Link Spam Alliances

Zoltán Gyöngyi

Stanford University
Computer Science Department
Stanford, CA 94305
zoltan@cs.stanford.edu

Hector Garcia-Molina

Stanford University
Computer Science Department
Stanford, CA 94305
hector@cs.stanford.edu

Abstract

Link spam is used to increase the ranking of certain target web pages by misleading the connectivity-based ranking algorithms in search engines. In this paper we study how web pages can be interconnected in a spam farm in order to optimize rankings. We also study alliances, that is, interconnections of spam farms. Our results identify the optimal structures and quantify the potential gains. In particular, we show that alliances can be synergistic and improve the rankings of all participants. We believe that the insights we gain will be useful in identifying and combating link spam.

1 Introduction

As search engines become ubiquitous tools of our everyday lives, individuals and businesses crave to see their web pages showing up frequently on the top of query results lists. The economic advantage of high search engine ranking led to the emergence of the dark art of *web spamming* [5]: some authors create web content with the main purpose of misleading search engines and obtaining higher-than-deserved ranking in search results.

Successful spamming attempts induce a bias in search results and decrease quality, as truly popular pages are replaced by artificially boosted spam documents. Counterbalancing the negative effects of an increasing volume of web spam represents a major challenge for today's web search engines [6].

Proceedings of the 31st VLDB Conference, Trondheim, Norway, 2005

Among the plethora of techniques used by spammers, one that deserves special attention is *link spamming*. Link spamming refers to the cases when spammers set up structures of interconnected pages, called *link spam farms*, in order to boost the connectivity-based ranking, most frequently the PageRank [9], of one or a small number of *target pages*. The issue of link spamming is important not only because it can render significant gains in the rankings of target pages, but also because many instances of it are very hard to detect.

In this paper we analyze how link spammers manipulate PageRank scores. We study the problem in two phases. First, we take a look at the ways in which a spammer can improve the ranking of a single target page. Then, we investigate how groups of spammers could collaborate by forming alliances of interconnected spam farms. For the latter scenario, we suppose that individual spammers already have their own spam farms. Such spammers might want to cooperate, either for mutual benefit, or based on a financial agreement. As we will see, with carefully devised interconnection of spam farms, cooperation could be reciprocally advantageous to all participants.

While recent analyses of PageRank's mathematical properties [1, 8] touch on the subject of link spamming, our paper represents a more detailed discussion dedicated exclusively to this subject.

It is important to mention that while our ultimate goal is to combat link spam, in this paper we only focus on studying various farm structures and alliances that can impact rankings. We briefly touch on the topic of combating link spam in Section 7, where we illustrate how our understanding of spam structures can lead to useful detection schemes.

One obvious question that arises is whether we help spammers by presenting our results. Our experience indicates that all the spamming techniques that we will present are already widely used by the large community of spammers. Our contribution here is simply to formalize these link spam structures, to quantify their impact on ranking, and to compare them against each other.

The rest of the paper is organized as follows. First, we offer an overview of PageRank, the commonly used ranking algorithm that we investigate from the perspective of link spamming. Then, we discuss our model for a spam farm, and derive the optimal internal structure of a farm based on the properties of PageRank. Sections 4 and 5 first focus on the optimal structure of two interconnected spam farms, then also analyze larger spam farm alliances. Next, we discuss how our findings apply to generic link spam structures. We conclude the paper with a summary of applicable link spam detection techniques.

2 Preliminaries

2.1 Web Model

In this paper we adopt the usual graph model for representing the web of interlinked hypertext documents. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be the web graph with vertices $\mathcal{V}$, representing web pages, and directed unweighted edges $\mathcal{E}$, representing hyperlinks between pages. Please note that we do not allow self loops (links on a page pointing to itself).

As we will see, pages without outlinks play an important role in our analysis. Such pages are usually referred to as *sink pages*.

It is common to associate with the web graph a *transition matrix* $\mathbf{T} = (T_{i,j})_{n \times n}$ defined as:

$$T_{i,j} = \begin{cases} 1/\mathrm{out}(i), & \text{if } (i,j) \in \mathcal{E} \\ 0, & \text{otherwise} \end{cases}$$

where $\mathrm{out}(i)$ is the *outdegree* of page i, that is, the number of links (edges) leaving i.

2.2 The PageRank Algorithm

Search engines usually combine the results of several *ranking algorithms* to produce the ordering of the pages returned as answers to a query. One of the best-known ranking algorithms is PageRank [9], which computes global importance scores for all web pages. Because the scores are determined based on the link structure of the web, PageRank is a natural target to link spamming. Our discussion will focus on link spam structures that target the PageRank algorithm. Next, we offer a short overview of PageRank.

Let us introduce a constant c called the *damping factor*. The scores computed by PageRank will correspond to the stationary distribution of a Markov chain [7] where:

1. The states represent web pages.

2. A transition from page i to page j occurs with a probability $c/\mathrm{out}(i)$ whenever one of the $\mathrm{out}(i)$ outgoing links of i points to j.

3. With probability $(1 - c)$, the transition from a page will be made uniformly at random to any web page. This latter case is called *random jump* or *teleportation*.

The traditional formulation of the PageRank problem is based on the eigensystem corresponding to a Markov matrix. For the purposes of this paper, we define the PageRank score vector $\mathbf{p}$ in a different way, as the solution of the matrix equation

$$\mathbf{p} = c\,\mathbf{T}'\,\mathbf{p} + \frac{1-c}{N}\,\mathbb{1}_N\,, \qquad (1)$$

where c is the damping factor, $\mathbf{T}'$ is the transposed transition matrix, N is the total number of web pages, while $\mathbb{1}_N$ is a vector consisting of N elements of 1. Hence, our formulation is based on a linear system, which not only yields the same relative scores for the pages as the traditional approach, but also has several additional advantages [8].

3 Single-Target Spam Farm Model

In the first part of the paper, we introduce our spam farm model, and investigate what link spam structure yields the highest PageRank of the target page. This sets the stage for the analysis of spam alliances (Section 4) and other link spam structures that deviate in some ways from the presented ones (Section 6).

3.1 Definition

As mentioned in Section 1, link spamming targets those ranking algorithms that derive the importance of a page from the link structure of the web. In order to boost the rankings of some of their pages, spammers often set up (large) groups of web pages with carefully devised interconnection structures. We will call the group of pages controlled by a spammer a *link spam farm*, or simply a *spam farm*.

