Web Spam Detection with Anti-Trust Rank

Vijay Krishnan
Computer Science Department
Stanford University
Stanford, CA 4305
vijayk@cs.stanford.edu

Rashmi Raj
Computer Science Department
Stanford University
Stanford, CA 4305
rashmi@cs.stanford.edu

ABSTRACT

Spam pages on the web use various techniques to *artificially* achieve high rankings in search engine results. Human experts can do a good job of identifying spam pages and pages whose information is of dubious quality, but it is practically infeasible to use human effort for a large number of pages. Similar to the approach in [1], we propose a method of selecting a seed set of pages to be evaluated by a human. We then use the link structure of the web and the manually labeled seed set, to detect other spam pages. Our experiments on the WebGraph dataset [3] show that our approach is very effective at detecting spam pages from a small seed set and achieves higher precision of spam page detection than the Trust Rank algorithm, apart from detecting pages with higher pageranks, on an average.

1. INTRODUCTION

The term Web Spam refers to the pages that are created with the intention of misleading a search engine [1]. In order to put the tremendous amount of information on the web to use, search engines need to take into account the twin aspects of relevance and quality. The high commercial value associated with a web page appearing high on the search results of popular search engines, has led to several pages attempting *adversarial IR* i.e. using various techniques to achieve higher-than-deserved rankings. There is also a huge Search Engine Optimization (SEO) and Adversarial IR industry involved with attempting to find out a search engine's scoring function and *artificially* making a webpage appear in the top results to various queries on a search engine. Though it is not difficult for a human expert to recognize a spam web page, it is a challenging task to automate the same, since spammers are constantly coming up with more and more sophisticated techniques to beat search engines.

It has been found that a good percentage of web pages are spam. Search Engine spamming can be divided into two broad categories: Term Spamming and Link Spamming. Term spamming refers to manipulating the text of web pages in order to appear relevant to queries. Term Spamming can be achieved by various techniques including repetition of few specific terms, dumping a large number of unrelated terms, weaving spam terms at random positions and gluing together sentences and phrases from different sources. Link spamming refers to creating link structures that boost page rank or hubs and authorities class. A very common example of link spamming is boosting page rank value of a page by creating *link farms*, where webpages mutually reinforce each others' pagerank. Link spamming also includes boosting rank value by putting links from accessible pages to the spam page, such as posting web links on publicly accessible blogs.

Recent work [1], addressed this problem by exploiting the intuition that good pages i.e. those of high quality are very unlikely to point to spam pages or pages of low quality. They propagate *Trust* from the seed set of good pages recursively to the outgoing links. However, sometimes spam page creators manage to put a link to a spam page on a good page, for example by leaving their link on the comments section of a good page. Thus, the trust propagation is *soft* and is designed to attenuate with distance. The Trust Rank approach thus starts with a seed set of trusted pages as the *teleport set* [2] and then runs a biased page-rank algorithm. The pages above a certain threshold are deemed trustworthy pages. If a page has a trust value below a chosen threshold value then it is marked as spam.

In our work, we exploit the same intuition, in a slightly different way. It follows from the intuition of [1] that it is also very unlikely for spam pages to be pointed to by good pages. Thus we start with a seed set of spam pages and propagate *Anti Trust* in the reverse direction with the objective of detecting the spam pages which can then be filtered by a search engine.

We find that on the task of finding spam pages with high precision, our approach outperforms Trust Rank. We also empirically found that the average page-rank of spam pages reported by Anti-Trust rank was typically much higher than those by Trust Rank. This is very advantageous because filtering of spam pages with high page-rank is a much bigger concern for search engines, as these pages are much more likely to be returned in response to user queries.

1.1 Our Contributions

- We introduce the Anti-Trust algorithm with an intuition similar to [1], for detecting untrustworthy pages.
- We show that it is possible to use a small seed set of manually labeled spam pages, and automatically detect several spam pages with high precision.
- We propose a method for selecting seed sets of pages to be manually labeled.
- We experimentally show that our method is very effective both at detecting spam pages as well as detecting spam pages with relatively high pageranks.

