Pagerank Increase under Different Collusion Topologies

Ricardo Baeza-Yates
ICREA-Univ.Pompeu Fabra
and University of Chile
ricardo.baeza@upf.edu

Carlos Castillo
Dept. of Technology
Universitat Pompeu Fabra
carlos.castillo@upf.edu

Vicente López
Cátedra Telefónica
Universitat Pompeu Fabra
vicente.lopez@upf.edu

Abstract

We study the impact of collusion –nepotistic linking– in a Web graph in terms of Pagerank. We prove a bound on the Pagerank increase that depends both on the reset probability of the random walk ε and on the original Pagerank of the colluding set. In particular, due to the power law distribution of Pagerank, we show that highly-ranked Web sites do not benefit that much from collusion.

1 Introduction

This paper studies the effects of different linking topologies in the ranking function induced by the Pagerank algorithm [13]. The Pagerank algorithm receives as input an adjacency matrix $L_{N\times N}$, where N is the number of Web pages, and renormalizes each row of L to sum 1, generating a transition matrix A. This transition matrix is slightly modified by adding a "random jump", i.e.: a transition from each node to each of the other nodes using the uniform transition matrix – a matrix U such that $u_{ij} = 1/N$.

$$P = (1-\varepsilon)A + \varepsilon U \tag{1}$$

The Pagerank algorithm calculates the probabilities p_i of the stationary state of the Markovian process induced by matrix P. That is, the eigenvector x corresponding to the largest eigenvalue (which in the case of this matrix is $\lambda_1 = 1$) of the matrix P:

$$P^T x = x$$

The Pagerank value of a page is used as an estimator of the quality of a Web page based on properties of the Web graph. The rationale for this estimation is that a high quality page is a page with many in-links coming from other high-quality pages.

This algorithm is expected to work much better than simply counting in-links, as it might be more resistant to what is called a "Sybil attack" [5]. A Sybil attack is an attempt of altering a recommendation system by creating multiple identities; in this context, this means creating multiple pages pointing to a single page.

The resistance of Pagerank to a Sybil attack comes from the fact that the pages created for the attack can only inherit the reputation they are currently receiving. However, in the case of Pagerank, the minimum is not zero, as even without in-links a page gets a minimum score of $\frac{\varepsilon}{N}$.

The strategy of creating many pages pointing to a single page is actually used on the Web, in fact, currently there are thousands or millions of Web pages created specifically for the objective of deceiving the ranking function of search engines [6]. "Because the Web environment contains profit seeking ventures, attention getting strategies evolve in response to search engine algorithms. For this reason, any evaluation strategy which counts replicable features of web pages is prone to manipulation"[13].

To the best of our knowledge, Pagerank by itself is not used as the sole indicator of quality by any of the larger search engines, but it is still an important part of the ranking function of some of them.

In this paper:

- We present an analysis for collusion under a more general case than the one presented in [15], this is, we consider the original links that the colluding set has.
- We prove that for a single page there is always something to win by colluding with other pages.
- We prove that the expected returns from collusion are lower for highly-ranked pages.

The rest of this paper is organized as follows: Section 2 summarizes previous work in this area, and Section 3 presents an analysis predicting the increase of Pagerank by using a collusion strategy. Section 4 validates these predictions in a synthetic graph, and Section 5 studies a real Web graph. Finally, Section 6 presents our conclusions and avenues for future work.

2 Previous Work

Several authors have observed the presence of spam pages on the Web. Fetterly *et al.* [7] showed that most of the outliers when observing statistics of Web page collections are machine-generated spam pages –these pages may be designed both to increase citation counts and to provide multiple "doorway" pages. Hence, the divergence between the expected and the observed values can be used in some cases to detect spam pages.

Eiron *et al.* [6] studied a 100-million page sample and found that 11 of the top 20 URLs by Pagerank were pornographic, and in all cases the specific technique used was taking the Pagerank from random teleportation in many pages and concentrate it into a single page by using links.