The initial link spam farm model that we adopt is based on the following rules:

1. Each spam farm has a single *target page*. The target page is the one that the spammer wishes to expose to a web user through a search engine. Therefore, the spammer focuses on boosting the ranking of the target page.

2. Each spam farm contains a fixed number of *boosting pages* that exist in order to improve the ranking of the target page, possibly by pointing to it. These boosting pages are under the spammer's full control. We assume that there is always an upper bound on the size of the spam farm (the number of boosting pages) because of the associated maintenance costs (domain registration fees, page hosting fees, hardware costs, invested time).

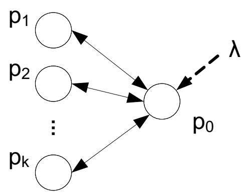

Figure 1: An optimal structure for a single spam farm with one target page.

3. It is also possible for spammers to accumulate links from pages outside the spam farm (for instance, by finding their way into a web directory, or an unmoderated bulletin board). We call these external links *hijacked links*, and the total PageRank that reaches the farm through these links is referred to as the *leakage*. Please note that the spammer does not have full control over the pages that contain hijacked links, i.e., can neither influence their PageRank scores significantly, nor determine where and how the scores get distributed through the outlinks. Therefore, the actual amount of leakage is fairly independent of the spammer's efforts—the spammer can at most struggle to hijack many links, preferably on pages that are suspected of having a high PageRank.

3.2 Structure

Let us consider a spam farm consisting of k boosting pages plus a target page. It is possible to identify an entire class of farm structures that yield the highest PageRank score for the target page. One optimal structure is presented in Figure 1. The k boosting pages point directly to the target, the target links back to each of them, and all hijacked links point to the target, providing the leakage λ.

First, let us take a look at what target score this structure yields. Then, we prove that this is the best target score one could achieve.

Theorem 1 *The PageRank score p_0 of the target page in Figure 1 is*

$$p_0 = \frac{1}{1-c^2}\left[c\lambda + \frac{(1-c)(ck+1)}{N}\right].$$

The proof for all our theorems can be found in the extended version of this paper [4].

3.3 Optimality

In this section we identify the class of spam farm structures that yield the highest target PageRank. Consider the generic spam farm in Figure 2, with a single target and k boosting pages. The pages of the farm are interconnected in an arbitrary manner. Spam pages

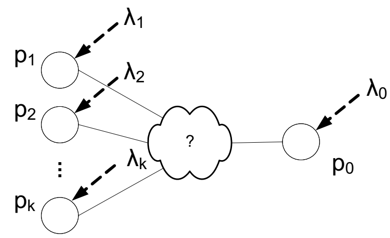

Figure 2: Generic farm structure used in Theorem 2.

may have outlinks pointing to pages outside the farm (although such links are omitted in the figure). Hijacked links point to pages in the farm so that the leakage to the target is $\lambda_0 \geq 0$, and to boosting page i is $\lambda_i \geq 0$. The total leakage is $\lambda = \lambda_0 + \cdots + \lambda_k$. Please note that while the spam pages may point to good pages, and thus possibly have some impact on the leakage, based on Assumptions 3 from Section 3.1, λ_i does not actually depend on the PageRank scores of spam pages.

We introduce two vectors $\mathbf{p}$ and $\boldsymbol{\lambda}$ for the PageRank and leakage of the *boosting* pages,

$$\mathbf{p} = \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_k \end{pmatrix} \qquad \boldsymbol{\lambda} = \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_k \end{pmatrix}.$$

With this notation, the matrix equation of the PageRank scores of the spam farm pages can be written as

$$\begin{pmatrix} p_0 \\ \mathbf{p} \end{pmatrix} = c\begin{pmatrix} \lambda_0 \\ \boldsymbol{\lambda} \end{pmatrix} + c\begin{pmatrix} 0 & \mathbf{e}' \\ \mathbf{f} & \mathbf{G} \end{pmatrix}\begin{pmatrix} p_0 \\ \mathbf{p} \end{pmatrix} + \frac{1-c}{N}\mathbb{1}_{k+1}, \quad (2)$$

where the row vector $\mathbf{e}'$ corresponds to the weights of the links from boosting pages to the target, $\mathbf{f}$ contains the weights of links from the target to the boosting pages, and $\mathbf{G}$ is the weight matrix capturing the connections among boosting pages.

Theorem 2 *The PageRank score p_0 of the target is maximal if the farm is structured so that $\mathbf{e} = \mathbb{1}_k$, $\mathbb{1}_k'\mathbf{f} = 1$, $\mathbf{G} = \mathbb{0}_{k\times k}$, and $\lambda_0 = \lambda$ and $\lambda_i = 0$, for $i = 1, \ldots, k$.*

In other words, p_0 is maximal if and only if

- all boosting pages point to and only to the target ($\mathbf{e} = \mathbb{1}_k$),
- there are no links among the boosting pages ($\mathbf{G} = \mathbb{0}_{k\times k}$),
- the target points to some or all boosting pages ($\mathbb{1}_k'\mathbf{f} = 1$), and
- all hijacked links point to the target ($\lambda_0 = \lambda$, and $\lambda_i = 0$ for $i = 1, \ldots, k$).

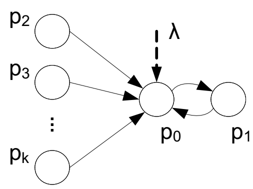

Figure 3: Another optimal structure for a single spam farm with one target page.

These constraints also imply that there are no outlinks pointing to external pages.

The farm structure in Figure 1 satisfies the properties required by Theorem 2. Similar structures will also satisfy the properties, as long as a proper subset of the target-to-boosting links are maintained. The extreme case when the target points to only one boosting page is shown in Figure 3.

3.4 Leakage

We have seen that in the optimal case the target accumulates PageRank from the boosting pages and through the hijacked links. In this section we show that the leakage can be thought of as an additional number of boosting pages. Therefore, we will not need a separate treatment of leakage in the rest of the paper.