2. PRELIMINARIES

2.1 Web Graph Model

The web can be modeled as a directed graph G = {V,E} whose nodes correspond to static pages (V) on the web, and whose edges correspond to hyperlinks (E) between these pages. The web graph (G) is massive containing billions of nodes and edges. In addition, G is *dynamic* or *evolving*, with nodes and edges appearing and disappearing over time.

In the web graph, each page has outgoing links referred to as *outlinks* and incoming links referred to as *inlinks*. The number of inlinks of a web page is called its *indegree* and the number of outgoing links is referred as *outdegree* of the page. Several studies on the analysis of the structure of web graph has shown that these links exhibit a power-law degree distribution. One study [12] models the structure of the web as a Bow-tie structure. In this model, the majority of the web pages are a strongly connected graph. Some pages do not have inlinks called *unreferenced* pages. Pages without any outlink are referred as *nonreferencing* pages. Also, pages that do not have either inlink or outlink are called as *isolated* pages. Mathematically, the graph structure can be encoded as a matrix where

$$G[i,j] = \begin{cases} 1 & if\ i\ connects\ to\ j \\ 0 & Otherwise \end{cases}$$

In addition, *transition matrix*(T) and *inverse transition matrix* (I) captures the outdegree and indegree of the web graph and they can be defined as:

Transition Matrix:

$$T[i,j] = \begin{cases} 1/outdegree(j) & if\ j\ connects\ to\ i \\ 0 & if\ j\ does\ not\ connect\ i \end{cases}$$

Inverse Transition Matrix:

$$I[i,j] = \begin{cases} 1/indegree(j) & i\ connects\ to\ j \\ 0 & if\ i\ does\ not\ connect\ j \end{cases}$$

2.2 Biased Page Rank

Page Rank [13] is one of the most popular link based methods to determine a page's global relevance or importance. Page rank assigns an importance score (page rank) proportional to the importance of other web pages which point to it. While page rank is a good approach to measure the relevance of a page, it is also vulnerable to adversarial IR, by way of link spamming, which can enable web pages to achieve higher than deserved scores.

Page rank r is defined as the first eigenvector of the matrix A where A is defined as follow:

$$A_{ij} = \beta T_{ij} + (1-\beta)/N$$

where T is the transition matrix,
N is the total number of web pages and
β is a decay factor and $0<\beta<1$.

Hence the page rank of a web page is sum of:

- scores which it gets from the pages pointing to it.
- A constant term which is the probability of *teleporting* to the page, which is same for all pages.

Another interpretation of page rank is based on the random surfer model. In this model, a random surfer picks a page on the web uniformly at random to start the walk. Suppose at time t, the random surfer is at page j. At time t+1, with probability β we randomly traverses along one of the outgoing links and walk to the new page and with probability 1-β, picks a page on the web at random with uniform probability and teleports to it. If the page has no outlinks, he teleports to a random page on the web with equal probability for all pages. The Page Rank of page P is the *steady state* probabilty that the random surfer is at page P. The same notion can be extended to many random surfers model where page rank of a page P is the fraction of random surfers that are expected to be at page p.

While page rank assigns a score proportional to generic popularity of a page, *biased page rank* or *topic-specific page rank* [2] measures the popularity within a topic or domain. Here the equivalent random surfer model is as follows. When the random surfer teleports, he picks a page from a set S of web pages which is called the teleport set. The set S only contains pages that are relevant to the topic (E.g., Open Directory (DMOZ) pages for a given topic). Corresponding to each teleport set S, we get a different rank vector r_S. In matrix representation:

$$A_{ij} = \begin{cases} \beta T_{ij} + (1-\beta)/|S| & if\ i\ to\ S \\ \beta T_{ij} & otherwise \end{cases}$$

where A is a stochastic matrix as before. Here, we have weight all pages in the teleport set S equally, but we could weight them differently if we wish.

