Some Preliminary Results from a Link-crawl of the European Union Research Area Web[1]

Viv Cothey

viv.cothey@wlv.ac.uk
University of Wolverhampton, Wolverhampton, UK.

Abstract
A constrained Web link crawler has been used to obtain a broad multi-national sample of the European Union Research Area Web. This preliminary sample confirms that the distribution of many Web parameters follows a power law. The power law exponent of the tail of the indegree frequency distribution is 1.25 +/- 0.01 which agrees with a previously published value for the Web. However the previous result cannot be recomputed. A new reliable scale independent Web indicator for national recognition is also proposed. This makes use of the inter-national indegree based on country code Top Level Domain names. It might therefore be regarded as an improved version of the Web Impact Factor (Ingwersen, 1998).

Introduction

The Web-page digraph, that is, the directed graph defined by the node and link structure of Web-pages and their associated hyperlinks, is of interest to many and in particular to both complexity theorists and social scientists including information scientists (see for example Thelwall, 2004). This digraph is, at least superficially, accessible to research. It provides a large scale example of a complex network the generative processes for which are still being investigated. In addition, studies of its graphical structure can provide insight into the comparative strength of the presence of individual nodes or collections of nodes in the graph as in Ingwersen's (1998) notion of Web impact.

The Web-page digraph is sampled by means of Web-crawling (see for example Burke (2002)). In principle this is a straightforward data collection procedure. Unfortunately this belies the practical challenges involved and as identified by Cothey (2004) reliable Web-crawling demands special care.

In this research-in-progress paper I report some analyses of a sample of the Web-page digraph from the European Union Research Area (EURA) Web. The EURA Web is defined as being the combined Web-space of 2120 domain names of EU research and development institutions that were identified by the European Indicators, Cyberspace and the Science-Technology-Economy System project (EICSTES, 2005) as being situated in one of the fifteen older member states of the EU (Aguillo, 2004).

The results presented here are necessarily preliminary. They include both the overall indegree distribution of the sample digraph and the "inter-national" link structure. This latter analysis provides the opportunity to present a reliable scale adjusted Web recognition factor that takes into account the scale effect of the power-law distribution which appears to pervade Web parameters.

The paper proceeds by explaining first the data collection and data analyses procedures; this is followed by the results which are then discussed.

Data-collection and analysis

The EURA Web was crawled using an incremental link-crawler. The seed workload was the collection of all the home-pages of the default Web servers in the 2120 domains defining the crawl-space. The crawl-space was also constrained lexically to exclude urls where there was more than one url-path component (that is where a url is of the form <scheme>://<server-name>/<path>/<file> and the <path> contains more than one "/" component delimiter). This constraint facilitates obtaining within a reasonable timescale a "broad" sample across the whole of the crawl space rather than a "deep" but narrow sample of just a few servers.

[1] This work was supported by a grant from the Common Basis for Science, Technology and Innovation Indicators part of the Improving Human Research Potential specific programme of the Fifth Framework for Research and Technology Development of the European Commission. It is part of the Web indicators for scientific, technological and innovation research (WISER) project, (Contract HPV2-CT-2002-00015).

Proceedings of ISSI 2005, pages 212–220.
Edited by P. Ingwersen and B. Larsen.

Cothey (2004) introduced the notion of link-crawling in order to classify differing approaches to crawling; a link-crawl does not exclude pages from the sampled Web-page digraph because of duplicated content. Since Web search engines rely on content-crawling, their databases do not, in principle, provide valid link structure information. An incremental crawler maintains a local summary of each page encountered so that, subject to a freshness parameter, the crawler can include this in the crawl-graph without imposing any further demand on either the network or Web-server. The freshness parameter was set to be 200 days.

The crawl-graph produced has 2,535,788 nodes. This includes nodes that are the termination of arcs to outside the crawl space. (The crawl-graph is available for download from <http://www.wiserweb.org/reports/geography/eura-crawlgraph-p1.20041007.xml.gz>.)

The sample Web-page digraph is a simplified subgraph of the crawl graph. That is, duplicated arcs between nodes are removed as are loops (or self-links). As well as being simplified the sample subgraph excludes nodes outside the crawl-space and unreachable nodes or so called dead-links. Ignoring the arcs that are dead-links improves the validity of the data which otherwise, for example, would be subject to bias caused by host reorganisation. The sample Web-page digraph has 388,883 nodes.

Analysis of the sample Web-page digraph determines the indegree for each node so that the indegree frequency distribution can be found. The power law nature of this distribution is then investigated. Similar analyses can be focussed on other parameters such as the file size.

A further analysis related to indegree considers the source nodes of the arcs as well as the quantity of arcs terminating at a given target node. Here the analysis involves the country code Top Level Domain (ccTLD) of the urls labelling the source and target nodes of each arc. The ccTLD is used as a proxy for country in order to estimate inter-country linkage within the Web-page digraph. For example the arc to the node <url:http://target.dk:8081/page.html> from the node <url:https://source.de/page.shtm> has a target ccTLD of "dk" and a source ccTLD of "de" which are the top level domain names controlled by the Danish and German Internet domain name registration authorities respectively. This arc is described as an ordered pair (de, dk).