Theorem 3 *For an optimal farm, a leakage of λ increases the target score by just as much as an additional number of $d\lambda$ boosting pages would, where d is a constant that depends on the farm structure.*

Please note that we are not expecting actual farms to lack leakage and be isolated from the rest of the web. Leakage is treated as additional boosting pages merely to simplify the exposition and our mathematical derivations. Our results can be easily generalized to the case when there is leakage.

3.5 Reachability

The structure presented in Figure 1 has the property that if the search engine's crawler reaches the target through at least one hijacked link, then the entire link farm becomes reachable. Thus, the entire farm gets crawled and indexed by a search engine and the boosting pages contribute to the score of the target indeed.

While reachability through hijacked links is important, there are also other ways in which one can make the crawler aware of specific pages. For instance, in order to make sure that the search engine's crawler reaches all the pages of a spam farm, one could use a separate domain for each of the pages. As search engines usually crawl all domains from the registrar databases, all the pages would get crawled and indexed this way.

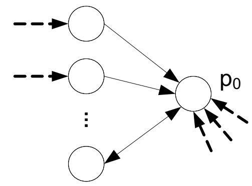

Figure 4: Making boosting pages reachable through hijacked links.

Also, it is possible to "sacrifice" some hijacked links and point to the boosting pages instead of the target. A corresponding spam farm is presented in Figure 4 (dashed lines represent hijacked links). These alternative approaches to reachability will become important later in our discussion, when we remove the links from target pages to boosting pages.

4 Alliances

The first part of this paper addressed the case of a single spam farm. In the second part of this paper, we turn our attentions to groups of spammers, each with an already built farm, and investigate how interconnecting their farms impacts the PageRank scores of target pages. As mentioned earlier in Section 1, these types of collaborations emerge on the web, either because they are mutually beneficial, or as a result of some financial agreement between a "client" and a "service provider."

First, in this section we derive formulas that quantify features of various alliance structures. Then, in Section 5 we use the derived formulas to study some collaboration scenarios of interest.

4.1 Alliances of Two

Let us first discuss ways in which we can combine two optimal farms. The farms have a single target page each, and have k and m ($k < m$) boosting pages, respectively. (As mentioned earlier, leakage is treated as being a fraction of the boosting pages.) Let $\bar{p}_0$ and $\bar{q}_0$ denote the (maximal) PageRank scores of the target pages when the farms are not interconnected:

$$\bar{p}_0 = \frac{ck+1}{(1+c)N} \qquad \bar{q}_0 = \frac{cm+1}{(1+c)N} \,.$$

Then, p_0 and q_0 will denote the scores of the targets when the two farms are interconnected in one way or another. We investigate three interconnection techniques next.

4.1.1 Shared Boosting Pages

There are a number of ways in which one could connect two spam farms. One way of doing it is just having all boosting pages point to both of the targets, as presented in Figure 5. In order to produce such a

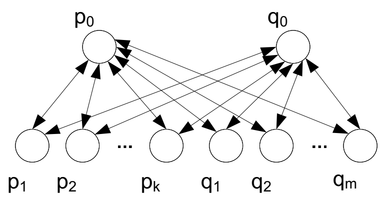

Figure 5: Two spam farms with all boosting pages pointing to both targets.

structure, both spammers have to add links from their boosting pages to the target of the other. Hence, a total number of $(k+m)$ new links has to be added.

What we achieve through this interconnection structure is two target pages with identical scores:

Theorem 4 *For the structure presented in Figure 5, $p_0 = q_0 = (\bar{p}_0 + \bar{q}_0)/2$.*

Accordingly, sharing boosting pages is clearly advantageous to the spammer with the smaller initial farm, as its target PageRank increases from $p_0 \propto k$ to $\bar{p}_0 \propto (k+m)/2 > k$.

On the other hand, sharing is inconvenient to the spammer with the larger initial farm, as the PageRank of its target decreases.

The net effect of sharing boosting pages is just equivalent to the scenario when there are two unconnected farms, and $(m-k)/2$ boosting pages simply get "moved" from the larger farm to the smaller one.

4.1.2 Connected Target Pages with Links to Boosting Pages

Instead of connecting all boosting pages to both targets, one could connect the two targets only, so that each would point to the other. In this case, the boosting pages in each of the two farms would still point to their respective target only. Also, targets would point back to the boosting pages in their own farms.

A simple analysis, similar to the one for Theorem 4, reveals that the effect achieved by this interconnection structure is exactly the same as in the case when all boosting pages are shared: both targets have the same score $(\bar{p}_0 + \bar{q}_0)/2$.

Please note that while all that we achieved is still a redistribution, rather than an increase of the target PageRank scores, this structure bears an advantage over the one presented in Section 4.1.1: the number of interconnecting links that have to be added is reduced from $(k+m)$ to only 2.

4.1.3 Connected Target Pages without Links to Boosting Pages

We can form a third possible structure by connecting the two target pages and removing all links to boosting pages, as shown in Figure 6. The corresponding

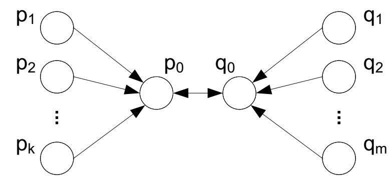

Figure 6: Two spam farms with interlinked target pages.

PageRank equations yield the target score p_0 (q_0 is symmetrical)

$$p_0 = \frac{ck + c^2 m}{(1+c)N} + \frac{1}{N} \, . \qquad (3)$$

The following theorem states that the scores of both target pages increase as compared to the maximum for unconnected farms.

Theorem 5 *For the structure presented in Figure 6, $(p_0 - \bar{p}_0) \propto m$ and $(q_0 - \bar{q}_0) \propto k$.*

The natural question that arises is how such an improvement was possible. A simple informal analysis reveals the reason.

Let us first take a look at the total PageRank of the three discussed alliances, that is, the sum of the PageRank scores of all target and boosting pages in each alliance. It turns out, that the total score is exactly the same in all three cases, being equal to $(k+m+2)/N$. It also turns out that in the general case there is always an upper bound on the total PageRank score of a structure with fixed connectivity to the rest of the web [1].

Now let us focus on the individual PageRank scores of the boosting pages. For the third structure, the PageRank of each boosting page is minimal. In contrast, for the first two structures the boosting pages have higher score, for the reason that the target pages have links pointing back to the boosting pages. By eliminating these links in the third structure, we avoided the distribution of a precious fraction of the total PageRank score to the boosting pages, which are irrelevant anyway.

