

BY ROHIT KAUL, YEOGIRL YUN, AND SEONG-GON KIM

Ranking Billions of Web Pages Using Diodes

BECAUSE OF THE WEB'S RAPID GROWTH¹³ and lack of central organization, Internet search engines play a vital role in assisting the users of the Web in retrieving relevant information out of the tens of billions of documents available.^{1, 6} With millions of dollars of potential revenue at stake, commercial Web sites compete fiercely to be placed prominently within the first page returned by a search engine. As a result, search engine optimizers (SEOs) developed various forms of search engine spamming (or spamdexing) techniques^{5, 8} to artificially inflate the rankings of Web pages. Link-based ranking algorithms,^{9, 10} such as Google's PageRank,² have been largely effective against most conventional spamming techniques.

However, PageRank has three fundamental flaws that, when exploited aggressively, can be proven to be its Achilles' heel: First, PageRank gives a minimum guaranteed score to every page on the Web; second, it rewards all incoming links as valid endorsements; and third, it imposes no penalty for making links to low-quality pages. SEOs can take advantage of these shortcomings to the extreme by employing an Artificial Web, a collection of an extremely large

number of computer-generated Web pages containing many links to only a few target pages. Each page of the Artificial Web collects the minimum PageRank and feeds it back to the target pages. Although the individual endorsements are small, the flaws of PageRank make it possible for an Artificial Web to accumulate sizable PageRank values for the target pages. The SEOs can even download a substantial portion of the real Web and modify only the destinations of the hyperlinks, thus circumventing any detection algorithms based on the quality or the size of pages. As the size of an Artificial Web can be comparable to that of the real Web, SEOs can seriously compromise the objectivity of the results that PageRank provides. Although some statistical measures can be employed to identify specific attributes associated with an Artificial Web and filter them out of search results,⁵ it is far more desirable to develop a new ranking model that is free of such exploits to begin with.

Affinity Index Ranking

The purpose of a ranking system is to assign a relative rank to a Web page according to the *quality* of the Web page. Although the quality of a Web page can mean many different things to different users, these qualities tend to make users prefer certain Web pages over others and express their endorsements in the form of hyperlinks. Therefore, it is reasonable to expect that when the vast link structure of the Web is analyzed properly, an objective measure to evaluate the overall quality of individual Web pages would emerge. However, the Web also contains a substantial amount of spam links that do not reflect the opinions of the general public. Therefore, we propose a new ranking model called Affinity Index Ranking (AIR) that approximates the relative quality (or rank) of each page as a measure of popularity or authority based on the endorsements of other pages, which are expressed in the form of hyperlinks. AIR computes the rank V of page i using a voting scheme with the following rules:

1. A link from page j to page i is considered as an *attempt* of endorsement by page j for page i . The attempted endorsement is valid only if the endorser has a higher rank than the beneficiary ($V_j > V_i$).

2. When the endorsement is validated, the amount of votes equal to the *difference of ranks* between the endorser, j , and the beneficiary, i , ($V_j - V_i$) are credited to the beneficiary.

3. The total amount of votes a page casts for others must be equal to the total amount of votes it receives.

4. All pages are connected to a universal sink⁴, whose rank is set at the minimum value of zero.

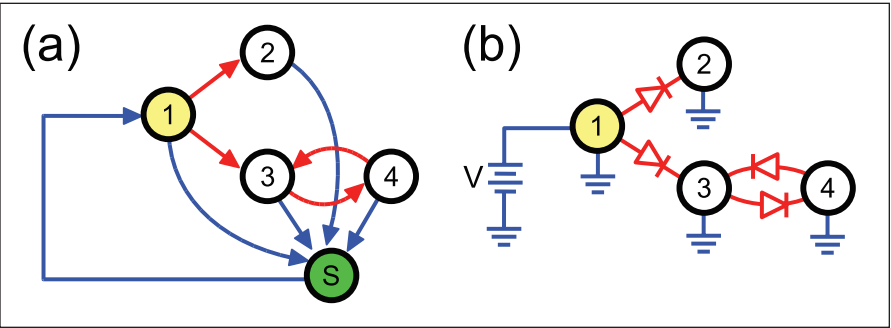
5. A few pages are chosen as paragons whose ranks are fixed at the maximum value V_{\max} .

