# Anti-Trust Rank: Fighting Web Spam

**Ms. Jyoti Pruthi[1] and Dr Ela Kumar[2]**

**[1]MCA, Manav Rachna College of Engineering,
Faridabad, India-121001**

**[2]SICT, Gautam Buddha University,
Gr Noida, India**

**Abstract**

The Web is both an excellent medium for sharing information as well as an attractive platform for delivering products and services. This platform is, to some extent, mediated by search engines in order to meet the needs of users seeking information. Search engines are the "dragons" that keep a valuable treasure: information [8]. Given the vast amount of information available on the Web, it is customary to answer queries with only a small set of results (typically 10 or 15 pages at most). Search engines must then rank Web pages, in order to create a short list of high-quality results for users. Web spam can significantly deteriorate the quality of search engine results. Thus there is a large incentive for commercial search engines to detect spam pages efficiently and accurately. Here we present the main techniques recently introduced for Web Spam detection.

***Keywords:*** *Web Graph Model, Biased Page Rank, Trust Rank, Anti Trust Rank.*

## 1. Introduction

*Web spam* refers to hyperlinked pages on the WorldWideWeb that are created with the intention of misleading search engines. With the search engines' increasing importance in the people's life, there are more and more attempts to mischievously influence the page rankings. This kind of action called web spamming is always illegal, since it misleads both search engines and users seriously. Web spamming, the practice of introducing artificial text and links into web pages to affect the results of searches. It is also a serious problem for users because they are not aware of it and they tend to confuse trusting the search engine with trusting the results of a search.

Furthermore, it has a negative economic and social impact on the whole web community. It has been found that a good percentage of web pages are spam. Spammers are playing tricks on the search engines by all means, for example, term spamming, link spamming, cloaking and redirection [1].

For example, a web site may spam the web by adding thousands of keywords to its home page, often making the text invisible to humans. A search engine will then index the extra keywords, and return the web page as an answer to queries that contain some of the keywords. Another web spamming technique is the creation of a large number of bogus web pages, all pointing to a single target page. Since many search engines take into account the number of incoming links in ranking pages, the rank of the target page is likely to increase, and appear earlier in query result sets. For instance, consider a cluster of web sites that link to each other's pages repeatedly. These links may represent useful relationships between the sites, or they may have been created with the express intention of boosting the rank of each other's pages. In general, it is hard to distinguish between these two scenarios.

For all the reasons we have mentioned, Web spam detection is a challenging problem. Since spammers are constantly coming up with more and more sophisticated techniques to beat search engines.

Many anti-spamming techniques have been proposed so far [2, 3, 4, 5, 6, 7]. Trust Rank [2] improves the PageRank by using good seeds. It can effectively demote the pages that adopt link spamming tricks. Baoning Wu and Brian D. Davison propose algorithms for detecting link farms automatically by first generating a desirable seed set and then expanding it [5]. In actual fact, almost all of these biased ranking algorithms employ a seed set and this set plays an important role in identifying web spam.

## 2. Related Work

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.

The taxonomy of web spam has been well defined by Zolt´an Gy¨ongyi, Hector Garcia-Molina [9]. 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 [10]. Other approaches rely on detecting anomalies in statistics gathered through web crawls [11]. 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 have been studies by Zolt´an Gy¨ongyi, Hector Garcia-Molina [12]. Baoning Wu, Brian D. Davison identifies link farm spam pages by looking for certain patterns in the webgraph structure.

In our work, we exploit the same intuition, in a slightly different way. 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 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 ages are much more likely to be returned in response to user queries.

## 3. Preliminaries

### 3.1 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 have shown that these links exhibit a power-law degree distribution.

One study [14] 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 non-referencing 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 Eq (i) where

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

In addition, transition matrix (T) Eq (ii) and inverse transition matrix (I) Eq (iii) captures the outdegree and indegree of the web graph and they can be defined as: Transition Matrix.