Our conclusion is that the third structure yielded higher target scores because of a better "housekeeping." The total PageRank being limited, it assured that boosting pages stay low, while all the rest of score gets properly distributed among the targets. In fact, it can be shown that this structure is the optimal one for two farms, in the sense that it maximizes the sum of target PageRank scores.

4.2 Web Rings

Now, as we know how to join two spam farms, it makes sense to try to extend our discussion to larger alliances.

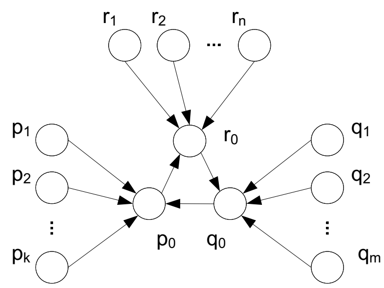

Figure 7: Three spam farms with target pages forming a ring.

We will call the subgraph of the target pages the *core* of the alliance. In the extreme case, the core of a single spam farm is the target page alone. From among the plethora of possible core structures for larger alliances (surveyed briefly in Section 4.4) in this paper we focus on two:

- Web rings represent the simplest way of interconnecting several target pages. Also, such structures are frequently encountered on the real web, and not necessarily in the context of spam only. Web ring structures have been popular among groups of authors interested in the same topic for long. In fact, web rings are one of the earliest forms in which web content was organized.
- Alliances with completely connected subgraphs of targets, or *complete cores*, are the extreme for a strongly connected group of targets.

We investigate each of these two structures in turn.

Our first way of connecting targets is by forming a ring, i.e., a cycle that includes all target pages. Figure 7 shows such a structure for three spam farms.

Solving the corresponding matrix equation yields the following PageRank score for the first farm's target:

$$p_0 = \frac{ck + c^2m + c^3n}{(1+c+c^2)N} + \frac{1}{N}\,. \tag{4}$$

For the more general case of F farms interconnected by forming a ring of the target pages, let us denote the score of each target page by t_i, and the number of boosting pages in each farm by b_i, where $i = 1, \ldots, F$. For this structure, the score of the first target will be

$$t_1 = \frac{\sum_{j=1}^{F} c^j b_j}{N \sum_{j=1}^{F} c^{j-1}} + \frac{1}{N}\,,$$

and, more generally, the PageRank score of target i will be

$$t_i = \frac{\sum_{j=i}^{F} c^{j-i+1} b_j + \sum_{j=1}^{i-1} c^{j+F-i+1} b_j}{N \sum_{j=1}^{F} c^{j-1}} + \frac{1}{N}\,. \tag{5}$$

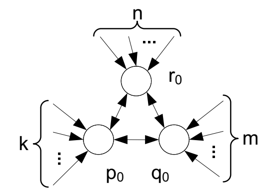

Figure 8: Three spam farms with target pages forming a complete core.

4.3 Alliances with Complete Cores

Beside web rings, connecting each spam farm with all the others is another way to move to larger structures from two collaborating spam farms.

Figure 8 shows the case when three spam farms collaborate by setting up a completely connected subgraph of the targets. (Please note that the links from targets to boosting pages are removed, just as for web rings. Boosting pages are not shown.)

Solving the corresponding matrix equation yields the following PageRank score for the first farm's target:

$$p_0 = \frac{2ck - c^2k + c^2m + c^2n}{(2+c)N} + \frac{1}{N}\,.$$

Again, the PageRank score of each target is greater than the maximum for unconnected farms. The additional score comes from the other target pages, and each other target's contribution is proportional to the number of boosting pages in that target's farm.

In the general case, we might have F farms with b_i boosting pages each, and target page scores t_i, where $i = 1, \ldots, F$. The PageRank scores of the targets are:

$$t_i = \frac{c(1-c)(F-1)b_i + c^2 \sum_{j=1}^{F} b_j}{(F+c-1)N} + \frac{1}{N}\,. \tag{6}$$

4.4 Other Core Structures

We have analyzed two possible ways of connecting the target pages of an alliance. While we will continue to focus on the presented two structures in the rest of this paper, it is important to emphasize that there are other ways to construct the core of an alliance. It is also important to emphasize that the analysis of these other structures is similar to what is presented for rings and alliances with complete cores. In this section we take a cursory look at an entire family of possible cores.

Let us consider the F target pages of the farms in an alliance. There are $2^{(F-1)}F$ possible ways to connect F nodes and form a directed graph without self-loops. However, not all of these possible graphs could act as an alliance core. In particular, the farms are actually allied only if the core is *weakly connected*, that is, the underlying *undirected* graph is connected. Moreover, the core is optimal (the sum of the target PageRank scores is maximal) only if it is *strongly connected*, that is, there is a directed path from each target page to every other.

In Sections 4.2 and 4.3 we have seen two alliance structures with optimal cores. But how many optimal cores exist for a specific F? What are the PageRank scores of the targets in each of them?

It turns out that answering the first question is not trivial. The number of strongly connected directed graphs of $F = 3, 4, 5, \ldots$ nodes is 18, 1606, 565080, ... This is Sloane's integer sequence A003030 [10] and we are not currently aware of any simple analytic generator function for it.

To answer the second question, we note the following: for each optimal core, it is possible to produce the equations that yield the target PageRank scores, exactly as we did in case of the ring and the complete core. In what follows, we attempt to provide a quantitative intuition of the possible outcomes through an example.

Consider the alliance formed of $F = 4$ spam farms, each having 100 boosting pages. As mentioned before, there are 1606 different optimal cores made up of 4 strongly connected target pages. In a simple experiment, we computed the target PageRank scores (with $c = 0.85$) for all the cores. Depending on the actual structure, each target PageRank can have one of 206 distinct values that range from $32.14/N$ to $165.07/N$. The values cover the range fairly uniformly. Hence, we conclude that it is possible to obtain roughly any distribution of PageRank scores among the targets by picking an appropriate core structure. The discussion on how to select a core that matches a specific distribution constitutes the topic of future research.

