Using Spam Farm to Boost PageRank

Ye Du
EECS Department, University of Michigan
2260 Hayward Ave, Ann Arbor, MI 48109-2121, USA
duye@umich.edu

Yaoyun Shi
EECS Department, University of Michigan
2260 Hayward Ave, Ann Arbor, MI 48109-2121, USA
shiyy@umich.edu

Xin Zhao
EECS Department, University of Michigan
2260 Hayward Ave, Ann Arbor, MI 48109-2121, USA
zhaoxin@umich.edu

ABSTRACT

Nowadays web spamming has emerged to take the economic advantage of high search rankings and threatened the accuracy and fairness of those rankings. Understanding spamming techniques is essential for evaluating the strength and weakness of a ranking algorithm, and for fighting against web spamming. In this paper, we identify the optimal spam farm structure under some realistic assumptions in the single target spam farm model. Our result extends the optimal spam farm claimed by Gyöngyi and Garcia-Molina through dropping the assumption that leakage is constant. We also characterize the optimal spam farms under additional constraints, which the spammer may deploy to disguise the spam farm by deviating from the unconstrained optimal structure.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrival

General Terms

Theory, Algorithms

Keywords

Link spamming, PageRank algorithm, Markov chain

1. INTRODUCTION

In the past decade, search engines such as Google, Yahoo, and MSN, etc. have played a more and more important role in our everyday lives. Therefore, web sites that show up on the top of query results lists have had an ever increasing economic advantage. This has given people the incentives to manipulate the search results by carefully designing the content or link structure of a web page —- web spamming [10]. The emergence of web spamming would undermine the reputation of a trusted information resource. A study in 2002 indicated that around 6 to 8 percent of the web pages in a search engine index were spam [7]. This number has increased to around 15 to 18 percent from 2003 to 2004 [11, 1]. This increasing tendency makes many researchers believe that web spamming will become a major challenge for web search [12].

There are two major categories of web spamming techniques: *term spamming* and *link spamming.* Term spamming boosts the ranking of the *target pages* by editing a page's textual content. For example, one can add thousands of irrelevant keywords as hidden fields to the target pages. A search engine will index those keywords and return the target pages as answers to queries that contain those keywords. Link spamming, on the other hand, manipulates the interconnected link structure of controlled pages, called a *link spam farm*, to boost the connectivity based ranking of the target pages higher than they deserve[9]. PageRank [4], the well known connectivity based ranking algorithm used by the leading search engine Google, is the most popular manipulating target for spammer. Compared to term spamming, link spamming is harder to detect as it can boost the ranking of the target pages without changing the content.

1.1 Single-Target Spam Farm Model

In order to boost the ranking of some web pages in the webgraph, the spammer often set up groups of web pages with carefully devised structure. The group of pages fully controlled by a spammer is called a *spam farm* while non spam farm pages are called *normal pages.* The simplest spam farm model is the single-target spam farm model [8], which has the following characteristics:

1. Each spam farm has a single target page and a fixed number of boosting pages.

2. The spammer wants to boost the target page by adding or deleting the outgoing links of the boosting pages and the target page.

3. It is possible for the spammer to accumulate links from webpages, such as public bulletins and blogs, outside the spam farm. These links and external pages are called *hijacked links* and *hijacked pages*, respectively. The total PageRank score that reaches the spam farm from the hijacked links is refer to as the *leakage.*

1.2 Related Work

As a first step to detect web spam, researchers need to identify various spamming techniques. Langville et al. introduced the problem of link spam analysis as future work

AIRWeb '07, May 8, 2007 Banff, Alberta, Canada.

in their comprehensive survey [14]. Bianchini et al. [3] studied how to design the structure of a webgraph that contains exactly N pages such that a page's PageRank score is maximized. However, in practice, spammer can not have control of all the web pages. When spammer only have control of a small fraction of the webgraph, the optimal link structure of the spam pages, especially links from external pages to the spam, is not addressed in [3].

Gyöngyi and Garcia-Molina [8] first introduced the single target spam farm model. In this model, the spammer wants to boost the PageRank score of the target page by manipulating the outgoing links of the target page and a set of boosting pages. They claimed to have identified a link structure that was *optimal* in maximizing the PageRank score of a single target page. However, we find that the optimality proof of the paper is flawed by assuming PageRank score flowed into spam farm was constant. Nevertheless, given the extremal nature of the optimal link structure, it is not surprising that their conclusion is very close to the correct answer. Moreover, the optimal spam farm structures are easy to detect [18]. The spammer can deviate from the optimal spam farm structure to disguise the spam farm. Unfortunately, this problem was not well addressed in [3, 8].

Adali *et. al.* [17] studied the optimal link structure under the assumption that the spammer only have control of the boosting pages, but not the target page. Moreover, the optimality of the *disguised attack* depends on the *forwarding value*, which has a flavor of PageRank score. In order to compute forwarding value, the spammer have to solve a system of linear equations like PageRank. Considering the size of the real web, this would require the spammer to have huge computation resources. Thus, such attack strategies are not very practical. A. Cheng and E. Friedman [5] quantitatively analyze PageRank score increase of the target page under optimal sybil attacks [3]. Moroever, although the definition of PageRank in [5] is significantly different from the standard one by Page *et. al.* [4], the result in [5] can be easily modified to the standard case by following exactly the same proof ideas.