Transition Matrix:

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

Inverse Transition Matrix:

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

### 3.2 Biased Page Rank

Page Rank [15] 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:

$$Aij = \beta Tij + (1 - \beta)/N \qquad \text{(iv)}$$

where T is the transition matrix,

N is the total number of web pages and
$\beta$ is a decay factor and $0\ \beta < 1$.
While page rank assigns a score proportional to generic popularity of a page, biased page rank or topic-specific page rank [16] 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. Corresponding to each teleport set S, we get a different rank vector.

In matrix Eq (v) representation:

$$Aij = \begin{cases} \beta Tij + (1 - \beta)/|S| & \text{if i to S} \\ \beta Tij & \text{otherwise} \end{cases} \qquad (v)$$

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.

## 4. Trust Rank

The Trust Rank algorithm is an approach to find differentiates trustworthy pages from spam pages [17]. 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 thresholds value is chosen and all pages below the threshold are marked as spam pages.

### 4.1 Inverse Page Rank

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.

### 4.2 Selecting the Seed Set of Spam pages

It was pointed out that there are two important issues in selecting the seed set of pages in the Trust Rank algorithm [17].

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.

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 [17] 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.

## 5. Antitrust Rank

Our approach is broadly based on the same approximate isolation principle [17], 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. Thus, for the purpose of our work we consider pages in the latter category as spam as well.

### 5.1 Selecting the Seed Set of Spam pages

We have similar concerns to [17], 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. We 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.

### 5.2 The AntiTrust Algorithm

1) Obtain a seed set of spam pages labeled by hand. Assign pages with high pageranks labeling by a human to get a seed set containing high pagerank pages.

Let N = {n, where n=0, 1, 2, ……}
n → spam page in the seed set

2) Compute T
Let S = matrix of binary webgraph
Then T = transpose of S or S'

3) Run the biased pagerank algorithm on the matrix T, with the seed set as the teleport set.

4) 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.

## 5.3 Example

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 example in figure 1, let's assume that seed set of spam pages is 1. Thus Anti-Trust would propagate to page 5, from which it would propagate to node 4 and subsequently to node 2 and then to 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. In the given example blue nodes represent spam pages and orange nodes represent the good pages.

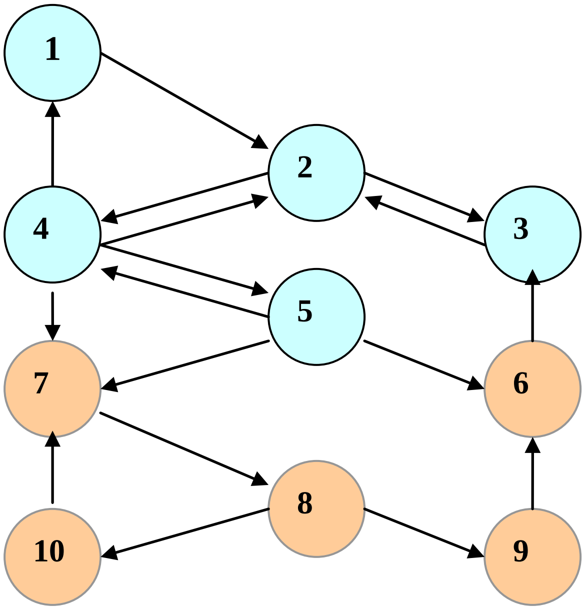


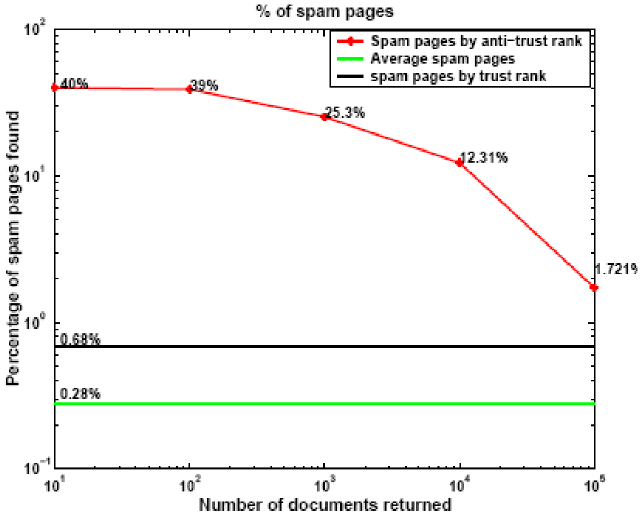
Fig.1 Web Graph with Spam and Good Pages

## 5.4 Results and Analysis

We ran our experiments on the WebGraph dataset, [18]. We chose data corresponding to a crawl of the "in" domain containing about 20 millions nodes and 400 million links. 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 [17] that this process is very time consuming and often hard to do in practice.