5 Alliance Dynamics

In the previous sections we derived a number of formulas that help us determine the target PageRank scores for different structures adopted in spam farm alliances. In this section we put our formulas at work, showing how our results could help us answer a number of practical questions of special importance. Among others, we seek answers to questions like: Why has one target in an alliance larger score than another? Does it make sense for a new farm to join an existing alliance? Does it make sense for a farm to leave an existing alliance in which it participates? How do additional boosting pages added to a farm influence its position within the alliance?

5.1 Being in an Alliance

Let us first look at what happens to some spam farms as soon as they form an alliance.

To illustrate, consider ten spam farms with target pages $t_1, \ldots, t_{10}$. The first farm has $b_1 = 1000$ boosting pages, the second has $b_2 = 2000$, and so on with

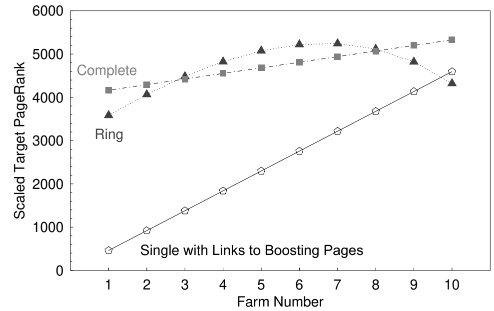

Figure 9: Scaled target PageRank scores with the farms connected in different ways.

the last having $b_{10} = 10000$ boosting pages. We discuss three scenarios. First, each farm could stay unconnected to the others, maximizing its target score by adopting the structure presented in Figure 1. Second, the farms could form an alliance with the targets connected in a ring: t_1 points to t_{10}, t_{10} to t_9, and so on until the cycle gets closed by t_2 pointing to t_1. Third, the farms could be interconnected so that the targets form a complete core.

Now let us take a look at the PageRank of each target in all three scenarios. Figure 9 presents the scores. The horizontal axis marks the ten farms. The vertical axis corresponds to the scaled PageRank scores of the targets. (We scaled the PageRank scores by multiplying them by N, the total number of web pages. This way, the obtained scaled scores are independent of the size of the web.) The three curves correspond to the three scenarios.

As we can see, for unconnected farms the target PageRank is linear in the farm size. If the targets form a complete core, each of the target PageRank scores increases with respect to the unconnected case. Moreover, the increase is so that the smallest farm gains the most additional PageRank and the largest gains the least. Even more intriguing, in case of the web ring some target scores increase while some others drop below the unconnected case. In particular, the target of the largest farm in the ring loses score.

Figure 10 helps us understand these phenomena. It shows the contributions of farm 1 to the PageRank scores of different target pages that are either in a ring, or form a complete core. The horizontal axis once again represents the farms. For each farm i, the vertical axis shows the fraction of the scaled PageRank of t_i that is due to the presence of farm 1 in the alliance.

Intuitively, Figure 10 shows what advantage of each farm draws from being connected to farm 1. Please note that for the complete core a larger fraction of the PageRank is preserved for farm 1's own target, and the other targets receive a considerably smaller, identical contribution. In comparison, in a web ring the contributions to itself and others are closer to each other, and decrease with the distance from farm 1.

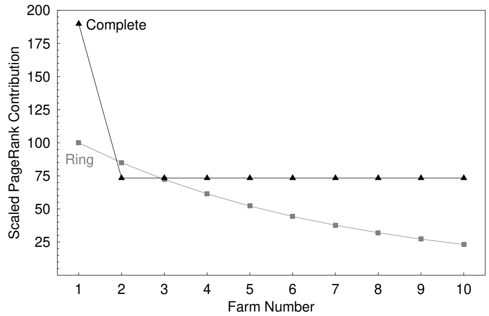

Figure 10: Scaled PageRank contribution of the first farm to the others.

Let us derive the formulas that yielded Figure 10. Please note that Equation 5 and 6 can be easily decomposed into independent terms corresponding to each farm in the alliance. Accordingly, for a web ring the *PageRank score contribution* $\mathrm{cr}(i,i)$ of farm i to its own target i is

$$\mathrm{cr}(i,i) = \frac{c(1-c)b_i}{(1-c^F)N} + \frac{1}{N} ,$$

while the contribution $\mathrm{cr}(i,j)$ of farm i to target j, $i \neq j$, is

$$\mathrm{cr}(i,j) = \frac{c^{\mathrm{d}(i,j)+1}(1-c)b_i}{(1-c^F)N} + \frac{1}{N} ,$$

where $\mathrm{d}(i,j)$ denotes the *distance*, or number of hops, on the ring between targets i and j. For instance, the distance between target t_3 and t_1 is 2, while between targets t_1 and t_3 it is 8.

Similarly, the self-contribution $\mathrm{cc}(i,i)$ for a complete core is

$$\mathrm{cc}(i,i) = \frac{c(1-c)(F-1)b_i + c^2 b_i}{(F+c-1)N} + \frac{1}{N} ,$$

and the contribution $\mathrm{cc}(i,j)$ of farm i to target j, $i \neq j$, is

$$\mathrm{cc}(i,j) = \frac{c^2 b_i}{(F+c-1)N} + \frac{1}{N} .$$

Indeed, in case of the ring the contribution to others depends on the distance, while for the complete core it is uniform. Also, it is easy to see that the total contribution of a farm is the same for both structures:

$$\sum_{j=1}^{F} \mathrm{cr}(i,j) = \sum_{j=1}^{F} \mathrm{cc}(i,j) = \frac{cb_i}{N} .$$

Please note that the total contribution made by a farm is proportional to the number of boosting pages. The contribution is independent of the interconnection structure between targets, as long as the targets are not sinks and only point to other targets.

5.2 Joining an Alliance

With the interplay of contributions in our mind, we may ask a new set of questions. First, consider an existing alliance and a new spammer who would like to join the alliance. Absent any payments, existing members of the alliance should allow the newcomer to join only if the PageRank scores of existing target pages increase. We would like to find out under what circumstances a new farm satisfies this criterion.

5.2.1 Web Rings

We first answer the previous question for web rings. For example, consider the case of adding a new farm (farm 3, with target PageRank r_0) to a ring of two farms 1 and 2 (with target PageRank scores p_0 and q_0, respectively). Using Equations 3 and 4, we can derive that the owner of farm 1 gains score by allowing farm 3 to join only if

$$\frac{ck + c^2 m + c^3 n}{(1+c+c^2)N} > \frac{ck + c^2 m}{(1+c)N} ,$$

hence,

$$n > \frac{k + cm}{1+c} .$$

That is, the sizes of farms 1 and 2 determine the minimum size of farm 3 above which it is beneficial for ring members to let the newcomer in. For instance, if $k = 20$ and $m = 10$, existing members should let the new farm join the ring only if it has at least $n = 16 > 15.4$ boosting pages.

