Predicting Fame and Fortune: PageRank or Indegree?

Trystan Upstill
Department of Computer Science
CSIT Building, ANU
Canberra, ACT 0200, Australia
trystan.upstill@cs.anu.edu.au

Nick Craswell and David Hawking
CSIRO ICT Centre
CSIT Building, ANU
Canberra, ACT 2601, Australia
nick.craswell@csiro.au, david.hawking@csiro.au

Abstract

Measures based on the Link Recommendation Assumption are hypothesised to help modern Web search engines rank 'important, high quality' pages ahead of relevant but less valuable pages and to reject 'spam'. We tested these hypotheses using inlink counts and PageRank scores readily obtainable from search engines Google and Fast.

We found that the average Google-reported PageRank of websites operated by Fortune 500 companies was approximately one point higher than the average for a large selection of companies. The same was true for Fortune Most Admired companies. A substantially bigger difference was observed in favour of companies with famous brands. Investigating less desirable biases, we found a one point bias toward technology companies, and a two point bias in favour of IT companies listed in the Wired 40. We found negligible bias in favour of US companies.

Log of indegree was highly correlated with Google-reported PageRank scores, and just as effective when predicting desirable company attributes. Further, we found that PageRank scores for sites within a known spam network were no lower than would be expected on the basis of their indegree. We encounter no compelling evidence to support the use of PageRank over indegree.

Keywords Information retrieval

1 Introduction

PageRank is the hyperlink-recommendation algorithm used by Google to measure page 'importance and quality'. PageRank is reported to be an important component of Google's ranking algorithm [2], although PageRank's true importance in ranking is a Google trade secret. PageRank is also provided directly to users via the Google Toolbar and Google Directory. When a page is visited, the toolbar lists its PageRank on a scale of 0 to 10, indicating 'the importance Google assigns to a page'[1]. When a directory category is viewed, the pages are listed in descending PageRank order and with a PageRank indicator next to each page, to 'tell you at a glance whether other people on the web consider a page to be a high-quality site worth checking out'[2].

Because PageRank is provided directly in these ways, we can analyse it as a direct indicator of quality, without needing to know whether or how it is used in Google ranking. If the PageRank is "advice" to the user, we can test whether that advice is useful and test it for bias. We also know that this advice was calculated by Google, using Google's 3.3 billion page crawl, ensuring our PageRanks are calculated by experts on a large crawl. Further, by comparing extracted Web in-link counts we can investigate whether PageRank is better advice than a less computationally expensive indegree score.

Our approach is to extract PageRank values from the Google Toolbar and ask a number of questions about them. Section 2 describes the PageRank calculation and examines whether PageRanks are correlated with indegree in a very large crawl. Section 5 examines what biases, useful and otherwise, are present in hyperlink-recommendation evidence. Finally, section 6 presents a discussion of the systematic effects observed in Google Toolbar PageRank.

2 PageRank and indegree

The link recommendation assumption is that by linking to a page an author is recommending it. Therefore a page with many incoming links has been highly recommended.

A simple measure of recommendation is a count of a page's incoming links (its indegree). A more sophisticated calculation is Google's PageRank [2]. PageRank is transmitted from the linking page to the link target, but the size of the contribution depends on the PageRank of the linking page. So a link from a page that has large PageRank, such as the Yahoo home page, contributes more than a link

[1] `http://toolbar.google.com/button_help.html`
[2] `http://www.google.com/dirhelp.html`.

Proceedings of the 8th Australasian Document Computing Symposium,
Canberra, Australia, December 15, 2003.

	Extracted from Google			
	link: indegree	contains 'indegree'	Page-Rank	Fast indegree
Min	0	0	0	0
Max	857 000	1 250 000	10	14 324 793
Mean	958	1 910	5.3	17 889
Median	82	112	5	319
Apple	87 500	237 000	10	2 985 141

Table 1: Values were extracted from Google and Fast for 5370 company home pages on 12/09/2003. Listed are range, mean, median and an example value (for `http://www.apple.com/`).

from a page with low PageRank (all other things being equal).

The PageRank formulation from Brin and Page [2] is:

$$PR(A) = d + (1-d)\sum_{i=1}^{N} PR(T_i)/C(T_i)$$

The PageRank of page A, with some constant damping factor (e.g. $d = 0.15$), is the sum of contributions from its N incoming links. The contribution from linking page T_i is its PageRank $PR(T_i)$ divided by its outdegree $C(T_i)$. If this formula is applied iteratively to all pages, the PageRank values will converge. Note, there is no reason that a page can't link to another page twice. In such cases we ignore the extra link(s), which means the indegree of a page is always equal to the number of pages which link to it.

Adding an additional link from page T_i to page A can only increase $PR(A)$, assuming T_i has some PageRank ($PR(T_i) > 0$). Further, we are assured $PR(T_i)$ is greater than zero in any case where d is greater than zero. This means that, as indegree increases PageRank should be expected to increase.

