Using Spam Farm to Boost PageRank

Ye Du

Joint Work with: Yaoyun Shi and Xin Zhao

University of Michigan, Ann Arbor

Roadmap

- Introduction: **Link Spam** and PageRank
- Critics about a Previous Work
- Optimal Spam Farm(OSF)
- Optimal Spam Farm under Constraints
- Conclusion and Future Work

Background: Web Spamming I

- What is PageRank?
  —— The connectivity-based webpages ranking algorithm used by Google
- Why to boost the PageRank?
  —— High ranking may bring economic advantages
- The activity of maliciously boosting webpages' ranking is called **web spamming.**

Background: Web Spamming II

- In 2002, around 6 to 8 percent of the web pages a search engine indexes were spam [FMN04].
- This number has increased to around 15 to 18 percent from 2003 to 2004 [GMP04].
- Web spamming is a major challenge for web search[HMS02]!

How to do web spamming [GM05]:

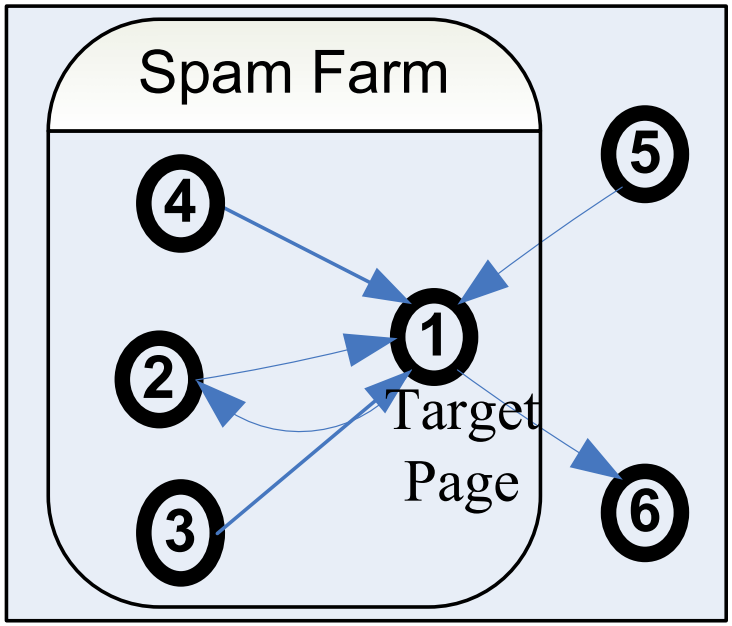
- Text Spam or Term Spam
- Link Spam

Background: Spam Farm Model

Single Target Spam Farm [GM05-VLDB]

- Each spam farm has a single **target page** and a fixed number of **boosting pages.**
- Spammers have full control of the outgoing links of the target page and the boosting pages.
- Spammers want to boost the target page by setting up the outgoing links of the boosting pages and target page.
- It is possible for spammers to accumulate links, which are called **hijacked links**, from pages outside the spam farm. But spammers only have partial control of such links. E.G. posting a link on another person's blog. We call the page with a hijacked link a **Hijacked Page.**

Background: Spam Farm Model



A Webgraph with Spam Farm

Key Questions

- Motivation: In order to fight against web spamming, the first step is to understand the techniques of it.
- Our work focuses on the **link spamming** on **PageRank** algorithm.
- Given a spam farm, how to set up
  - the outgoing links(add or delete) of target page
  - the outgoing links(add or delete) of boosting pages
  - the hijacked links(add)

  such that the PageRank score of the target page is maximized?
- A spam farm is **optimal** if it can maximize the PageRank score of the target page.

Roadmap

- Introduction: Link Spam and **PageRank**
- Critics about a Previous Work
- Optimal Spam Farm(OSF)
- Optimal Spam Farm under Constraints
- Conclusion and Future Work

PageRank Algorithm I

Basic idea: [BP98] given the webgraph $G = (V, E)$, we define a regular Markov chain M on this webgraph. The PageRank vector is the stationary distribution of M.

Step 1: suppose A is the adjacency matrix.