3. TRUST RANK

The Trust Rank algorithm proposed in [1], is an approach to find differentiate trustworthy pages from spam pages. The algorithm involves running a biased pagerank algorithm with the teleport set being a manually labeled set of trustworthy pages. This work exploits the intuition that good pages are unlikely to point to spam pages. Thus the approach looks to propagate *Trust* along forward link, attenuating with distance. Running the biased pagerank as mentioned achieves this effect. Finally, a thresohold value is chosen and all pages below the threshold are marked as spam pages.

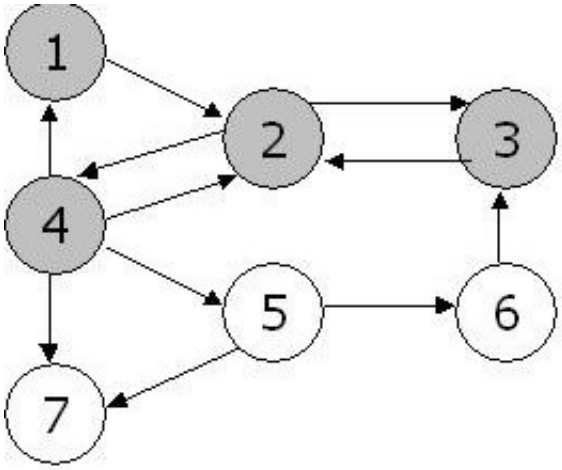

Figure 1: A toy-webgraph with good and spam pages used to illustrate the intuition behind Trust Rank and our Anti-Trust Rank algorithm. The white nodes represent good pages, while the others represent spam pages. Instances of edges from good pages to spam pages(node 6 to node 3) are relatively rare.

3.1 Inverse PageRank

Inverse page-rank is computed by reversing the in-links and out-links in the webgraph. In other words, it merely involves running pagerank on the transpose of the web graph matrix. Thus, a high inverse page-rank indicates that one can reach a huge number of pages in a few hops along outlinks starting with the given page. Thus, this metric was found to be useful in selecting a seed set of pages in the Trust Rank algorithm.

3.2 Selecting the Seed Set of Spam pages

It was pointed out in [1] that there are two important issues in selecting the seed set of pages in the Trust Rank algorithm

- It is important to choose pages in the seed set, which are *well connected* to other pages and can therefore propagate trust to many pages quickly. Since the Trust Rank approach makes trust flow along the outlinks of a pages, it was therefore important to choose pages that had a large number of outlinks. To generalize the notion of being able to reach a large number of pages with a small number of hops, [1] pointed out that pages with a high inverse pagerank would do very well at satisfying this criterion.
- It is generally more important to ascertain goodness of pages with higher pageranks, since these pages will typically appear high in search query results. It was observed in [1] that choosing pages with high pageranks would be more useful towards this goal, since the pages pointed to by high page rank pages are likely to have high pagerank themselves.

Thus, choosing pages with high inverse pagerank is good for the first goal, while choosing pages with a high pageranks is good for the second. An appropriate tradeoff could be done.

4. ANTI-TRUST RANK

Our approach is broadly based on the same *approximate isolation* principle [1], i.e it is rare for a good page to point to a bad page. This principle also implies that the pages pointing to spam pages are very likely to be spam pages themselves. The Trust Rank algorithm started with a seed set of trustworthy pages and propagated *Trust* along the outgoing links. Likewise, in our Anti-Trust Rank algorithm, *Anti-Trust* is propagated in the reverse direction along incoming links, starting from a seed set of spam pages. We could classify a page as a spam page if it has Anti-Trust Rank value more than a chosen threshold value. Alternatively, we could choose to merely return the top n pages based on Anti-Trust Rank which would be the n pages that are most likely to be spam, as per our algorithm.

Interestingly, both Trust and Anti-Trust Rank approaches need not be used for something very specific like detecting link spam alone. The *approximate isolation* principle can in general enable us to distinguish *good* pages from the not-so-good pages such as pages containing pornography and those selling cheap medication. Thus, for the purpose of our work we consider pages in the latter category as spam as well.