Therefore we might expect PageRank to be correlated with indegree. Since it is easy to create links, we might also suspect it is easy to artificially increase the PageRank of a particular page by increasing its indegree. If PageRank does feature strongly in the Google ranking, such manipulation could be a useful tool for search engine optimisers and spammers. This is particularly true if $d > 0$, but we might also observe correlation with $d = 0$.

3 Method

Our approach is to identify a set of interesting URLs, for example a set of company homepages, and obtain the PageRank and indegree of each URL. We then analyse distributions and correlations. We also split the URLs into subsets such as Fortune 500 and Wired 40, to test how well PageRank and indegree predict inclusion in such lists.

We extracted our PageRank and indegree scores from search engines Google and Fast (`http://www.alltheweb.com/`). This section describes the extraction process, which can be applied to any given set of URLs. For example we apply it to company homepage URLs, spam homepage URLs and site crawl URLs. As already noted, extracting such data from the engines themselves ensures we are obtaining data from commercially important crawls comprising billions of pages. The corresponding difficulty lies in extracting accurate information from the engines.

PageRanks were extracted from the Google Toolbar by visiting pages and noting the interaction between the toolbar and Google servers. The extraction process was straightforward, although we note that PageRank has been normalised and transformed. Reported scores lie in the range 0 to 10. If PageRanks were distributed according to a power law [7] or similar, the vast majority of pages would have toolbar PageRank of 0 or 1. Observed toolbar PageRanks have a quite different distribution (Figure 3). We suspect the normalisation and transformation is intended to improve PageRank's usefulness as advice to users, and we evaluate the extracted PageRank values as such.

Unfortunately extracting reliable indegree data proved more difficult. We faced three main problems with link counting methods, and no method was perfect: (a) counting pages which simply mention a URL rather than linking to it, (b) not anchoring the match, so that our count for `http://www.apple.com` includes pages with `http://www.apple.com.au` and `http://www.apple.com/quicktime/` and (c) under reporting the figure, for example by systematically ignoring links from pages with PageRank less than 4.

We tried three methods for accessing indegree estimates for URL X (Table 1). The first was Google query `link:X` which reportedly has problem (c)[3]. The second was to find pages in Google which contain URL X. This solution was suggested by the Google Team[4] but we found it has problems (a) and (b), and also seems to only return pages which contain the URL in visible text.

Our third method was the Fast query `link:X -site:X`. We included the `-site:X` because the method has problem (b), and in our experience adding a `-site:X` will exclude a large number of intra-site links to pages other than the homepage (we are only interested in counting homepage links).

[3] `http://www.webmasterworld.com/forum80/254.htm`

[4] `http://slashdot.org/comments.pl?sid=75934&cid=6779776`

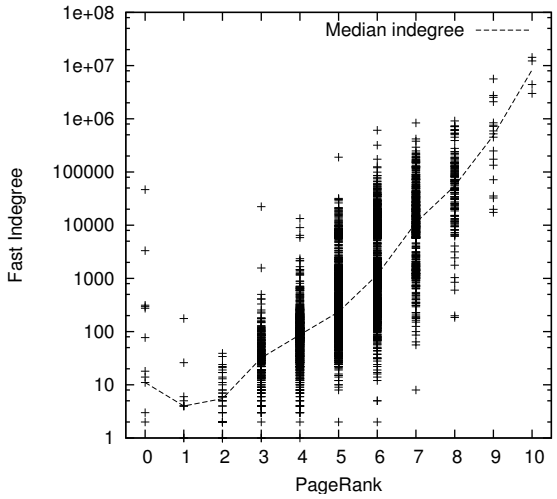

Figure 1: Company home pages. For our 5370 company home pages, Toolbar PageRank and log of Fast indegree have a correlation of 0.767 (Pearson r). This high confidence of correlation is achieved despite the relatively large spread of PageRank zero pages. Such pages may have been missed by the Google crawler or indexer, or might be penalised according to Google policy.

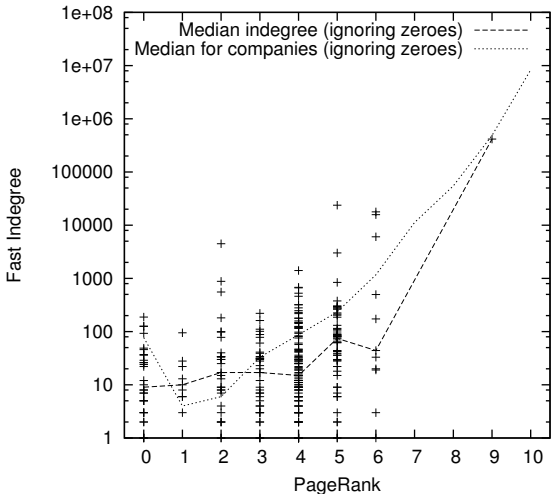

Figure 2: Spam backlinks. Our 280 spam pages achieve good PageRank without needing massive numbers of inlinks. In some cases, they achieve good PageRank with few links. Pages with PageRank 6 had a median indegree of 1168 for companies and 44 for spam pages.