Zhang *et al.*[15] study the following collusion strategy: pick a series of nodes with adjacent rankings, remove all their out links and add links from each node to the node before and after it in the list of nodes sorted by Pagerank. They also prove the following upper bound on any collusion strategy; let x_{orig} be the original Pagerank of a page, and x_{new} the Pagerank this page obtains after it colludes is:

$$\frac{x_{new}}{x_{orig}} < \frac{2}{\varepsilon}$$

They also prove that this bound is near $1/\varepsilon$ if $M \ll N$, as a typical value for ε is 0.15, the amplification factor is roughly 7. They do not take into account the starting Pagerank of the colluding set. We prove a tighter bound that shows that colluding works mostly for pages with low starting Pagerank.

Meyer [11] proved that if the second eigenvalue of an irreducible Markov chain is small, then the chain is not overly sensitive to small variations. Haveliwala and Kamvar [9] proved that the second eigenvalue of P is $1-\varepsilon$, therefore, a large ε produces a more stable matrix. Ng *et al.* [12] prove a similar result using a different approach: as long as ε is not too small, small variations in the matrix do not generate large variations of Pagerank.

Agogino and Ghosh [1] studied a reinforcement learning method for automatically finding a linking strategy for increasing the combined Pagerank of a set of domains. Their strategy relies on a utility function that considers the impact of every learner in the total Pagerank achieved by the colluding group.

Clausen [3] studied the cost of an attack on Pagerank considering that creating a new Web site requires a payment. In the same paper there is an analysis on how to lump pages together for Pagerank calculation, disregarding the internal link structure of each group.

In a recent article, Gyöngyi and Garcia-Molina [8] study optimal structures for "link spam farms" and combinations of them. A spam farm is an arrangement of links with the objective of increasing the ranking of a single target page. They prove that the optimal structure for a spam farm is a series of pages pointing to and only to the target page, while the target page points to some of all of them. In this optimal structure, if there are other external, "hijacked" links, they should also point to the target page.

3 Impact of Collusion in Pagerank

A group of nodes can collude to get a higher Pagerank by manipulating the out-links of the group. We will assume that the group's objective is to maximize its *total* Pagerank value.

Let N be the total number of nodes in Web graph G, M the number of colluding nodes in sub-graph G' – we will assume $M \ll N$. Let x be the total Pagerank of the colluding nodes, so $1-x$ is the total Pagerank of the rest of the graph. All the links in this graph are shown in Figure 1.

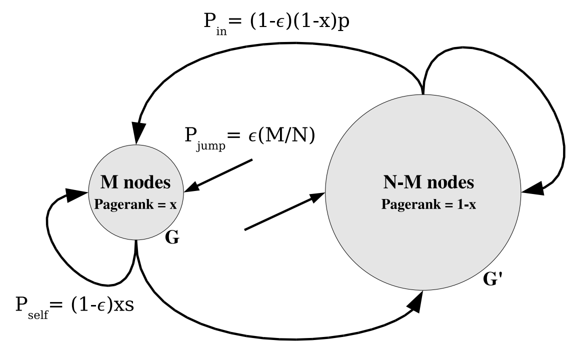

Figure 1: Variables used in the analysis.

The total Pagerank entering the colluding nodes PR_{in} is given by the sum of three terms representing links from random jumps, links from the non-colluding nodes and internal links between colluding nodes. This is the same approach taken by Clausen [3] to "lump" a set of pages into a single node for Pagerank computation.

$$Pagerank_{in} = P_{jump} + P_{in} + P_{self}$$

For calculating P_{in}, we first take the sum over all nodes pointing to the colluding set, instead of on all the links:

$$\begin{aligned} P_{in} &= \sum_{(a,b):(a,b)\in E_{in}} \frac{PR_a}{deg(a)} \\ &= \sum_{a:a\in G-G'} PR_a p(a) \end{aligned}$$

Where $p(a)$ is the fraction of links from node a pointing to the colluding set G', and it can be zero if no link from node a points to G'.