4.1 Selecting the Seed Set of Spam pages

We have similar concerns to [1], with regard to choosing a seed set of spam pages. We would like a seed set of pages from which *Anti-Trust* can be propagated to many pages with a small number of hops. We would also prefer if a seed set can enable us to detect spam pages having relatively high pageranks. In our approach, choosing our seed set of spam pages from among those with high pagerank satisfies both these objectives.

Pages with high pagerank are those from which several pages can be reached in a few hops if we go backward along the incoming links. Thus this helps in our first objective. Also, having high pagerank pages in our seed set makes it somewhat more probable that the spam pages we detect would also have high pageranks, since high pageranks pages often get pointed to by other pages with high pagerank. We therefore select our seed set of spam pages from among the pages with high pagerank. This helps us nail our twin goals of fast reachability and detection of spam pages with high pagerank.

4.2 The Anti-Trust Algorithm

- Obtain a seed set of spam pages labeled by hand. Assign pages with high pageranks for labeling by a human in order to get a seed set containing high pagerank pages.
- Compute T to be the Transpose of the binary webgraph matrix.
- Run the biased pagerank algorithm on the matrix T, with the seed set as the teleport set.
- Rank the pages in descending order of pagerank scores. This represents an ordering of pages based on estimated Spam content. Alternatively, set a threshold value and declare all pages with scores greater than the threshold as spam.

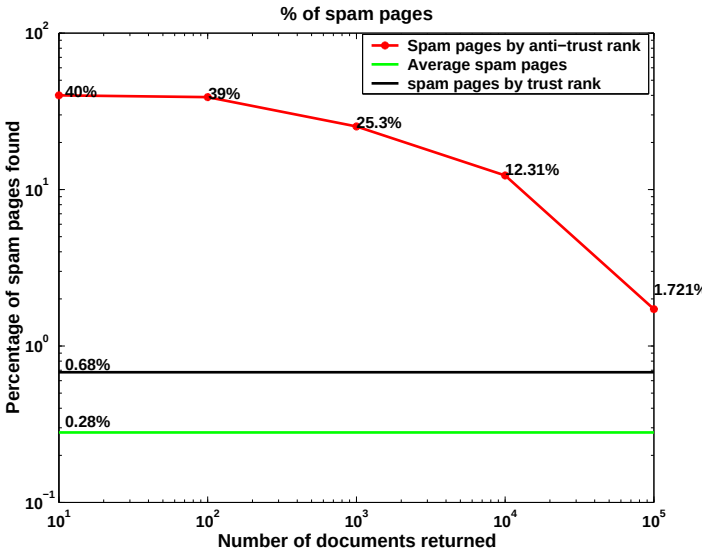

Figure 2: Comparison of the precisions of Anti-Trust Rank and Trust Rank at various levels of recall, against the naive baseline of total percentage of spam documents in the corpus. It can be seen what Anti-Trust Rank does significantly better than Trust Rank which is in turn clearly better than the naive baseline.

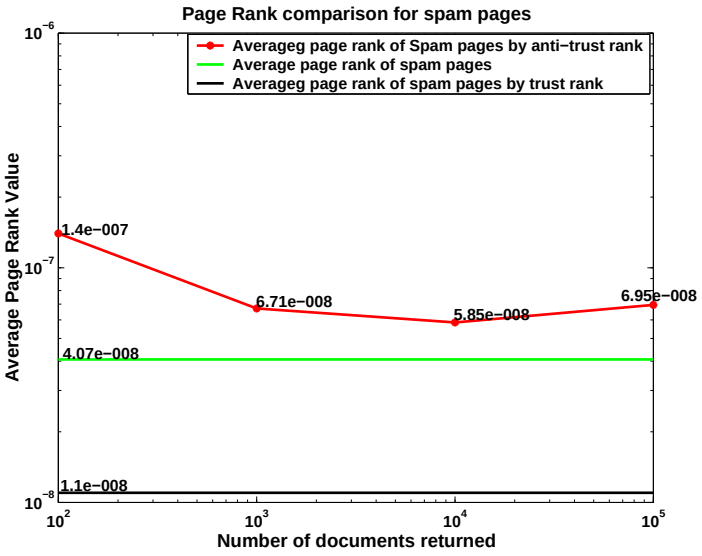

Figure 3: Comparison of the page ranks of spam pages returned by Anti-Trust Rank and Trust Rank at various levels of recall, against the baseline of average page rank of spam pages in the corpus. It can be seen that while Anti-Trust Rank returns spam pages with higher-than-average page ranks, Trust Rank returns spam pages with clearly lower-than-average page ranks.