Example: A= $\begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$

Step 2: normalize matrix A to get P

Example: P= $\begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 \end{pmatrix}$ We call a page without outgoing links **dangling pages.**

PageRank Algorithm II

Step 3: $\overline{P}$ can be created by replacing rows of $\mathbf{0}^T$ in P with vector $(\frac{1}{N}, \frac{1}{N}, ..., \frac{1}{N})$

Example:$\overline{P}$= $\begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix}$

Step 4: transition Matrix of Markov Chain[a]: $M = c\overline{P} + \frac{(1-c)}{N}E$

Example: $M = \begin{pmatrix} 0.05 & 0.475 & 0.475 \\ 0.475 & 0.05 & 0.475 \\ 0.3333 & 0.3333 & 0.3333 \end{pmatrix}$

[a]the transition graph of M is a complete graph

PageRank Algorithm III

- M is regular.
- **Theorem 1 (KS60)** *For a finite state regular Markov chain* P
  - *There exists a unique stationary distribution* π
  - $$B = \lim_{k \to \infty} P^k = \begin{pmatrix} \pi_1 & \pi_2 & \dots & \pi_N \\ \pi_1 & \pi_2 & \dots & \pi_N \\ \pi_1 & \pi_2 & \dots & \pi_N \\ \pi_1 & \pi_2 & \dots & \pi_N \end{pmatrix}$$
- *PageRank* π is the stationary distribution of the Markov chain M *i.e.* $\pi M = \pi$ and $\sum_i \pi_i = 1$
- In the previous example, the PageRank vector is $(0.292, 0.292, 0.416)$

Roadmap

- Introduction: Link Spam and PageRank
- Critics about a Previous Work
- Optimal Spam Farm(OSF)
- Optimal Spam Farm under Constraints
- Conclusion and Future Work

Optimal Spam Farm in [GM05-VLDB]

In their seminal work, Gyöngyi *et al.* [GM05-VLDB] claimed that the PageRank score of the target page can be maximized, if and only if:

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.