Now, let:

$$
\begin{aligned}
p &= \frac{\sum_{a:a\in G-G'} PR(a)p(a)}{\sum_{a:a\in G-G'} PR(a)} \\
&= \frac{\sum_{a:a\in G-G'} PR(a)p(a)}{1-x}
\end{aligned}
$$

So p is a weighted average of $p(a)$ over $G-G'$, in which the weights are the Pagerank values of the nodes in $G-G'$. The important issue is that p cannot be controlled by the colluding nodes, and will remain constant whatever strategy is used. We can now write the equation for P_{in} as:

$$P_{in} = (1-\varepsilon)(1-x)p$$

For P_{self} we make a similar replacement. If $s(b)$ is the fraction of links from node $b \in G'$ pointing to the colluding set G', then let:

$$
\begin{aligned}
s &= \frac{\sum_{b:b\in G'} PR(b)s(b)}{\sum_{a:a\in G'} PR(b)} \\
&= \frac{\sum_{b:b\in G'} PR(b)s(b)}{x}
\end{aligned}
$$

So s represents a weighted average of $s(a)$ over G', and this yields:

$$P_{self} = (1-\varepsilon)xs$$

Now we can write the equation for the sum of the Pagerank of the colluding nodes as:

$$Pagerank_{in} = \varepsilon\frac{M}{N} + (1-\varepsilon)(1-x)p + (1-\varepsilon)xs$$

Solving the stationary state $Pagerank_{in} = x$ yields:

$$x_{orig} = \frac{\frac{\varepsilon M}{N} + (1-\varepsilon)\,p}{(p-s)(1-\varepsilon)+1}$$

The only thing the colluding nodes can do is to link more internally than externally. This means that $s \to s'$, with $s' > s$, and the ratio between the resulting Pagerank and the original Pagerank is:

$$\frac{x_{new}}{x_{orig}} = 1 + \frac{s'-s}{p-s'+\frac{1}{1-\varepsilon}} \tag{2}$$

A trivial observation is that if $s' > s$ then:

$$\frac{x_{new}}{x_{old}} > 1$$

That is, there is always something to win by colluding with other nodes. In particular, colluding by forming a clique means $s' \to 1$, and the ratio between the resulting Pagerank x_{new} and the original Pagerank is:

$$\frac{x_{new}}{x_{old}} = 1 + \frac{1-s}{p+\frac{\varepsilon}{1-\varepsilon}} \tag{3}$$

This ratio is inversely proportional to the Pagerank that originally entered the colluding nodes. Therefore, if the colluding set has a high connectivity at the beginning, the returns from colluding will be poor, and viceversa.

For instance, if the starting set has very few, or no in-links from the rest of the graph, $p = 0$, and at the beginning $s = 0$ (originally all the out-links went to the rest of the graph), then:

$$\frac{x_{new}}{x_{old}} = \frac{1}{\varepsilon} \approx 7$$

If the starting Pagerank is very good, and has all the links from the non-colluding set we have $p = 1$, but has no internal links at the beginning, then:

$$\frac{x_{new}}{x_{old}} = 2-\varepsilon$$

However, this situation is very unlikely, because if the colluding set has all the in-links from the rest of the graph, then it should also have some links from itself. In fact, if we assume that in the original situation, the fraction of links going to the colluding nodes was the same for all nodes in the graph, then $s = p$, and the change in Pagerank value is given by the following equation:

$$\frac{x_{new}}{x_{old}} = \frac{1}{p(1-\varepsilon)+\varepsilon} \tag{4}$$

The graph of Pagerank change for $\varepsilon = 0.15$ and varying values of p is shown in Figure 2. We can see that while the starting fraction of links received remains roughly below 1% the returns are still maximum.

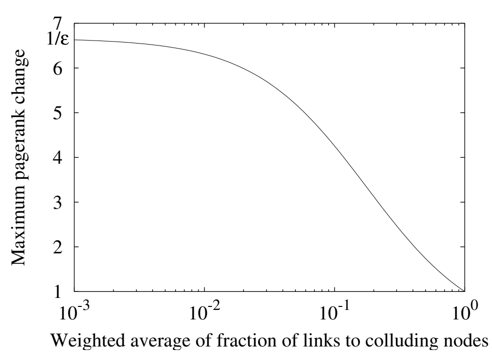

Figure 2: Expected change of Pagerank values under different starting conditions p, using $\varepsilon = 0.15$.

Focus on a single page In this section we have discussed the issue of increasing the *average Pagerank* of a set of pages. This is not the same as increasing the Pagerank of a single page as in the link spam farm structures studied in [8]. A simple, brute-force strategy of creating M (unlinked) pages, each with a single link pointing to a target page yields a Pagerank for the latter of at most:

$$
\begin{aligned}
x_{brute-force(M)} &= \frac{\varepsilon}{N} + (1-\varepsilon)\varepsilon\frac{M}{N} \\
&= \frac{\varepsilon + \varepsilon M - \varepsilon^2 M}{N} \\
&\approx \varepsilon\frac{M}{N} \qquad (M \gg 1; \varepsilon \ll 1)
\end{aligned}
$$

This is, an individual page can get a increase of Pagerank larger than in our analysis, but the average amplification of all the pages in the colluding set will be as described by Equation 4.