The number of inter-country arcs between nodes provides a measure of the inter-national recognition of the target node within the graph. This is saying that the more inter-country sources that link to a particular Web-page target, the more that Web-page is inter-nationally recognised. Ingwersen (1998) put forward a similar rationale in support of the Web Impact Factor (WIF). His computational technique is based on querying a search engine in order to infer indegree; Smith (1999) discusses further some of the challenges that arise when using search engines to compute a WIF.

In the analysis here a fifteen node digraph (one for each ccTLD within the EURA Web) is constructed from all the arcs such as (de, dk) in the above example. This digraph is called here the EURA inter-national recognition digraph.

Similar analyses can be undertaken at different levels of recognition. For example, one may be interested in inter-institutional recognition or inter-server recognition rather than inter-national recognition.

Results

Indegree and outdegree distributions

The power law distribution of several Web parameters has been identified, that is

$$P(x) \approx x^{n}$$

where P(x) is the probability of occurrence of the parameter with value x and n is the power-law exponent. For example, the size of Web pages appears to be distributed as a power law as evidenced by the log-log plot shown in Figure 1 where the tail approximates to a straight line.

In particular the tail of the frequency distribution of both the indegree and outdegree in the simplified Web-page graph has been found to follow a power law (Albert, Jeong & Barabási, 1999). Figures 2 and 3 illustrate the tails of the indegree and outdegree distributions of the sample Web-page digraph respectively. For ease of analysis the plots shown are those of the complementary distribution, P(X>x) (Crovella, Taqqu, & Bestavros, 1998). The tail of the distribution is given by ignoring x<10 while in the legend p=1 signifies a crawl constrained to one path component.

