Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2016) 24: 2106 – 2123
© TÜBİTAK
doi:10.3906/elk-1401-202

A new algorithm for detection of link spam contributed by zero-out link pages

Ravi Kumar PATCHMUTHU[1,*], Ashutosh KUMAR SINGH[2], Anand MOHAN[2]
[1]Department of Electrical and Computer Engineering, Curtin University, Miri, Malaysia
[2]National Institute of Technology, Kurukshetra, Haryana, India

Received: 22.01.2014 • **Accepted/Published Online:** 05.07.2014 • **Final Version:** 15.04.2016

Abstract: Link spammers are constantly seeking new methods and strategies to deceive the search engine ranking algorithms. The search engines need to come up with new methods and approaches to challenge the link spammers and to maintain the integrity of the ranking algorithms. In this paper, we proposed a methodology to detect link spam contributed by zero-out link or dangling pages. We randomly selected a target page from live web pages, induced link spam according to our proposed methodology, and applied our algorithm to detect the link spam. The detail results from amazon.com pages showed that there was a considerable improvement in their PageRank after the link spam was induced; our proposed method detected the link spam by using eigenvectors and eigenvalues.

Key words: Link spam, zero-out link/dangling pages, PageRank, adjacency matrix, transition probability matrix, jump probability matrix/Google matrix, eigenvector, eigenvalues

1. Introduction

Link spam is a method of deliberately manipulating the search engine result pages (SERPs) in an unethical manner and a major challenge in the area of web mining. It is also called spamdexing by Gyöngyi and Garcia-Molina [1], i.e. using spamming techniques to improve the index of web pages in the search engine rankings. According to Henzinger et al. [2], web spam is one of the most important challenges for web search engines. Web spam became increased after the introduction of e-commerce in the late 1990s and the dependency on search engines in web information retrieval. Generally, web users only look at the first few pages of search engine results. This is one of the reasons why commercial and business companies push their web sites to appear at the top of search engine results. There is also financial gain for the companies when more visitors visit their web sites. Moreover, web users believe that their search engine results are authentic. The order of web pages appearing in the SERPs is the main reason for spamming in web information retrieval. The intention of web spammers is to mislead search engine ranking algorithms by promoting certain pages to the top SERPs, and consequently misleading the web users with irrelevant information. This kind of misleading can affect the creditability of search engines in web information retrieval.

There are many ways spamming can be achieved. Content spamming and link spamming are two popular techniques used in web information retrieval. Link spamming is a type of spamming used to improve the ranking of certain web pages by having illegitimate links. Link spam is the most effective way of achieving web spam. As the internet grows in an exponential way, web spamming also grows accordingly. Web spammers are looking for every opportunity to induce spamming on the web. One such opportunity is the zero-out link pages in the

*Correspondence: ravi2266@gmail.com

web. Zero-out link pages are the pages in the web without outgoing links, but they may have one or more incoming links. According to [3], zero-out link pages accounts for more than one-fourth of the web's pages and the percentage of zero-out link pages keeps increasing. Zero-out link pages can also be called hanging/dangling pages [4,5]. Zero-out link pages receive a share of the rank from other pages and they do not propagate their rank to other pages. These zero-out link pages are one of the potential targets for spammers because in link structure based ranking algorithms the ranking of web pages is decided only by the number of incoming links and not by the content of the web pages.

This paper proposes a method to detect link spam in the form of irreducible closed subsets contributed by zero-out link pages in the web. Link structure-based ranking algorithms like PageRank [6], hyperlink-induced topic search [7], and stochastic approach for link structure analysis [8] can be affected by this kind of link spam. Among these three ranking algorithms, PageRank is the most affected one because it is the only algorithm that is being used commercially in the Google search engine for ranking web pages. PageRank is a query-independent and link-dependent ranking algorithm used in the famous Google search engine to rank web pages. PageRank computes the importance or relevance of a web page by counting the number of pages that are linking to it (called as "back links"). If a back link comes from an "important" page, then that back link is given a higher weighting than those back links coming from unimportant pages. In a simple way, a link from one page to another page may be considered as a vote. However, not only the number of votes a page receives is considered important, but also the "importance" or the "relevance" of the ones that cast these votes (e.g., a web site is getting a back link from Google.com or Yahoo.com).

This kind of spamming can be a threat to the integrity of the PageRank algorithm. The next section briefly reviews related works in the detection of web spam. Section 3 describes the web model and the mathematical notation used in this paper. The proposed methodology, algorithm, and an example are discussed in section 4. Experiments and results are shown in section 5, and the last section concludes the paper.

2. Related works

There are two kinds of spamming according to Gyöngyi and Garcia-Molina [9]. They are link spamming and term spamming. Link spamming is a kind of spamming where the link structure of the web sites can be altered by using link farms [10,11]. A link farm is a heavily connected set of pages, created explicitly with the purpose of deceiving a link based search engine's ranking algorithm. Term spamming includes content - and meta spamming. Gyöngyi et al. [12] introduces the concept of spam mass and measures the impact of link spamming on a page's ranking. Zhou and Pei [13] introduced effective detection methods for link spam target pages using page farms. Nikita and Jiawei [14] did a survey on web spam detection and compared different spam detection methods. The current paper proposes a method to detect link spam contributed by zero-out link pages.

Bianchini et al. [5] worked on the role of dangling pages and their effect on the PageRank. They introduced the notion of energy, which simply represents the sum of PageRanks for all the pages in a given web site. Eq. (1) below shows the energy balance, which makes it possible to understand the way different web communities interact with each other and helps to improve the ranking of certain pages.

$$E_I = |I| + E_I^{in} - E_I^{out} - E_I^{dp} \tag{1}$$

Let G_1 be a sub graph that represents the energy of a web site. In the above energy balance equation, $|I|$ denotes the number of pages of G_1, E_I^{in} is the energy that comes to G_1 from other sites. E_I^{out} is the energy

that goes out from G_1, which is an energy loss, i.e. hyperlinks going out from G_1 decrease the energy. E_I^{dp} is the energy lost in the dangling pages. Thus, the presence of dangling or zero-out link pages in a web site triggers energy loss. According to [5], in order to maximize energy, one should not only pay attention to the references received from other sites, but also to the dangling pages, and to the external hyperlinks. Dangling pages can be manipulated by spammers to boost the PageRank of web sites. This paper provides a method to detect link spam contributed by zero-out link pages.