The rank of a page must be adjusted until the requirement of Rule 3 is met. If a page pledges more votes to other pages than the sum of the votes it receives from incoming endorsements, the page must reduce its rank to increase the incoming votes and decrease the outgoing votes invoking Rule 2. The rank must be raised if the page is receiving more incoming votes than outgoing votes. Of course, these changes will cause all pages connected to this page through links to be out of balance, forcing them to adjust their ranks. The updating process will continue until all pages satisfy Rule 3. The paragons are supplied with a sufficient amount of votes needed to sustain their endorsement commitments. Typically, the paragons correspond to Web pages that are widely accepted as extremely high-quality and trustworthy, such as www.yahoo.com. By using a sufficient number of paragons, the dependence on a few particular paragons can be avoided. Since the maximum rank only determines the scale of rank values, we can set $V_{\max} = 100$ without any loss of generality.

Equivalent Electronic Circuit Model

Although AIR's voting scheme appears to be complex, a simple and elegant solution can be obtained by using an equivalent electronic circuit (EEC), as illustrated in Figure 1. While networks of resistors have been used to study community structures in social networks^{4, 14} and small-world networks,^{11, 12} we use an EEC of diodes. In an EEC, the rank of a Web page is defined as the electrical potential of the correspond-

Figure 1. A connectivity graph (a) of a small portion of the Web containing four web pages and its EEC (b).



ing node. The value of the votes of endorsement for page i by page j is the current I_{ji} flowing from node j to node i given by Ohm's law [Rule 2]:

(1)

$$I_{ji} = g_{ji} \cdot (V_j - V_i)$$

where g_{ji} is the conductance of an ideal diode with unit conductance representing the link from page j to page i [Rule 1]:

(2)

$$g_{ji} = 1, \text{ if the link exists and } V_j > V_i$$

or

$$g_{ji} = 0 \text{ otherwise}$$

Optionally, the conductance to the universal sink, g_0 , can have a value different from those of regular links.

The rank V_i of page i is determined by Kirchhoff's current law that requires the sum of all currents entering node i must be equal to the sum of all currents leaving node i [Rule 3]:

(3)

$$\sum_j I_{ji} = \sum_j I_{ij}$$

All nodes are connected to ground [Rule 4] and an ideal voltage source maintains a constant potential at the paragon nodes [Rule 5]. Since the goal of AIR is to compute the relative rank of Web pages using an EEC, all quantities will be treated as dimensionless.

By combining Equations (1)-(3), V_i can be expressed in terms of the potential of its neighboring nodes:

(4)

$$V(i) = \sum_j \tilde{g}_{ji} V_j + \sum_j \tilde{g}_{ij} V_j$$

where

(5)

$$\tilde{g}_{ij} = \frac{g_{ij}}{\sum_j g_{ji} + \sum_j g_{ij}}$$

is the scaled conductance. Equation (4) can be solved using an iterative method such as Jacobi method^{3, 15} while keeping the universal sink at a minimum potential $V_{\min} = 0$, and the paragon nodes at their prescribed potentials. The complexity of the computation of AIR, like PageRank, depends on the number and the structure of hyperlinks, although additional iterations are required for AIR due to the dependence of the conductance of diodes on the rank of pages.

Although PageRank can also be viewed as a voting scheme,² there are several critical differences between the two models. First, unlike PageRank that regards *all* links as valid endorsements, AIR views all links as diodes, and thus respects only *downward* links [Rule 1]. Second, the assessment of the endorsement value of a link [Rule 2] is different. In AIR, it 'costs' more to endorse low-quality pages than high-quality pages. Consequently, if a page makes many links to low-quality pages, the page's electrical potential, or rank, will drop to meet the requirement of Rule 3. This forces Web page authors to be more judicious in making hyperlinks and properly punishes pages using linkspam. Third, unlike PageRank, there is no minimum rank given to all pages. Every page must earn its rank through endorsements from other pages.

