Do TREC Web Collections Look Like the Web? *

Ian Soboroff
National Institute of Standards and Technology
Gaithersburg, MD
ian.soboroff@nist.gov

Abstract

We measure the WT10g test collection, used in the TREC-9 and TREC 2001 Web Tracks, and the .GOV test collection used in the TREC 2002 Web and Interactive Tracks, with common measures used in the web topology community, in order to see if these collections "look like" the web. This is not an idle question; characteristics of the web, such as power law relationships, diameter, and connected components have all been observed within the scope of general web crawls, constructed by blindly following links. The .GOV collection is a fairly straightforward 18GB crawl of sites in the `.gov` domain. In contrast, WT10g was carved out from a much larger crawl specifically to be a web search test collection within the reach of university researchers. Do such collections retain the properties of the larger web? In the case of WT10g and .GOV, yes.

1 Introduction

A critical requirement of a retrieval test collection is that it match the task. When the collection in question is a web collection, the issue expands to cover not only the content of the pages, but the broader hypertext structure of the collection as a whole. Since it is impossible to conduct repeatable retrieval experiments as we understand them on the "live web", several static web test collections have been built and used by the retrieval community in the past few years.

Bailey et al. [1] describe the construction of WT10g, the Web Track test collection used for TREC-9 and TREC 2001. This collection is about 10GB in size, and contains 1.69 million web pages. Their goal was to create a testbed for "realistic and reproducible" experiments on web documents with traditional, distributed and hyperlink-based retrieval algorithms. They began with VLC2, a 100GB subset of a 1997 crawl by the Internet Archive. From this they selected documents using a process designed to maximize inter-server connectivity, retain as many pages as possible from each server represented, incorporate documents likely to be relevant to a wide variety of queries, and exhibit a realistic distribution of server sizes. This process is described in detail in [1]. They measured the properties of the resulting collection according to mean in- and out-links per server, fraction of connected servers in the collection, and server "relevance", measured using a large query set.

One question that they did not answer was, does WT10g look like the World Wide Web? To answer that, we first need to understand more about what the web looks like. Singhal and Kaszkiel [4] looked at average in- and out- links, within and across hosts, between the smaller

*A shorter version of this paper "Does WT10g Look Like the Web?", appeared as a poster in SIGIR 2002

WT2g corpus and their own large crawl. They concluded that linkage in WT2g was inadequate for web experiments. However, the mean is a poor statistic to describe the power-law distributions of links on the web; average linkage is dominated by the many pages with few links and gives little insight into the topology.

Broder et al. [2] analyzed two large web crawls of about 200M pages each done by Altavista in 1999, and compared their structure to two important earlier studies. They looked at the distributions of in-links and out-links in their crawls, illustrating that these distributions obey power laws with exponents close to those observed in other studies. Further, using breadth-first traversals from a large sample of starting points they sketched out the high-level structure of the web in what has become the well-known "bow-tie model". These characteristics seem to hold for the web in general, however, Pennock et al. [3] found that category-specific subsets of the web can deviate strongly from power law scaling.

In order to show that WT10g indeed does resemble the web in many important ways, we measured the collection's link graph using the yardsticks of Broder et al. We show that while WT10g is small, structurally it does resemble larger web crawls that have been studied. This is an important result, because a primary criticism of web test collections is that they are inherently too small to be realistic testbeds of the web. These metrics can also be used to tune the construction methods of future test collections.

Lastly, a new web test collection, ".GOV", recently made its debut in the TREC 2002 Web Track. This collection is built around a straightforward crawl performed in January 2002, and as such contains much more recent web data. We provide the first analysis of the .GOV collection structure and compare it to WT10g.

2 Power-law distributions

Broder et al. found that the distributions of links in their crawls followed a power-law, that is, that the probability that a node has (in- or out-) degree d is proportional to $1/x^d$ for some $d > 1$. The exponents in their crawls was 2.1 for in-degree, and 2.72 for out-degree. Figure 1 shows the degree distributions in WT10g. These graphs are very similar to those found by Broder et al. In particular, notice the linear shape in the log-log plot, the messy tails for those few pages of very high degree, and that out-links diverge from the fitted curve at very low degree. The power-law exponents are 2.03 for in-degree, and 2.49 for out-degree. We are missing some spikes that they found and attributed to a spammer.