Haveliwala and Kamvar [15] conducted research on the second eigenvalue of the Google matrix and the irreducible closed subset, and they mathematically proved the relationship between the second eigenvector and link spam, concluding that the second eigenvalues are artifacts of certain structures in the web graph. Wang et al. [4] addressed a problem called "zero-one gap" in the PageRank algorithm and developed the DirichletRank algorithm, which eliminated the "zero-one gap" problem and proved that their algorithm is more resistant to link spamming than the PageRank algorithm. According to [4], the probability of jumping to a random page is 1 in the case of a dangling or zero-out link page, whereas the probability of a single-out link page drops to 0.15 in most of the cases. There is a big gap between 0 and 1 out link. This gap is referred as "zero-one gap". This "zero-one gap" allows a spammer to manipulate PageRank to achieve spamming. The DirichletRank proposed in [4] not only solves the "zero-one gap" problem, but also the zero-out link problem. Ipsen and Selee [16], Langville and Meyer [17], de Jager and Bradley [18], Gleich et al. [19], and Singh et al. [20] also conducted research on dangling pages and presented methods to compute PageRank. They only proposed methods to handle zero-out link pages, but they did not explore how the zero-out link pages contribute towards link spam.

3. Web model and mathematical notation

Let $G = (V, E)$ be a directed graph with vertices or nodes V and directed edges E. G can be called as web graph [21] in which V is a set of web pages and E is a set of hyperlink connections between pages. For any particular edge $e \in E$, e^i and e^t represent the initial and terminal vertex, respectively. The following are the definitions and notations used in this paper.

Definition 1 *The in-degree (id) of a page i is the number of incoming links $id(i) = \sum_i E_{ji}$.*

Definition 2 *The out-degree (od) of a page i is the number of outgoing links $od(i) = \sum_i E_{ij}$.*

Definition 3 *An adjacency matrix can be created by using the formula in Eq. (2). An element A_{ij} is equal to 1 if page i has a link to page j; it is equal to 0 otherwise. It can also be called a connectivity/link matrix.*

$$A_{ij} = \begin{cases} 1 & \text{if } (i,j) \in E \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

The generalized $n\,times\ n$ adjacency matrix A for a directed graph is shown below:

$$A = \begin{bmatrix} a_{1,1} & a_{2,1} & \cdots & a_{n,1} \\ a_{1,2} & a_{2,2} & \cdots & a_{n,2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1,n} & a_{2,n} & \cdots & a_{n,n} \end{bmatrix}$$

Definition 4 *A transition probability matrix is defined as* $P_{ij} = E_{ij}/od(i)$ *when* $deg(i) > 0$. *For* i *it is a stochastic column, which means that the* i^{th} *column elements sum to 1. The zero-out link pages will not comply in a stochastic system. The following formula can also be used to find the transition probability matrix:*

$$p_{i,j} = \begin{cases} a_{i,j} / c_j & if\ c_j \neq 0 \\ 0 & if\ c_j = 0 \end{cases} \tag{3}$$

In Eq. (3), $a_{i,j}$ is the connection from page j to page i. If there is a connection, then $a_{i,j} = 1$, otherwise $a_{i,j} = 0$, c_j is the column sums of the pages, and $c_j = \sum i\, a_{ij}$. That is c_j is the *od* of page j. The generalized form of the transition probability matrix is given below:

$$P = \begin{bmatrix} p_{1,1} & p_{2,1} & \cdots & p_{n,1} \\ p_{1,2} & p_{2,2} & \cdots & p_{n,2} \\ \vdots & \vdots & \ddots & \vdots \\ p_{1,n} & p_{2,n} & \cdots & p_{n,n} \end{bmatrix}$$

Definition 5 *A jump probability (JP) matrix can be created by adding a damping factor* d *in the transition probability matrix. It is used to simulate the random web surfer model (explained in section 3.1). It is also called a Google matrix. It can be defined as follows:*

$$JP = dP + \frac{1-d}{n}E \tag{4}$$

In Eq. (4), P *is the transition probability matrix,* d *is the damping factor, which is usually set at 0.85,* n *is the number of nodes in the graph, and* E *is the* n *times* n *matrix of all ones.*

Definition 6 *A set of states* T *is an irreducible closed subset of the Markov chain corresponding to the transition probability matrix* (P) *if and only if* T *is a closed subset, and no other subset of* T *is a closed subset [15].*

Theorem 1 *Let* P *be a transition probability matrix, and the sum of any column in a column matrix be 0; then that element can be referred to as a dangling or zero-out link.*

Proof If $\sum_i E_{ij} = 0$, then that corresponding page of the element can be referred to as a dangling or zero-out link page. □

3.1. PageRank and the Markov chain

The Markov chain can be used in any system where there is a transition from one state to another [22]. Imagine a random surfer surfing the web, going from one page to another by randomly choosing an outgoing link from one page to go to the next. This can sometimes lead to dead ends, i.e. pages with no outgoing links, cycles around a group of interconnected pages. For a certain fraction of time the surfer chooses a random page from the web; this theoretical random walk is known as the Markov chain or Markov process. The limiting probability

that an infinitely dedicated random surfer visits any particular page is its PageRank. The following example shows a sample transition matrix (column matrix) of a 3-state Markov chain [23]:

$$P_{ij} = \begin{bmatrix} 1/4 & 1/2 & 1/2 \\ 1/2 & 0 & 1/4 \\ 1/4 & 1/2 & 1/4 \end{bmatrix}$$

In the above transition probability matrix, P_{ij}, the probability of moving from one state to another state, can be easily seen. For example, $P_{3,2} = 1/4$, i.e. the probability of moving from state 2 to state 3 is only 25%. Markov chains are used to predict the probability of an event.

4. Proposed methodology

Let us consider a given graph $G(V, E)$ that has both zero-out link pages and nonzero-out link pages. There are two steps in this methodology to induce link spam. The first step is used to identify a target page, say T, and remove all the forward links of the target page T to make it a hanging page. In Eq. (5), T is the target page and it can be used to remove all the forward links of the target page T.

$$\sum_j T_{ij} = 0 \tag{5}$$

$$p_{i,j} = \begin{cases} a_{i,j} / c_j & if\ c_j \neq 0 \\ 1 & if\ c_j = 0 \end{cases} \tag{6}$$

The second step involves using Eq. (6), i.e. all the zero-out link pages that are getting an incoming link from the target page are connected back to the target page.