1.3 Our Results

In this paper, first, we show that the result about optimal spam farm structure in [8] is flawed by assuming constant leakage. In particular, the hijacked links pointing to the boosting pages may be helpful to boost the target page, too(ref. figure 1). This is the main difference between our result and the result of Gyöngyi and Garcia-Molina. The link structure of the hijacked pages is worth special study. Because compared to the target page and the boosting pages that are created by the spammer, the hijacked pages may have diversified contents, associate with different domains and thus are harder to be detected. Moreover, while the *link auction* business, which sells the outgoing links of some webpages with relative high rankings by auction, is emerging, it becomes a great resource for the hijacked pages. Therefore, a careful study of the hijacked links is important for designing spam farm structure in link spamming.

We characterize the optimal spam farm by using sensitivity analysis of Markov chain. Under realistic assumptions, the optimal spam farm(ref. figure 2) should have the following features: 1) The boosting pages point to and only to the target page; 2) The target page points to and only to some of the generous pages, which are the web pages that only point to the target page; 3) Spammer accumulates as many hijacked links as possible.

In an optimal spam farm, the boosting pages, as well as all the hijacked links that the spammer is able to accumulate, must point to the target page. This structure is easy to detect. In order to disguise the spam farm, the spammer may deviate from the optimal one. Thus we also characterize optimal spam farms under some realistic constraints. We show that in the optimal spam farm, if the target page must point to some non spam farm pages, the target page should point to all the generous pages(ref. figure 3); if some of the boosting pages can not directly point to the target page, they should point to some of the generous pages(ref. figure 4); if the hijacked links can not directly point to the target page, spammer should accumulate as many hijacked links pointing to the boosting pages as possible(ref. figure 5).

The rest of the paper is organized as follows. Section 2 introduces PageRank algorithm and the mathematical foundation of this paper. Section 3 revisits the optimal spam farm structure in [8]. Next, Section 4 characterizes the optimal spam farm under some realistic assumptions and section 5 characterizes a set of optimal spam farms under some natural constraints. We conclude in Section 6 and suggest some future directions.

2. PRELIMINARIES

In this section, we briefly describe PageRank algorithm and the sensitivity analysis of Markov chain, which is the main mathematical tool of this paper.

2.1 The PageRank Algorithm

We follow [14, 4] to define PageRank. Let $G = (V, E)$ [1] be a directed graph with vertex set V and edge set E. We assume that there is no self-loop in G. Let $N = |V|$, and for a vertex $i \in V$, denote by $\mathrm{out}(i)$ the out-degree of i. The *transition matrix* of G is $T = [T_{ij}]_{1 \le i,j \le N}$:

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

Denote by $e \in \mathbb{R}^N$ the all 1 row vector $(1, 1, \cdots, 1)$, and by $E \in \mathbb{R}^{N \times N}$ the all 1 matrix. Let $\bar{T}$ be identical to T except that if a row in P is all 0, it should be replaced by e/N.A page without outgoing links is called a *dangling* page. For some constant c, $0 < c < 1$, the transition matrix for the PageRank Markov chain is

$$P = c\bar{T} + (1 - c)E/N.$$

The PageRank π is the stationary distribution, i.e.,$\pi P = \pi$, of the above Markov chain M. Our definition of PageRank score is different from the definition in [8], where the PageRank score $\bar{\pi}$ is defined as $\bar{\pi} = c\bar{\pi}T + \frac{(1-c)}{N}e$. However, the two definitions yield the same relative PageRank scores [14, 8]. This relation can be represented as $\pi = \alpha\bar{\pi}$, where α is a constant. When the constraints $\sum_i \pi_i = 1$ and $\sum_i \bar{\pi}_i = 1$ are enforced, the two definitions will induce exactly the same PageRank score for each page.

[1] We assume that there is no self loop in the webgraph. All our results can be easily extended by following the same proof ideas in this paper if self loop is allowed.

2.2 Sensitivity Analysis of Markov chain

Our theoretical foundation consists of one theorem addressing the mean first passage time of Markov chain [13], two theorems about fundamental matrix of Markov chain [13] and one theorem of the monotone property of Markov chain [6] . Due to the space constraint, we only present the theorem statements. Interested readers can refer to the standard textbooks such as [13, 6, 2] for details. We fix a Markov chain of N states and of which the transition matrix is P.

Definition 1. *The* mean first passage time from i to j, *denoted by* m_{ij}*, is the expected number of steps entering State* j *starting from State* i.

Theorem 1. *[13] Let* P *be the transition matrix of a regular Markov chain. We have the following facts:*

1. *For any two states* i *and* j, $m_{ij} = 1 + \sum_{k \neq j} p_{ik} m_{kj}$;
2. *For any state* i*, the stationary distribution* $\pi_i = \frac{1}{m_{ii}}$;
3. *For any two states* $i \neq j$*, changing the transition probabilities of* j *to any other states does not change* m_{ij}.