Power laws of hyperlink degree have been found in nearly every study of a web crawl, through a wide variety of crawl sizes. In contrast, WT10g is a subset of a web crawl carefully chosen to incorporate whole servers and highlight inter-site links, but without regard to the overall link distribution.

The WT10g distribution includes lists of in-links and out-links within the collection. We found in the course of our experiments that the WT10g in-links file is not consistent. If the out-links file is transposed to create a set of in-links, the resulting set contains 109 links not present in the WT10g in-links file. There is one truncated line in the in-links file, easily identified because it ends in a partial document identifier. We are unsure if the missing in-links can be attributed to further line truncations or some other reason. Consequently, the results reported here are are from our own list of in-links, constructed by transposing the out-links file included with WT10g so that the link graphs are internally consistent. The author's in-links file, and the script that generated it,

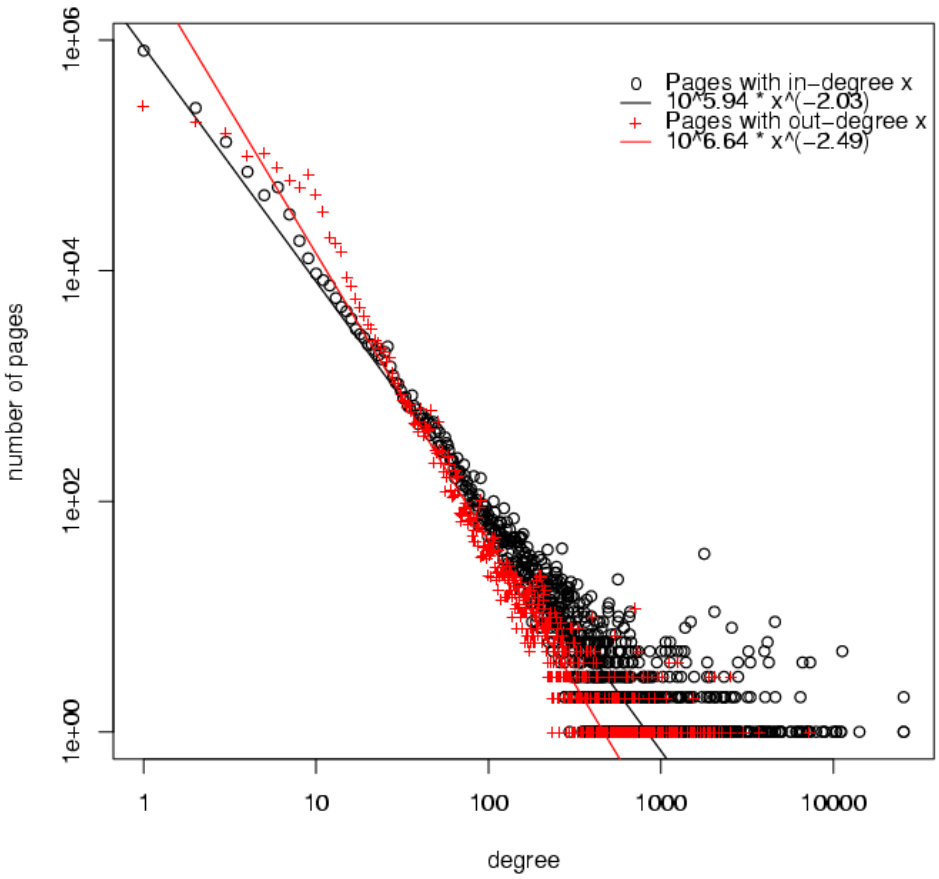

Figure 1: In- and out-degree distributions in WT10g.

are available upon request.

3 Connected components

Broder et al. also examined strongly- and weakly-connected components of the link graphs of their crawls. A strongly connected component (or, "strong component") of a graph G is a subgraph G' such that every node in G' is reachable from every other node in G' by following forward links through the graph. A weak component is the equivalent structure in an undirected graph; in our web graphs, this means taking the union of in-links and out-links into consideration when finding connected components. Figure 2 shows the distributions of strong and weak components in WT10g.