A Counter Example

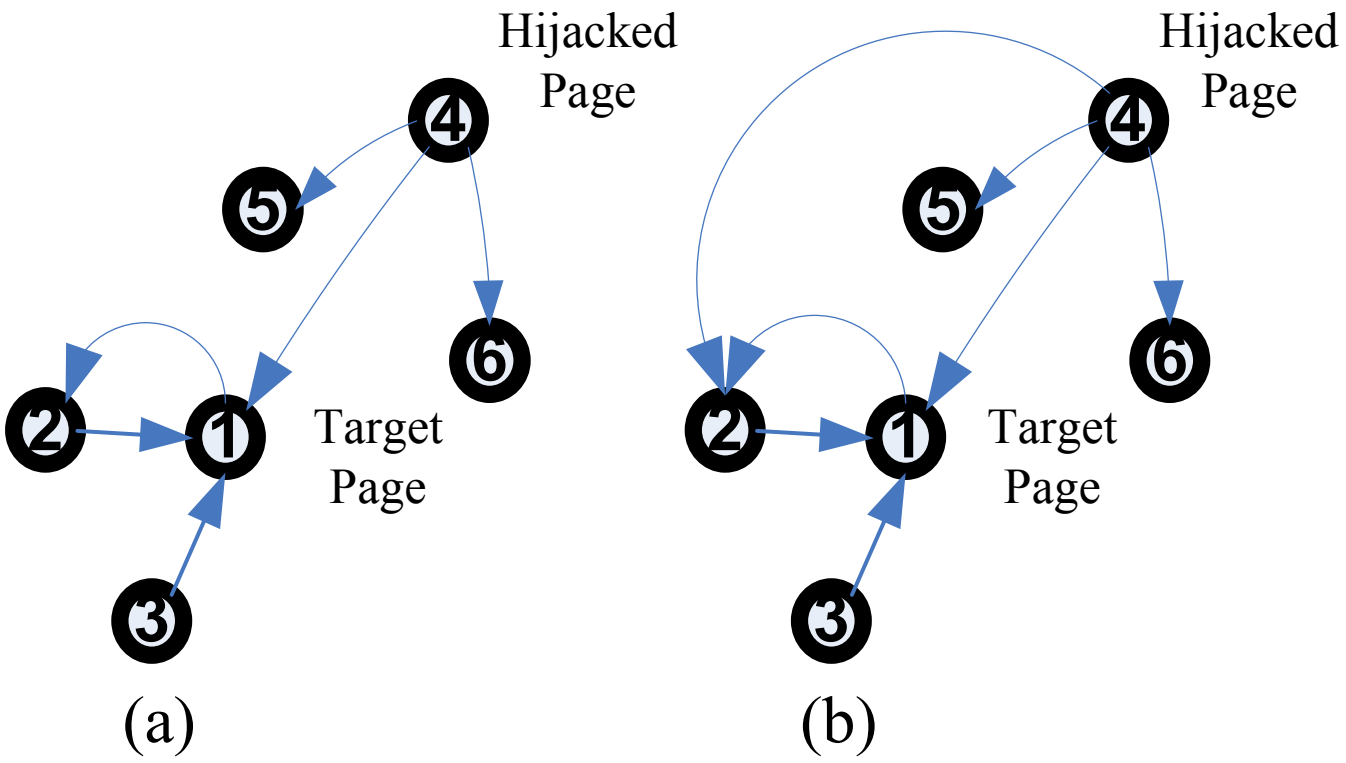


Figure	PageRank score of the target page
a	0.4223
b	0.4245

Constant Leakage Assumption

- *Leakage* is the PageRank score flowed into the spam farm from hijacked pages.
- The previous result is based on the constant leakage assumption.
- In practice, the hijacked pages could be online bulletins, blogs and link auctioned pages. The spammer may add multiple links on the hijacked pages.
- And PageRank is a global quantity.
- Therefore, it is not reasonable to assume that leakage is constant.

Another Drawback of Previous Work

- This special structure can be easily detected by search engine.
- In practice, spammers may like to sacrifice PageRank to disguise the spam farm.
- In order to make the tradeoff between boosting the PageRank and disguising the spam farm, spammers should have a deep understanding of the following question.

”What is the effect of adding or deleting a specific link?”

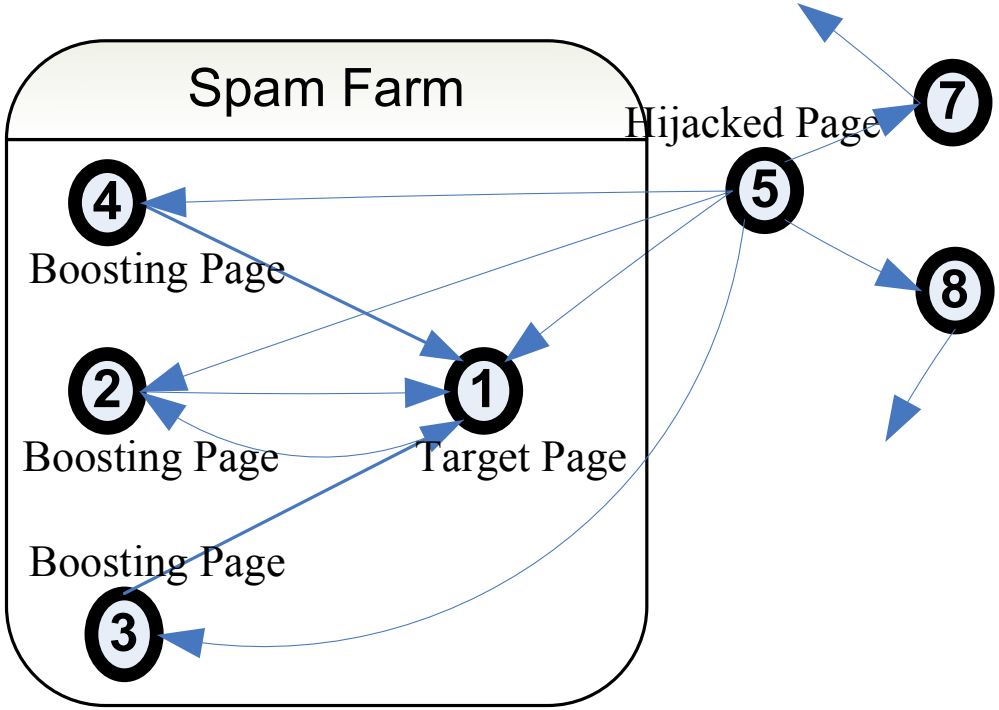
Roadmap

- Introduction: Link Spam and PageRank
- Critics about a Previous Work
- Optimal Spam Farm(OSF)
- Optimal Spam Farm under Constraints
- Conclusion and Future Work