Applications of AIR

Figure 2(a) depicts one of the simplest link structures containing three nodes in a chain formation connected to a

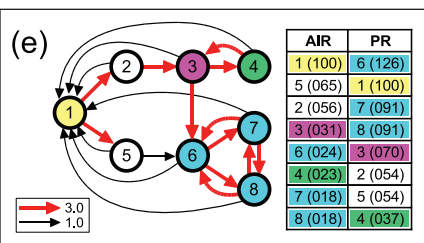
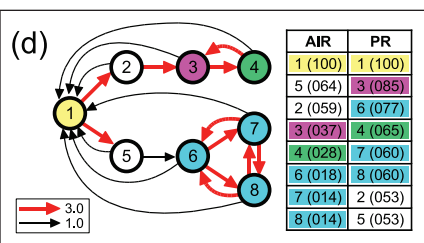
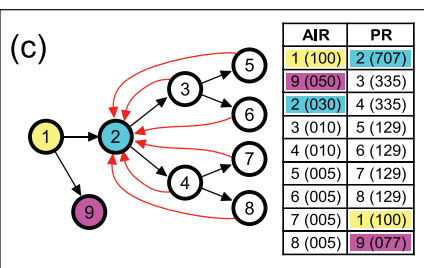
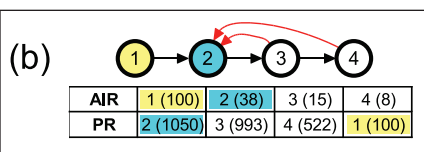
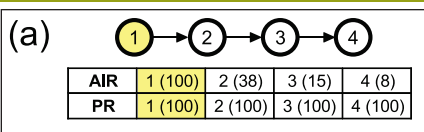
paragon node. The rank values of AIR decrease for nodes with larger click distances from the paragon node. In contrast, PageRank assigns identical rank values to all nodes in the chain. We note that node 3 (node 4) is two (three) clicks away from the paragon while node 2 has a direct connection from the paragon. Since it is assumed that paragons represent pages of high-quality and popularity, it is quite reasonable to assign a higher rank value to node 2 than node 3 or 4. This is also justified in terms of the behavior of a typical user of the Web. Most users visit one of their trusted Web sites (paragon nodes) first and follow links to reach other less-known pages. Internet users are often “impatient” and “lazy,” and it is therefore less likely for them to reach node 4 than node 2 or 3. In this sense, Figure 2(a) demonstrates that AIR measures the “closeness,” or affinity index of each page to paragons, hence the name.

These results can be compared with those of the link structure shown in Figure 2(b), which has an identical structure with Fig. 2(a), except for the two additional feedback links from node 3 and 4 to node 2. According to Rule 1, AIR disregards these links and the rank values of AIR are not affected by them. In contrast, PageRank dutifully boosts the rank of node 2 and its underlings. In fact, nodes 2-4 have much higher values of PageRank than the paragon that endorses them.

The structure in Figure 2(c) further illustrates the advantage of AIR over PageRank. Node 2 employs a linkspam technique to generate many child pages (nodes 3-8) and feedback links. The technique succeeds in manipulating PageRank: nodes 3-8 return all votes that come from node 2 plus the collective sum of the minimum scores given to them back to node 2. Consequently, node 2 receives a much higher PageRank than the paragon. As node 2 rises in ranking, its underlings also get higher ranks. This feedback continues until self-consistency is achieved. On the other hand, this technique fails to manipulate AIR. Not only does AIR prevent node 2 from getting a higher rank, but AIR also penalizes node 2 for linking to many low-quality pages: node 2 has a substantially lower rank than node 9.

Figure 2(d) depicts a microcosm of the Internet containing an Artificial

Figure 2: Comparison of Affinity Index Ranking (AIR) and PageRank (PR). (a) A chain structure. (b) A chain structure with feedback links. (c) A tree structure with feedback links. (d) A link structure containing an Artificial Web. (e) A legitimate way to improve ranking in AIR.