All three types of indegree estimate were correlated with each other (Pearson $r > 0.7$) and produced very similar plots. However, we decided that to eliminate any potential biases, it is most interesting to compare Google PageRank to the indegrees from Fast, which has an independent crawl of similar magnitude (3.1 billion compared to Google's 3.3 billion).

4 PageRank-indegree correlation

The investigation of PageRank and indegree correlation on the WWW was conducted over a set of company home pages and a set of known spam pages.

4.1 Company homepages

We identified 8 329 company listings from three US stock exchanges: the New York Stock Exchange (NYSE), Nasdaq and the American Stock Exchange (AMEX). These were obtained from the Nasdaq Web site, and included company name, symbol and description. Then using the company information service at `http://quote.fool.com/` we identified 5 370 unique company homepage URLs. These were almost always the root page of a host `http://hostname.com/` without any file path (only 14 URLs had some path). These are considered to be the company home pages, even though in some cases the root page is a Flash animation or another form of redirect. The company information service also provided us with an industry for each stock e.g. 'Real Estate'.

The strong correlation between PageRank and log of indegree for company home pages is depicted in Figure 1. To better understand the differences between indegree and PageRank we performed an analysis of who were the "winners" and "losers" from the PageRank calculation. Winners in the PageRank calculation have high PageRanks even though they have low indegree (the bottom right quadrant of the figure), whilst losers have high indegree but receive a low PageRank (top left quadrant). Some anomalies were observed due to errors in indegree calculations (e.g. `www.safeway.com` had PageRank of 6 with indegree 0). However these cases were rare and uninteresting. After weeding out cases where Fast numbers disagreed with our other two indegree estimates, we still found extreme cases where indegree and PageRank were at odds (Table 2).

In some cases the discrepancies shown in the table are very large. For example, ESS Technology (`http://www.esstech.com`) lost out, achieving only PageRank of 3 despite having 22,357 indegree. On the other hand, Akamai (`http://www.akamai.com`) achieved a PageRank of 9 with only 17,359 links. Unfortunately, the promotions and demotions of sites relative to their indegree ranking do not seem to indicate a more accurate assessment by PageRank of the site's quality or importance.

4.2 Spam pages

One claimed benefit of PageRank over indegree is that it is less susceptible to link spam [3]. To test this claim we identified 399 backlinks of a search engine optimiser company using Google, and extracted Toolbar PageRanks. After extracting Fast

Stock	URL	Industry	PageRank	Fast Indegree
AAPL	http://www.apple.com	Computers	10	2985141
YHOO	http://www.yahoo.com	Internet Services	9	5620063
AKAM	http://www.akamai.com	Internet Services	9	17359
EBAY	http://www.ebay.com	Consumer Services	8	737792
BDAL	http://www.bdal.com	Advanced Medical Supplies	8	199
GTW	http://www.gateway.com	Computers	7	170888
JAGI	http://www.janushotels.com	Lodging	7	64
FLWS	http://www.1800flowers.com	Retailers	6	38254
KB	http://www.kookminbank.co.kr	Banks	6	5
IO	http://www.i-o.com	Oil Drilling	5	235
FFFL	http://www.fidelityfederal.com	Savings & Loans	5	34
USNA	http://www.usanahealthsciences.com	Food Products	4	13353
RSC	http://www.rextv.com	Retailers	4	6
ESST	http://www.esstech.com	Semiconductors	3	22347
CAFE	http://www.selectforce.net	Restaurants	3	3
MCBF	http://www.monarchcommunitybank.com	Savings & Loans	2	6
WEFC	http://www.wellsfinancialcorp.com	Savings & Loans	2	1
PTNR	http://investors.orange.co.il	Wireless Communications	1	176
HMP	http://www.horizonvascular.com	Medical Supplies	1	5
VCLK	http://www.valueclick.com	Advertising	0	46659

Table 2: Extreme cases, where PageRank and indegree disagree. We eliminated cases where the Fast was in disagreement with our two Google indegrees, and still found cases where having 1000 times more links gives 1 point less in PageRank.

indegrees, we eliminated those with indegree of 0, and plotted the remaining 280 points in Figure 2. These pages are largely content-free, having been created to funnel traffic and PageRank towards the search engine optimiser's customers.

If PageRank were spam-resistant, we would expect high indegree spam pages to have low PageRank. Such a case would be placed in the top left quadrant of our scatter plots. However, for our 280 spam pages the effect is minimal, and in some cases the opposite. For example, our medians for a PageRank score of 6 were 1168 for company homepages and 44 for spam pages. Spam pages tended to achieve PageRank of 6 seemingly with fewer incoming links than legitimate companies.