Definition 2. *The* fundamental matrix Z *of the transition matrix* P *of a Markov chain is defined as:*

$$Z \stackrel{\text{def}}{=} (I - (P - B))^{-1}.$$

Here $B \stackrel{\text{def}}{=} \lim_{k \to \infty} P^k$.

Two fundamental results about the fundamental matrix are:

Theorem 2. *[13] The fundamental matrix of a regular Markov chain with transition matrix* P *always exists, and further more,*

$$Z = I + \sum_{k=1}^{\infty} (P - B)^k.$$

Theorem 3. *[6] Let* P *and* $\tilde{P}$ *be the transition matrices of two Markov chains and* $\tilde{P} = P + \Delta$*. Suppose* $\tilde{\pi}$ *and* π *are the stationary distributions of* $\tilde{P}$ *and* P *while* Z *is the fundamental matrix of* P*. We have the following facts:*

1. $\tilde{\pi}$=$\tilde{\pi}\Delta Z$+π;
2. Z *is diagonally dominant over columns, that is,* $z_{jj} \geq z_{ij}$ *for all* i *and* j*. Furthermore, for all* i *and* j, $j \neq i$, $z_{jj} - z_{ij} = m_{ij}\pi_j$;

Chien et al. [6] proves the following useful monotone property of Markov chain.

Theorem 4. *Let* P *be the transition matrix of a finite state regular Markov chain and let* i *and* j *be arbitrary states of* P*. Let* Δ *be a matrix that is zero everywhere except in row* i*, the* (i, j) *entry is the only positive entry, and* $\tilde{P} = P + \Delta$ *is also the transition matrix of a regular Markov chain. Let* $\tilde{\pi}$ *denote the stationary distribution of* $\tilde{P}$*. Then* $\tilde{\pi}_j > \pi_j$.

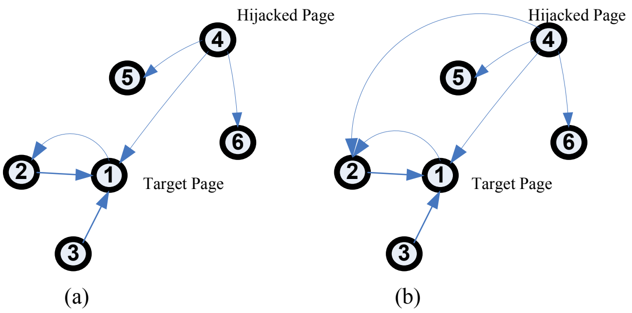

Figure 1: Counter Example

3. THE OPTIMAL SPAM FARM STRUCTURE REVISITED

In the seminal paper on the single target spam farm model [8], Gyöngyi et al. claimed that the linking structure of a spam farm would be optimal if and only if all the following conditions are satisfied:

1. all boosting pages point to and only to the target page;
2. there are no links among the boosting pages;
3. the target page points to some or all of the boosting pages;
4. all hijacked links point to the target page.

Unfortunately, the result is not always true. A counter example is given in Figure 1. In this example, we set $c = 0.85$. The webgraph has 6 nodes (i.e. web pages). Node 1 is the target page, Node 2 and 3 are the boosting pages while node 4 is the hijacked page. Figure 1(a) shows the initial link structure of the webgraph, in which hijacked page 4 only points to the target page 1. The link structure of webgraph in figure 1(a) meets all conditions of the optimal spam structure in [8], inducing PageRank score 0.4223 for the target page. Now, if we add an additional link from hijacked page 4 to boosting page 2, the resulting structure, as shown in figure 1(b), violates condition 4 of optimal spam farm claimed in [8] but gives target page 1 PageRank score 0.4245, which is higher than the previous case.

The reason why the optimal structure in [8] breaks down is: Gyöngyi et al.'s proof lies in the mistaken assumption that leakage, the flow of PageRank scores into the spam farm, is constant. This value is obviously not a constant and may depend substantially on the link structure of spam farm. The reasons are: first, although Gyöngyi et al.'s assume that the spammer does not have full control of the hijacked pages, it is still possible that the spammer can add multiple links on the hijacked page; thus leakage should depend on the number of hijacked links. Actually, in practice, the hijacked pages could be the bulletin or online blogs. The webpages in link auctions could also be used as hijacked pages. Therefore, in both cases, the spammer can add multiple links on the hijacked pages. Second, even if the link structure of the hijacked pages is fixed, the leakage could still be a variable. This is because PageRank is a global property and any change of local structure(such as the links of the target page and the boosting pages) can influence the global distribution of PageRank. In all, the assumption that the leakage is constant is mistaken.

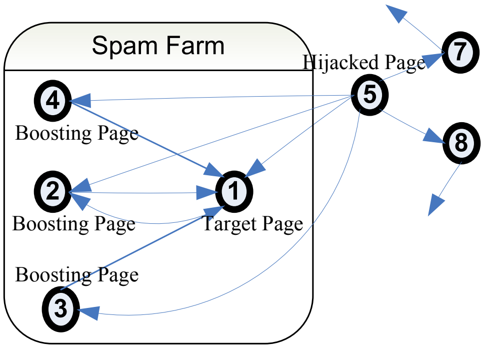

Figure 2: Optimal Spam Farm

Actually, by assuming the leakage is constant, the proof in [8] implicitly assumes that there is no links from the target page or the boosting pages to the hijacked pages.(Otherwise, the leakage is a variable.) Moreover, if we drop the constant leakage assumption, the proof in [8] does not work anymore. Thus it requires us to develop new techniques to deal with the case when leakage is a variable. In the next section, we introduce the main tools for our analysis, which is fundamentally different from Gyöngyi et al.'s approach.