In general, it is beneficial to append a new farm at the end of the ring of F farms (i.e., between t_F and t_1) if the following inequality is satisfied:

$$b_{F+1} > \frac{\sum_{i=1}^{F} c^{i-1} b_i}{\sum_{i=1}^{F} c^{i-1}} . \tag{7}$$

As we can see, the lower bound on farm size is a weighted mean of the farm sizes already in the alliance. Moreover, the weights depend on the position where the new farm is to be inserted.

It is interesting to follow how the insertion point influences minimum size. For instance, consider Figure 11, which shows the minimum size of a new farm as a function of the insertion point. The horizontal axis shows the farm in the ring before which the new one would be inserted. For instance, if the farm after the new one is 3 then the new farm would be inserted between farms 2 and 3, pointing to t_2. The vertical axis shows the minimum size as required by Equation 7. For instance, if the new farm is inserted before farm 1 the minimum size is only 4216, whereas if inserted before farm 7, the minimum size is 6167.

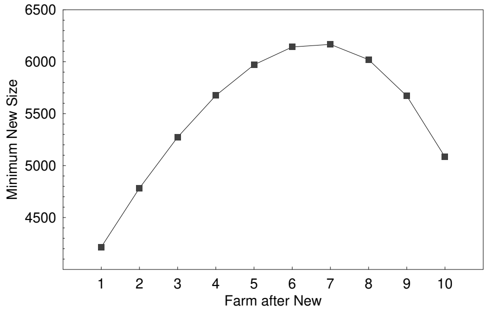

Figure 11: Minimum farm size as function of the insertion point in a ring.

5.2.2 Alliances with Complete Cores

Let us also investigate when it is beneficial to let a newcomer join an existing complete-core alliance. Unfortunately, it turns out that the answer for completely connected targets is not as straightforward as it is for rings. In this case, a newcomer with b_{F+1} boosting pages is welcome (i.e., it increases every existing target's score) only if it satisfies the following inequalities for $i = 1, \ldots, F$:

$$\frac{(1-c)Fb_i + c\sum_{j=1}^{F+1} b_j}{F+c} \geq \frac{(1-c)(F-1)b_i + c\sum_{j=1}^{F} b_j}{F+c-1}.$$

Simplifying the terms we get an inequality for each b_i:

$$b_{F+1} \geq \frac{\left(\sum_{j=1}^{F} b_j\right) - (1-c)b_i}{F+c-1}.$$

Fortunately, a closer look at the inequalities reveals that it is enough to satisfy one of them in order to also satisfy all the rest:

Theorem 6 *The inequality corresponding to the smallest farm already in the alliance determines alone the minimum size of the newcomer farm.*

It follows that the lower bound on the number of boosting pages for the newcomer is given by the inequality

$$b_{F+1} \geq \frac{\left(\sum_{j=1}^{F} b_j\right) - (1-c)\min_{i=1}^{F}\{b_i\}}{F+c-1}. \tag{8}$$

From this result, we can find a convenient approximate lower bound to the size of the new farm. Let us introduce $\eta \geq 1$ so that ηb^* corresponds to the arithmetic mean of farm sizes in the alliance. Then, the previous inequality can be written as

$$b_{F+1} \geq \frac{F\eta b^* - (1-c)b^*}{F+c-1} = \frac{F\eta + c - 1}{F+c-1} b^*.$$

As $F \gg (c-1)$, we can safely assume that

$$\frac{F\eta + c - 1}{F+c-1} \gtrapprox \eta.$$

Hence, if the new farm satisfies Equation 8, it also satisfies $b_{F+1} \geq \eta b^*$, and the current average farm size ηb^* is very close to (but below) the lower bound on the new farm's minimum size.

To illustrate the previous results, consider the alliance of two interconnected farms with $k = 20$ and $m = 10$ boosting pages. It makes sense to accept a third and form a complete-core alliance if

$$\begin{cases} t_3 \geq 15.4054 \text{ for farm with 10 boosting pages,} \\ t_3 \geq 14.5946 \text{ for farm with 20 boosting pages.} \end{cases}$$

Therefore, the third farm should have at least 16 boosting pages.

5.3 Leaving an Alliance

We may also ask: When does it make sense for a farm that is part of an existing alliance to split off from the alliance and continue to exist as a stand-alone farm instead? We have seen in Figure 9 that target t_{10} had a lower PageRank in a ring than it would have had if it were alone. Our intuition is that the contribution of farm 10 to the others is too large, and it does not receive enough contribution from the others in return. Let us formalize this intuition by deriving the appropriate inequalities for rings and alliances with complete cores.

5.3.1 Web Rings

A farm should leave an alliance if the PageRank of its target is lower than it would be when the farm were unconnected to others, and had an optimal internal structure as shown in Figure 1. The corresponding inequality for the first farm in a ring is

$$\frac{cb_1 + 1}{(1+c)N} \geq \frac{\sum_{i=1}^{F} c^i b_i}{N\sum_{i=1}^{F} c^{i-1}} + \frac{1}{N},$$

with the solution

$$b_1 \geq \frac{c^F - c(1-c^2)\sum_{i=1}^{F-1} c^i b_{i+1}}{c^2 - c^F}. \tag{9}$$

For instance, farm 1 should have 11389 boosting pages for it to make sense to leave the ring. On the other hand, the limit for farm 10 is 9091. As its size is 10000, which is above the limit, the PageRank of farm 10's target is lower than it would be if the farm were unconnected.

5.3.2 Alliances with Complete Cores

Similarly, it makes sense for farm 1 to leave an alliance with complete core if the following inequality is satisfied:

$$\frac{cb_1 + 1}{(1+c)N} \geq \frac{c(1-c)(F-1)b_1 + c^2\sum_{i=1}^{F} b_i}{F+c-1} + \frac{1}{N}.$$

The solution is

$$b_1 \geq \frac{F + c - 1 + (1 + c) \sum_{i=1}^{F} b_i}{c^2(F - 2)}. \qquad (10)$$

Here the differences between the minimum sizes for the various farms are less than they were for web rings, as the contributions get distributed more uniformly. For instance, the limit for farm 1 is 14693, while for farm 10 is 12445. As none of the farms reaches the size limit in Figure 9, it makes sense for all of them to stay in the alliance.