It is possible that any pages which did fall in the top left quadrant have already been excluded from Google. However, this still shows that Google can not rely entirely on PageRank for eliminating spam. This is not surprising, if we consider the extreme case: A legitimate page such as an academic's home page might have indegree of 10, while a search engine optimiser has massive resources to generate link spam from thousands or millions of pages. PageRank alone can probably not give high PageRank to the academic and low PageRank to the spammer's customer, given the indegree correlations we have observed. This is particularly true if every spam page has a nonzero PageRank ($d > 0$). Google must rely on blacklists and other methods for suppressing link spam.

In summary, we find a reasonable correlation between PageRank and log indegree for company homepages, and find that spam pages exhibit a similar correlation. We are yet to find a task for which PageRank is superior to indegree as an indicator of quality or authority.

5 Hyperlink-recommendation biases

Here we examine both positive and negative hyperlink-recommendation biases. Our investigation of positive biases examines whether hyperlink-recommendation identifies home pages, and whether it can, to some degree, recognise page quality and importance. Our investigation of negative biases looks at whether the use of hyperlink recommendation evidence introduces an unintended bias towards American, technology-oriented companies.

5.1 The home page bias

A web server might have hundreds of pages ([5] reported an average of 289 pages), of which only one is the root page. In this section we consider PageRank variation between pages in a site. We crawled 100 pages each from 20 uniformly selected company web sites. The sites were selected to have a range of homepage PageRanks and to include 10 Fortune 500 companies. This gave the overall PageRank distribution shown in Figure 3.

Figure 4 shows the PageRank distributions for 8 typical examples from our 20 crawls. In almost every case, the home page has the highest PageRank. In every case, some pages were inferior to the homepage according to PageRank advice. This is not surprising as links from one server to another usually target the root page of the target server. In fact, targeting deeper pages has even led to lawsuits

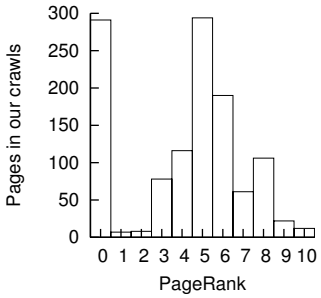

Figure 3: Combined PageRank distribution for twenty 100 page crawls. These pages are more representative of the general Web page population than our other sets, because the others are all homepages. It is not a power law distribution, suggesting some sort of transformation has been applied to toolbar PageRanks. The zero PageRanks most likely pertain to pages in our crawls which were not in Google crawls.

in some cases [6]. We would therefore expect a site to receive most of its externally supplied PageRank through its home page.

This home page bias can be thought of as helpful for search engines. A home page is designed to be the 'front door' of a site. A search engine which is biased towards returning home pages is therefore more likely to let users enter through the site's preferred entry point. However, homepage bias is even greater for indegree, where there is no propagation from high indegree home pages to lower tier pages. So if homepage bias is required, counting indegree might be preferable. URL features which prefer shorter URLs and/or directory default pages might be an even better indicator [4, 7].

From the user's point of view, it might seem mystifying that, for example, the Apple home page is rated 10 but its 'PowerBook G4 15-inch' page is rated 7. Is the Toolbar implying that the product is less important or of lower quality? Is it useful to tell users to give such advice about deeper pages in general?

In summary, the homepage bias of PageRank may be useful in search, but for that purpose indegree and URL-based indicators are even better. For Toolbar users, watching the PageRank decline during browsing could cause confusion if the user is attempting to stick with the most highly rated pages. We conclude that PageRank may be more useful in indicating differences between companies, than for rating different pages within a site.

5.2 Bias towards large, famous companies

Having considered intra-site PageRank effects, we now consider inter-site comparisons using our company homepages. This is similar to work by Amento [1] except we use more sites and our link