Web. The legend indicates that the red arrows are assumed to be equal to three black links. The Artificial Web (nodes 6-8) is strongly inter-connected and separated from the main portion of the Web (nodes 1-5), except for a few weak links represented by a link from node 5 to node 6. While PageRank yields to the manipulation of the Artificial Web and elevates the top pages in the Artificial Web (nodes 6-8) above many high-quality Web pages in the real Web (nodes 2 and 5), AIR identifies these bogus Web pages correctly and ranks them at the bottom of the list. The more Web pages that an Artificial Web has, the heavier

the penalty that AIR imposes.

Figure 2(e) illustrates a legitimate way for Web sites to improve their ranks in AIR. The additional link from node 3 to node 6 signifies that the Web site represented by nodes 6-8 improves its quality and becomes a part of the main Web by receiving strong endorsements from many high-quality pages, represented by node 3. AIR now assigns a higher rank to node 6 than node 4. It is instructive to note that node 3 suffered a reduction in its rank by endorsing node 6. Considering the fact that the link from node 3 to node 6 is the only difference between Figure 2(d) and Figure 2(e), it appears that node 3 is being ‘punished’ for doing the ‘right thing’—endorsing a high-quality page. In fact, this pattern can be found in most of the previous examples: node 5 of Figure 2(d) and Figure 2(e), for example, has a higher rank than node 2 by making fewer links. This seemingly ‘unfair’ feature is actually one of the most important characteristics of AIR. In AIR, votes of endorsement always flow downward from the paragons at the top to the universal sink at the bottom. No reverse flow is allowed. Only a superior can promote its subordinate at the cost of its rank. Either an outgoing link from a page has absolutely no effect on its rank (link to a subordinate) unless the page is a paragon whose rank is fixed at the maximum value. One can never increase the rank of a page that already has the highest rank in a group by making more internal links among themselves. This means that AIR uses links for analysis but it is no longer vulnerable to link manipulation. This feature of AIR discourages the frivolous use of links and returns links to their originally intended purposes: providing paths to useful information and facilitating navigation between Web pages. Webmasters will be forced to choose their links carefully, ending practices such as indiscriminate link exchanges.

Resilience to Link Manipulation

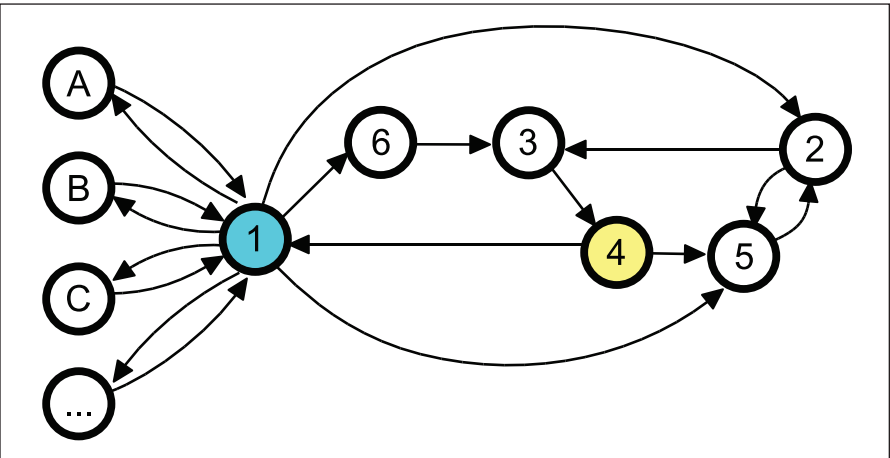
To compare the effectiveness of the AIR algorithm, we demonstrate the resilience of AIR to link manipulation as compared to PageRank and other newer link analysis techniques specifically designed to target link spam, TrustRank⁷ and DiffusionRank.¹⁶ We

use the same graph used by Yang *et al.* 16 wherein a node is manipulated by adding new child nodes, creating a link exchange network as shown in Figure 3. The results from PageRank (PR), TrustRank (TR), DiffusionRank (DR) and AIR are then compared.