5.4 Adding More Boosting Pages

Another situation that might arise is when a spammer participating in an alliance wishes to add more boosting pages to its own farm. Such increase in the number of boosting pages increases the contribution of that farm to its own target and all others in the alliance, following the patterns shown in Figure 10. Obviously, the more new boosting pages are added, the closer the farm gets to the limit for leaving, as stated in Inequalities 9 and 10. The question is, given the current size of farm i, how many pages need to be added to farm i before it is better off on its own?

We can find the answer to this question by taking a look at the difference between the minimum size as determined by Inequalities 9 and 10 and the current size of the farm. Figure 12 presents the corresponding results for our 10 spam farms connected either in a ring or in an alliance with complete core. The horizontal axis shows farm numbers, while the vertical axis represents the minimum number of additional boosting pages a spammer should add, so that the target PageRank after leaving the alliance would be higher than the PageRank when staying in the alliance. For instance, if farm 3 with a current number of 3000 boosting pages would receive approximately 10000 more boosting pages, it could achieve higher target PageRank by splitting off from a complete-core alliance than staying within it. Please note that farm 10 is already above the limit for the ring structure. In such an instance, the spammer might want to leave the alliance, drop some boosting pages, or charge the others for the "loss" incurred due to staying in the alliance.

6 Generalized Link Spam Structures

Our analysis so far has focused on optimal spam farms and how they can be interconnected. However, the use of optimal structures makes it easier to detect spam farms (see Section 7), so spammers might try to deviate from the best structures, even if the rankings of their target pages decrease somewhat. Still, to avoid losing too much PageRank, spammers may not want to deviate too much. This means that "real" structures will still *resemble* the ones we have studied.

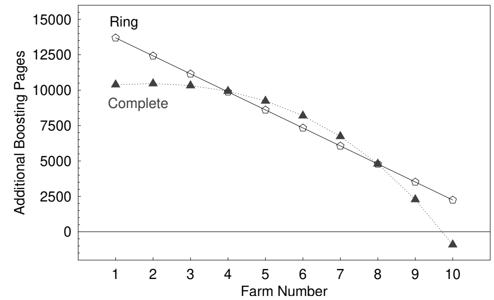

Figure 12: Additional boosting pages required before leaving.

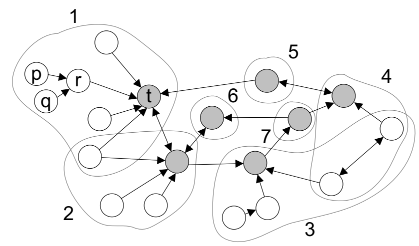

Figure 13: A spam alliance with irregular structure.

To illustrate, consider the graph in Figure 13. What seems to be an irregular, convoluted structure at first, is in fact an alliance of seven spam farms, and can be analyzed as such.

For instance, we can identify several special boosting structures in the figure. The group of pages $\{p, q, r\}$ is one such structure that boosts target t. Boosting structures can always be modeled through an equivalent number of simple boosting pages with an only link pointing to the target. For instance, the contribution of the group $\{p, q, r\}$ in Figure 13 is equivalent to that of $2 \cdot 0.85 + 1 = 2.7$ simple boosting pages. The total boosting target t gets is equal to that produced by $b_1 = 5.2$ simple boosting pages. After accounting for all boosting structures, we find that the total boosting effect for the entire alliance is equivalent to that of $b = 11.55$ simple boosting pages.

We also discover that the target pages (gray nodes) form an optimal core. Accordingly, the total target PageRank is $(cb + 7)/N$.

Thus, the structures encountered in practice can be modeled by equivalent optimal structures. In particular, the effect of complex boosting structures can be modeled easily through an equivalent number of simple boosting pages, as illustrated in the previous example. Leakage can also be incorporated as if it were some additional boosting pages, as mentioned in Section 3.4. The structure that interconnects the target pages may be a ring or a complete core, or one of the graphs discussed in Section 4.4. In conclusion, we believe that the insights obtained for the regular spam farms and alliances also hold for the generalized link spam structures.

7 Countermeasures

In this paper we have studied spam farms and alliances from the point of view of the spammer: What are the optimal structures? How can they be interconnected? What are the costs and benefits when spammers collaborate? We have argued that understanding the spammer's side can provide essential insights for combating spam. After all, how can one fight spam without knowing what one is up against?

Of course, understanding farms and alliances does not automatically solve the spam detection problem. As a matter of fact, detection is in its infancy, and as one develops better tools for combating spam, spammers adapt and devise more resistant schemes. In this closing section we briefly summarize some of the spam detection techniques that have been developed to date, and we argue that understanding the spammer's side played an important role in developing these techniques.

The basic idea for detecting link spam is to identify, directly or indirectly, structures like the ones we have studied in this paper. While the presence of these structures does not necessarily mean link spamming, it does indicate potential candidates. We next outline three schemes used in counteracting some forms of link spamming.

1. In practice, large spam farms are often machine-generated and have very regular structures. A number of techniques are available to detect such instances of link spam. For example, Fetterly *et al.* [2] analyze the indegree and outdegree distributions of web pages. Most pages have in- and outdegrees that follow a Zipfian distribution. Occasionally, however, one encounters substantially more pages with the exact same in- or outdegrees as expected according to the distribution. The authors find that the vast majority of such outliers are spam pages that belong to large farms.

2. A common feature of the alliances presented in this paper is that target pages are very effective at harnessing the boosting provided by other pages. For instance, the two target pages of the alliance in Section 4.1.3 have a total PageRank score $p_0 + q_0 = [c(k+m)+2]/N$, most of it coming from the $(k+m)$ boosting pages. At the same time, the contribution of the boosting pages is only

$$c\left(\sum_{i=1}^{k} p_i + \sum_{j=1}^{m} q_j\right) = \frac{c(1-c)(k+m)}{N} .$$

The ratio between the two sums is of order

$$\frac{p_0 + q_0}{\sum_{i=1}^{k} p_i + \sum_{j=1}^{m} q_j} = O\left(\frac{1}{1-c}\right) , \qquad (11)$$

that is, the target pages *amplify* the contribution of the boosting pages by a factor of approximately $1/(1-c)$. This effect is achieved through the strong interconnection between the target pages, and can also be observed for the other optimal alliances that we presented.