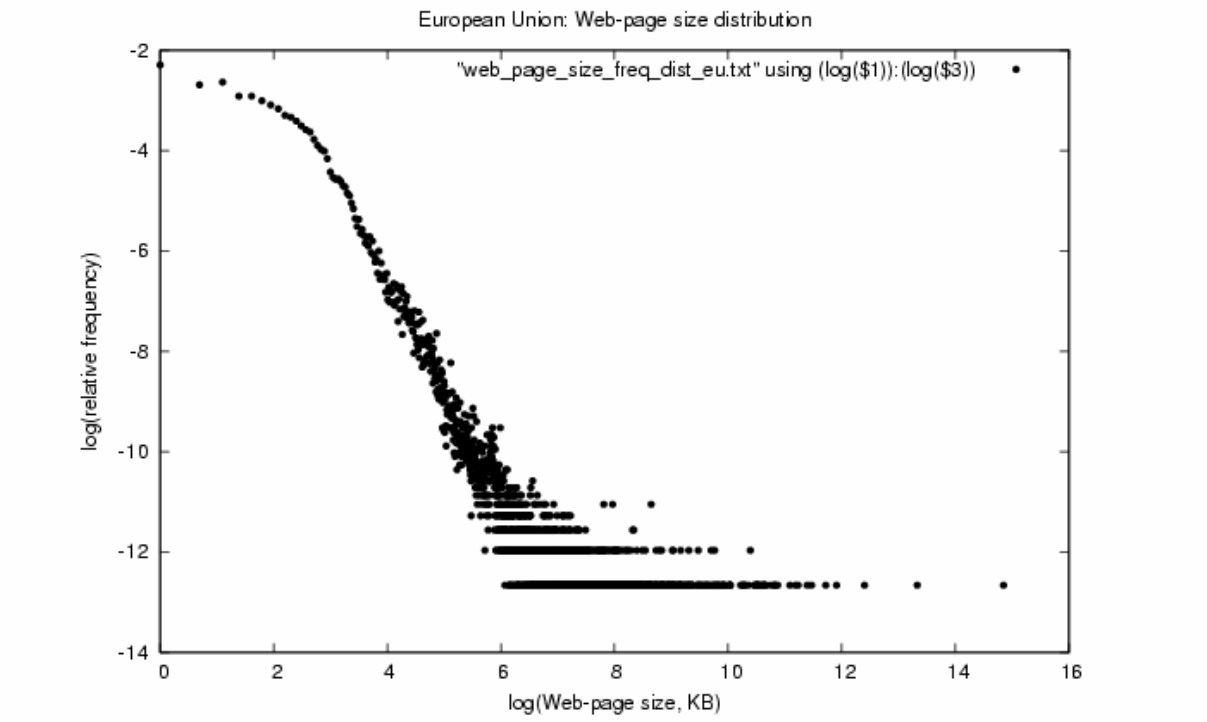

Figure 1: Probability distribution of Web-page size

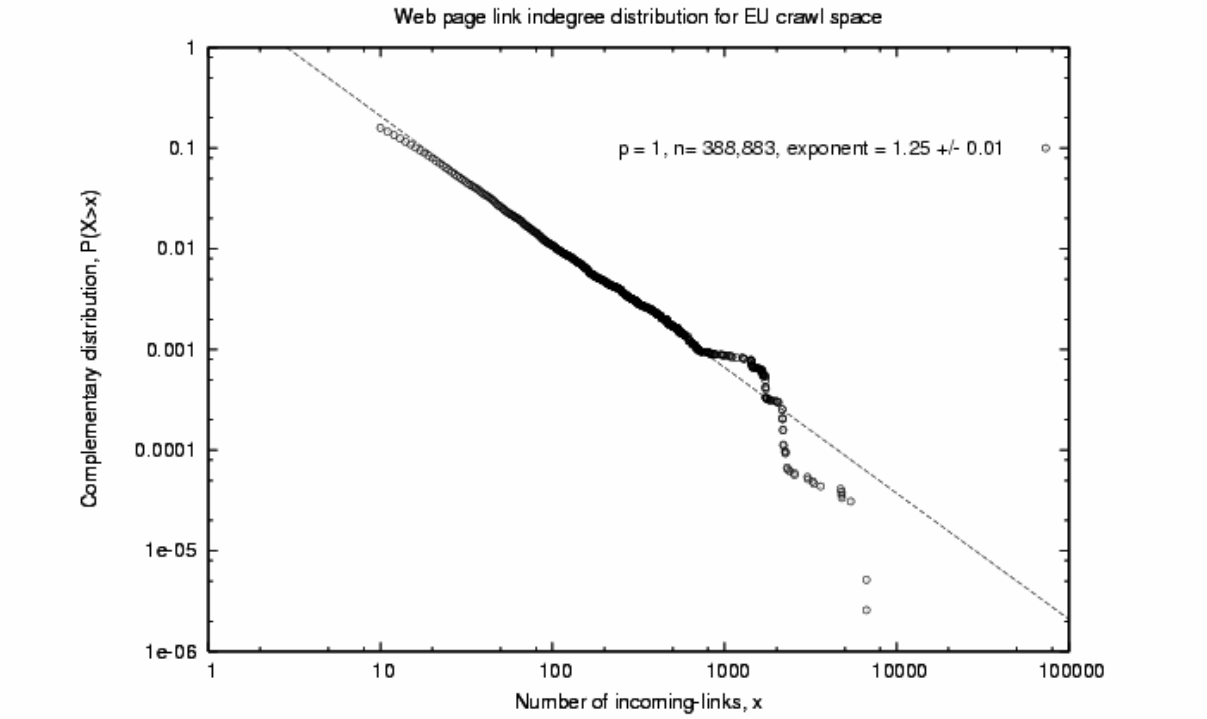

Figure 2: Complementary distribution of indegree

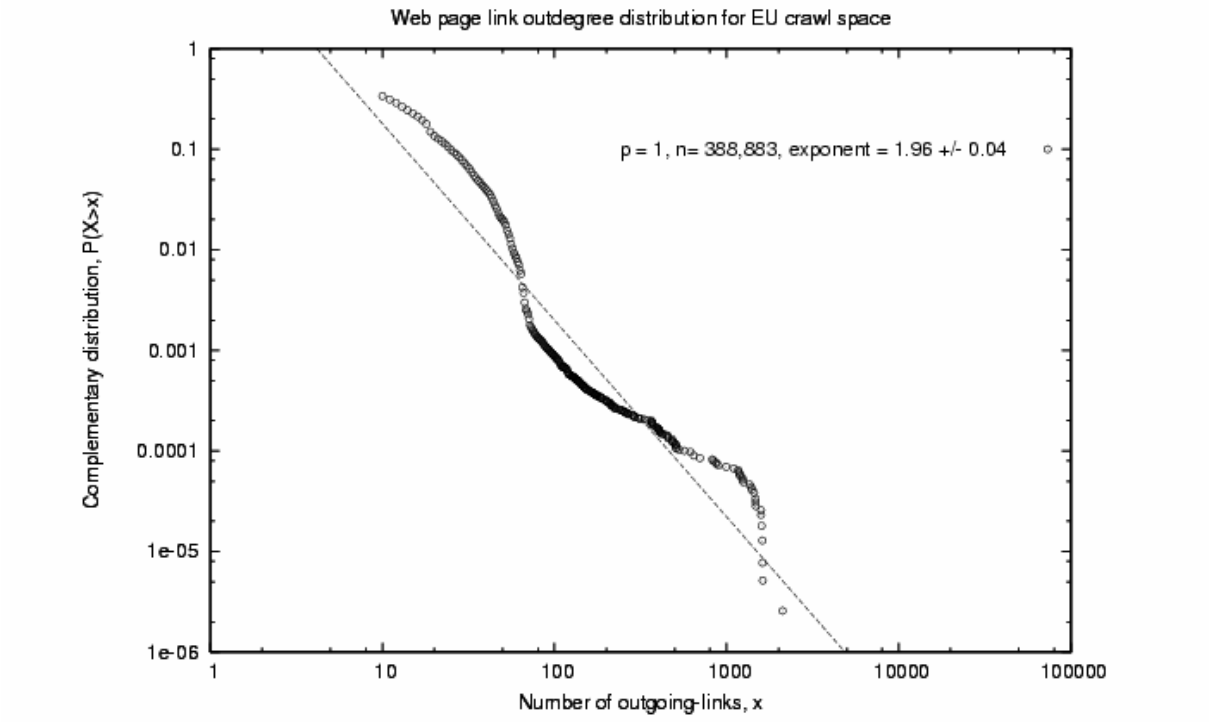

Figure 3: Complementary distribution of outdegree

The empirical power law exponent for these distributions is determined using a least squares method. For the indegree distribution this is found to be about 1.25.

Inter-national recognition

Table 1 provides an overall summary of the sample Web-page digraph analysed by ccTLD. Since not all of the url labels for the 388,883 nodes have a EURA ccTLD then the total number of Web-pages represented is only 360,676. There are 25,721 servers and 25,571 hosts; each host is defined by its unique host name and can run a server on several ports e.g. host:80, the default http port, and host:8080.

Table 1: Summary of the sample Web-page digraph by ccTLD

"Country"	Σweb-servers	Σhosts	ΣWeb-pages	avg. Web-pages/server
(.de) Germany	6457	6417	97960	15.2
(.uk) Great Britain	4949	4927	74584	15.1
(.fr) France	2592	2578	21110	8.1
(.es) Spain	1391	1385	19664	14.1
(.it) Italy	2261	2245	30273	13.4
(.nl) The Netherlands	1355	1350	15031	11.1