4 Experiments with a Synthetic Web Graph

We obtained a synthetic graph using a generative model described by Kumar *et al.* [10]:

- Nodes are added one at a time.
- Each time a node is added, d links are added. For adding a link, the source and destination nodes are chosen as follows:
  - With probability β the source node is chosen at random, and with probability $1-\beta$ the source node is chosen with probability proportional to the current out-degree of nodes.
  - With probability α the destination is chosen at random, and with probability $1-\alpha$ the destination is chosen with a probability proportional to the current in-degree of nodes.

We used $d = 7$, $\alpha = 0.2$ and $\beta = 0.45$, parameters experimentally determined by Pandurangan *et al.* [14] that produce graphs simultaneously fitting the distributions of in-degree, out-degree and Pagerank to the values observed in real Web graphs. The parameters for the power-law in the center part of the distributions are -2.1 for in-degree and Pagerank, and -2.7 for out-degree.

It is very important to remove the disconnected nodes from the resulting graph, as they affect the Pagerank normalization factor. This is specially critical for groups of pages with very low starting ranking, as they will certainly include the disconnected nodes, and those nodes

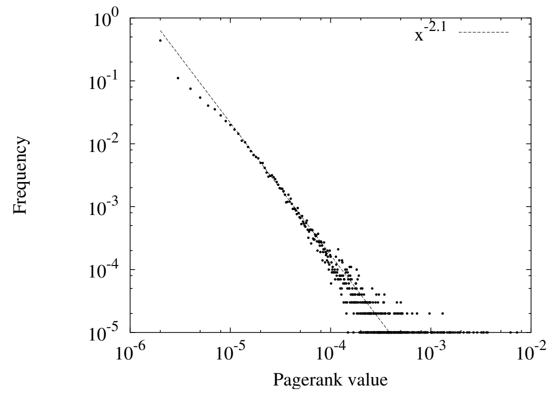

Figure 3: Distribution of Pagerank values in the synthetic graph.

will become connected after the collusion, modifying the number of nodes involved in the total Pagerank calculation. Using the generative model described above, we created a 125,000-nodes graph and then removed all the disconnected nodes to obtain a connected graph of roughly 106,000 nodes. We also made preliminary experiments with a 10,000-nodes graph and the results were very similar.

Instead of sampling according to the number of nodes, we sampled according to the amount of Pagerank. We divided the nodes in the Web graph into 10 segments, each segment having $1/10^{th}$ of the total Pagerank. Note that because of the distribution of Pagerank, shown in Figure 3, these segments represent sets exponentially decreasing in size.

Inside each segment, we picked a group of $M = 100$ nodes at random –except in the last segment, as the top 10% of Pagerank was found in only 50 nodes. We labeled these groups $1 \ldots 10$.

In the following, we denote by **Pagerank values** the actual probabilities given by the Pagerank algorithm, this is, the resulting values of the vector x. We denote by **ranking** the order in which a page appears when pages are sorted by Pagerank values. This number is normalized so 0 is the last page and 1 is the top page by Pagerank value.

The original Pagerank values of the pages inside each group, as well as the group averages, are shown in Figure 4.

As the distribution of the Pagerank values is very skewed, the distribution of the rankings inside each of these groups appears as shown in Figure 4. Note that most of the pages in group 1 have the same Pagerank value, so their ranking is distributed uniformly in the 0.0-0.5 interval, meaning that the bottom 50% of the pages have in total just 10% of the Pagerank.