4. CHARACTERIZATION OF OPTIMAL SPAM FARM

In this section, we shall give the characterization of optimal spam farm under some realistic assumptions. First, we introduce two definitions.

DEFINITION 3. *A webgraph is called a realistic webgraph if*

1. *The number of web pages N is large enough such that $2c\frac{N-1}{N} > 1$;*
2. *The number of dangling pages is at least 2.*

In practice, c takes value from 0.8 to 0.9 [15, 4] while Google indexes around 6 billion web pages, many of which are dangling pages. Therefore the definition of realistic webgraph is very natural.

DEFINITION 4. *If a web page only points to the target page, we call it a generous page.*

Then we characterize the optimal spam farm as shown in Figure 2 in the following theorem.

THEOREM 5. *In a realistic webgraph, we assume that each hijacked page already points to a set of non generous pages such that at least two of them are not hijacked pages and do not point to any generous page. Then a spam farm is optimal iff*

1. *The boosting pages point to and only to the target page;*
2. *The target page points to and only to some of the generous pages;*
3. *The hijacked pages point to the target page and all the boosting pages.*

In Theorem 5, the assumptions about the hijacked pages are realistic. Because in the real world, the hijacked pages are most likely to be online bulletins or blogs. Those pages probably point to a number of normal web pages that are neither generous pages nor hijacked pages. Moreover, the contents of the hijacked pages are not likely to be relevant to spam farm pages. Based on the belief that one page points to another if they are relevant, it is reasonable to assume that at least two of the web pages, which the hijacked pages point to, do not point to generous pages.

Although we give the optimal spam farm in the above theorem, spammer may not achieve the maximum PageRank score for the target page in practice. The reasons are two folds. First, there are enormous bulletin pages and blogs that can be used as hijacked pages and it is impossible for spammer to add hijacked links to all of them. But, according to the proof of Theorem 5, in order to maximize the PageRank score of the target page, spammer should hijack as many pages that satisfy our assumption as possible. Secondly, adding a link from a hijacked page to a normal page may boost the target page, too. However according to the definition of single target spam farm model, the hijacked links should point to spam farm pages. Thus we do not consider such case in our theorem.

In our proof of Theorem 5, we find out the optimal structure by optimizing outgoing links of the target page and the boosting pages as well as the hijacked links step by step. This proof is totally different from proof in [8], which solves the optimization problem as a whole. Consequently, our method will provide more insights about the effect of adding or deleting links compared to the method of Gyöngyi and Garcia-Molina. In the following proof, let t be the target page, g be a generous page, d be a dangling page, b be a boosting page and h be a hijacked page.

First we study the outgoing links of the boosting pages in the optimal spam farm.

LEMMA 1. *In the optimal spam farm, the boosting pages should point to and only to the target page.*

PROOF. First we claim that the boosting pages should point to the target page in the optimal spam farm. We will prove this claim in two cases. The first case is that the boosting page b has nonzero out degree. When adding (b, t) into E, $\forall k \neq t$, if edge $(b, k) \in E$, p_{bk} decreases from $c\frac{1}{l} + (1-c)\frac{1}{N}$ to $c\frac{1}{l+1} + (1-c)\frac{1}{N}$, where l is the out degree of page b before adding the link (b, t); if edge $(b, k) \notin E$, p_{bk} does not change. At the same time, p_{bt} increases from $(1-c)\frac{1}{N}$ to $c\frac{1}{l+1} + (1-c)\frac{1}{N}$. The second case is that b has zero out degree. When adding the edge (b, t) to E can increase p_{bt} from $\frac{1}{N}$ to $c + (1-c)\frac{1}{N}$. $\forall k \neq t$, p_{bk} decreases from $\frac{1}{N}$ to $\frac{1-c}{N}$. According to theorem 4, adding the edge (b, t) can increase the PageRank score of the target page t. Therefore the boosting pages should point to the target page in the optimal spam farm.

Next we claim that the boosting pages can not point to any other pages besides the target page in the optimal spam farm. We would prove this claim by contradiction. If b points to t and some other pages $k_1, k_2, ..., k_l$ other than t, we will delete all the links from b to $k_1, k_2, ..., k_l$ and see what happens. Before the deletion, $p_{bk_1}, ..., p_{bk_l}$ and p_{bt} should be $c\frac{1}{l+1} + (1-c)\frac{1}{N}$; after deletion, $p_{bk_1}, ..., p_{bk_l}$ would be $(1-c)\frac{1}{N}$ and p_{bt} would be $c + (1-c)\frac{1}{N}$, it is obvious that p_{bt} increases and $p_{bk_1}, ..., p_{bk_l}$ decreases. According to theorem

4, the deletion operations can increase the PageRank score of the target page t. Therefore the boosting pages can not point to any other pages besides the target page.