(.se) Sweden	2113	2111	27334	13.0
(.at) Austria	641	630	15010	23.4
(.be) Belgium	764	748	9471	12.4
(.dk) Denmark	727	724	13159	18.1
(.fi) Finland	864	858	14991	17.4
(.pt) Portugal	917	911	14234	15.5
(.ie) Ireland	275	273	4224	15.4
(.gr) Greece	396	395	3557	9.0
(.lu) Luxembourg	19	19	74	3.4

Typically about fourteen Web-pages are located at each server. It should be noted here that the data is collected by a lexically constrained crawler so that only pages having urls with up to a single path

component are reported. It should also be noted that the distribution of pages across servers and hosts is governed by a power law as illustrated by the log-log plot of the probability distribution of Web-pages per host shown in Figure 4. In consequence statistics such as average Web-pages per server are problematic due to the heavy tail nature of the frequency distribution.

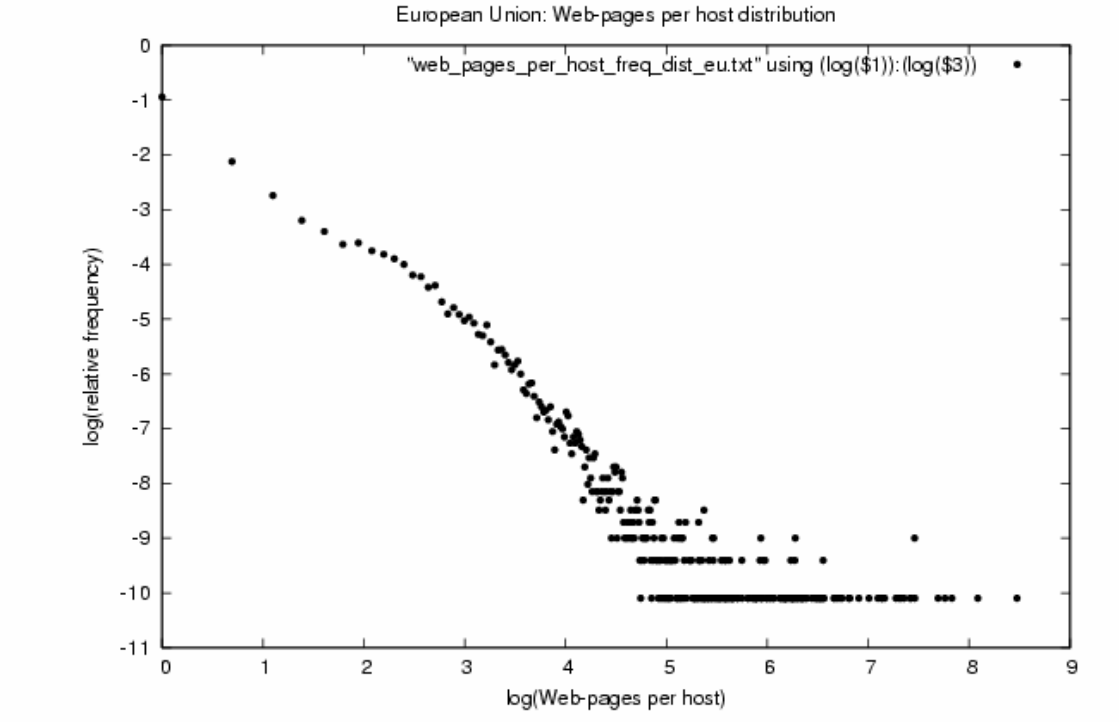

Figure 4: Probability distribution of Web-pages per host

The fifteen nodes of the EURA inter-national recognition digraph are given in Table 2. The "countries" are ranked by their aggregate inter-country indegree. The matrix underlying Table 2 is reported as Table A1 in the Appendix.