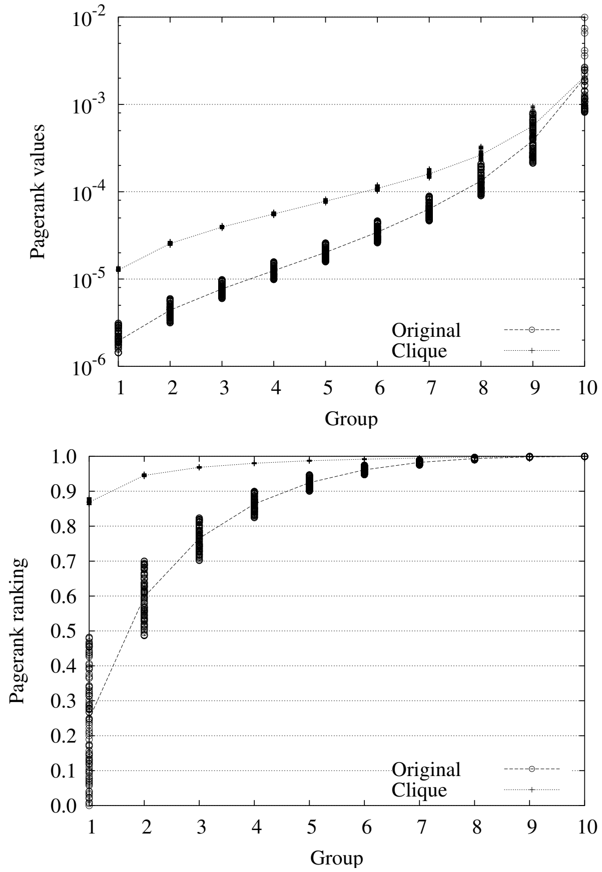

Figure 4: Pagerank values and rankings, in both the original and the modified graphs, using a clique inside each group.

4.1 Collusion via a complete sub-graph

The first strategy we tested was to create a clique (a complete sub-graph) inside each group. Unlike the experiments by [15], we did not remove any outgoing links, as that is very easy to detect and can be penalized by search engines. The rationale is that if within a group the number of internal links outnumbers the number of external links then that group will preserve its Pagerank.

Figure 4 compares the Pagerank values before and after the collusion.

Figure 5 plots the variation in both Pagerank values and ranking after collusion. Clearly the bound $1/\varepsilon$ is too coarse for medium to highly ranked pages, in those cases, the starting incoming links should be considered, as in Equation 4.

Note that a factor of 3 in the Pagerank ranking is a very large increase for a page. As shown in Figure 5, all of the pages from group one get a new ranking in the top 10% by colluding. Moreover, all of the pages inside each group get roughly the same ranking.

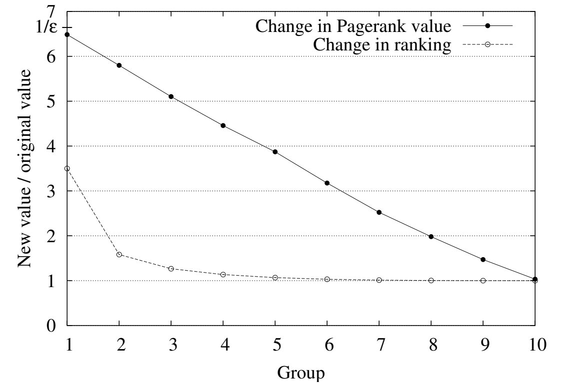

Figure 5: Relative variation of Pagerank value and ranking using a clique inside each group. Link spamming only yields returns for pages with low starting Pagerank values.

4.2 Collusion via a partial sub-graph

It might not be necessary to create all links, just enough links to keep most of the Pagerank inside the colluding set. We tested varying amounts of linking in the synthetic graph. The amplification factor obtained by colluding is shown in Figure 6.

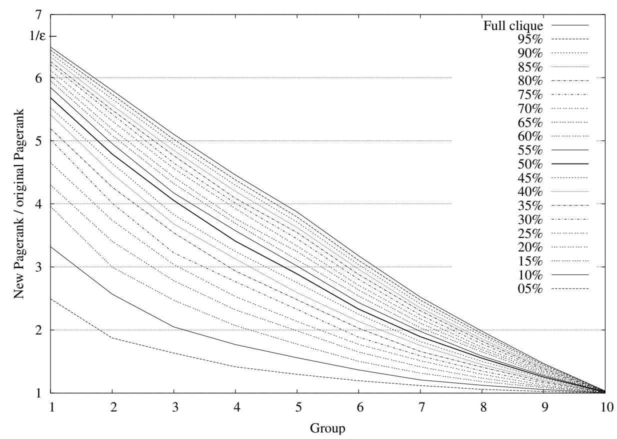

Figure 6: Pagerank change under varying amounts of internal links.