After applying Eqs. (5) and (6), the graph can induce an irreducible closed subset that can create link spam and promote ranks for the target page. Additionally, this method works only when graph G contains two irreducible closed subsets. These irreducible closed subsets can absorb lots of energy and they do not propagate their energy out. That is why the pages in the irreducible closed subsets are having higher ranks than other pages. The link spam can be detected by studying the eigenvectors and the eigenvalues, and particularly the second eigenvector. Our proposed method can detect link spam contributed by zero-out link pages.

5. Eigen vectors

Eq. (5) about the zero-out link pages, $\sum_j T_{ij} = 0$, and definition 6 from section 3 (irreducible closed subset) are important for the second eigenvalue.

Theorem 1 refers to a zero-out link or dangling node, i.e. in a closed Markov chain a node can get an incoming link and no outgoing link from that node. In link spamming, the dangling nodes can be connected back to the target node.

In the real web there may be many irreducible closed subsets and those irreducible closed subsets and zero-out link pages help us to understand more about the second eigenvalue.

The first eigenvector is nothing but the PageRank values [24] of the JP matrix, which can be calculated by Eq. (7).

$$JP\, g^{(1)} = \lambda_1\, g^{(1)} \tag{7}$$

In Eq. (7), $g^{(1)}$ is the distribution of the visiting frequency of each page in the random web surfer model; $g^{(1)}$ is the unique dominant eigenvector corresponding to the dominant eigenvalue $\lambda_1 = 1$. To show that $\lambda_1 = 1$ exists and is unique, the Perron–Frobenious theorem [25] can be used for the Markov matrix JP.

A matrix is irreducible if its graph shows that every node is reachable from every other node [15,17]. An irreducible Markov chain with a primitive transition matrix is called an aperiodic chain [15]. As mentioned before, $\lambda_1 = 1$ is the dominant eigenvalue and the corresponding eigenvector is the PageRank vector g^1. The second largest eigenvalue, λ_2, is always less than λ_1, i.e. $\lambda_1 = 1 > \lambda_2$.

Theorem 2 *The second eigenvector g^2 of JP is orthogonal to e: $e^T g^2 = 0$.*

The proof for theorem 2 is given in [15]. Here, e is the vector of all ones. From theorem 2, $e^T g^2 = 0$, therefore the second eigenvector of JP is only depending on P in Eq. (4), as given above in definition 5 of section 3.

Theorem 3 *The second eigenvalue of JP, $\lambda_2 = d$ if P has at least two irreducible closed subsets.*

The proof for theorem 3 is given in [15] and the results are shown in the experimental section. Theorem 3 has the following inferences for the PageRank algorithm.

PageRank convergence: The power method used by PageRank has the conversion rate equal to $\lambda_2 / \lambda_1 = d$.

Stability of PageRank algorithm: According to [15], when the eigengap, i.e. $|\lambda_1|$- $|\lambda_2|$, is greater, a more stable stationary distribution of the Markov chain occurs.

Spam Detection: The eigenvectors corresponding to $\lambda_2 = d$ are artifacts of certain structures in the web [15]. This can help us to detect link spamming.

According to Bianchini et al. [5], Langville and Meyer [17], and Boldi et al. [26], when the value of d is higher, an accurate PageRank will be produced. When the value of d is lower, a faster convergence and a more stable distribution will occur. The initial value of d used by Google is 0.85 and the best value of d is also 0.85 as suggested by other researchers [15,17,24]. Hence, we also used 0.85 as the value for d in our experiment. By studying the eigenvalues and the eigenvectors, link spam can be detected in the ranking process. The first eigenvector is nothing but the PageRank vector using the power method.

6. The power method

The power method is the most simple and popular method to solve large system problems such as finding the eigenvalues and eigenvectors of a matrix. When we apply the power method to the JP matrix in Eq. (4), the convergence of the method for diagonalizable matrices is proven, provided $|\lambda_1| > |\lambda_2|$.

If the JP matrix is diagonalizable, then there are n independent vectors of JP. Let the eigenvectors be g^1, ... g^n, then g^1, ... g^n form the basis of T^n. The initial vector $v^{(0)}$ can be written as:

$$v^{(0)} = a_1 g^1 + a_2 g^2 + ... + a_n g^n \tag{8}$$

In Eq. (8), $a_1,\ldots,a_n$ are scalars and multiply both sides of Eq. (8) by JP^k, producing:

$$\begin{aligned} JP^k\, v^{(0)} &= JP^k(a_1\, g^1 + a_2\, g^2 + \ldots + a_n\, g^n) \\ &= a_1\, JP^k\, g^1 + a_2\, JP^k\, g^2 + \ldots + a_n\, JP^k\, g^n \\ &= a_1\, \lambda_1^k\, g^1 + a_2\, \lambda_2^k\, g^2 + \ldots + a_n\, \lambda_n^k\, g^n \\ &= a_1\, \lambda_1^k \left(g^1 + \sum_{j=2}^{n} \frac{a_j}{a_1} \left(\frac{\lambda_j}{\lambda_1} \right)^k g^j \right) \end{aligned}$$

If $|\lambda_1| > |\lambda_2| \geq \ldots |\lambda_n|$, then λ_1 can be called a dominant eigenvalue. For example, $\left(\frac{\lambda_j}{\lambda_1}\right)^k \to 0$ and if $a_1 \neq 0$, $JP^k v^{(0)} \to a_1\, JP^k\, g^1$. The power method normalizes the product $JPv^{(k-1)}$ and it converges to g^1. Here, each iteration is a single matrix-vector multiplication, which can be performed more efficiently than a matrix-matrix multiplication. The convergence factor is determined by the second most dominant term, $a_2 \left(\frac{\lambda_2}{\lambda_1}\right)^k g_2$ and the rate of convergence is equal to $|\lambda_2|/|\lambda_1|$. The algorithm used for creating our program is given below.