Similar to AIR, TrustRank and DiffusionRank use a set of reference nodes to “seed” the initial scores. Node 4 in Figure 3 is selected as the trusted node (the paragon in AIR). Node 1 is then manipulated by adding child nodes in multiples of 2. We monitor the change in two quantities with the addition of new nodes in order to evaluate the susceptibility of different algorithms to link spam: (1) the relative score of node 1, and (2) the rank of node 1. For consistency with previous literature, we use the decay factor of 0.85 for PageRank, TrustRank and DiffusionRank computation. We also use the heat coefficient of 1 for DiffusionRank. In all cases, we use 100 iterations to converge the rank scores. For AIR, we set $g_0 = 0.5$, which provides ratio of the scores between nodes 1 and 4 (0.47) within the range computed by the other three algorithms (PR = 0.56, TR = 0.42, DR = 0.45) on the unmanipulated graph.

The result of this experiment, summarized in Figure 4, shows that AIR is the most resilient to link spam, followed by DiffusionRank, TrustRank and PageRank which is the most susceptible to manipulation. We also noted that AIR is the only algorithm that actually reduces the score of the manipulated page. Another observation we made is that even though TrustRank score of node 1 increases at

Figure 3: An example graph used to test the resilience of ranking algorithms against link manipulation. Node 4 is chosen to be the paragon or the trusted node. Node 1 is manipulated to boost its rank by adding child nodes in multiples of 2.



a slower rate than PageRank, node 1 attains the rank of 1 after 16 child nodes have been added (Figure 4(b)). The DiffusionRank score grows very slowly largely in part because the recommended value of heat diffusion coefficient is very small. On the contrary, the rank of node 1 in AIR always drops to 5 with in a reasonable range (0.1 – 1.0). Note that node 1 can never drop to the last rank since the only source of rank for node 6 is node 1, and hence node 6 will always have a lower rank than node 1.

To further demonstrate the effectiveness of the AIR algorithm in suppressing spamming Web pages, we performed a large scale realistic experiment with an actual Web database at Become.com. As the Web is crawled, we store the downloaded pages on a randomly chosen machine out of a set of distributed computers that contain

over 3.5 billion Web pages and 40 billion hyperlinks in total.^a We chose two separate sets of top one million pages that are ranked either by AIR or PageRank algorithm. Among the top one million pages, we counted the number of pages that contained keywords that are known to be heavily spammed in their URL. Since most of these keywords contain terms that are too graphic, such as “sex” and “porn”, we refer to them only as *sw1*, *sw2*, ..., *swN*.^b Figure 5 shows that the AIR algorithm is more

a At the time of this study, the database contained 3,567,188,889 web pages and 40,450,027,040 hyperlinks. Each link was represented by its source page ID and destination page ID. When multiple machines are used for parallel computation, each page ID requires 5 bytes of data (1 byte for the machine ID where the page is stored and 4 bytes for the page offset). Therefore, the total connectivity graph for our data set required more than 200 GB of disk space.

b The list of keywords used for this study is available upon request.

Figure 4: Comparison of AIR with other rank algorithms. (a) Relative score of the manipulated node (node 1 in Figure 3) as a function of the number of child nodes. (b) Rank of the manipulated node. The number of child nodes is given in logarithmic scale of 2.