In most cases, adding just 50% of the links yields high returns, and in the case of the group with the lower starting Pagerank, even 30% of the links results in an increase of Pagerank by a factor of 5.

4.3 Other collusion strategies

We also tested collusion strategies involving $O(M)$ internal links instead of $O(M^2)$ as is the case with cliques. The two topologies we studied were a star and a ring, as depicted in Figure 7.

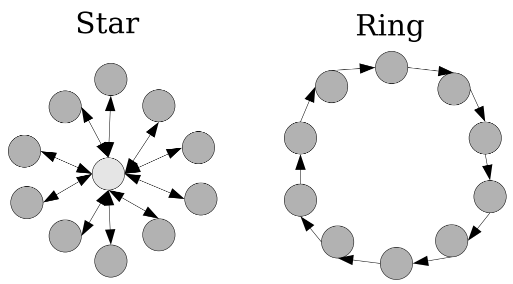

Figure 7: Studied linking topologies.

In the case of the star, to avoid a positive bias in the choice of the center of the star, we picked a new node originally without in-links as the center of the star. The comparison between the results under these two topologies is shown in Figure 8. There is a slight advantage of forming a star instead of a ring for the lowly-ranked sites, but for the other groups both strategies yield similar returns, and those are much lower than in the case of cliques.

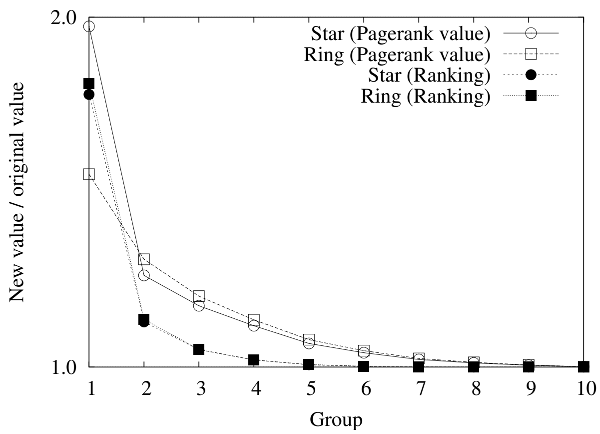

Figure 8: Pagerank under other linking topologies; both the star and ring topologies yield much lower returns than forming a clique.

5 Experiments with a Real Web Graph

We started with a collection of 16 million pages from Spanish Web sites obtained during 2004. In this case, we are interested in complete Web sites instead of individual pages, so we first converted multiple links between pages in different Web sites, into a single link between two Web sites. Two sites s_1, s_2 are linked iff there is at least one page on site s_1 pointing to a page in s_2.

We obtained a graph with $310,486$ Spanish sites and $3,037,913$ directed links between them.

We calculated the Pagerank (or Hostrank [4]) values of each Web site, and the corresponding ranking induced by this value. Note that this is not the same as the sum of the page-wise Pagerank for each page in the Web site, because there may be multiple links between two Web sites [2]. The distribution of values for the Hostrank is shown in Figure 9.

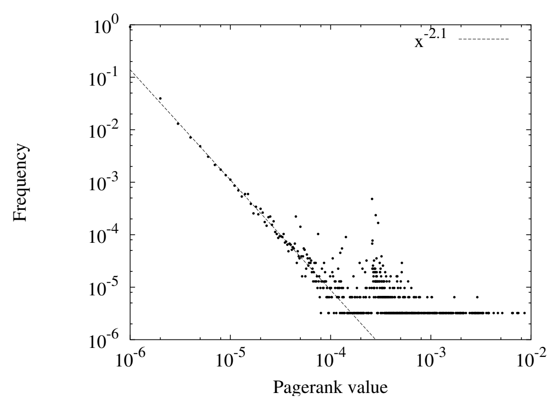

Figure 9: Distribution of Hostrank values in the Spanish Web sites graph. There are a number of Web sites already colluding.

Comparing this with the distribution of Pagerank in the synthetic graph, shown in Figure 3, we can see that while both exhibit a Zipf's law with roughly the same parameter, in the real Web there is a significant number of outliers. Manual inspection of these outliers showed that most of them are Web sites that can be considered as spam, for instance, we found several groups of dozens of Web sites with names such as `http://cityname.company.es/`, in which `cityname` is the name of a Spanish city and `company` is a tour operator or hotel company.