Finally putting together the two claims proves the lemma. □

Lemma 1 implies that in the optimal spam farm, the boosting page should be a generous page. However, please note that not every generous page is a boosting page, since a normal page in the webgraph may only point to the target page too.

Next, we study the outgoing links of the target page in the optimal spam farm. The basic idea is that according to Theorem 1, if we want to maximize the PageRank score of target page t, we need to minimize the mean first passage time m_{tt}. Since the mean first passage time will play an important role in our proofs, the following two lemmas would address the mean first passage times of generous pages and dangling pages to the target page. Please recall that m_{ij} stands for the mean first passage time from page i to page j while g and t stand for the generous page and the target page respectively.

Lemma 2. *For any page k, $m_{gt} \leq m_{kt}$. Moreover, $m_{gt} = m_{kt}$ iff k is a generous page.*

Proof. According to Theorem 1,

$$m_{gt} = 1 + \sum_{i \neq t} p_{gi} m_{it} \tag{1}$$

$$m_{kt} = 1 + \sum_{i \neq t} p_{ki} m_{it} \tag{2}$$

According to the PageRank algorithm, for any page $i \neq t$, $p_{gi} = \frac{1-c}{N} \leq p_{ki}$. It is obvious that $m_{gt} \leq m_{kt}$. Moreover, $m_{gt} = m_{kt}$ iff for any page $i \neq t$, $p_{gi} = p_{ki}$, which implies that k is a generous page too. □

Lemma 3. $m_{dt} - m_{gt} \geq c\frac{N-1}{N} m_{gt}$.

Proof. According to equations 1 and 2, we can get

$$m_{dt} - m_{gt} = \frac{c}{N} \sum_{k \neq t} m_{kt} \tag{3}$$

Since for any k, $m_{kt} \geq m_{gt}$, we can get

$$m_{dt} - m_{gt} \geq c\frac{N-1}{N} m_{gt}$$

□

Based on the above two lemmas, we can characterize the outgoing links of the target page in the optimal spam farm.

Lemma 4. *In a realistic webgraph, the target page should point to and only to some of the generous pages in the optimal spam farm.* [2]

Proof. According to Theorem 1, if we want to maximize the PageRank score of the target page t, we need to minimize the mean first passage time m_{tt}. We would prove this lemma by proving the following three claims.

First, we claim that in the optimal spam farm, t have nonzero out degree. Because if t has zero out degree,

$$m_{tt} = 1 + \frac{1}{N} \sum_{i \neq t} m_{it}$$

However, if t points to a generous page g,

$$\begin{aligned} \widetilde{m_{tt}} &= 1 + c \cdot m_{gt} + \frac{1-c}{N} \sum_{i \neq t} m_{it} \\ &= 1 + c(m_{gt} - \frac{1}{N} \sum_{i \neq t} m_{it}) + \frac{1}{N} \sum_{i \neq t} m_{it} \end{aligned}$$

According to the assumption of a realistic webgraph, there are at least two dangling pages d_1 and d_2. Therefore

$$N \cdot m_{gt} - \sum_{i \neq t} m_{it} = \sum_{i \neq t, d_1, d_2} (m_{gt} - m_{it}) + (3m_{gt} - m_{d_1 t} - m_{d_2 t})$$

According to Lemma 2, $m_{gt} \leq m_{it}$; furthermore, according to Lemma 3 and the assumption that $2c\frac{N-1}{N} > 1$, then $3m_{gt} < m_{d_1 t} + m_{d_2 t}$. Therefore $N \cdot m_{gt} - \sum_{i \neq t} m_{it}$ is negative. Consequently, $\widetilde{m_{tt}} < m_{tt}$. It implies that in the optimal spam farm, t can not have zero out degree.

Next we claim that in the optimal spam, the target page can not point to non generous page. Suppose the set of pages t points to is $\mathcal{K}$. Then

$$m_{tt} = 1 + c\frac{\sum_{i \in \mathcal{K}} m_{it}}{|\mathcal{K}|} + \frac{1-c}{N} \sum_{i \neq t} m_{it}$$

If t only points to generous pages,

$$\widetilde{m_{tt}} = 1 + c \cdot m_{gt} + \frac{1-c}{N} \sum_{i \neq t} m_{it}$$

According to Lemma 2, when $\mathcal{K}$ contains non generous pages, $\widetilde{m_{tt}} < m_{tt}$. It implies that in the optimal spam farm, the target page can only point to generous page.

At last, we claim that in the optimal spam farm, when the target page only points to generous page, the number of generous pages the target page points to does not matter. If t points to $q \in \mathcal{N}$ generous pages,

$$m_{tt} = 1 + \frac{c \cdot q}{q} m_{gt} + \frac{1-c}{N} \sum_{i \neq t} m_{it}$$

If t points to $q + 1$ generous pages,

$$\widetilde{m_{tt}} = 1 + \frac{c \cdot (q+1)}{q+1} m_{gt} + \frac{1-c}{N} \sum_{i \neq t} m_{it}$$

It is obvious that $\widetilde{m_{tt}} = m_{tt}$. Therefore, when the target page only points to generous page, the number of generous pages the target page points to does not matter.