Table 2: EURA countries by Web-page indegree

"Country"	Σindegree	ΣWeb-pages	avg. indegree/Web-page
(.de) Germany	4888	97960	0.05
(.uk) Great Britain	3173	74584	0.04
(.fr) France	2064	21110	0.10
(.es) Spain	1324	19664	0.07
(.it) Italy	1276	30273	0.04
(.nl) The Netherlands	1022	15031	0.07
(.se) Sweden	802	27334	0.03
(.at) Austria	710	15010	0.05
(.be) Belgium	581	9471	0.06
(.dk) Denmark	554	13159	0.04
(.fi) Finland	516	14991	0.03
(.pt) Portugal	422	14234	0.03
(.ie) Ireland	361	4224	0.09
(.gr) Greece	281	3557	0.08
(.lu) Luxembourg	10	74	0.14

The relationship between the number of Web-pages and the inter-country indegree for the EURA Web "countries" is also illustrated in Figure 5. Conceptually the ratio indegree/Web-page is an external national WIF (Ingwerson, 1998).

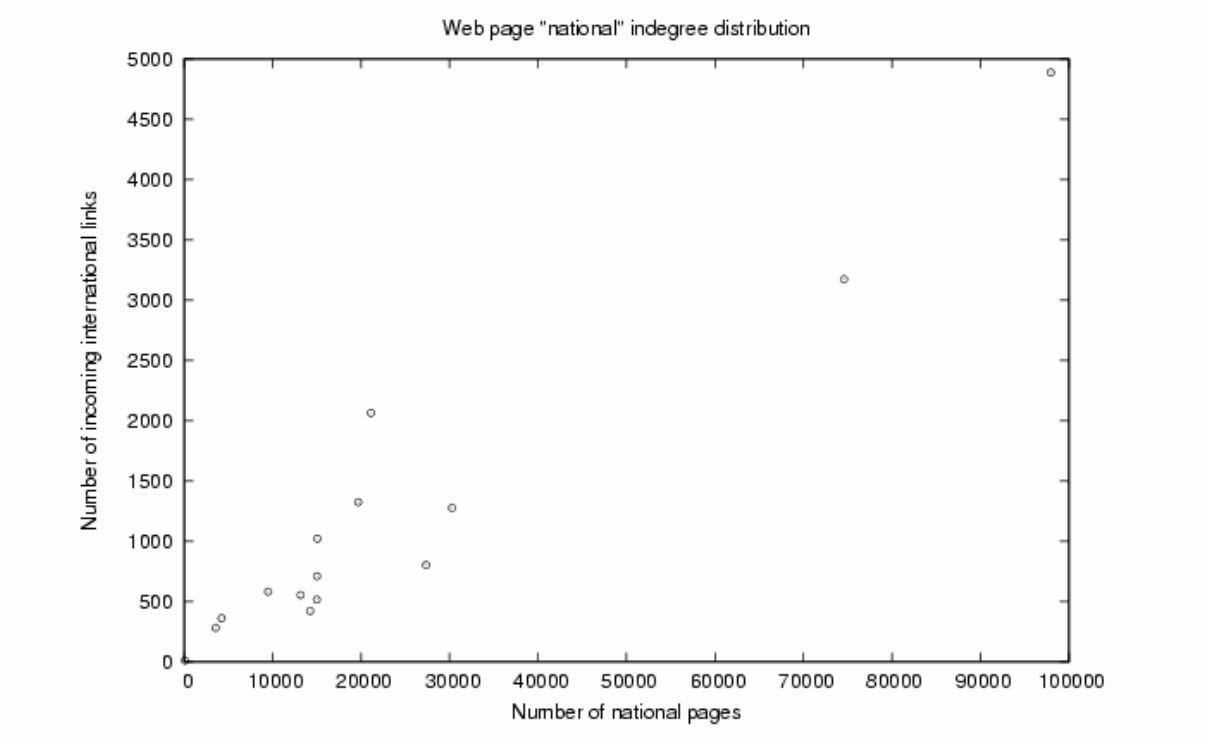

Figure 5: Inter-national indegree by national Web-pages

It is evident from this that, as might be thought obvious, a country's inter-national indegree is influenced by the number of its "country" pages that are available as linking targets. This relationship also appears to be governed by a power law although in this instance the exponent is positive. The log-log plot of the relationship is shown in Figure 6.