6.1. Algorithm

1) Let *G(V, E)* be a directed graph with a set of vertices and edges.
2) Let *SG* be a subgraph with zero-out link pages and nonzero-out link pages.
3) Call PageRank.
4) Select a random target page, say T, and follow the steps.
   a) $\sum_j T_{ij} = 0$ (Remove all the outgoing links from target page T).
   b) Look for all the zero-out link pages having an incoming link from target page T and connect back all those zero-out link pages back to target page T.
5) Check the subgraph *SG* having two irreducible closed subsets.
6) Create an adjacency matrix for the sub graph *SG* by using the following formula:

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

7) Create a transition probability matrix P for the sub graph by using the following formula:

$$p_{i,j} = \begin{cases} a_{i,j} \,/\, c_j & if\ c_j \neq 0 \\ 0 & if\ c_j = 0 \end{cases}$$

8) Create the JP matrix using the following formula:

$$JP = dP + \frac{1-d}{n} E$$

In the above formula, the value of d is 0.85, P is the probability matrix created in step 7, E is the $ntimesn$ vectors of all ones, and n is the number of nodes in the subgraph.

9) Call the PageRank to calculate the new page rank values.

10) Call the eigenfunction to calculate the eigenvalues and eigenvectors from the JP matrix.

11) Check the second eigenvector for whether the target page T is in the irreducible closed subset by using the following formula:

$v_i \neq 0$ if page $i \in$ irreducible closed subset;

$v_i = 0$ if page $i \notin$ irreducible closed subset.

PageRank Procedure:

a. Create a sparse transition probability matrix.

b. Calculate the PageRank using the power method by assigning the damping factor, number of iterations.

c. Create bar chart for PageRanks.

d. Return.

6.2. Example

Let us take a sample web graph $G1$ with 8 pages and 12 edges, as shown in Figure 1. We have used color coding in this graph. Pages with blue color (pages 3, 4, 5, and 6) are nonzero-out link pages. The page with orange color (page 8) is a zero-out link page. The sum of column 8 is zero (using theorem 1) and this indicates that page 8 is a zero-out link page. Pages with green color (pages 1 and 2) are irreducible closed subsets. Pages 1 and 2 have high PageRanks (0.25 and 0.27) among the 8 pages because of the irreducible property and they do not propagate their score to other pages. Let us say that page 7 (shown in red color) is our target page for link spam. Based on definitions 1 and 2, the *id* and *od* were computed and are shown in Table 1.

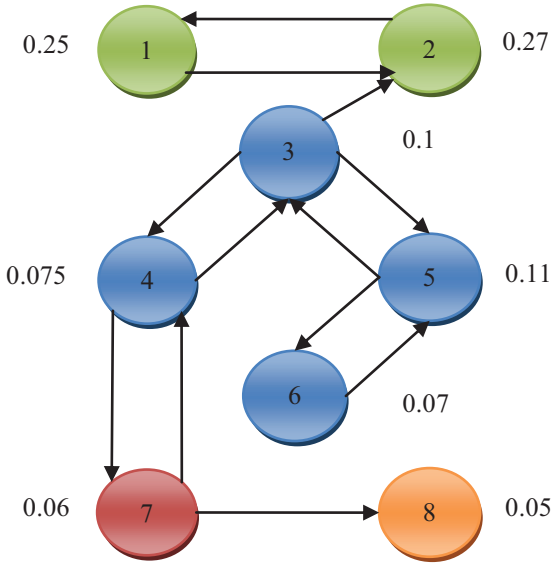

Figure 1. Sample web graph G1 with PageRank values.

Table 1. In-degree and out-degree list for G1.

Page	*id*	*od*	Zero-out link / nonzero-out link
1	1	1	Nonzero-out link
2	2	1	Nonzero-out link
3	2	3	Nonzero-out link
4	2	2	Nonzero-out link
5	1	2	Nonzero-out link
6	1	1	Nonzero-out link
7	1	2	Nonzero-out link
8	1	0	Zero-out link

The adjacency matrix (column matrix) A is generated as per definition 3 for the graph $G1$ in Figure 1 and is shown below. The last column represents the *od* of page 8 and the last row represents the *id* of page 8. The sum of the columns in matrix A gives the *od* and the sum of the rows gives the *id*.

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

In the above adjacency matrix A, the 8th column represents the zero-out link for page 8 by having all zero entries (Theorem 1).

The transition probability matrix P is computed as per definition 4 for the graph $G1$ in Figure 1 and is shown below:

$$P = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1/3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 1/3 & 0 & 0 & 0 & 1/2 & 0 \\ 0 & 0 & 1/3 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1/2 & 0 \end{bmatrix}$$

This transition probability matrix P is not stochastic because column 8 is not summing up to 1. This is due to the zero-out link pages. The handling the zero-out link pages is described in section 4. The JP matrix is described in the proposed method section. In Figure 1, the PageRank values are shown for each page. For example, the PageRank of page 1 is 0.25 and the PageRank of page 2 is 0.27 (highest among the 8 pages).

The web graph $G1$ in Figure 1 was modified according to our proposed methodology and is shown in Figure 2. Consider page 7 as the target page. In the first step all the outgoing links from page 7 are removed

and the zero-out link page 8 is connected back to the target page 7 as shown in Figure 2. Now this modified graph *G2* has two irreducible closed subsets, pages 1 and 2, and pages 7 and 8. The PageRank results are shown in Figure 2 after the link spam is induced. The target page rank increased from 0.06 to 0.185 (more than 3 times) due to our proposed methodology (link spam). Furthermore, notice that the ranks of pages 7 and 8 increased to 0.185 and 0.179 respectively, due to the irreducible property as shown in Figure 2.

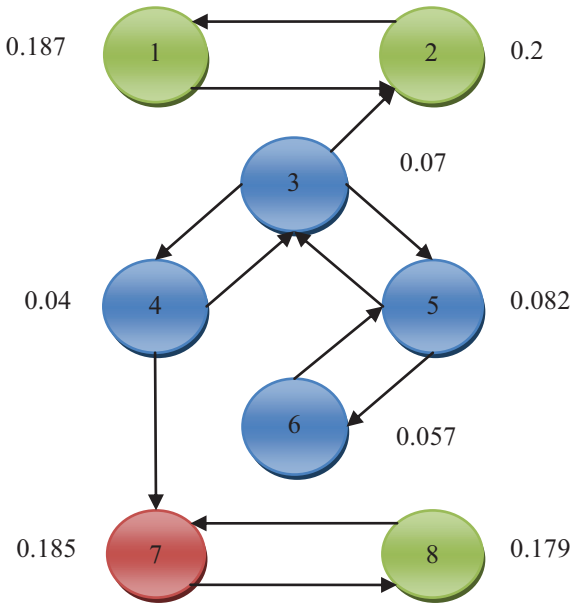

Figure 2. Modified web graph G2 with PageRank values after spam.