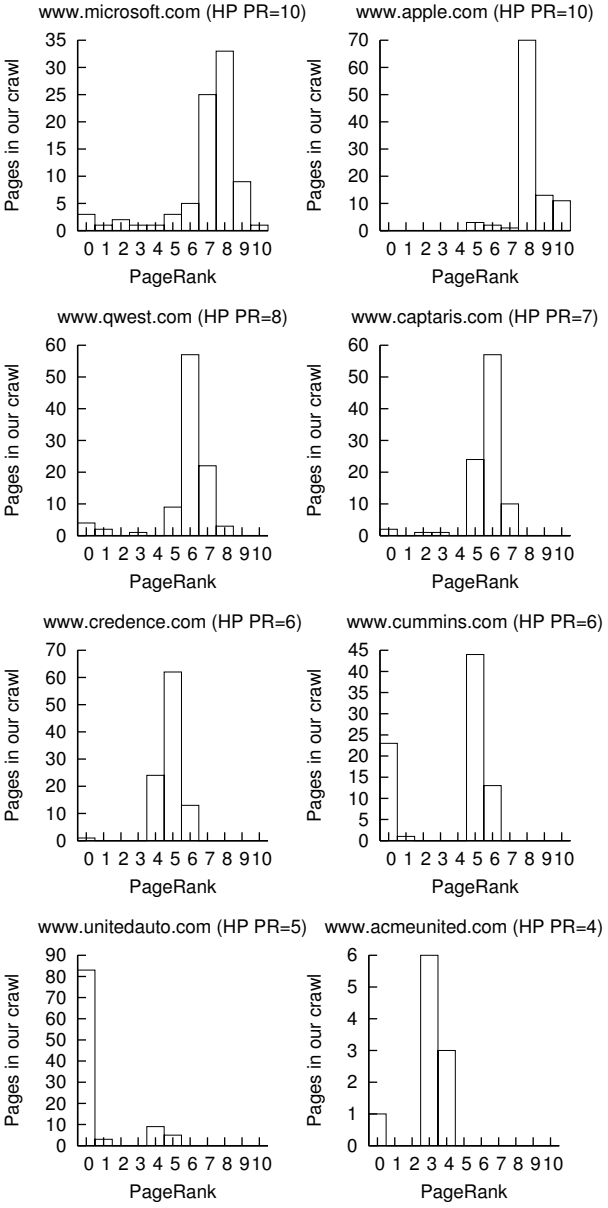

Figure 4: PageRank distributions within sites. The PageRank advice to users is usually that the homepage is the most important or highest quality page, and other pages are less important or of lower quality.

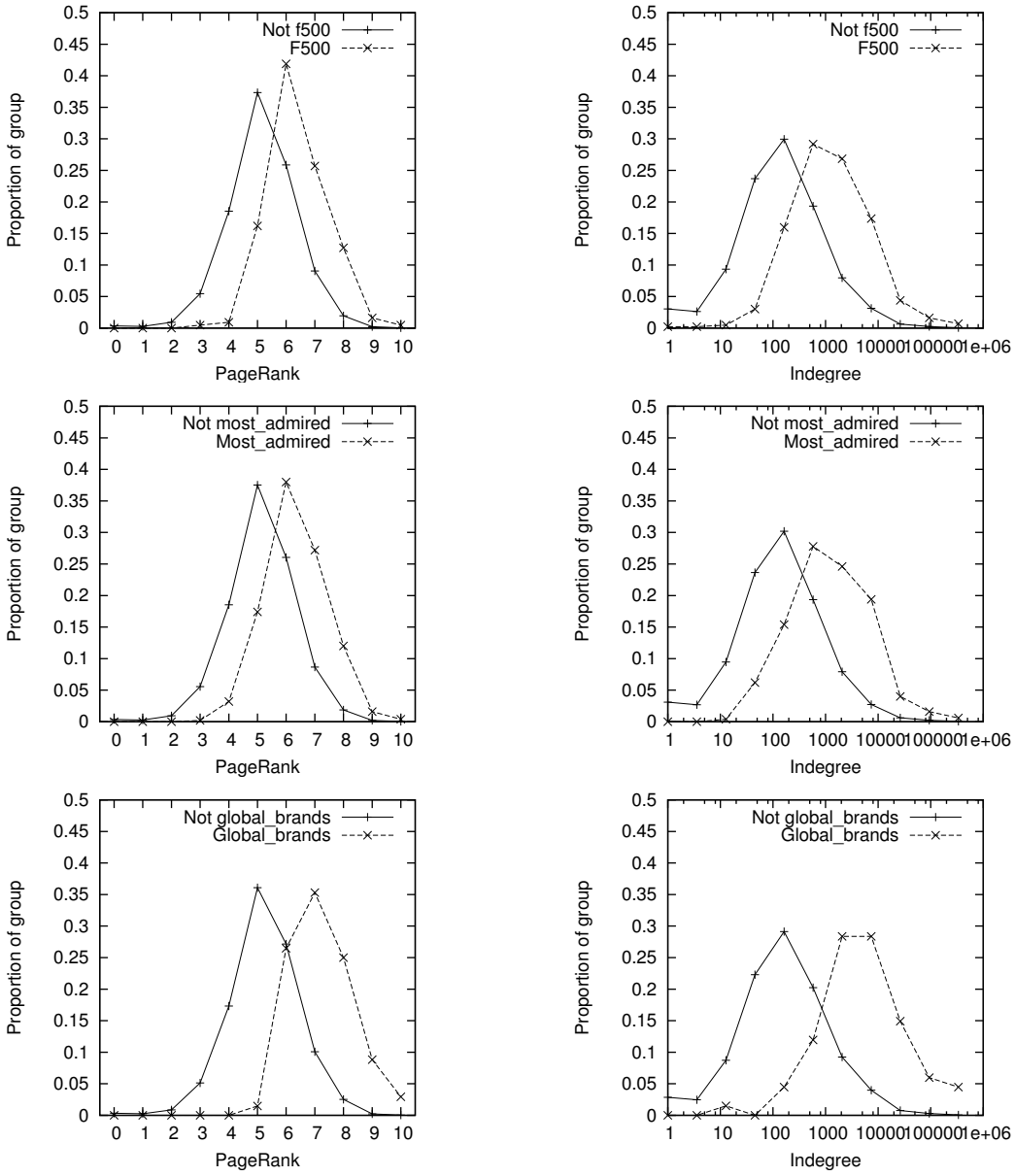

Figure 5: Useful effects. Companies in Fortune 500, Fortune Most Admired and Business Week Top 100 Global Brands lists tend to have higher PageRank. The effect is most strong for companies with well known brands. On the right, we see similar effects are present in indegree.

measures come from much larger (Google and Fast) crawls. Also, Amento evaluated the quality of sites, whereas we are testing based on qualities of the companies behind the sites.