Finally, putting together the three claims proves the lemma. □

The last step to prove Theorem 5 is to study the hijacked links. The proof of Lemma 1 implies that in the optimal spam farm, the hijacked pages should point to the target page. The key question is whether the hijacked pages should point to the boosting pages besides the target page. In order to answer this question, we first prove the following lemma.

Lemma 5. *Suppose a hijacked page h already points to the target page t and a set of non generous pages $\mathcal{K}$, adding the link (h, g) where g is a generous page can boost the target page iff $\sum_{k \in \mathcal{K}} m_{kt} > (|\mathcal{K}| + 1) m_{gt}$.*

[2] If self loop is allowed, the target page should only point to itself in the optimal spam farm.

Proof. Let π_t and $\widetilde{\pi}_t$ be the PageRank score of the target page before and after adding the link (h, g). According to Theorem 3, we can get

$$\widetilde{\pi}_t - \pi_t = \widetilde{\pi_h} \Delta_{h*} Z_{*t}$$

When adding the link (h, g), $\delta_{hg} = -\sum_{i \neq g} \delta_{hi}$ and $\forall i \neq g, \delta_{hi} \leq 0$, then

$$\Delta_{h*} Z_{*t}$$

$$= \sum_{i \neq g, (h,i) \in E} \delta_{hi}(z_{it} - z_{gt})$$

$$= \sum_{i \neq g, t, (h,i) \in E} \delta_{hi}((z_{tt} - z_{gt}) - (z_{tt} - z_{it})) + \delta_{ht}(z_{tt} - z_{gt})$$

Since $z_{tt} - z_{it} = m_{it}\pi_t$, we can get

$$\begin{aligned} \Delta_{h*} Z_{*t} &= \pi_t \delta_{ht} (\sum_{i \neq b, t, (h,i) \in E} (m_{gt} - m_{it}) + m_{gt}) \\ &= \pi_t \delta_{ht} ((|\mathcal{K}| + 1) m_{gt} - \sum_{k \in \mathcal{K}} m_{kt}) \end{aligned}$$

Because $\delta_{ht} < 0$ and $\pi_t > 0$, we know that $\widetilde{\pi}_t > \pi_t$ iff $\sum_{k \in \mathcal{K}} m_{kt} > (|\mathcal{K}| + 1) m_{gt}$. □

Lemma 1 and 5 give a necessary and sufficient condition for the optimal link structure of the hijacked pages. However, spammer may not have any knowledge about the mean first passage time. Therefore, in the following lemma, we will address the link structure for those hijacked pages that satisfy some realistic assumptions. The statement of this lemma has nothing to do with mean first passage time.

Lemma 6. *In a realistic webgraph, suppose a hijacked page h already points to a set of non generous pages $\mathcal{K}$; moreover at least two web pages in $\mathcal{K}$ do not point to any generous page. In the optimal spam farm, h should point to the target page and all of the boosting pages.*

Proof. Suppose $k_1, k_2 \in \mathcal{K}$ do not point to any generous page. We claim that $m_{k_1t} - m_{gt} \geq c\frac{N-1}{N} m_{gt}$(the same inequality holds for k_2). We prove this claim by considering two cases. If k_1 has zero out degree, Lemma 3 proves that $m_{k_1t} - m_{gt} \geq c\frac{N-1}{N} m_{gt}$. If k_1 has non zero out degree, based on equation 1 and 2, we can get

$$\begin{aligned} m_{k_1t} - m_{gt} &= \sum_{i \neq t} (p_{k_1i} - \frac{1-c}{N}) m_{it} \\ &\geq c \centerdot m_{gt} \end{aligned}$$

Therefore $m_{k_1t} - m_{gt} \geq c\frac{N-1}{N} m_{gt}$.

The proof of Lemma 1 implies that in the optimal spam farm, all the boosting pages are generous pages and h should point to the target page. When h already points to the target page, Lemma 5 tells us that adding a hijacked link from h to a generous page g can further boost the target page iff $\sum_{k \in \mathcal{K}} m_{kt} > (|\mathcal{K}| + 1) m_{gt}$. Based on our assumption, we know that

$$\begin{aligned} \sum_{k \in \mathcal{K}} m_{kt} &\geq |\mathcal{K}| m_{gt} + 2c\frac{N-1}{N} m_{gt} \\ &> (|\mathcal{K}| + 1) m_{gt} \end{aligned}$$

Therefore, h should point to all of the boosting pages. The lemma is proved. □

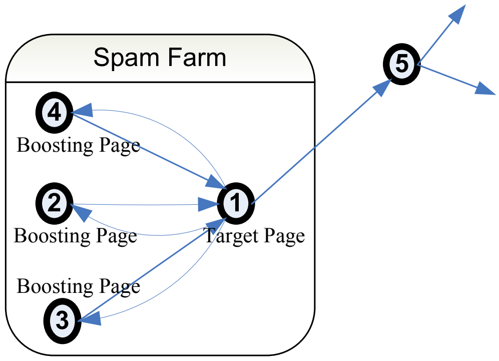

Figure 3: Optimal spam farm when the target page points to non generous pages

At last, if we summarize Lemma 1, 4 and 6, it would give us a unique configuration of the optimal spam farm. This will complete the proof of Theorem 5.