The adjacency matrix (A) for the modified graph $G2$ is shown below:

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

In the above adjacency matrix (A), page 8 (the last column) is no more a zero-out link page because it is connected back to page 7, as per our proposed method in Eq. (6).

The transition probability matrix (P) (column matrix) for graph $G2$ in Figure 2 can be developed by using the formula in Eq. (3):

$$P = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1/3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 1/3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/3 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

The above probability matrix P has the following problems:

- The matrix P does not model the random jump to another page (1-d). The first eigenvectors are not necessarily unique because the matrix P is reducible (because of pages 1, 2, 7, and 8).
- The computation of the first eigenvector becomes difficult because of the reducibility of the matrix.
- This matrix is not stochastic (in our example P is stochastic because there are no other zero-out link pages in the graph, but in the real web that is not the case).

All of the above problems, i.e. reducibility, random surfer model, and a stochastic matrix, are addressed in the JP matrix. The JP matrix can be obtained by using the following formula:

$$JP_{i,j} = \begin{cases} da_{i,j} \,/\, c_j + (1-d)/n & if\, c_j \neq 0 \\ 1/n & if\, c_j = 0 \end{cases} \tag{9}$$

This is the same as Eq. (4) below:

$$JP = dP + \frac{1-d}{n}E$$

When we apply Eq. (4) to the probability matrix (P) we get the JP matrix as follows:

$$JP = \begin{bmatrix} 0.019 & 0.869 & 0.019 & 0.019 & 0.019 & 0.019 & 0.019 & 0.019 \\ 0.869 & 0.019 & 0.302 & 0.019 & 0.019 & 0.019 & 0.019 & 0.019 \\ 0.019 & 0.019 & 0.019 & 0.444 & 0.444 & 0.019 & 0.019 & 0.019 \\ 0.019 & 0.019 & 0.302 & 0.019 & 0.019 & 0.019 & 0.019 & 0.019 \\ 0.019 & 0.019 & 0.302 & 0.019 & 0.019 & 0.869 & 0.019 & 0.019 \\ 0.019 & 0.019 & 0.019 & 0.019 & 0.444. & 0.019 & 0.019 & 0.019 \\ 0.019 & 0.019 & 0.019 & 0.444. & 0.019 & 0.019 & 0.019 & 0.869 \\ 0.019 & 0019 & 0.019 & 0.019 & 0.019 & 0.019 & 0.869 & 0.019 \end{bmatrix}$$

We used MATLAB (Version R2012b) to calculate the eigenvectors and eigenvalues for the JP matrix. The following are the eigenvectors and eigenvalues:

$$v = \begin{bmatrix} 0.4725 & \mathbf{0.5000} & 0.4566 & 0.2280 & 0.2280 & 0.4564 & -0.7071 & 0.0167 \\ 0.5005 & \mathbf{0.5000} & 0.3867 & 0.0777 & -0.0777 & -0.3864 & 0.7071 & 0.0167 \\ 0.1831 & 0.0000 & -0.3880 & -0.6052 & -0.6053 & -0.3882 & 0.0000 & 0.0000 \\ 0.0996 & 0.0000 & -0.1526 & -0.5914 & 0.5914 & 0.1527 & 0.0000 & 0.0000 \\ 0.2191 & 0.0000 & -0.5044 & 0.1789 & -0.1789 & 0.5047 & 0.0000 & 0.0000 \\ 0.1409 & 0.0000 & -0.2979 & 0.2625 & 0.2625 & -0.2981 & 0.0000 & 0.0000 \\ 0.4667 & \mathbf{-0.5000} & 0.2285 & 0.1140 & 0.1140 & 0.2282 & -0.0024 & -0.7069 \\ 0.4439 & \mathbf{-0.5000} & 0.2698 & 0.3346 & -0.3346 & -0.2696 & 0.0024 & 0.7069 \end{bmatrix}$$

The above v is the eigenvector produced by the MATLAB for the graph in Figure 2. The first column is the first eigenvector, which is nothing but the PageRank values of pages 1 through 8. The second column refers to the second eigenvector. The right eigenvector $v^{(i)}$ of JP, i.e. $v^{(i)} = (v_1, ..., v_n)$ has the following properties:

$$\begin{cases} v_j \neq 0 \; if \text{node } j \in \text{irirducibleclosedsubset} \\ v_j = 0 \text{ if node} j \notin \text{irreducibleclosedsubset} \end{cases} \tag{10}$$

Eq. (10) shows that the second eigenvector will have a nonzero value if a page is in an irreducible closed subset; otherwise they will have zero values, as can be seen in the second column of v. This second eigenvector indicates that the pages in the irreducible closed subset contribute to link spam.

You can observe that the two irreducible closed subsets (pages 1 and 2, and pages 7 and 8) have nonzero values (0.5000 and 0.5000, and –0.5000 and –0.5000, respectively) and the other pages have zero values. This indicates that irreducible closed subsets contribute to link spam. Zero-out link pages play an important role in forming the irreducible closed subset and in turn contribute to link spam. From our experiment, the PageRank order of the target page (page 7) is increased from 7 to 3. The eigenvalues for the JP matrix are shown below:

$$e = \begin{bmatrix} 1.0019 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.8500 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.7197 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.2897 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -0.2897 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -0.7197 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -0.8500 & 0 \\ 0000000 - 0.8500 & & & & & & & \end{bmatrix}$$

The second eigenvalue from the above eigenvalues e is 0.85, which is same as the damping factor we used for the JP matrix. According to [15], if the transition probability matrix has at least two irreducible closed subsets, then the second eigenvector of the Google matrix or JP matrix is $\lambda_2 = d$ (theorem 3). Our sample experiment result also produced $\lambda_2 = d$ (0.85). The detail results are shown in the experimental section.

7. Experiment and results

The first task in our experiment is to prove how link spam can increase the PageRank values. We created a PageRank program using MATLAB (R2012b) to calculate the rank before spam and after spam. We modified the basic PageRank algorithm by Moler [27] to include our proposed method. The program was tested on an Intel i7 Processor (1.70 Ghz) with 6GB RAM. To begin with, we used our PageRank program for the sample graph in Figure 1 (i.e. the graph before link spam); PageRanks were calculated for the 8 pages; the output is shown in Figure 3 as a bar graph. Our target page for the link spam was page 7 and it was ranked number 7. The order of rank for the 8 pages from high to low was the following: page 2, 1, 5, 3, 4, 6, 7, and 8.