Fortune 500 companies are those with the highest revenue, based on publicly available data, as listed by Fortune Magazine (`http://www.fortune.com/`). Likewise, Fortune Magazine maintains a Most Admired company list, rated by peer review. The Business Week Top 100 Global Brands lists the most valuable brands from around the world, based on publicly available marketing and financial data. These three lists give us good examples of large, famous companies, relative to our general population of 5370 companies.

Figure 5 shows that Fortune 500, Fortune Most Admired and Business Week top brand companies tended to have higher PageRank than other companies. However, there are examples of non-F500 companies with PageRank 10 such as `http://www.adobe.com`. At the other end of the spectrum, the Zanett group `http://www.zanett.com` has a F500 rank of 363, but PageRank of 3. This puts it in the bottom 349 of 5371 companies, based on toolbar advice.

The home pages of Fortune 500 and Most Admired companies receive, on average, one extra PageRank point. Further, Business Week Top Brand companies receive, on average, two extra PageRank points. Similar findings were observed for indegree.

These findings give some credit to the Google claim that PageRank indicates importance and quality, since the companies are important and this might also be associated with the quality of their sites or their offerings. Without making too many value judgments, we would also add that they are large and famous companies. Perhaps link recommendation is most associated with fame (or infamy).

Since indegree is an equally good indicator, we once again wonder whether there is any reason to favour PageRank over indegree.

5.3 Country and technology biases

We hypothesize that there may be some unintentional bias introduced through the use of hyperlink-recommendation evidence. A bias towards certain regions or industry groups may create anomalies that are confusing to Web search users.

To test country bias we retrieved the base countries for all companies using an information service provided by `blah.com`. Figure 6 shows that we did not find a pro-American bias. However, we should note that all companies studied are listed in American stock exchanges. Further, as we are including a smaller, regional stock exchange (AMEX) we might be unfairly biasing non-US companies by comparing large international (globally listed) companies against smaller (regionally listed) American companies. Perhaps if we compared the Australian Stock Exchange (ASX) companies to similarly sized companies from the American Stock Exchange we may get a more accurate picture in this respect. We leave this for future work.

		PageRank	
Industry	Companies	Range	Mean
Internet Services	29	3–9	6.66
Publishing	58	4–9	6.66
Airlines	25	3–8	6.48
Office Equipment	7	5–8	6.43
Entertainment	14	4–8	6.36
Software	306	3–10	6.35
Computers	86	4–10	6.29
Consumer Electronics	18	5–8	6.17
Automobile Manufacturers	7	4–8	6.14
Diversified Technology Services	46	4–8	6.02
...			
Steel	34	3–7	4.68
Coal	6	4–5	4.67
Clothing & Fabrics	54	2–7	4.63
Oil Companies	132	1–8	4.60
Pipelines	25	3–6	4.56
Banks	433	0–8	4.55
Real Estate	174	2–7	4.55
Precious Metals	38	0–6	4.47
Marine Transport	12	3–6	4.42
Savings & Loans	146	0–6	4.08

Table 3: PageRanks by industry. The Internet services and Publishing industries, with 29 and 58 companies respectively, had the highest mean PageRank.

We investigated two measures of technology bias (i.e. bias towards companies which produce technology or are heavy users of it). First, using the industry information taken from `http://quote.fool.com/`, we identified companies in industries involving: computer software, computer hardware, or the Internet. Figure 6 shows a bias towards technology-oriented companies. Our second test of technology bias uses the 2003 Wired 40 list of technology-ready companies [5]. This demonstrates an even greater pro-technology bias (same figure).

In summary, technology-oriented companies receive an extra PageRank point on average, whilst Wired 40 companies receive two extra PageRank points on average. American companies were not observed to be systematically favoured.