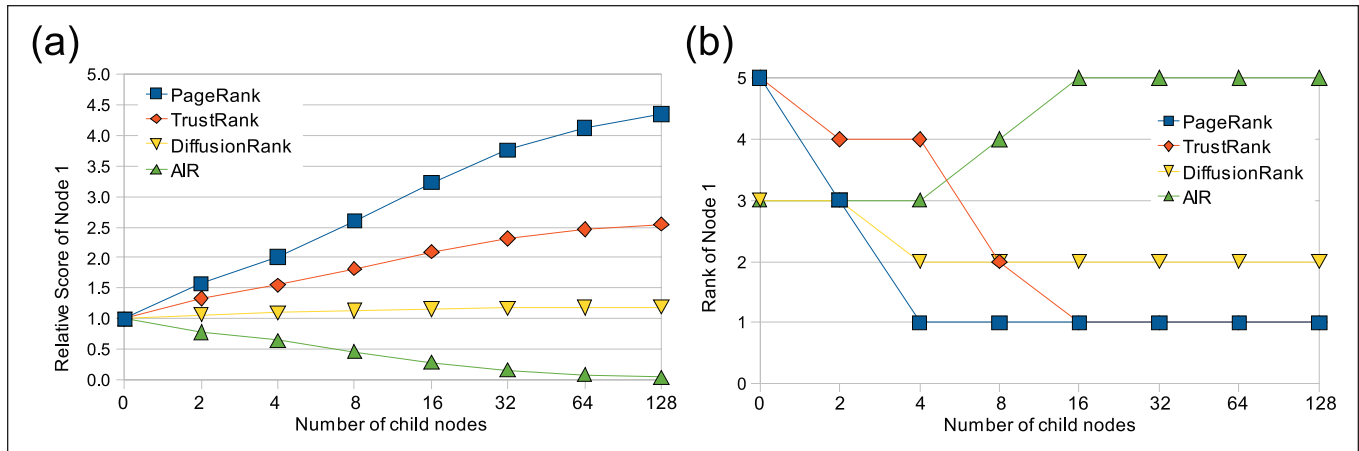
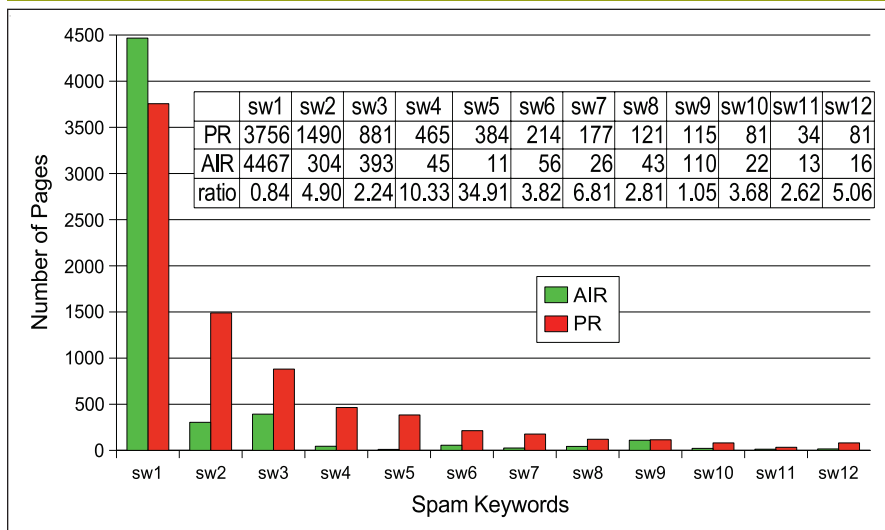


Figure 5: Suppression of spamming web pages with AIR system. The number of web pages containing spam keywords in their URLs among the top one million pages ranked by PageRank and AIR algorithms. The table shows the actual numbers of pages and their ratios.



The authors thank Marcin Kadluczka for stimulating discussions in developing the original concept of AIR.

Rohit Kaul (rohit@become.com) is a Principal Engineer at Become, Inc., 1300 Crittenden Lane, Suite 403, Mountain View, CA 94043, USA.

Yeogirl Yun (yeogirl@wisenu.co.kr) is the Founder and Associate Professor of Physics at Mississippi State University, Mississippi State, MS, where he is also the Associate director of the Center for Computational Sciences in the High Performance Computing Collaboratory.

Seong-Gon Kim (kimsg@hpc.msstate.edu) is an Associate Professor of Physics at Mississippi State University, Mississippi State, MS, where he is also the Associate director of the Center for Computational Sciences in the High Performance Computing Collaboratory.