We however solve 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. 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. As per this heuristic, 0.28 % was spam pages.



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.

## 6. Conclusion

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.

## Acknowledgment

I express my sincere gratitude and acknowledgement towards Dr. Ela Kumar, Associate Professor, who guided me. It was her constant support and inspiration without which my efforts would not have taken this shape. I sincerely thank her for this, and seek her support for all my future endeavors.

## References

[1] Z. Gyongyi and H. Garcia-Molina. Web spam taxonomy. In *First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)”*, 2005.

[2] Z. Gy¨ongyi, H. Garcia-Molina, and J. Pedersen. Combating web spam with trustrank. In *vldb’2004: Proceedings of the Thirtieth international conference on Very large data bases*, pages 576–587. VLDB Endowment, 2004.

[3] V. Krishnan and R. Raj. Web spam detection with anti-trust rank. In AIRWeb’06, August 2006.

[4] B.Wu and K. Chellapilla. Extracting link spam using biased random walks from spam seed sets. In AIRWeb ’07: Proceedings of the 3rd international workshop on Adversarial information retrieval on the web, pages 37–44, New York, NY, USA, 2007. ACM.

[5] B.Wu and B. D. Davison. Identifying link farm spam pages. In WWW ’05: Special interest tracks and posters of the 14th international conference on World Wide Web, pages 820– 829, New York, NY, USA, 2005. ACM.

[6] B. Wu, V. Goel, and B. D. Davison. Propagating trust and distrust to demote web spam. In Proceeding of Models of Trust for the Web (MTW), May 2006.

[7] L. Zhang, Y. Zhang, Y. Zhang, and X. Li. Exploring both content and link quality for anti-spamming. In CIT ’06: Proceedings of the Sixth IEEE International Conference on Computer and Information Technology, page 37, Washington, DC, USA, 2006. IEEE Computer Society. Marco Gori and Ian Witten. The bubble of web visibility. Commun. ACM, 48(3):115–117, March 2005.

[8] Web Spam Taxonomy. Zolt´an Gy¨ongyi, Hector Garcia-Molina. First International Workshop on Adversarial Information Retrieval on the Web (at the 14th International World Wide Web Conference), Chiba, Japan, 2005.

[9] 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.

[10] 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.

[11] Link Spam Alliances. Zolt´an Gy¨ongyi, Hector Garcia-Molina. . 31st International Conference on Very Large Data Bases (VLDB), Trondheim, Norway, 2005.

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

[13] 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.

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

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

[16] Combating Web Spam with Trust Rank. Z. Gy¨ongyi, H. Garcia-Molina and J. Pedersen. In VLDB 2004.

**Jyoti Pruthi** is pursuing PhD (computer science).She received her MPhil (computer science) degree in 2008. She holds an MCA from MDU, Rohtak. She is currently working as Lecturer with Manav Rachna College of Engineering, Faridabad, MCA Department. Earlier, she worked with CSSL, Gurgaon as a Software Engineer. Her area of specialization includes Search Engine, Spamming & Anti-Spamming Techniques.

**Dr Ela Kumar** PhD (Artificial Intelligence) from University of Delhi. She holds an M.Tech (CSE) and B.E (ECE) from IIT, Roorkee. She is currently working as Associate Professor with Gautam Buddha University, Gr Noida. Earlier, she worked with CSSL, Gurgaon as a Software Engineer. Her current areas of specialization include Search Engine, Spamming & Anti-Spamming Techniques. She holds to her credit around 19 Years of experience of teaching and research. Besides publishing 30 research papers in journals of repute, she has written 3 books.