The second program, which included our proposed methodologies, was applied to the graph in Figure 2; the output is shown in Figure 4 as a bar graph. In Figure 4, the order of rank for the 8 pages became the following: page 2, 1, 7, 8, 5, 3, 6, and 4. The target page 7 order went up from 7 to 3. Just connecting a zero-out link page back to the target page significantly increased the rank of the target page. When many zero-out link pages on the web are connected to a target page for the purpose of link spam, the PageRank score can significantly increase.

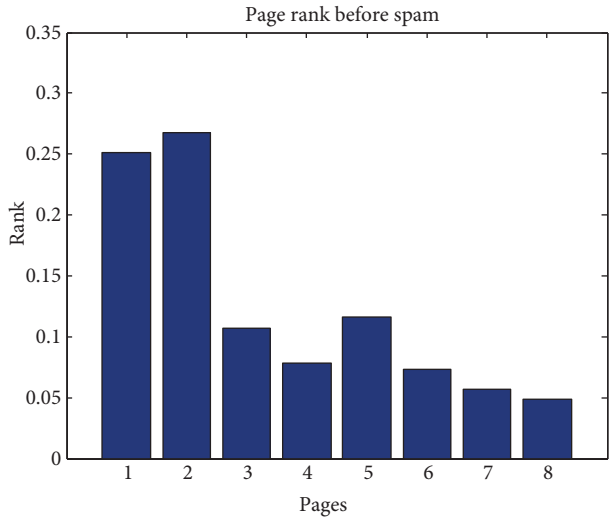

Figure 3. PageRank results before spam.

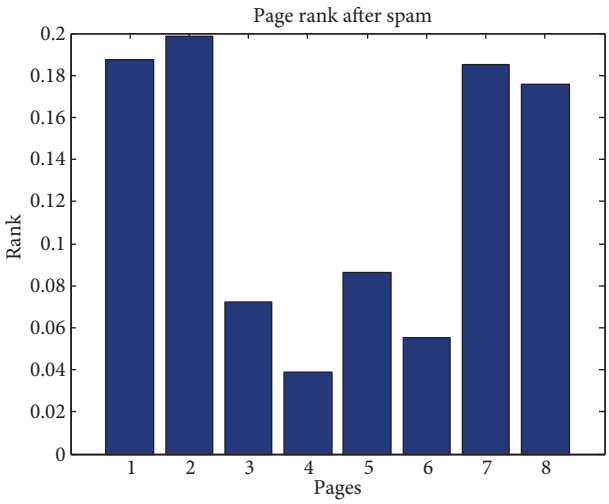

Figure 4. PageRank results after spam.

Table 2. The top 10 web sites in the world (Source: Alexa.com).

Rank	Website Name	URL	In-links
1	Facebook	www.facebook.com	8,296,430
2	Google	www.google.com	4,656,505
3	YouTube	www.youtube.com	3,802,453
4	Yahoo!	www.yahoo.com	1,804,470
5	Baidu.com	www.baidu.com	304,348
6	Amazon.com	www.amazon.com	1,148,899
7	Wikipedia	www.wikipedia.org	2,171,478
8	QQ.com	www.QQ.com	445,248
9	Windows Live	www.live.com	134,048
10	Taobao.com	www.taobao.com	163,653

8. Experiments with Amazon.com

To prove this further, we conducted experiments with live data from the internet. Table 2 shows the top 10 web sites in the world and their incoming links.

Due to the huge size of the web and the computational complexity, we took only one site (amazon.com) from the top 10 for our experiment. First, using the surfer program from MATLAB [27], we downloaded the pages. Due to size complexity, we show only the first 50 pages from amazon.com in the adjacency matrix in Figure 5. Table 3 shows the list of the first 50 pages in amazon.com.

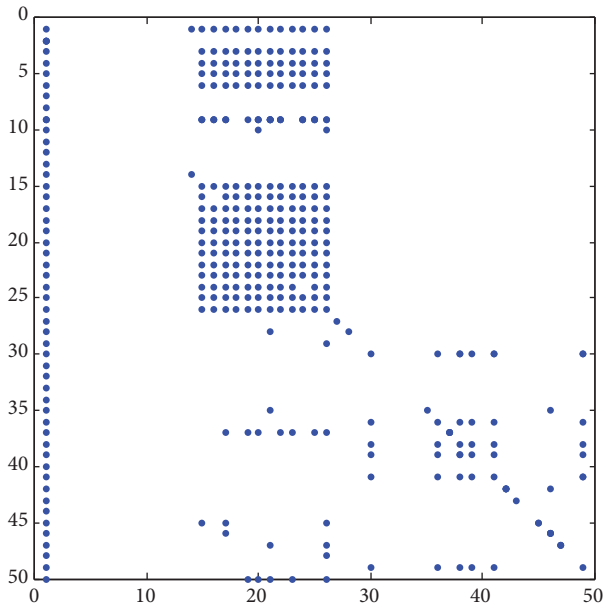

Figure 5. Adjacency matrix for amazon.com for the first 50 pages.

Table 3. A list of the first 50 pages from amazon.com.

Page No	Pages
1	'http://www.amazon.com'
2	'http://www.amazon.com.br'
3	'http://www.amazon.ca'
4	'http://www.amazon.cn'
5	'http://www.amazon.fr'
6	'http://www.amazon.de'
7	'http://www.amazon.in'
....	.....
49	'http://www.look.com'
50	'http://www.myhabit.com'

Due to computational complexity, we applied our proposed methodology on only the first 20 pages from amazon.com (some images and pictures are omitted). Our target page for the link spam was page 15. The transition probability matrix P (column matrix) (shown below) was produced after link spam was introduced. It is a sparse matrix, as can be seen below. Generally, the transition probability matrix for the real web is a sparse matrix. The JP matrix is not shown here due to the huge size.

$$P = \begin{bmatrix}
0 & 1/3 & 0 & 0 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1/2 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1/3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1/2 & 0 & 0 & 0 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1/3 & 0 & 0 & 0 & 0 & 1/4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1/2 & 0 & 1/2 & 0 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1/4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1/4 & 0 & 0 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1/4 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1/2 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/2 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1/2 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}$$