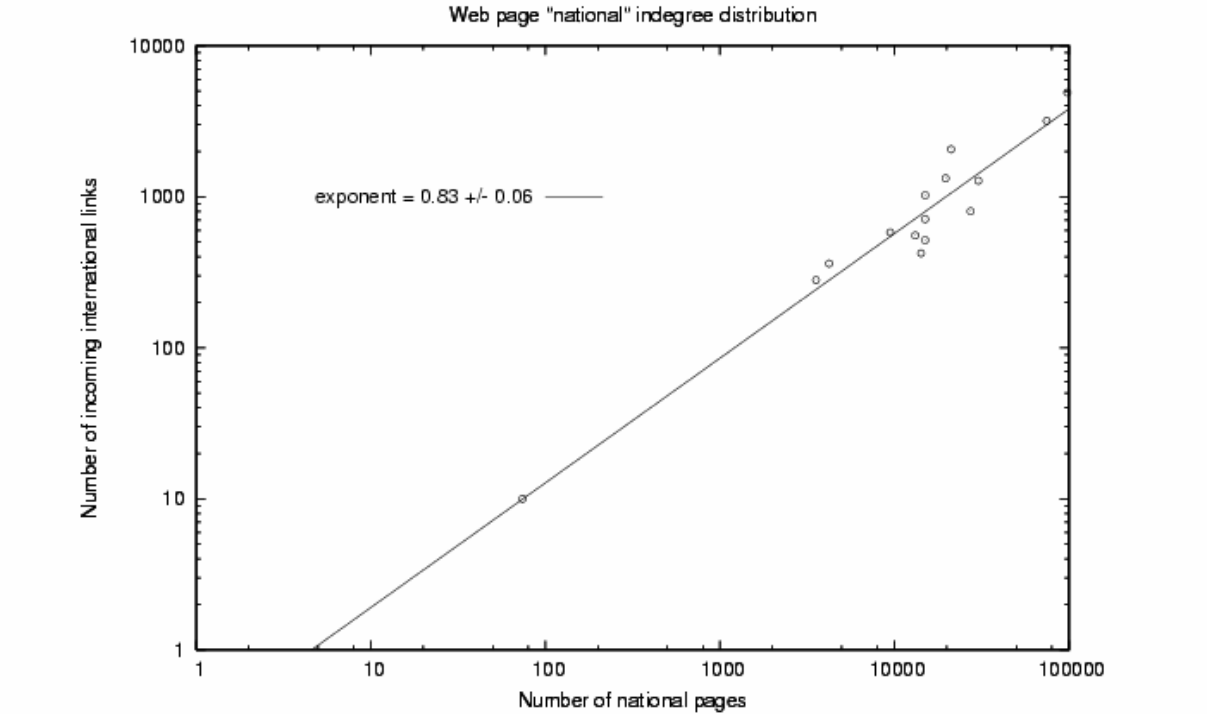

Figure 6: Power law relationship between inter-national indegree and national Web-pages

The correlation here is 0.97 while the power law exponent is 0.83.

The best fit straight line function here provides the means to model the relationship between the number of Web-pages and the inter-national indegree.

Discussion

The general form of the complementary distribution of Web-page indegree reported above is as expected and corroborates earlier empirical work by Mossa, Barthélémy, Stanley, & Amaral (2002) which used the University of Notre Dame crawl data. These data are available as a summarized simplified digraph (Jeong, 2005) although no details of the crawl procedure have been provided. In order to facilitate comparison the Notre Dame digraph has been reprocessed in the same way as the EURA sample Web-page digraph reported above. Figure 7 illustrates the complementary distribution of indegree for the Notre Dame digraph and is analogous to Figure 2.