We modified this graph with the strategies we have discussed so far and computed the new Pagerank values after each strategy. The objective of these strategies is to increase the ranking of a small set of 242 sites (0.08% of the total number of sites). The target sites for this experiments were obtained from the directory of an agency certifying the quality of Spanish Web sites, and are expected to be sites that adhere to certain standards of coding, content, etc.

The average ranking of the Web sites in the selected group is 0.75. Note that this only takes into account the Pagerank value, while the quality of a site may come from very different factors.

Table 1: Linking strategies.

Strategy	Average ranking
Disconnect group	0.75
Normal	0.77
Central site	0.82
Ring, alphabetical	0.93
Ring, inverted	0.93
Star	0.96
Clique	0.99

Table 1 lists the linking strategies used. We started by disconnecting all the links between the participating nodes, that yields a minimum of ranking without collusion at all. After that, we returned to the normal situation. Then we added a central site that lists all the participating sites, then a ring of all the sites in alphabetical order, and then an inverted ring. Finally, we also added a star and a clique. Figure 10 lists the resulting rankings under different strategies.

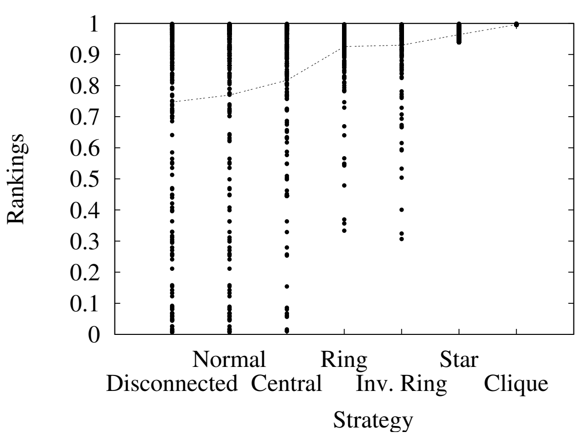

Figure 10: Relative rankings under different strategies. Each dot represents a site.

Clearly, creating a complete subgraph is the best strategy, as all of the sites in the group get very high positions. We noted that the star strategy gets a similar average to the ring strategy, but much lower variability.

Finally, we explored the possibility of adding less than 50% of the links of a complete subgraph. In Figure 11, a varying amount between 5% and 50% of the links in the complete subgraph are added randomly.

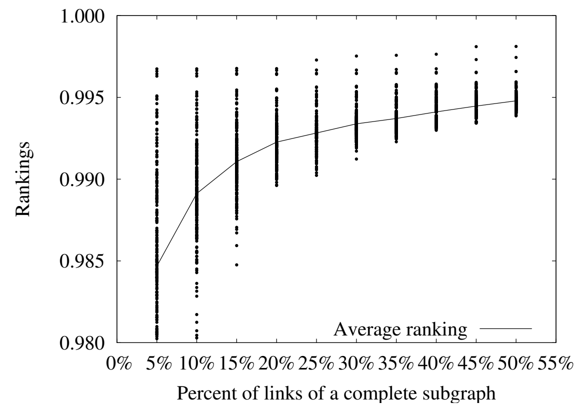

Figure 11: Relative rankings when adding a fraction of the links of a complete subgraph.

We observe that even adding 5% of the links of a complete subgraph, i.e.: each of the Web sites in the group links to 5% of the other sites (in this case, roughly 10 Web sites each one), then the average position (0.985) is higher than all the strategies based on other topologies. After adding about 20% of the links of the complete subgraph, the gains increase linearly with the number of links.

6 Conclusions and Future Work

While any group of nodes can increase their Pagerank by forming a tightly-connected sub-graph of the Web, the increase they obtain by doing so is inversely related to their starting Pagerank. This means that the Pagerank algorithm is particulary vulnerable to Sybil attacks from the nodes with low Pagerank. As the distribution of Pagerank is very skewed, even a modest increase in Pagerank value may imply a large increase in the ranking of a page.

Collusion strategies have never been studied in a more microscopic scale. For instance, we noted slight differences when creating a ring of pages in two different orderings. There is an optimum ordering for forming a ring, and we are interested in studying different strategies under a limited "budget" in terms of links.