5. OPTIMAL SPAM FARM UNDER CONSTRAINTS

As shown in Theorem 5, in the optimal spam farm, the target page only points to generous pages while the boosting pages only point to the target page. This structure is easy to detect. In order to disguise the spam farm, the spammer may require that the target page should point to some non generous pages or the boosting pages should not directly point to the target page or the hijacked pages should not directly point to the target page. What are the optimal spam farm structures under those constraints is an interesting question.

First, we characterize the optimal spam when the target page is required to point to some non generous pages in the following theorem.

Theorem 6. *If the target page t is required to point to a set of pages $\mathcal{K}$, a spam farm is optimal only if*

1. *The boosting pages point to and only to the target page;*

2. *The target page points to a set of pages $\mathcal{K} \bigcup \mathcal{L}$ such that $(\sum_{k \in \mathcal{K}} m_{kt} + \sum_{l \in \mathcal{L}} m_{lt})/(|\mathcal{K}| + |\mathcal{L}|)$ is minimized, where $\mathcal{L} \subseteq V$.*

Proof. The proof of this theorem directly follows from the proof of Theorem 5. □

In order to design the optimal spam farm under the above constraint, spammer needs to find out the set of web pages $\mathcal{L}$ such that the average mean first passage time of web pages in $\mathcal{K} \bigcup \mathcal{L}$ to the target page is minimized. This requires spammers has the knowledge about the mean first passage time of the webgraph. Given the limited computing resources of a spammer, it is a nontrivial task for him to find out the set $\mathcal{L}$. However, Theorem 6 implies that in the optimal spam farm, the target page should point to all the generous pages, which is shown in Figure 3.

Next, we characterize the optimal spam farm, as shown in Figure 4, when some of the boosting pages can not directly point to the target page in the following theorem.

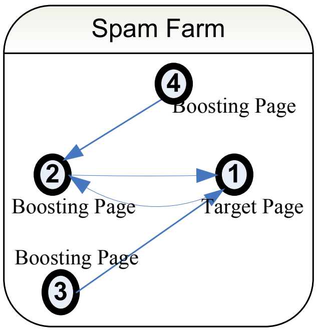

Figure 4: Optimal spam farm when some boosting page can not point to the target page

Theorem 7. *In a realistic webgraph, suppose $\mathcal{B}$ is the set of boosting pages and a subset of it $\overline{\mathcal{B}} \subset \mathcal{B}$ can not directly point to the target page, then a spam farm is optimal only if*

1. *For any page in $\mathcal{B} \backslash \overline{\mathcal{B}}$, it points to and only to the target page;*
2. *For any page in $\overline{\mathcal{B}}$, it points to some of generous pages;*
3. *The target page points to and only to some of generous pages.*

Proof. For the pages in $\mathcal{B} \setminus \overline{\mathcal{B}}$, Lemma 1 implies that they should point to and only to the target page.

For the pages in $\overline{\mathcal{B}}$, first we claim that it can not have zero out degree in the optimal spam farm. Because if $b \in \overline{\mathcal{B}}$ has zero degree, adding the link (b, g) where g is a generous page can boost the target page t iff

$$\sum_{k \neq t} m_{kt} - N \centerdot m_{gt} > 0$$

Given the assumption of a realistic webgraph, similar to the proof of the first claim in Lemma 4, we know that b should have nonzero out degree in the optimal spam farm.

Next, we claim that b can not point to non generous pages in the optimal spam farm. Because when b has nonzero out degree, deleting a link from b to another page j(but still keeps b has non zero out degree) can boost the target page iff

$$\sum_{k \neq j, (b,k) \in E} (m_{kt} - m_{jt}) < 0$$

According to Lemma 2, for a generous page g, m_{gt} is minimum. Combined with the previous necessary and sufficient condition, it implies that b can not point to non generous pages and the number of generous pages b points to does not matter if b only points to generous pages. Consequently, b should point to some of the generous pages in the optimal spam farm.

For the target page, Lemma 4 implies that it should point to and only to some of the generous pages. □

Please note that in the above two characterizations, we ignore the hijacked links to avoid repeatedness. Because according to the proof of Theorem 5, the structure of the hijacked links in the optimal spam is quite independent of the structure of outgoing links of the target page and the boosting pages. Therefore, when the hijacked links need to be taken into consideration, we can follow almost the same analysis as Theorem 5 to find out the structure of the hijacked links in the optimal spam farm.

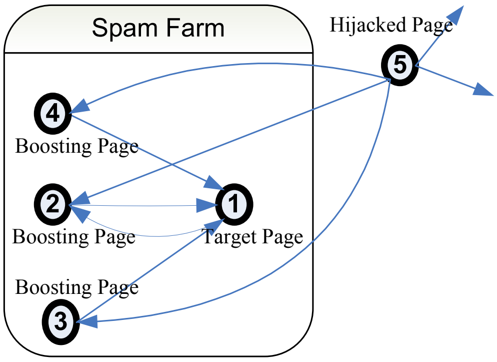

Figure 5: Optimal spam farm when the hijacked pages can not point to the target page

Finally, we characterize the optimal spam farm, as shown in Figure 5, when the hijacked pages can not directly point to the target page in the following theorem.