4.3 Example to illustrate Anti-Trust Rank computation

Initially, the Anti-Trust Rank value is equally distributed among all the pages of seed set. The subsequent Anti-Trust Rank computation is simply the Inverse-Page rank computation with the teleport set chosen to be our seed set.

In the above example in figure 1, Lets assume that seed set of spam pages is 1. Thus Anti-Trust would propagate to page 4, from which it would propagate to node 2 and subsequently to node 3. As it can be expected, the Anti-Trust rank would constantly attenuate with distance from the seed set, as a result of which the good nodes would get relatively low Anti-Trust scores, in the given example.

5. EXPERIMENTS

5.1 Dataset

We ran our experiments on the WebGraph dataset, [3]. We chose data corresponding to a 2002 crawl of the "uk" domain containing about 18.5 millions nodes and 300 million links.

5.2 Evaluation Metric

Clearly, the only perfect way of evaluating our results is to manually check if the pages with high Anti-Trust score are indeed spam pages and vice-versa. It was observed in [1] that this process is very time consuming and often hard to do in practice.

We however circumvented this problem by coming up with a heuristic which in practice selects spam pages with nearly 100% precision and also a recall which is a reasonable fraction of the set of true spam pages, on our dataset.

The Heuristic: We compiled a list of substrings whose presence in a URL almost certainly indicated that it was a spam page, on our dataset. As one would expect, our list contained strings like viagra, casino and hardporn. Thus, this heuristic enables us to measure the performance of our Anti-Trust Rank algorithm and compare it against the Trust Rank algorithm with a good degree of reliability. It seems reasonable to expect that the relative scores obtained by the spam detection algorithms with the evaluation being heuristic based would be representative of their actual performance in spam detection, since our heuristic has a pretty reasonable recall and is independent of both the Trust Rank and Anti-Trust Rank algorithms and would not give the algorithms we are looking at, an unfair advantage.

As per this heuristic, out of the 18,520,486 pages, 0.28 % i.e. 52,285 were spam pages.

5.3 Choosing the Seed Set

We chose the top 40 pages based on page rank from among the URLs that got flagged as spam by our heuristic. For comparing with Trust-Rank we picked the top 40 pages based on inverse page rank, among the pages marked non-spam by our heuristic. We also manually confirmed that the seed sets were indeed spam in the former cases and trustworthy pages in the latter case. We also studied the effect of increasing the seed set size in Anti-Trust rank. We found that we could benefit substantially from a larger seed set. Also we used the common α value of 0.85 i.e. the probability of teleporting to a seed node was 0.15.