As future work, we would like to study other forms of ranking calculation such as a two-level ranking scheme that ranks entire Web sites and then Web pages.

While this article is mainly descriptive, we are also interested in developing ways of detecting deceptive linking practices to improve reputation algorithms. There is not a trivial answer to this problem. Finding regular structures [7] may not be enough as spammers can randomize their link spam farms. Measuring the ratio between the total Pagerank of a group of pages and the Pagerank they receive externaly [15] may detect groups of pages that are strongly linked among them for legitimate reasons. An open question is if the current linking practices used amongst "good sites" should be used –and accepted– or not.

References

[1] Agogino, A., and Ghosh, J. Increasing pagerank through reinforcement learning. In *Proceedings of Intelligent Engineering Systems Through Artificial Neural Networks* (St. Louis, Missouri, USA, November 2002), vol. 12, ASME Press, pp. 27–32.

[2] Baeza-Yates, R., and Castillo, C. Relating Web characteristics with link based Web page ranking. In *Proceedings of String Processing and Information Retrieval SPIRE* (Laguna San Rafael, Chile, November 2001), IEEE CS Press, pp. 21–32.

[3] Clausen, A. The cost of attack of PageRank. In *Proceedings of the international conference on agents, Web technologies and Internet commerce (IAWTIC)* (Gold Coast, Australia, July 2004).

[4] Dill, S., Kumar, R., Mccurley, K. S., Rajagopalan, S., Sivakumar, D., and Tomkins, A. Self-similarity in the web. *ACM Trans. Inter. Tech. 2*, 3 (2002), 205–223.

[5] Douceur, J. The Sybil Attack. In *Proceedings of the first International Peer To Peer Systems Workshop (IPTPS)* (Cambridge, MA, USA, January 2002), vol. 2429 of *Lecture Notes in Computer Science*, Springer, pp. 251–260.

[6] Eiron, N., McCurley, K. S., and Tomlin, J. A. Ranking the web frontier. In *Proceedings of the 13th international conference on World Wide Web* (New York, NY, USA, 2004), ACM Press, pp. 309–318.

[7] Fetterly, D., Manasse, M., and Najork, M. Spam, damn spam, and statistics: Using statistical analysis to locate spam Web pages. In *Proceedings of the seventh workshop on the Web and databases (WebDB)* (Paris, France, June 2004).

[8] Gyöngyi, Z., and Garcia-Molina, H. Link spam alliances. Tech. rep., Stanford University, California, March 2005.

[9] Haveliwala, T., and Kamvar, S. The second eigenvalue of the Google matrix. Tech. rep., Stanford University, 2003.

[10] Kumar, R., Raghavan, P., Rajagopalan, S., Sivakumar, D., Tomkins, A., and Upfal, E. Stochastic models for the web graph. In *Proceedings of the 41st Annual Symposium on Foundations of Computer Science (FOCS)* (Redondo Beach, CA, USA, 2000), IEEE CS Press, pp. 57–65.

[11] Meyer, C. D. Sensitivity of the stationary distribution of a Markov chain. *SIAM Journal on Matrix Analysis and Applications 15*, 3 (1994), 715–728.

[12] Ng, A. Y., Zheng, A. X., and Jordan, M. I. Link analysis, eigenvectors and stability. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)* (Seattle, Washington, USA, 2001), pp. 903–910.

[13] Page, L., Brin, S., Motwani, R., and Winograd, T. The Pagerank citation algorithm: bringing order to the web. Tech. rep., Stanford Digital Library Technologies Project, 1998.

[14] Pandurangan, G., Raghavan, P., and Upfal, E. Using Pagerank to characterize Web structure. In *Proceedings of the 8th Annual International Computing and Combinatorics Conference (COCOON)* (Singapore, August 2002), vol. 2387 of *Lecture Notes in Computer Science*, Springer, pp. 330–390.

[15] Zhang, H., Goel, A., Govindan, R., Mason, K., and Roy, B. V. Making eigenvector-based reputation systems robust to collusion. In *Proceedings of the third Workshop on Web Graphs (WAW)* (Rome, Italy, October 2004), vol. 3243 of *Lecture Notes in Computer Science*, Springer, pp. 92–104.