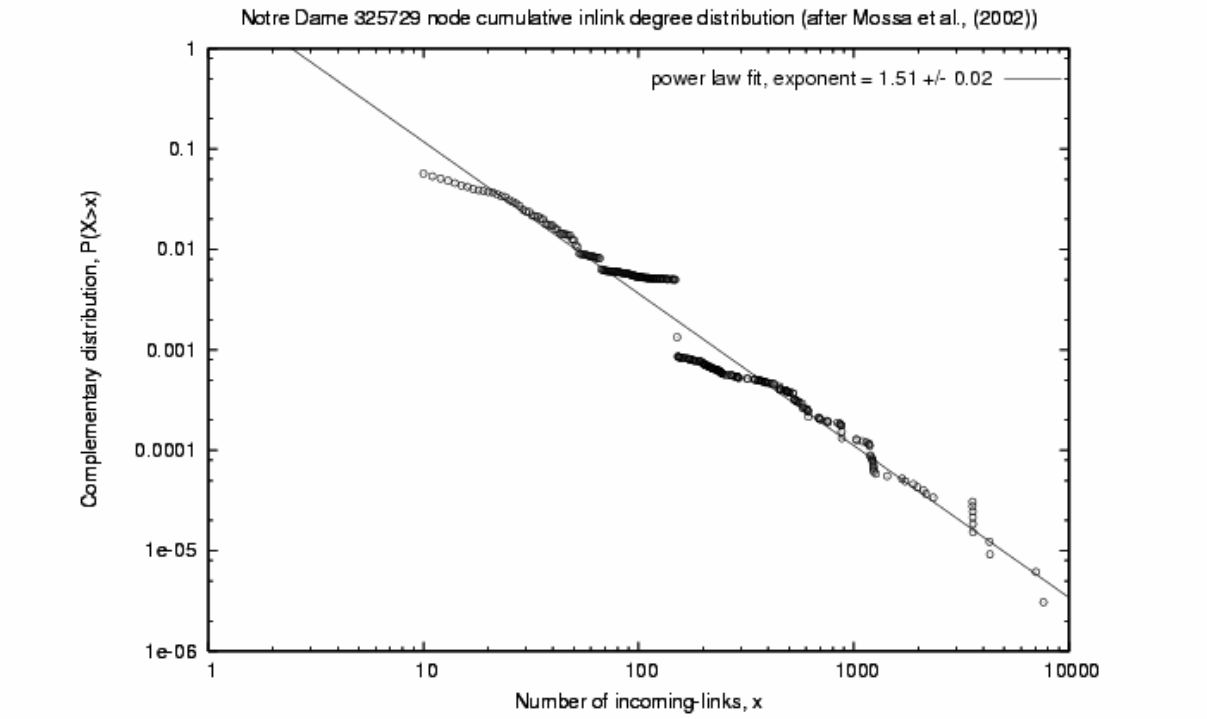

Figure 7: Complementary distribution of indegree from the Notre Dame data

The value of the power law exponent computed by Mossa, Barthélémy, Stanley, & Amaral (2002) is 1.25 which is in agreement with the corresponding value for the EURA sample digraph, also 1.25. However the reprocessed Notre Dame data yields the value 1.51. Whether the two samples are in agreement is thus uncertain. In any event, as yet it is not clear whether these are two estimates of a single uniform value applicable to the Web, or they represent a characteristic of the Web-page digraph that distinguishes different subgraphs (but in this comparison may be the same!).

Two other features illustrated in both Figures 2 and 7 are the "bumps" and "drops-off" that occur. In both distributions between 1000 and 10,000 there is a bump close to an almost vertical section. Donato, Laura, Leonardi & Millozzi (2004) also note this phenomenon but they only speculate at an explanation.

At the extreme the complementary distribution drops-off the fitted power law model. This divergence has been considered by Newman (2001), Mossa, Barthélémy, Stanley, & Amaral (2002) and by Boguñá, Pastor-Satorras & Vespignani (2004). It may or may not be explained by a finite size effect.

The value of a country's indegree in the EURA inter-national recognition digraph given in Table 2 is a candidate parameter for a Web indicator of national recognition or strength of presence within the EURA Web. However the gross value tends to reflect the number of target pages available within each country. This suggests that these values should be normalized in some way in order to identify whether or not there is a surplus or deficit in inter-national recognition regardless of size (number of Web-pages). The existence of the power law relationship illustrated in Figure 6 means that a simple normalisation "per Web-page" is not appropriate. Instead a normalisation procedure that takes account of the effects of scale is required in order to produce a scale independent indicator (Katz, 2000); in this instance there is an inverse Matthew effect in that the growth in inter-national indegree is less than

pro-rata to the growth in the number of country pages. A doubling of size (number of pages) produces less than a doubling of indegree.

The normalisation procedure uses the power law model (the straight line in Figure 6) to estimate the expected inter-country indegree given the number of country Web-pages present. The scale adjusted Web recognition factor is given as the variance ratio between the actual inter-country indegree and the expected inter-country indegree. That is, the scale adjusted Web recognition factor is the variance given in Table 3.

Table 3: EURA countries by scale adjusted Web recognition factor

"Country"	indegree (actual)	national pages	indegree (expected)	variance (actual:expected)
(.fr) France	2064	21110	1056.3	2.0
(.es) Spain	1324	19664	996.2	1.3
(.de) Germany	4888	97960	749.0	1.3
(.ie) Ireland	361	4224	280.0	1.3
(.nl) The Netherlands	1022	15031	798.1	1.3
(.gr) Greece	281	3557	242.9	1.2
(.be) Belgium	581	9471	545.1	1.1
(.uk) Great Britain	3173	74584	2993.6	1.1
(.lu) Luxembourg	10	74	9.9	1.0
(.it) Italy	1276	30273	1422.4	0.9
(.at) Austria	710	15010	797.2	0.9
(.dk) Denmark	554	13159	715.1	0.8
(.fi) Finland	516	14991	796.3	0.6
(.se) Sweden	802	27334	1307.4	0.6
(.pt) Portugal	422	14234	763.0	0.6

Table 3 contrasts with Table 2. It shows that "Luxembourg" is recognised appropriately given its size (actual 10, expected 9.9) while "France" receives nearly twice the actual inter-national recognition (2064) compared with what would be expected (1056.3) given how many .fr Web-pages there are in the sample digraph. On the other hand "Great Britain's" high ranking position in Table 2 is considerable reduced following the scale adjustment. The scale adjusted Web recognition factor which is scale independent shows that the previous high ranking of "Great Britain" is almost entirely due to the large number of .uk national Web-pages in the sample digraph.