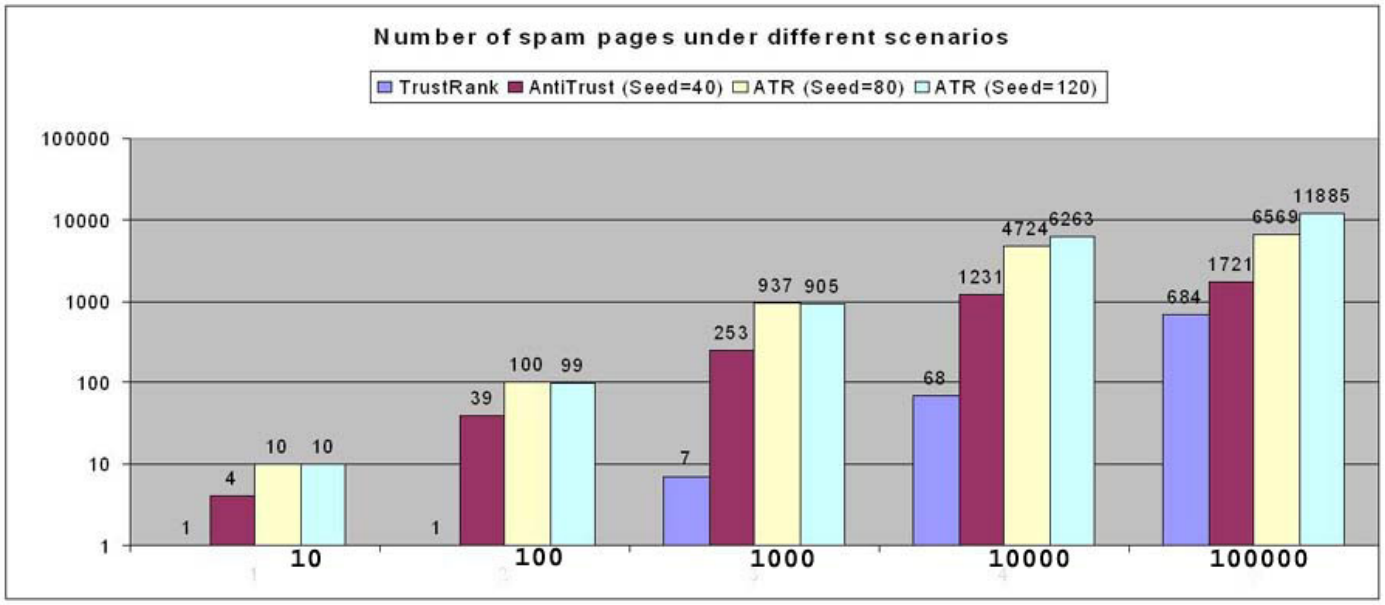

Figure 4: Comparison of the performance of Trust Rank with a seed set of 40 pages against Anti-Trust rank with 40, 80 and 120 pages respectively. The X-axis represents the number of documents selected having the highest Anti-Trust and lowest Trust scores. The Y-axis depicts, how many of those documents were actually spam(as measured by our heuristic). We observe that Anti-Trust rank typically has a much higher precision of reporting spam pages than Trust rank. Also, Anti-Trust rank benefits immensely with increasing seed-set size.

5.4 Results and Analysis

From figure 2, we can see that both Anti-Trust Rank and Trust Rank are significantly better than the naive baseline corresponding to a random ordering of the pages, for which the precision of reporting spam would merely be the percentage of spam pages in the corpus. However we also see that Anti-Trust rank typically does much better than Trust Rank at different levels of recall.

This is intuitive because Trust Rank is capable of reporting with high confidence that the pages reachable in short paths from its seed set are trustworthy, while it cannot be expected to say anything with high confidence about pages that are far away from the seed set. Likewise our Anti-Trust Rank approach is capable of reporting with high confidence that the pages from which its seed set can be reached in short paths are untrustworthy.

Also, from figure 3, we find that the average rank of spam pages returned by Trust Rank is even lower than the average page rank of all spam pages. Anti-Trust rank however manages to return spam pages whose average page rank is substantially above the overall average page rank of all spam pages. The ratio of average page ranks of spam pages reported by Anti-Trust Rank and Trust rank was over 6:1 for different levels of recall. Thus, Anti-Trust rank has the added benefit of returning spam pages with high page rank, despite the fact that it has a significantly higher precision than Trust Rank at all levels of recall that we explored.

This is intuitive because, by starting with seed spam pages of high page rank, we would expect that walking backward would lead to a good number of spam pages of high page rank.

Figure 4 compares the performance of Trust Rank against Anti-Trust rank with an equal seed size of 40 and also show performance of Anti-Trust Rank with larger seed sets of 80 and 120 respectively. It shows the precisions achieved by Trust Rank and Anti-Trust Rank at various levels of recall such as 10, 100, 1000, 10000 and 100000 web pages. We find that apart from achieving better precision of spam page detection than Trust Rank for the same seed set size, increasing the seed set size in Anti-Trust rank can lead to dramatic improvement in performance. W

An analysis of success of these algorithms in picking trustworthy pages would not be very useful. This is because our corpus has over 99% trustworthy pages, and it would be very hard to conclude anything about the performance of these algorithms given that they would all attain a precision of well over 99% and would differ merely by a tiny fraction of a percent.