6 Discussion of systematic effects

Our experiments show that PageRank is correlated with log indegree (even for a collection of spam pages). Given the extra cost involved in computing PageRank, the correlation raises serious doubts about the benefit of using PageRank over indegree. In our data, any useful effect attributed to PageRank is also present in indegree, with both measures performing equally well when identifying

[5] available at `http://www.wired.com/wired/archive/11.07/40main.html`

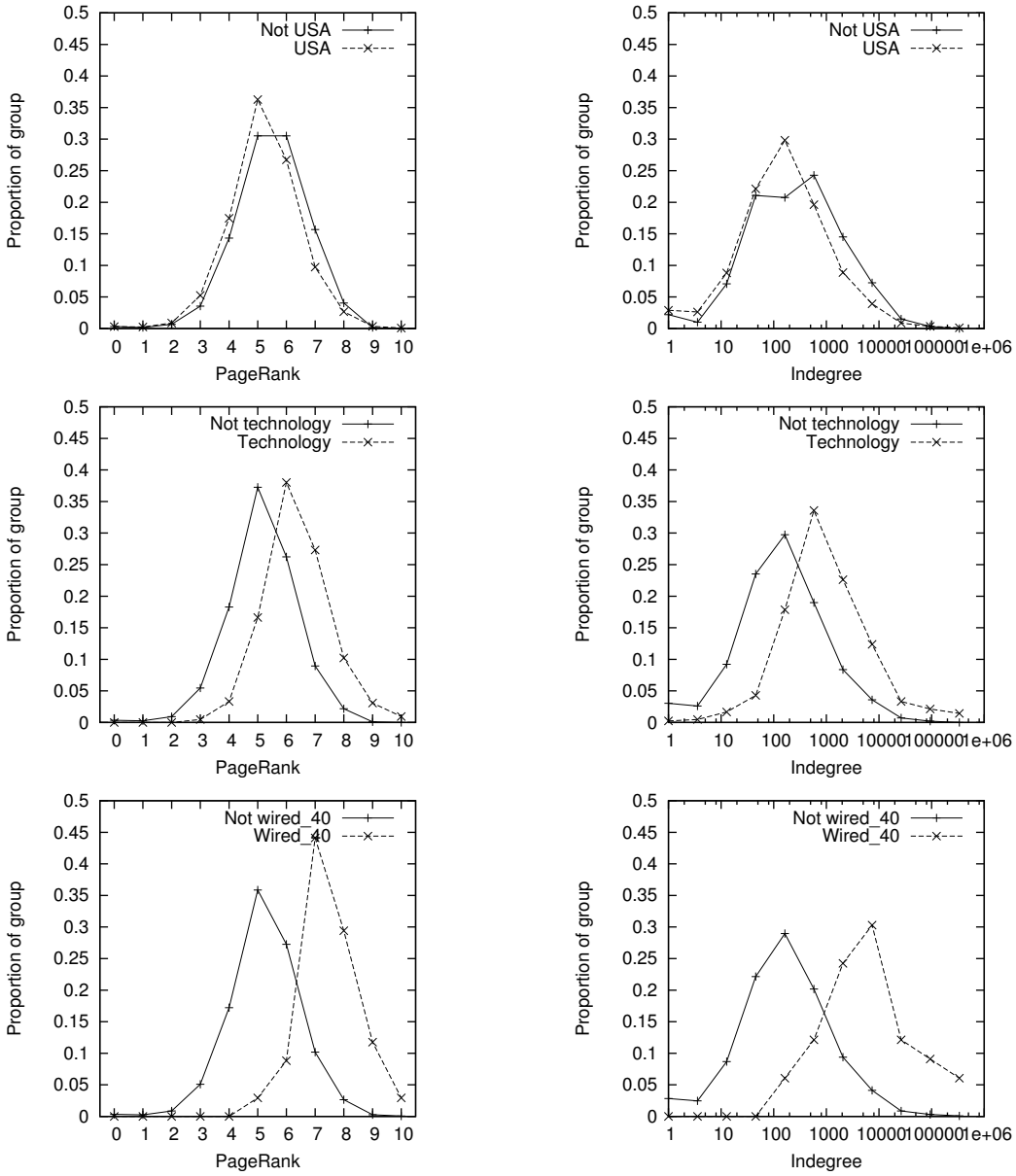

Figure 6: Less useful effects. We did not find a strong PageRank bias towards US companies. However, companies in the Internet Services, Software and Computers industries had higher PageRank, as did those in the Wired 40. Having a strong bias towards US and technology companies is most useful if you are interested in technology and from the US. Given the increasing global reach of the Web, and the increasing ease of access for non-technical users, such biases are helping a smaller and smaller proportion of the Web user population. On the right, we see similar plots for indegree.

Fortune 500, Most Admired and Global Brands listed companies. In cases such as those in Table 2, we cannot conclude that PageRank was clearly improving upon indegree. Our findings, compounded with previous research observing no performance improvements through the use of PageRank for common search tasks [7], gives us no reason to recommend the use of PageRank over indegree.

We have identified a number of systematic effects in hyperlink-evidence on the Web:

1. Within a site, the homepage usually has the highest PageRank.
2. Comparing company sites, we found that PageRank is biased by, on average, one point towards:
   (a) Companies with large revenue (by Fortune 500 membership)
   (b) Admired companies (by Fortune Most Admired membership)
   (c) Technology-oriented companies (by Industry type)
3. Further, we found that PageRank is biased by two points on average towards:
   (a) Companies with famous brands (by Business Week Top Brands)
   (b) Companies prepared for the new economy (by Wired 40 listing)

The same patterns were observed for indegree.