We used our rank program developed in MATLAB to calculate the PageRank. Figure 6 shows the PageRank results for the first 20 pages of amazon.com before link spam was introduced. Table 4 shows the summary of the results before and after link spam. Figure 7 shows the PageRank results after link spam was introduced. Our target page 15 (http://amazonlocal.com) is shown in bold. This is not the actual PageRank of amazon.com. It is one of the pages in amazon.com and the PageRanks are based on our proposed method. The PageRank for the target page 15 increased. Before spam the order of the target page was 4. After link spam, the PageRank

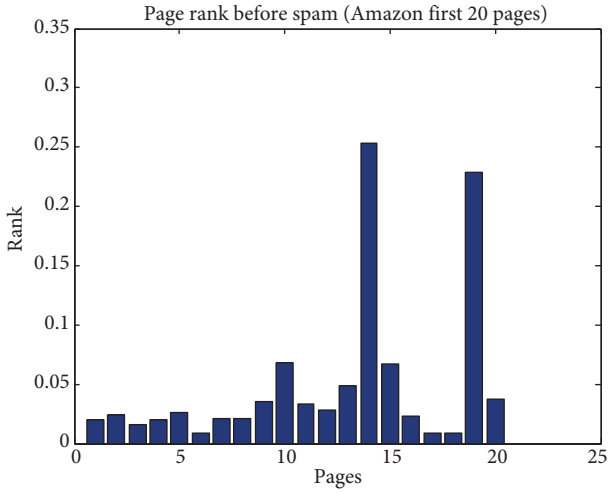

Figure 6. PageRank results before spam for amazon.com.

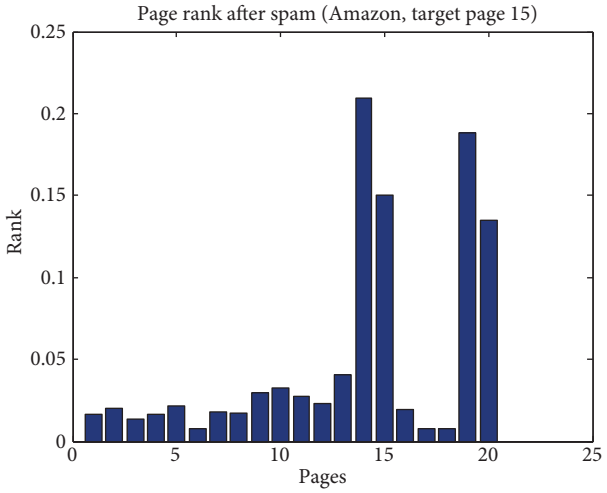

Figure 7. PageRank results after link spam for amazon.com.

of page 15 more than doubled and the order of the target page was promoted to 3, as shown in Table 4. A comparison graph before link spam and after link spam is shown in Figure 8.

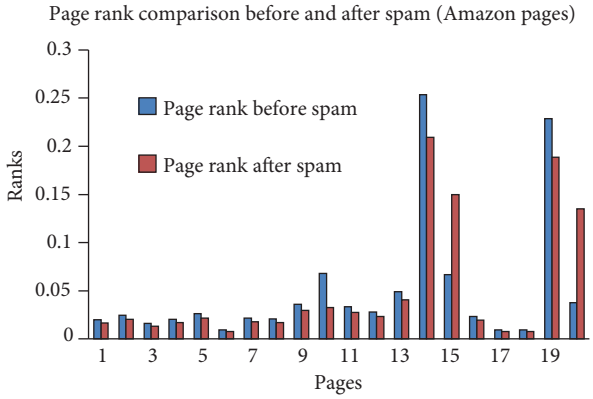

Figure 8. PageRank comparisons before and after spam.

Table 4. Experimental results showing the PageRank and second eigenvectors and eigenvalues.

Page	Amazon Pages	PageRank Before Spam	PageRank After Spam	Second Eigenvector	Second eigenvalue
1	'http://www.amazon.com.br'	0.0199	0.0164	0.0000	0
2	'http://www.amazon.ca'	0.0243	0.0201	0.0000	**0.85**
3	'http://www.amazon.cn'	0.016	0.0132	0.0000	0
4	'http://www.amazon.fr'	0.0203	0.0167	0.0000	0
5	'http://www.amazon.de'	0.0263	0.0217	0.0000	0
6	'http://www.amazon.in'	0.0091	0.0075	0.0000	0
7	'http://www.amazon.it'	0.0214	0.0176	0.0000	0
8	'http://www.amazon.co.jp'	0.0205	0.0169	0.0000	0
9	'http://www.amazon.es'	0.0358	0.0295	0.0000	0
10	'http://www.amazon.co.uk'	0.0679	0.0326	0.0000	0
11	'http://www.6pm.com'	0.0335	0.0276	0.0000	0
12	'http://www.abebooks.com'	0.0279	0.023	0.0000	0
13	'http://www.afterschool.com'	0.0489	0.0403	0.0000	0
14	'http://fresh.amazon.com'	0.2537	0.2093	**0.5000**	0
15	**'http://amazonlocal.com'**	**0.0668**	**0.1499**	**−0.5000**	0
16	'http://www.amazonsupply.com'	0.0233	0.0192	0.0000	0
17	'http://aws.amazon.com'	0.0091	0.0075	0.0000	0
18	'http://askville.amazon.com'	0.0091	0.0075	0.0000	0
19	'http://www.audible.com'	0.2286	0.1886	**0.5000**	0
20	'http://www.beautybar.com'	0.0375	0.1349	**−0.5000**	0

Next, our program produced the eigenvalues and eigenvectors for the JP matrix. The first eigenvector is nothing but the PageRank values (after spam). Table 4 shows the second eigenvector and the second eigenvalue for the first 20 pages of amazon.com along with the PageRank values.

The important observation here is the second eigenvector, which shows that pages 14 and 19, and pages 15 and 20 are in an irreducible closed subset. As per Eq. (10), they have nonzero values, while all the other pages have zero values. This clearly proves that page 15 (target page) in the irreducible closed subset contributes to

link spam and this can be detected using the second eigenvector. Our proposed method made the zero-out link page form an irreducible closed subset and the second eigenvector was used to detect this kind of link spam.

9. Conclusion

This paper explored the contribution of zero-out link pages in link spam and proposed a method to form link spam using zero-out link pages and to detect link spam using eigenvectors and eigenvalues. We took the PageRank algorithm of Google as our base algorithm and included it in our algorithm to calculate the rank of web pages before spam and after spam.