6. RELATED WORK

The taxonomy of web spam has been well defined by [4]. There are many pieces of work on combating link spam. The problem of trust has also been studied in other distributed fields such as P2P systems [5]. Similar ideas have also been used to identify email spam [6]. Other approaches rely on detecting anomalies in statistics gathered through web crawls [9]. Approaches such as [10], focus on higher-level connectivity between sites and between top-level domains for identifying link spams. The data mining and web

mining community has also worked on identifying link farms. Various farm structures and alliances that can impact ranking of a page has been studies by [7]. [11] identifies link farm spam pages by looking for certain patterns in the webgraph structure.

7. CONCLUSION AND FUTURE WORK

We have proposed the Anti-Trust Rank algorithm, and shown that it outperforms the Trust Rank algorithm at the task of detecting spam pages with high precision, at various levels of recall. Also, we show that our algorithm tends to detect spam pages with relatively high pageranks, which is a very desirable objective.

It would be interesting to study the effect of combining these both the Trust Rank and Anti-Trust Rank methods especially on data containing a very high percentage of spam pages. It would also be interesting to attempt combining these link-based spam detection techniques with techniques that take text into account, such as text classifiers trained to detect spam pages.

Acknowledgements

We would like to thank Zoltán Gyöngyi, Dr. Anand Rajaraman and Dr. Jeffrey D. Ullman for helpful discussions. We would also like to express our gratitude to Paolo Boldi and Sebastiano Vigna whose compressed WebGraph dataset, with its useful Java API's made it very convenient for us to run experiments on a significant sized subgraph of the web.

8. REFERENCES

[1] Combating Web Spam with Trust Rank. Z. Gyöngyi, H. Garcia-Molina and J. Pedersen. In VLDB 2004.

[2] Topic-sensitive Page Rank. Taher Haveliwala. In WWW 2002.

[3] The WebGraph dataset. Online at: *http://webgraph-data.dsi.unimi.it/*

[4] Web Spam Taxonomy. Zoltán Gyöngyi, Hector Garcia-Molina. First International Workshop on Adversarial Information Retrieval on the Web (at the 14th International World Wide Web Conference), Chiba, Japan, 2005.

[5] The EigenTrust algorithm for reputation management in P2P networks. S. Kamvar, M. Schlosser, and H. Garcia-Molina. In Proceedings of the Twelfth International Conference on World Wide Web, 2003.

[6] Improving Spam Detection Based on Structural Similarity. Luiz H. Gomes, Fernando D. O. Castro, Rodrigo B. Almeida,Luis M. A. Bettencourt, Virgilio A. F. Almeida, Jussara M. Almeida

[7] Link Spam Alliances. Zoltán Gyöngyi, Hector Garcia-Molina. . 31st International Conference on Very Large Data Bases (VLDB), Trondheim, Norway, 2005.

[8] The Connectivity Sonar: Detecting Site Functionality by Structural Patterns. E. Amitay, D. Carmel, A. Darlow, R. Lempel and A.Soffer. In 14th ACM.

[9] Spam, Damn Spam, and Statistics. Dennis Fetterly, Mark Manasse and Marc Najork. Seventh International Workshop on the Web and Databases (WebDB 2004), June 17-18, 2004, Paris, France.

[10] Links to Whom: Mining Linkage between Web Sites. K. Bharat, B. Chang, M. Henzinger, and M. Ruhl. In 2001 IEEE International Conference on Data Mining, Nov. 2001.

[11] Identifying Link Farm Spam Pages. Baoning Wu, Brian D. Davison. WWW 2005, May 1014, 2005, Chiba, Japan.

[12] PageRank Computation and the Structure of the Web: Experiments and Algorithms. A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins and J. Wiener. Proc. WWW9 conference, 309-320, May 2000.

[13] The PageRank citation ranking: Bringing order to the web. L. Page, S. Brin, R. Motwani and T. Winograd. Technical Report, Stanford University, 1998.