Since homepages of relevant sites are often useful search results, the homepage bias present in PageRank may be useful. However, the homepage bias is also present in indegree, and may be best achieved by discriminating based on URL depth. In terms of the toolbar's advice, it is not clear whether PageRank (or indegree) drop off for deeper pages is a desirable effect. In fact, it may be preferable to display a constant indicator in the toolbar when navigating within a website. Further, quality may be better indicated using measures other than link recommendation, such as whether companies are listed on the stock exchange, present in online directories and/or highly recommended by peer review[6].

The bias towards high-revenue, admired and famous companies can be seen to be consistent with the goal of hyperlink-recommendation algorithms. The fact that hyperlink measures more strongly recommend sites operated by companies with highly recognised brands suggests that recognition is a key factor. This is intuitively obvious, as a a website can only be linked to by authors who know of its existence.

Favouring high-recognition sites in search results or directory listings helps searchers by bringing their to bear their existing knowledge. A list of relevant websites already known to the searcher inspires confidence in the value of the list. Similarly, a site which appears in the list but which doesn't belong there can be easily ignored if it is well known to the searcher.

Consider the Google Directory category for Australian health insurance[7]. Viewed alphabetically the top two entries are the Web sites "Ask Ted" and "Australian Health Management Group". Viewed in PageRank order the top two are "Medibank Private" and "MBF Health Insurance". Even if the user does not agree that these are the best results, it is better to return results which the user will immediately recognise.

An important but less beneficial side effect of using hyperlink-recommendation algorithms is the inherent bias towards technology-oriented companies. There are a number of query terms whose common interpretation may be lost through heavy use of hyperlink-recommendation algorithms. For example, using Google there are a number of general queries where technology interpretations are ranked higher than their non-technology interpretations: 'opera', 'album', 'java', 'jakarta', 'png', 'putty', 'blackberry', 'orange' and 'latex'. The strong technology bias may be an artifact of having a largely technology-oriented demographic building web pages. Many Web authors are technical and believe that Jakarta is a Java programming project, but many Web users believe that Jakarta is the capital of Indonesia! As the demographics of Web users change, returning an obscure technology-related result will become less and less desirable.

7 Conclusion

Our experiments report a high correlation between PageRank and log indegree on the WWW. Given the similarity between indegree and PageRank we find no reason to use the more computationally expensive PageRank over indegree. Page quality as represented by PageRank in the Google Toolbar, in the context of company home pages and in certain search engine optimiser webs, would be just as useful if based on indegree. This finding, in combination with previous PageRank failures, casts serious doubt on the usefulness of PageRank over indegree.

[6] using scores from a service such as `http://www.alexa.com`

[7] `http://directory.google.com/Top/Regional/Oceania/Australia/Business_and_Economy/Financial_Services/Insurance/Health/`

We found hyperlink-recommendation algorithms do provide some indication for site home pages. However, while home page bias may be useful in web ranking, in the context of the Google toolbar it is a potentially confusing effect. The investigation of whether Web users understand hyperlink-recommendation scores for non-homepages remains for future work.

In our company homepage experiments, we found that important and well-known companies are favoured by both PageRank and indegree. Such bias can lead to rankings which are more easily understood by the searcher. A less desirable bias, towards technology-oriented companies, indicates a need for recommendation methods which more closely match user expectations. Such methods, which could take into account individual differences, or simply Web user demographics, remain for future work.

References

[1] Brian Amento, Loren G. Terveen and William C. Hill. Does "authority" mean quality? Predicting expert quality ratings of Web documents. In *Proceedings of SIGIR 2000*, pages 296–303. ACM, 2000.

[2] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Proc. 7th International World Wide Web Conference*, Brisbane, Australia, May 1998. `www7.scu.edu.au/programme/fullpapers/1921/com1921.htm`.

[3] Taher H. Haveliwala. Topic-sensitive pagerank. In *Proceedings of the eleventh international conference on World Wide Web*, pages 517–526. ACM Press, 2002.

[4] Wessel Kraaij, Thijs Westerveld and Djoerd Hiemstra. The importance of prior probabilities for entry page search. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 27–34. ACM Press, 2002.

[5] S. Lawrence and C. L. Giles. Accessibility of information on the Web. *Nature*, Volume 400, pages 107–109, 1999.

[6] Richard A. Spinello. An ethical evaluation of web site linking. *ACM SIGCAS Computers and Society*, Volume 30, Number 4, pages 25–32, 2000.

[7] Trystan Upstill, Nick Craswell and David Hawking. Query-independent evidence in home page finding. *ACM Transactions on Information Systems (TOIS)*, Volume 21, Number 3, pages 286–313, 2003.