet Technol.* 2(3): 205–223. ISSN: 1533-5399. DOI: 10.1145/572326.572328. URL: http://doi.acm.org/10.1145/572326.572328.

Donato, D., S. Leonardi, S. Millozzi, and P. Tsaparas. 2005. “Mining the inner structure of the Web graph.” In: *WebDB*. 145–150. URL: http://www.research.yahoo.net/files/donato2008mining.pdf.

Fetterly, D., M. Manasse, and M. Najork. 2004. “Spam, damn spam, and statistics: Using statistical analysis to locate spam web pages”. *Proc. WebDB’04*: 1–6. URL: http://dl.acm.org/citation.cfm?id=1017077.

Hall, W. and T. Tiropanis. 2012. “Web evolution and Web Science”. *Computer Networks*. 56(18): 3859–3865. ISSN: 1389-1286. DOI: http://dx.doi.org/10.1016/j.comnet.2012.10.004. URL: http://www.sciencedirect.com/science/article/pii/S1389128612003581.

Hirate, Y., S. Kato, and H. Yamana. 2008. “Web structure in 2005”. In: *Algorithms and models for the web-graph*. Springer. 36–46. URL: http://harbormist.com/proxy1/SyncHandler.ashx/www.cis.upenn.edu.sync/~mkearns/teaching/NetworkedLife/web2005.pdf.

Lehmberg, O., R. Meusel, and C. Bizer. 2014. “Graph structure in the web: aggregated by pay-level domain”. In: *Proceedings of the 2014 ACM conference on Web science*. ACM. 119–128.

Li, L., D. L. Alderson, J. Doyle, and W. Willinger. 2005. "Towards a Theory of Scale-Free Graphs: Definition, Properties, and Implications". *Internet Math.* 2(4): 222–262. DOI: 10.1080/15427951.2013.865686. URL: http://www.tandfonline.com/doi/abs/10.1080/15427951.2005.10129111#.Va-MYvntkc8.

Malevergne, Y., V. Pisarenko, and D. Sornette. 2005. "Empirical distributions of stock returns: between the stretched exponential and the power law?" *Quantitative Finance*. 5(4): 379–401. DOI: 10.1080/14697680500151343. URL: http://dx.doi.org/10.1080/14697680500151343.

Malevergne, Y., V. Pisarenko, and D. Sornette. 2009. "Gibrat's law for cities: uniformly most powerful unbiased test of the Pareto against the lognormal". *Swiss Finance Institute Research Paper Series*. 09-40. URL: http://EconPapers.repec.org/RePEc:chf:rpseri:rp0940.

Marchiori, M. and V. Latora. 2000. "Harmony in the small-world". *Physica A: Statistical Mechanics and its Applications*. 285(3-4): 539–546. URL: http://arxiv.org/abs/cond-mat/0008357.

Meusel, R., S. Vigna, O. Lehmberg, and C. Bizer. 2014. "Graph structure in the web - revisited: a trick of the heavy tail". In: *Proc. of the companion publication of the WWW'14*. International World Wide Web Conferences Steering. 427–432. URL: http://www2014.kr/wp-content/uploads/2014/05/companion_p427.pdf.

Page, L., S. Brin, R. Motwani, and T. Winograd. 1998. "The PageRank Citation Ranking: Bringing Order to the Web". *Tech. rep.* No. SIDL-WP-1999-0120. Stanford Digital Library Technologies Project, Stanford University. URL: http://ilpubs.stanford.edu:8090/422/.

Pandurangan, G., P. Raghavan, and E. Upfal. 2002. "Using Pagerank to characterize web structure". *Computing and Combinatorics*: 330–339. URL: http://link.springer.com/chapter/10.1007/3-540-45655-4%5C_36.

Serrano, M., A. Maguitman, M. Boguñá, S. Fortunato, and A. Vespignani. 2007. "Decoding the structure of the WWW: A comparative analysis of Web crawls". *TWEB*. 1(2): 10. URL: http://complex.ffn.ub.es/ckfinder/userfiles/files/a10-serrano.pdf.

Spiegler, S. 2013. "Statistics of the Common Crawl Corpus 2012". *Tech. rep.* SwiftKey. URL: https://docs.google.com/file/d/1_9698uglerxB9nAglvaHkEgU-iZNm1TvVGuCW7245-WGvZq47teNpb_uL5N9/.

Vigna, S. 2013. "Fibonacci Binning". *CoRR*. abs/1312.3749. URL: http://arxiv.org/abs/1312.3749.

Willinger, W., D. Alderson, and J. C. Doyle. 2009. "Mathematics and the Internet: A source of enormous confusion and great potential". *Notices of the AMS*. 56(5): 586–599. URL: http://authors.library.caltech.edu/15631/.

Zhu, J. J. H., T. Meng, Z. Xie, G. Li, and X. Li. 2008. "A teapot graph and its hierarchical structure of the Chinese web". *Proc. WWW'08*. WWW '08: 1133–1134. DOI: 10.1145/1367497.1367692. URL: http://doi.acm.org/10.1145/1367497.1367692.