Theorem 8. *In a realistic webgraph, suppose the hijacked pages already point to some non generous pages and the hijacked pages can not directly point to the target page, then a spam farm is optimal iff*

1. *The boosting pages point to and only to the target page;*
2. *The target page points to and only to some of the generous pages;*
3. *The hijacked pages point to all of the generous pages.*

A similar analysis as the proof of Theorem 7 can show correctness of Theorem 8. Because of the space constraints, we omit the proof here.

6. CONCLUSIONS AND FUTURE WORK

Identifying spamming techniques is the first step to combat web spam. In this paper, we characterized the optimal spam farm structure under some realistic assumptions in the single target spam farm model. Our result extends the conclusion of Gyöngyi and Garcia-Molina [8] by dropping the constant leakage assumption. Moreover, we characterized the optimal spam farms under some natural constraints, which may be deployed by spammer to disguise the spam farm. We believe that the sensitivity analysis of Markov chain is a fundamental tool to design and analyze link spamming. Furthermore, our techniques are not only useful to boost the PageRank score, but also helpful to analyze other ranking methods based on the stationary distribution of Markov chain, such as *Invariant Ranking method* [16].

One natural open problem of our work is to estimate the mean first passage time of a Markov chain. With the knowledge of mean first passage time, spammer can design more sophiscated spam farms that disguise the detection of search engine. Another direction could be that using mean first passage time information to identify spam farm structure.

7. ACKNOWLEDGEMENTS

This work was supported in part by NSF grant 0347078. Y. Du would like to thank H.V. Jagadish for introducing him the problem of link spamming in the single target spam farm model. The authors also acknowledge the anonymous reviewers for their valuable suggestions to improve this paper.

8. REFERENCES

[1] T. Sarls A. Benczr, K. Csalogny and M. Uher. Spamrank - fully automatic link spam detection. In *Proc. Int'l Workshop Adversarial Information Retrieval on the Web*, 2005.

[2] David Aldous and James Allen Fill. Reversible markov chains and random walks on graphs. www.stat.berkeley.edu/ aldous/RWG/book.html.

[3] Monica Bianchini, Marco Gori, and Franco Scarselli. Inside pagerank. *ACM Trans. Inter. Tech.*, 5(1):92–128, 2005.

[4] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. In *WWW7: Proceedings of the seventh international conference on World Wide Web 7*, pages 107–117, Brisbane, Australia, 1998.

[5] Alice Cheng and Eric Friedman. Manipulability of pagerank under sybil strategies. 2006. First Workshop on the Economics of Networked Systems.

[6] Steve Chien, Cynthia Dwork, Ravi Kumar, Daniel R. Simon, and D. Sivakumar. Link evolution: Analysis and algorithms. *Internet Mathematics*, 1(3):277–304, 2003.

[7] Dennis Fetterly, Mark Manasse, and Marc Najork. Spam, damn spam, and statistics: using statistical analysis to locate spam web pages. In *WebDB '04: Proceedings of the 7th International Workshop on the Web and Databases*, pages 1–6, New York, NY, USA, 2004. ACM Press.

[8] Zoltán Gyöngyi and Hector Garcia-Molina. Link spam alliances. In *VLDB '05: Proceedings of the 31st international conference on Very large data bases*, pages 517–528. VLDB, 2005.

[9] Zoltán Gyöngyi and Hector Garcia-Molina. Spam: It's not just for inboxes anymore. *IEEE Computer Magazine*, 38(10):28–34, October 2005.

[10] Zoltán Gyöngyi and Hector Garcia-Molina. Web spam taxonomy. In *First International Workshop on Adversarial Information Retrieval on the Web*, 2005.

[11] Zoltán Gyöngyi, Hector Garcia-Molina, and Jan Pedersen. Combating web spam with TrustRank. In *Proceedings of the 30th International Conference on Very Large Databases*, pages 576–587. Morgan Kaufmann, 2004.

[12] Monika R. Henzinger, Rajeev Motwani, and Craig Silverstein. Challenges in web search engines. *SIGIR Forum*, 36(2):11–22, 2002.

[13] John G. Kemeny and J Laurie Snell. Finite markov chains, 1960. D. Van Nostrand Company.

[14] A. Langville and C. Meyer. Deeper inside pagerank. *Internet Mathematics*, 1(3):335–380, 2005.

[15] Andrew Y. Ng, Alice X. Zheng, and Michael I. Jordan. Link analysis, eigenvectors and stability. In *IJCAI*, pages 903–910, 2001.

[16] Ignacio Palacios-Huerta and Oscar Volij. The measurement of intellectual influence. *Econometrica*, 72(3):963–977, 2004.

[17] Tina Liu Sibel Adali and Malik Magdon-Ismail. Optimal link bombs are uncoordinated. *In Proceeding of AIRWeb*, 2005.

[18] Baoning Wu and Brian 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 Press.