Our Results

Theorem 2 *Under realistic assumptions, a spam farm is optimal iff*

1. *the link structures of the boosting pages and target page are the same as [GM05-VLDB];*
2. *The hijacked pages point to the target page and all the boosting pages.*



Outgoing Links of Boosting Pages

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

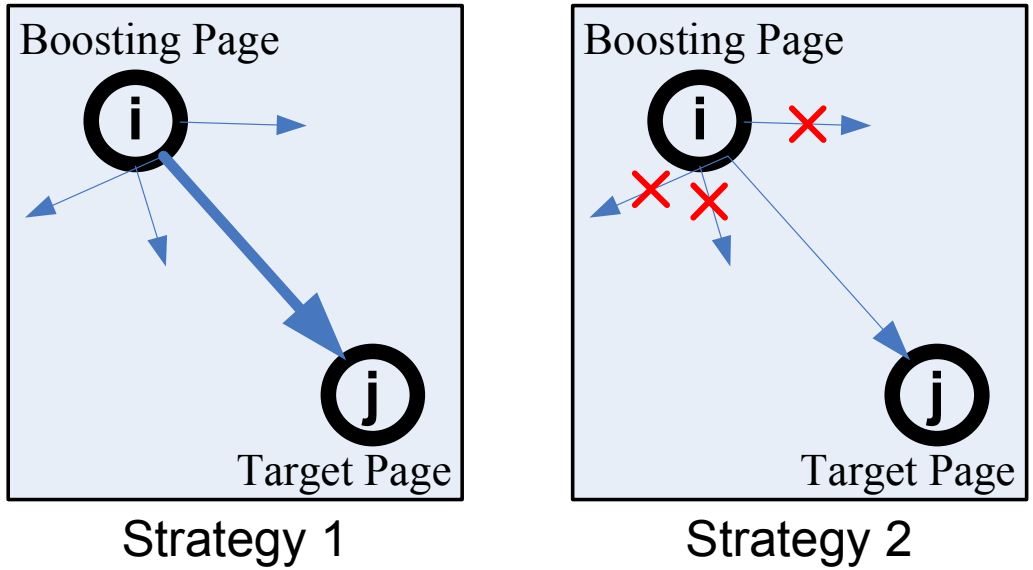


Figure 2: Strategy 1 and 2

Proof of Lemma 1

The proofs of lemma 1 follow directly from:

Theorem 3 (CDKSS03) *Let P be the transition matrix of a finite state regular Markov chain and let i, j be arbitrary states of P. Let Δ be a matrix that is zero everywhere except in row i, $(i, j) > 0$ is the only positive entry, and such that $\widetilde{P} = P + \Delta$ is also the transition matrix of a regular Markov chain. Let $\widetilde{\pi}$ denote the stationary distribution of $\widetilde{P}$. Then $\widetilde{\pi}_j > \pi_j$.*

Outgoing Links of Target Page

Definition 1 *If a web page only points to the target page, we call it a generous page.*

Lemma 2 *For any page k, $m_{gt} \leq m_{kt}$. Moreover, $m_{gt} = m_{kt}$ iff k is a generous page, where m_{gt} is the hitting time from g to t.*

Lemma 3 *In a realistic webgraph, the target page should point to and only to some of the generous pages in the optimal spam farm.*

Proof idea:

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

Outgoing Links of Hijacked Pages

Lemma 4 *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 $\sum_{k \in \mathcal{K}} m_{kt} > (|\mathcal{K}| + 1) m_{gt}$.*

Proof idea:

Theorem 4 *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*,* $j \neq i$*,* $z_{jj} - z_{ij} = m_{ij}\pi_j$;