By doing this, we explored the zero-out links and the mathematical model behind the Google search engine (the adjacency matrix, transition probability matrix, Google or JP matrix, Markov chain, eigenvectors, and eigenvalues).

Two important findings stand out from the current research. The first one is the significant role played by the zero-out link pages in forming irreducible closed subsets, and the second one is the detection of link spam formed by zero-out links. These irreducible closed subsets absorb a lot of energy and get a high PageRank because they do not propagate their ranks to other pages. If more and more zero-out link pages are connected to an irreducible closed subset, an efficient link spamming can be achieved. This can affect the ranking order of SERPs, and in turn the search engine creditability will be affected. We also discovered that the second eigenvector of the Google matrix can detect irreducible closed subsets.

We had a couple of problems in the experiment. The first one was finding the second eigenvector due to a poor convergence rate of the nonunique values of the second eigenvector. Furthermore, we had problems in the large matrix due to computational complexity.

We took live pages from amazon.com and conducted the experiment. We simulated our proposed methodology in the amazon.com web pages and did the ranking. Our experiment gave the same result as our example of the proposed methodology. If web site developers or search engine optimization professionals create web sites without zero-out link pages or fix the zero-out link pages, this kind of link spamming can be controlled.

One future area of research could be the effects of zero-out link/dangling pages on search engine optimization.

References

[1] Gyöngyi Z, Garcia-Molina H. Link Spam Alliances. In: The 31st International Conference on Very Large Databases (VLDB); 2005; Trondheim, Norway: ACM. pp. 517-528.

[2] Henzinger MR, Motwani R, Silverstein C. Challenges in web search engines. Journal of ACM SIGIR 2002; 36: 11-22.

[3] Eiron N, McCurley KS, Tomlin AJ. Ranking the Web Frontier. In: The 13th International conference on WWW; 17–22 May 2004; New York, USA: pp. 309-318.

[4] Wang X, Tao T, Sun JT, Shakery A, Zhai C. DirichletRank: solving the zero-one-gap problem of PageRank. ACM T Inform Syst 2008; 26: 10.

[5] Bianchini M, Gori M, Scarselli F. Inside PageRank. ACM T Internet Techn 2005; 5: 92-128.

[6] Brin S, Page L, Motwani R, Winograd T. The PageRank Citation Ranking: Bringing Order to the Web. Technical Report 1999-0120. Stanford, CA, USA: Computer Science Department, Stanford University, 1999.

[7] Kleinberg J. Authoritative sources in a hyper-linked environment. J ACM 1999; 46: 604-632.

[8] Lempel R, Moran S. SALSA: the stochastic approach for link-structure analysis. ACM T Inform Syst 2001; 19: 131-160.

[9] Gyöngyi Z, Garcia-Molina H. Web spam taxonomy. In: The 1st International Workshop on Adversarial Information Retrieval on the Web; 10–14 May 2005; Chiba, Japan: pp. 39-47.

[10] Baeza-Yates R, Castillo C, Lst opez V. PageRank increase under different collusion topologies. The 1st International Workshop on Adversarial Information Retrieval on the Web; 10–14 May 2005; Chiba, Japan: pp. 17-24.

[11] Zhang H, Goel A, Govindan R, Mason K, Van Roy B. Making eigenvector-based reputation systems robust to collusion. In: The 3rd Workshop on Web Graphs (WAW). Lecture Notes in Computer Science, Vol. 3243; 2004; Rome, Italy: Springer. pp. 92-104.

[12] Gyöngyi Z, Berkhin P, Garcia-Molina H. Link spam detection based on mass estimation. The 32nd International Conference on Very Large Data Bases; 12–15 September 2006; Seoul, Korea: ACM. pp. 439-450.

[13] Zhou B, Pei J. Link spam target detection using page farms. ACM Transactions on Knowledge Discovery from Data (TKDD) 2009; 3: 13.

[14] Nikita S, Jiawei H. Survey on web spam detection: principles and algorithms. ACM SIGKDD Explorations Newsletter 2011; 13: 50-64.

[15] Haveliwala TH, Kamvar SD. The Second Eigenvalue of the Google Matrix. Technical Report 2003-20. Stanford, CA, USA: Stanford University, 2003.

[16] Ipsen ICF, Selee TM. PageRank computation, with special attention to dangling node. Society for Industrial and Applied Mathematics 2007; 29: 1281-1296.

[17] Langville AN, Meyer CD. Deeper Inside PageRank. Internet Mathematics 2003; 1: 335-380.

[18] de Jager DV, Bradley JT. PageRank: splitting homogeneous singular linear systems of index one. In: The 2nd International Conference on the Theory of Information Retrieval: Advances in Information Retrieval Theory; 10-12 September 2009; Cambridge, UK. Berlin, Germany: Springer. pp. 17-28.

[19] Gleich DF, Gray AP, Greif C, Lau T. An inner-outer iteration for computing PageRank. SIAM J Sci Comput 2010; 32: 349-371.

[20] Singh AK, Kumar PR, Goh AKL. Efficient methodologies to handle hanging pages using virtual node. Cybernet Syst 2011; 42: 621-635.

[21] Broder A, Kumar R, Maghoul F, Raghavan P, Rajagopalan S, Stata R, Tomkins A, Wiener J. Graph structure in the web. Comput Netw 2000; 33: 309-320.

[22] Gao B, Liu TY, Ma Z, Wang T, Li H. A general Markov framework for page importance computation. In: The 18th Conference on Information and Knowledge Management; 2–6 November 2009; Hong Kong, China: ACM. pp. 1835-1838.

[23] Kumar PR, Goh AKL, Singh AK, Application of Markov chain in the PageRank algorithm. Pertanika Journal of Science and Technology 2013; 21: 541-554.

[24] Langville AN, Meyer CD. A survey of eigenvector methods of web information retrieval. SIAM 2005; 47: 135-161.

[25] Meyer CD. Matrix Analysis and Applied Linear Algebra. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2000.

[26] Boldi P, Vigna S, Santini M. PageRank as the function of the damping factor. In: The 14th International Conference on World Wide Web; 2005; Chiba, Japan: pp. 557-566.

[27] Moler C. Experiments with MATLAB. Natick, MA, USA: MathWorks, Inc., 2011.