© 2009 ACM 0001-0782/09/0800 \$10.00

effective than PageRank in suppressing the spam Web pages and in promoting higher-quality pages. Even in the case of the first keyword (“sex”) where the AIR system selected more pages than the PageRank algorithm, a closer inspection reveals that most of the Web pages chosen by AIR are not from spamming Web sites, but are from legitimate Web sites related to sex education and women’s health, such as iVillage.com. On the contrary, the PageRank algorithm selected a large number of pages from unequivocally pornographic sites that employ the most egregious forms of spamming techniques.

Conclusions

We introduced the AIR algorithm using a simple voting scheme. The model was implemented by converting the Web to an equivalent electronic circuit with diodes. We showed that the AIR system, employing the concept of paragons and a universal sink, was very effective in recognizing a group of tightly linked low-quality pages. We further showed that the AIR system readily recognized the Artificial Web and actively punishes the entire group algorithmically without any manual intervention. In a direct comparison with the widely adopted PageRank algorithm, we demonstrated that the AIR system provided not only more intuitive and objective rankings but also more effective measures to guard against most existing search engine spamming techniques.

References

1. Arasu, A., Cho, J., Garcia-Molina, H., Paepcke, A., and Raghavan, S. Searching the Web. *ACM Transactions on Internet Technology* 1, 1, (June 2001) 2–43.
2. Brin, S., and Page, L. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems* 30, (1-7) (1998) 107–117.
3. Bronshtein, N. I., and Semendyayev, A. K. *Handbook of Mathematics*, 892. Springer-Verlag, New York, 3rd edition, 1997.
4. Faloutsos, C., Mccurley, S. K., and Tomkins, A. Fast discovery of connection subgraphs. In *Proceedings of the 2004 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, NY, (2004), 118-127.
5. Fetterly, D., Manasse, M., and Najork, M. Spam, damn spam, and statistics: using statistical analysis to locate spam Web pages. In *Proceedings of the 7th International Workshop on the Web and Databases*, ACM, NY, (2004) 1-6.
6. Gulli, A., and Signorini, A. The indexable Web is more than 11.5 billion pages. In *Special interest tracks and posters of the 14th International Conference on World Wide Web*, ACM, NY, (2005), 902-903.
7. Gyongyi, Z., Garcia-Molina, H., and Pedersen, J. Combating Web spam with TrustRank. In *Proceedings of the 30th VLDB Conference*, 2004, 576-587.
8. Henzinger, M., Motwani, R., and Silverstein, C. Challenges in Web search engines. *SIGIR Forum* 36, 2, 2002.
9. Kleinberg, J., and Lawrence, S. The structure of the Web. *Science*, 294:1849, 2001.
10. Kleinberg, M. J. Authoritative sources in a hyperlinked environment. *Journal of the ACM* 46, 5, (1999), 604–632.
11. Korniss, G., Hastings, B. M., Bassler, E. K., Berryman, J. M., Kozma, B., and Abbott, D. Scaling in small-world resistor networks. *Physics Letters A*, 350:324, 2006.
12. Lopez, E., Buldyrev, V. S., Havlin, S., and Stanley, S. H. Anomalous transport in scale-free networks. *Physical Review Letters* 94, 24, 2005.
13. Lyman, P., Varian, R. H., Swearingen, K., Charles, P., Good, N., Lamar Jordan, L., and Pal, J. How much information? 2003; <http://www.sims.berkeley.edu/howmuch-info-2003>, 2003.
14. Newman, M. E. J., and Girvan, M. Finding and evaluating community structure in networks. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)* 69, 2, 2004.
15. Press, H. W., Flannery, P. B., Teukolsky, A. S., and Vetterling, T. W. *Numerical Recipes in FORTRAN: The Art of Scientific Computing*. Cambridge University Press, Cambridge, U.K, 1992, 864-866.
16. Yang, H., King, I., and Lyu, M. R. DiffusionRank: A possible penicillin for Web spamming. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, NY, (2007), 431-438.