Conclusion

These preliminary results appear to validate the aspect of the data-collection procedure which responds to the need to sample the Web across a broad inter-national front rather than on a more "depth" orientated domain focussed basis.

The indegree analysis builds on existing work and extends our empirical knowledge of the Web. It identifies two specific areas for further attention. Firstly how variable is the exponent value of the indegree power law distribution and what if any is the source of this variability? Secondly, what is the explanation for the bumps and drop-off in the distribution?

The scale adjusted Web recognition factor improves upon the WIF in two important ways. It does not use Web search engine queries. It is therefore immune to the vagaries, bias and questions over longer term stability and viability that affect commercial search engines. Secondly, because of the scale adjustment, the Web recognition factor is scale independent. It can therefore be used to compare the strength of the presence of collections of Web-pages where there are big differences in the size of these collections.

Since the data collection procedure is valid and reliable then the scale adjusted Web recognition factor is proposed as a scale independent Web indicator. In the particular instance presented this is a national indicator of the strength of the presence within the EURA Web of a country's Web-pages. However the approach is generalisable to other collections of Web-pages.

Establishing this Web indicator will require further work to determine its stability over time and also its sensitivity to changing the crawl policies of the data collection procedure, for example the constraints imposed on the crawler.

References

Aguillo, I. (2004). Personal communication to the author.

Albert, R., Jeong, H. & Barabási, A-L. (1999). Diameter of the World Wide Web. *Nature*, 401, 9 Sept. 130-131.

Boguñá, M., Pastor-Satorras, R., & Vespignani, A. (2004). Cut-offs and finite size effects in scale-free networks. *European physical journal B*, 38, 205-209.

Burke, S. M. (2002). *Perl & LWP*. Farnham: O'Reilly.

Cothey, V. (2004). Web-crawling reliability. *Journal of the American Society for Information Science and Technology*, 55(14), 1228-1238.

Crovella, M., Taqqu, M. & Bestavros, A. (1998). Heavy tailed probability distributions in the World Wide Web. In R. Adler, R. Feldman & M. Taqqu (Eds.), *A practical guide to heavy tails: statistical techniques and applications* (pp. 3-26). Berlin: Birkhäuser.

Donato, D., Laura, L., Leonardi, S., & Millozzi, S. (2004). Large scale properties of the Webgraph. *European physical journal B*, 38, 239-243.

EICSTES (2005). European indicators, cyberspace and the science-technology-economy system. Retrieved 27 January 2005 from: http://www.eicstes.org/websites.asp

Ingwersen, P. (1998). The calculation of Web impact factors. *Journal of documentation*, 54(2), 236-243.

Jeong, H. (2005). World-Wide-Web data. Retrieved 27 January 2005 from: http://www.nd.edu/~networks/database/www/www.dat.gz

Katz, S. (2000). Scale independent indicators and research assessment. *Science and public policy*, 27(1), 23-36

Mossa, S., Barthélémy M., Stanley H. E., & Amaral L. A. N. (2002). Truncation of power law behavior in "sale-free" network models due to information filtering. *Physical review letters*, 88(13), 138701/1-138701/4.

Newman, M. E. J. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences of the United States of America*, 98(2), 404-409.

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

Thelwall, M., (2004). Link analysis: an information science approach. London: Elsevier

Appendix

Table A1: EURA countries Web-page international indegree matrix

	All	Germany	G Britain	France	Spain	Italy	The N'lands	Sweden	Austria	Belgium	Denmark	Finland	Portugal	Ireland	Greece	Luxembourg
Germany	4888		816	490	330	256	227	157	244	136	79	141	100	99	87	6
G Britain	3173	2407		375	215	237	181	134	76	74	100	88	78	100	62	0
France	2064	244	298		121	173	56	64	33	66	63	32	31	22	10	2
Spain	1324	179	206	100		73	48	31	22	54	21	19	33	8	13	0
Italy	1276	193	325	136	158		67	67	58	53	26	22	40	20	30	1
The N'lands	1022	898	229	56	50	50		38	27	54	37	21	16	7	7	0
Sweden	802	168	359	93	66	62	85		39	25	104	80	19	15	7	0
Austria	710	232	61	65	41	46	25	12		11	18	12	21	11	11	0
Belgium	581	68	219	182	46	49	52	10	67		19	11	12	17	8	1
Denmark	554	62	101	62	26	62	38	67	2	12		25	4	7	3	0
Finland	516	67	119	46	31	61	48	139	31	14	36		9	15	8	0
Portugal	422	61	127	48	130	46	44	23	15	28	22	13		18	10	0
Ireland	361	20	78	19	17	17	5	9	8	2	1	5	14		3	0
Greece	281	20	49	18	15	19	3	4	6	4	0	5	9	2		0
Luxembourg	10	3	0	3	0	0	0	0	0	4	0	0	0	0	0	