These graphs also follow a power law (exponents 1.8 for SCCs, 1.56 for WCCs) similar to the distributions found in the Altavista crawls. Our largest weak component contains 91% of the pages in WT10g, the same fraction as in the Altavista crawls. The largest strong component encompasses 29.4% of all the pages in the collection, compared to 28% in the Altavista crawls. The similarity

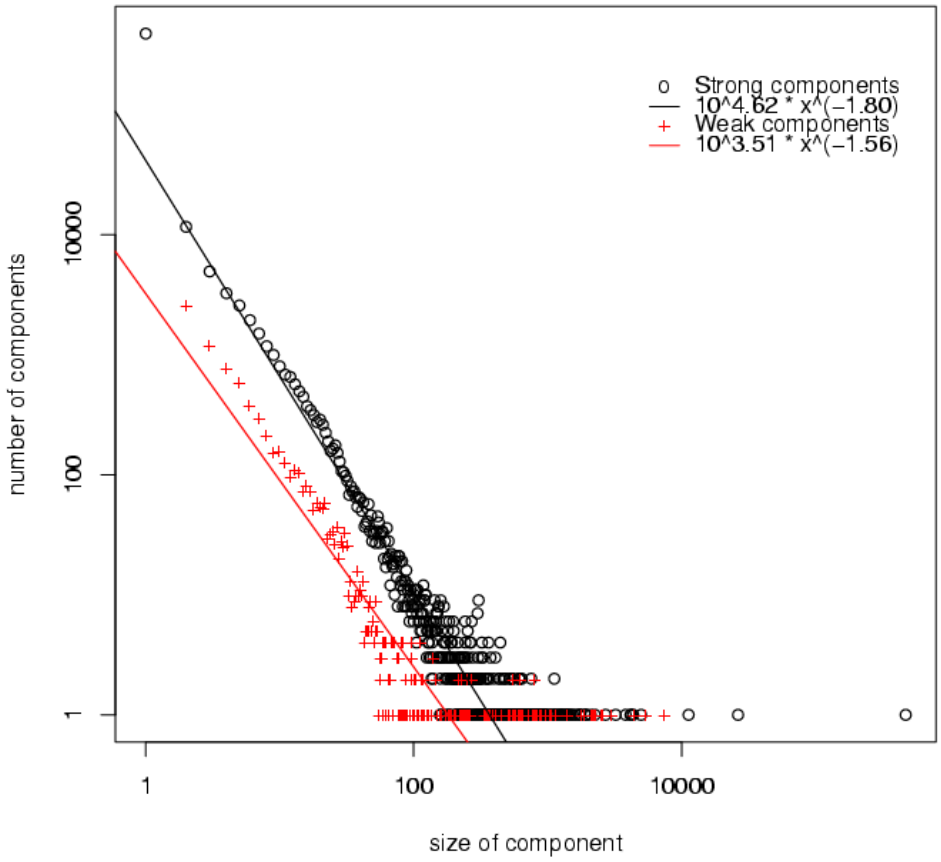

Figure 2: Distributions of strongly and weakly connected components in WT10g.

in largest component coverage is striking, but the smaller exponents in the WT10g distributions indicate a more gradual falling-off of component sizes. This probably reflects the tendency of WT10g to favor entire servers while at the same time having many fewer pages overall than the Altavista crawls.

4 Exploring with BFS

The third and most interesting component of Broder's study was designed to probe the dichotomy in coverage between the largest weak and strong components: if 91% of the collection is connected by undirected links, but only 29.4% by browseable links, what happened to all the other pages? If nothing else, it means that understanding the web to have uniformly small diameter is inaccurate; obviously, some pages are only reachable from certain places in the web, and a large fraction are all reachable from each other within a short distance. To explore this phenomenon, they conducted breadth-first searches backward and forward from random starting nodes, noting the depth of each traversal. We did the same for 500 random starting points.

Our findings again mirror those from the Altavista crawls. The traversal depths are sharply bimodal: either they would stop after reaching a small set of pages (often, fewer than 100), or they would balloon to a huge node set (roughly 740,500 following in-links, 926,500 for out-links). For about 30% of the start nodes, both directions would balloon; 30% would balloon following in-links only, and 10% following out-links. Following Broder's analysis, we find a bow-tie in WT10g with an IN set leading into the large SCC of 270,059 pages, an OUT set of pages reachable from the SCC of 456,059 pages, and 261,828 TENDRILS pages. WT10g's OUT set is larger than IN, compared to the Altavista crawls, where the sets were of roughly equal size. We hypothesize that the strategy in WT10g of selecting by server in order of size is biased somewhat toward SCC+OUT pages.