Based on the previous observation, Zhang *et al.* [11] provide a method for identifying strongly interconnected groups of web pages. For any group of web pages H, they define the *amplification factor* $\mathrm{Amp}(H)$, which is just the the ratio between the total PageRank of the pages in the group and the contribution received from other pages outside the group, as illustrated in Equation 11. If the amplification factor of a group is close to $1/(1-c)$, it is said that the pages in the group are *colluding*. Since the target pages of spam alliances collude, the corresponding large amplification factors reveal them.

3. Another observation that we can make about spam alliances is that most of the target PageRank scores are accumulated through boosting. Accordingly, boosting pages contribute their minimal score (which is due to the random jump) to increase the ranking of the target(s). We can measure the magnitude of the boosting effect as follows. Consider for instance the farm structure in Figure 1. The PageRank score of the target is

$$p_0 = \frac{1}{1-c^2}\left[c\lambda + \frac{(1-c)(ck+1)}{N}\right] .$$

Now, if one "cuts off" the random jump going to the pages in the farm, the target score is only $\tilde{p}_0 = \frac{1}{1-c^2}c\lambda$, and the difference $(p_0 - \tilde{p}_0)$ is large. Thus, for target pages that benefit from significant boosting, the ratio $(p_0 - \tilde{p}_0)/p_0$ is large. On the other hand, for web pages that do not benefit from boosting, the ratio is close to zero.

Based on our understanding of farms and alliances, and using this observation, we have developed a new spam detection scheme [3]. The method combines two scores for each web page i: the regular PageRank p_i and a biased PageRank $\tilde{p}_i$, for which the random jump is "cut off" (set to zero) for all but some known non-spam pages. The ratio $(p_i - \tilde{p}_i)/p_i$, is called the *relative spam mass* $\mathrm{Mass}(i)$ of page i, and used to identify the target pages of the largest spam farms. In our experiments (using the full August 2003 index of the AltaVista search engine) we have found that on the order of 95% of the sites identified by our detection scheme (25000 out of 31 million sites) are actual link spam target sites of very large farms.

The techniques we have outlined are useful, but are still far from perfect. Solution 1 often fails to identify non-regular farm structures (like the one shown in Figure 13), which are typical of more sophisticated (and higher-ranking) spammers. Solution 2 identifies *any* colluding group of pages, which may or may not be spam. (For example, the colluding pages could simply be weblogs frequently referencing each other.) Thus, it is not a spam detection technique *per se*, though it could have a pivotal role in spam detection. Solution 3 is effective in detecting instances of significant boosting, but, for example, it fails to detect target pages that obtain most of their scores through leakage. On the positive side, both Solutions 2 and 3 are effective as spammers deviate from the optimal structures in an effort to conceal their farms and alliances, as discussed in Section 6.

Incidentally, Solutions 2 and 3 could be used together: first, relative mass can help spotting out some pages of a spam farm, then the amplification factor can be used to identify neighboring pages that together render a very effectively organized (highly colluding) link spam structure.

The presented solutions identify only some of the pages of a farm or alliance, typically the core. Other techniques, such as the spectral analysis of the coreference matrix, could then be used to reveal the other connected spam pages.

8 Conclusions

The analysis that we have presented shows how the PageRank of target pages can be maximized in spam farms. Most importantly, we find that there is an entire class of farm structures that yield the largest achievable target PageRank score. All such optimal farm structures share the following properties:

- All boosting pages point to and only to the target,
- All hijacked links point to the target,
- There are some links from the target to one or more boosting pages.

We have investigated how spammers with originally unconnected farms could cooperate and set up alliances that increase the target PageRank scores. We presented the optimal alliance for two farms, and introduced two possible structures for larger alliances, one with the targets forming a rings and another with the targets forming a complete core. Our major finding is that alliances could further improve the PageRank of each target in the alliance; the distribution of target PageRank scores depends on the way the targets are interconnected.

We have also analyzed the dynamics of alliances, determining under what conditions should new farm be added, or should current members leave an existing alliance.

As argued, a first, critical step in combating link spam is understanding what one is up against. We believe that our analysis of spam farms and alliances provides a solid understanding of some spamming techniques, and could lead to effective schemes for combating link spam.

References

[1] M. Bianchini, M. Gori, and F. Scarselli. Inside PageRank. *ACM Transactions on Internet Technology*, 5(1), 2005.

[2] D. Fetterly, M. Manasse, and M. Najork. Spam, damn spam, and statistics. In *Seventh WebDB Workshop*, 2004.

[3] Z. Gyöngyi, P. Berkhin, H. Garcia-Molina, and J. Pedersen. Web spam detection based on mass estimation. Technical report, Stanford University, 2005. `http://infolab.stanford.edu/~zoltan/publications.html`

[4] Z. Gyöngyi and H. Garcia-Molina. Link spam alliances. Technical report, Stanford University, 2005. `http://infolab.stanford.edu/~zoltan/publications.html`

[5] Z. Gyöngyi and H. Garcia-Molina. Web spam taxonomy. In *First International Workshop on Adversarial Information Retrieval on the Web*, 2005.

[6] M. Henzinger, R. Motwani, and C. Silverstein. Challenges in web search engines. *ACM SIGIR Forum*, 36(2), 2002.

[7] J. G. Kemény and J. L. Snell. *Finite Markov Chains*. Springer-Verlag, New York, 1976.

[8] A. Langville and C. Meyer. Deeper inside PageRank. *Internet Mathematics*, 1(3), 2004.

[9] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford University, 1998. `http://dbpubs.stanford.edu/pub/1999-66`

[10] N. Sloane. On-line encyclopedia of integer sequences. `http://www.research.att.com/~njas/sequences/`.

[11] H. Zhang, A. Goel, R. Govindan, K. Mason, and B. V. Roy. Making eigenvector-based reputation systems robust to collusion. In *Third Workshop on Algorithms and Models for the Web Graph*, 2004.