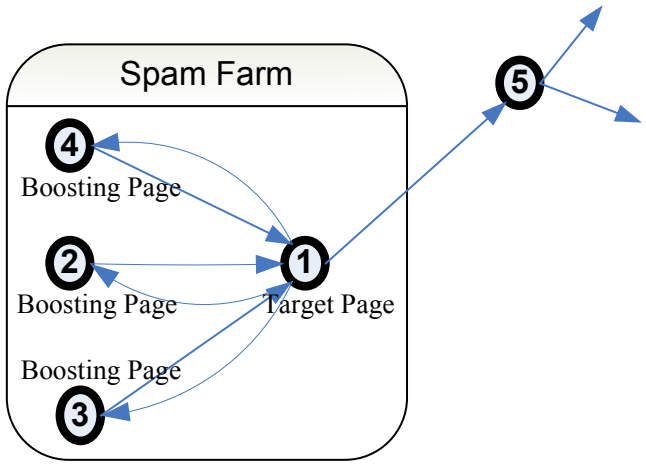
Roadmap

- Introduction: Link Spam and PageRank
- Critics about a Previous Work
- Optimal Spam Farm(OSF)
- Optimal Spam Farm under Constraints
- Conclusion and Future Work

OSF Under Constraints I

Theorem 5 *If the target page t is required to point to a set of pages $\mathcal{K}$, a spam farm is optimal only if*

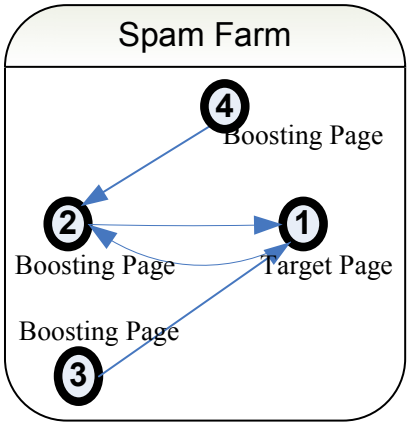
1. *The boosting pages point to and only to the target page;*
2. *The target page points to a set of pages $\mathcal{K} \bigcup \mathcal{L}$ such that $(\sum_{k \in \mathcal{K}} m_{kt} + \sum_{l \in \mathcal{L}} m_{lt})/(|\mathcal{K}| + |\mathcal{L}|)$ is minimized, where $\mathcal{L} \subseteq V$.*



OSF Under Constraints II

Theorem 6 *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*

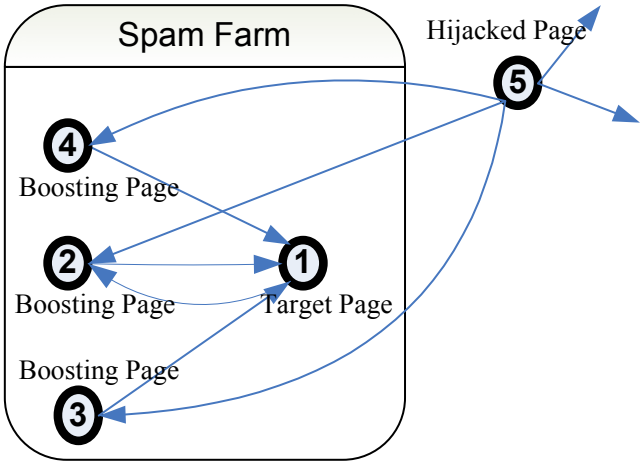
1. *For any page in $\mathcal{B} \setminus \overline{\mathcal{B}}$, it points to and only to the target page;*
2. *For any page in $\overline{\mathcal{B}}$, it points to some of generous pages;*
3. *The target page points to and only to some of generous pages.*



OSF Under Constraints III

Theorem 7 *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.*



Roadmap

- Introduction: Link Spam and PageRank
- Critics about a Previous Work
- Optimal Spam Farm(OSF)
- Optimal Spam Farm under Constraints
- Conclusion and Future Work

Conclusion and Future Work

- We extend a previous work in [GM05-VLDB] by throwing the constant leakage assumption.
- We characterize the optimal spam farm and its variations under different constraints.
- Our work is based on Markov chain and its perturbation theory.
- In the next step, people can investigate how the *hitting time* information can be used to design and detect link spam.

Questions

Questions?