5 The .GOV Collection

After TREC 2001, the organizers of the Web Track set about constructing a new web test collection. They wanted to address concerns about the validity of conducting further experiments on 1997 web data. Furthermore, the track was interested in looking at a domain-specific collection. The track organizers and participants agreed that US Government web pages, a large and very diverse collection of information and services relatively free of copyright and distribution restrictions, would make an ideal testbed.

Following much discussion on what the new collection should contain, Charlie Clarke of the University of Waterloo crawled 95GB, approximately one million pages, of the `.gov` domain in January 2002 using systems hosted at Virginia Tech. The crawl collected not only HTML web pages, but also associated images and linked Microsoft Word, Postscript, and PDF files. The lack of images in WT10g has long been a problem for relevance assessment; furthermore, images would be needed by the Interactive Track who wished to use the collection as well. The inclusion of other document formats reflected the wealth of textual information on the current web that is not in HTML; indeed, more than half the crawl (54GB) was PDF.

Nick Craswell at CSIRO processed the crawl into what came to be called the .GOV collection.[1] Documents were truncated to 100KB, assembled into bundles, and assigned document identifiers. ASCII text was extracted from Word, Postscript, and PDF files using freely-available tools, in order to provide a standard baseline text for participants to use. The resulting collection contains 1.2 million textual documents and is 18.1GB in size. Images and binaries are available from CSIRO separately from the text portion of the collection.

Note that aside from being new and domain-specific, the .GOV collection was constructed differently than WT10g in two important ways. First, rather than creating a collection by selecting whole sites from a larger crawl, .GOV is essentially the crawl itself with minimal post-processing. Second, about 200,000 documents in .GOV are non-HTML pages, linked to by web pages but not themselves linking to anything. While PDF files can contain hyperlinks, in practice this is not common, and in any event the text extraction process would destroy any hyperlinks present.

5.1 Link Degree in .GOV

Figure 3 shows that the distributions for link degrees in .GOV are very similar to those in WT10g and in larger crawls. Out-links have a steeper curve (exponent 2.713) than in WT10g, but in fact this curve is actually closer to the distribution in the Altavista crawls. .GOV also seems to exhibit

[1] see `http://www.ted.cmis.csiro.au/TRECWeb/govinfo.html`

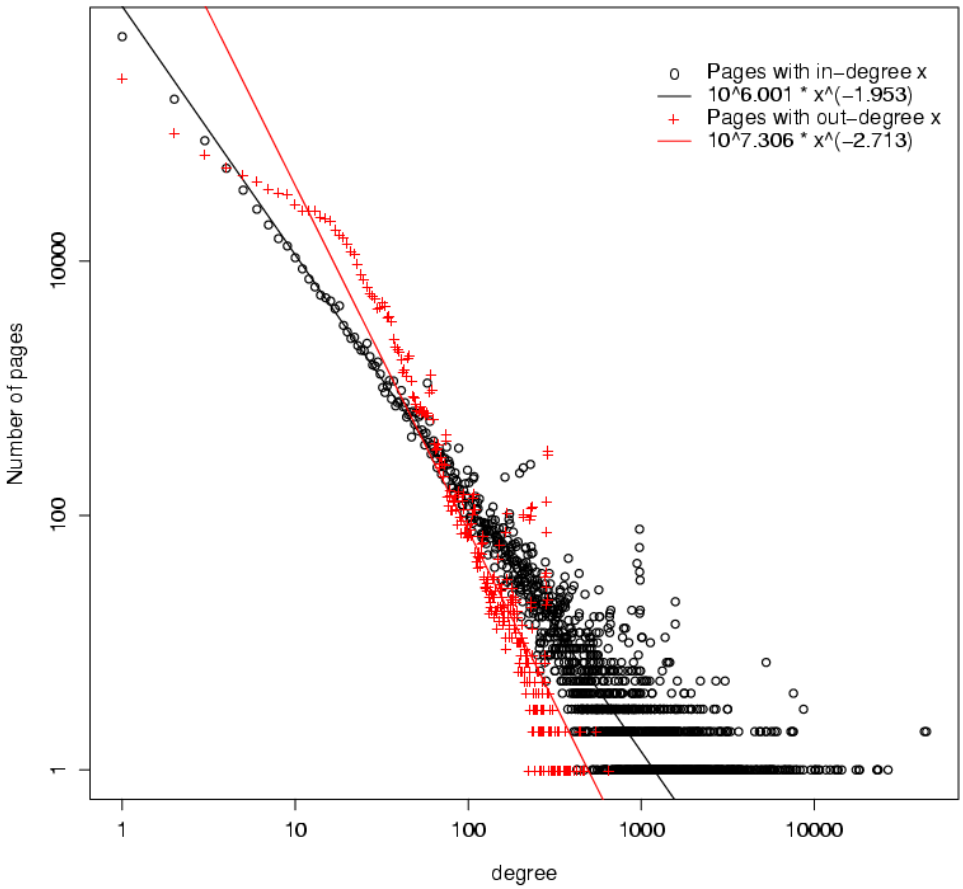

Figure 3: In- and out-degree distributions in .GOV.

some spiking behavior in the middle ranks, observed in the Altavista study but not in WT10g. These spikes come from very large hubs, and likely reflect the fact that .GOV is closer to an actual crawl than WT10g.

5.2 Connected Components in .GOV

The connected components distributions for .GOV, shown in Figure 4, are much different than those of WT10g. Whereas in WT10g, 91% of pages are connected in a single weak component, the large weak component in .GOV contains all but 154 pages! The strong components in the graph do follow a power law distribution, but the largest strong component contains 73.2% of the collection. This indicates either that the `.gov` domain is extremely well connected by large hub sites, or that the crawl was initiated from one or a few large hubs. In email conversations prior to crawling, it was proposed to seed the crawl from several thousand `.gov` discovered in an earlier Waterloo crawl; however, the effect of a large seed set might be minimal if all the seeds were in the strong component.

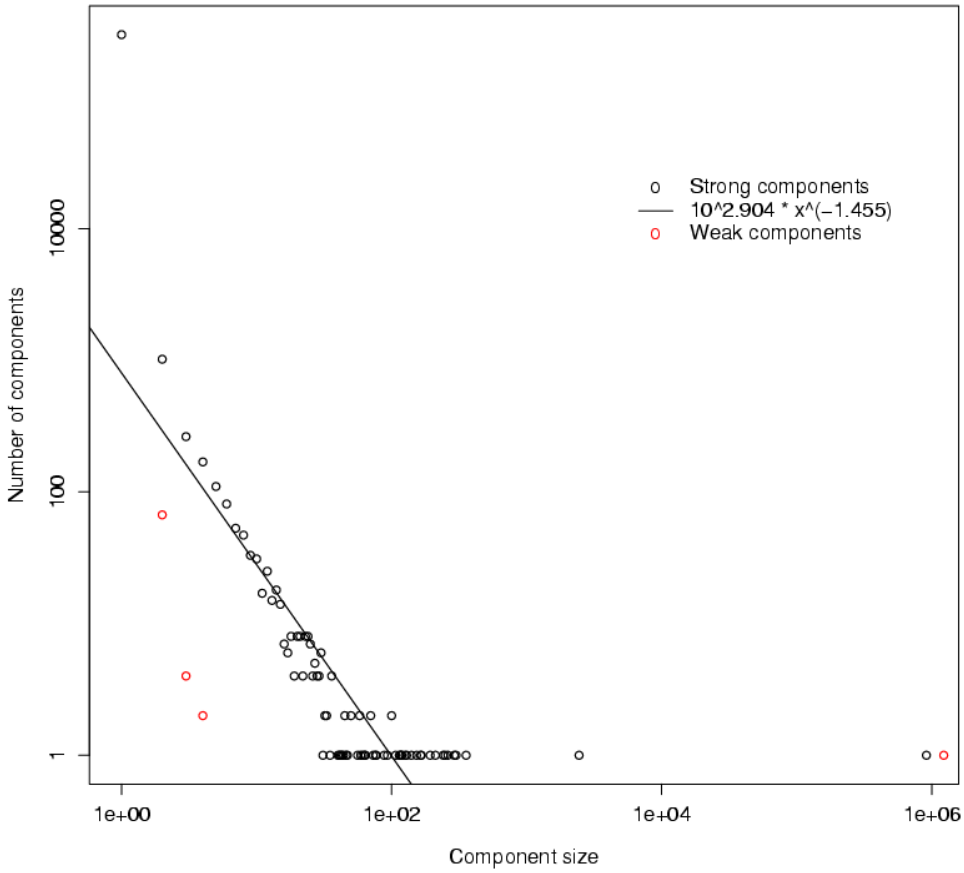

Figure 4: Distributions of strongly and weakly connected components in .GOV.

5.3 Bow-Tie Structure in .GOV

We completed our initial study of .GOV by following breadth-first traversals in the web graph from random starting pages, as we did for WT10g. Following the "bow-tie" analysis, we found

SCC	912,887 pages
IN	144,637
OUT	151,790
TENDRILS	17,570
DISCONNECTED	154

IN and OUT are nearly equal in size, supporting our earlier hypothesis that WT10g has a disproportionately larger OUT set as a result of the site selection process. When Bailey's procedure selects an entire server to include in WT10g, and when that server is part of the SCC, the non-SCC pages contained in that server are more likely to be pages pointed to from the SCC, rather than pages leading into the SCC. Only if the server is selected from IN are we likely to add more IN pages. In contrast, the .GOV crawl balances IN and OUT as is seen in larger crawls, simply because all pages were included.

The relatively small size of IN/OUT compared to SCC is again indicative of the unusually high connectivity in the crawl. Had we initiated the crawl in parallel from several disjoint starting points, we would see more balance between IN/OUT and SCC as a larger portion of the crawl was only weakly connected.

6 Conclusion

A frequent criticism of the test collection methodology in IR as applied to web search, is that the collections are not realistic, and thus conclusions do not generalize to the web. We have shown in fact that WT10g and .GOV, two web test collections used in TREC, structurally resemble much larger web crawls.

The WT10g collection was constructed by selecting whole servers from a larger crawl so as to encourage inter-server connectivity while incorporating complete sites. As a result, the collection is slightly biased towards pages linked to from the strong component of the larger crawl, at the expense of sites which can't be reached by browsing from the SCC, but which nevertheless link into it. This bias might be reduced by changing the selection algorithm to include more small sites. This would increase the chances of including sites in the IN set, without unbalancing the rest of the collection.

The .GOV collection is an entire crawl performed in the beginning of 2002 within the specific domain of US Government web sites.[2]

As an "unedited" crawl, .GOV reflects a balance of pages leading to and from the central component which has been observed in other larger crawls. However, the .GOV collection is much more closely connected than WT10g, with nearly all pages contained within a single weak component. This indicates either that the crawl was seeded from a small set of large hubs, or that the seeds were all within the strong component, or that the `.gov` domain is much more highly connected than the larger web.

Which is better? .GOV clearly has the advantage of being new, domain-specific, and reflecting a wider range of content formats than were prevalent in the 1997 web. It is also clear that we can manipulate the degree of connectivity using the WT10g selection algorithm on a larger crawl. We cannot yet conclude whether the connectivity of .GOV reflects the domain, or has a measurable impact on retrieval effectiveness.

Structure is only one aspect of the web which should be reflected in our test collections. The collection-building process has been refined over the years to ensure that the data resembles the task, and that a diversity of user information needs are represented by the search topics. With the advent of web test collections, hypertext structure must also be considered. There are other characteristics that differentiate the web from our earlier collections, adversarial content ("spam") being chief among them. To our knowledge, no one has yet measured spam quantity in a test collection.

Acknowledgments

We are grateful to the attendees of SIGIR 2002, and in particular David Hawking and Andrei Broder, for their insightful and illuminating comments.

[2] Actually, several US states and some cities have `.gov` domains, so the collection is not entirely within the Federal Government.

References

[1] Peter Bailey, Nick Craswell, and David Hawking. Engineering a multi-purpose test collection for web retrieval experiments. *Information Processing and Management*, to appear.

[2] Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener. Graph structure on the web. In *Proceedings of the 9th International WWW Conference*, pages 309–320, Amsterdam, The Netherlands, May 2000.

[3] David M. Pennock, Gary W. Flake, Steve Lawrence, Eric J. Glover, and C. Lee Giles. Winners don't take all: Characterizing the competition for links on the web. *Proceedings of the National Academy of Sciences*, 99(8):5207–5211, April 2002.

[4] Amit Singhal and Marcin Kaszkiel. A case study in web searching using TREC algorithms. In *Proceedings of the 10th International World Wide Web Conferenece*, pages 708–716, Hong Kong, May 2001.