Manipulability of PageRank under Sybil Strategies

Alice Cheng * Eric Friedman †

Abstract

The sybil attack is one of the easiest and most common methods of manipulating reputation systems. In this paper, we quantify the increase in reputation due to creating sybils under the PageRank algorithm. We compute explicit bounds for the possible PageRank value increase, and we use these bounds to estimate the rank increase. Finally, we measure the effect of sybil creation on nodes in a web subgraph. We find that the resulting rank and value increases agree closely with the analytic values.

1 Introduction

Ranking systems are an important tool in a wide range of online settings, such as online shopping (Amazon, eBay), as a means of inferring reputation of sellers or goods; in the peer-to-peer setting, to weed out untrustworthy or freeloading users; and the area of online search, as a means of ranking webpages.

However, many ranking systems are vulnerable to manipulation, and users often have incentives to cheat. A higher ranking may offer an economic benefit - for example, a study of the eBay reputation system found that buyers are willing to pay a premium of 8% for buying from sellers with high reputation [11]. Another example is in online search, where websites, in order to gain web traffic, use the services of online companies which help sites improve their search engine rankings.

PageRank is currently one of the most widely used reputation systems. It is applied in peer-to-peer networks in the EigenTrust algorithm [7], and in web search, as the foundation for the Google search algorithm [9]. Although PageRank has proven to be a fairly effective ranking system, it is easily manipulable by a variety of strategies, such as collusion or the sybil attack [12, 5].

We focus primarily on the sybil attack, described by Douceur [4]. In this attack, a single user creates several fake users - called sybils - who are able to link to (or perform false transactions with) each other and the original user. For example, in the web, a user can create new webpages and manipulate the link structure between them. In many online settings, new identities are cheap to create, and it may be difficult to distinguish between sybils and real users. In the case of PageRank, users have already been observed performing sybil-like strategies, such as forming link farms [5].

It is easy to see that PageRank is vulnerable to sybil attacks. However, as we showed in earlier work, almost all practical reputation systems are vulnerable to sybil attacks [3]. It may be unrealistic to restrict one's attention only to sybilproof reputation systems, and reputation systems may vary widely in their exploitability. For example, the indegree reputation function (where a user's reputation value is his indegree) is easily exploitable - a user may increase his indegree to any desired value by creating sybils. On the other hand, a reputation function based on maximum flow is not sybilproof with respect to rank, but is more difficult to manipulate. Thus, it becomes important to gauge the degree of vulnerability of different reputation systems. In order to systematically compare PageRank to other reputation systems, we develop a method of estimating the potential PageRank rank and value improvement of a node in a web-like graph.

In this paper, we begin this research program with a formal and experimental analysis of the vulnerability of PageRank to sybil attacks. We provide analytic estimates of this vulnerability, which only depend on the overall PageRank distribution in the graph and then check the tightness of our analysis on empirical web graph data. We find a very close agreement and are led to believe that our estimates can be applied to estimate the vulnerability of PageRank on web-like graphs.

*Center for Applied Mathematics, Cornell University, alice@cam.cornell.edu

†School of Operations Research & Industrial Engineering and Center for Applied Mathematics, Cornell University, ejf27@cornell.edu

[0]Work supported by NSF. ITR-0325453

2 Related Work

Our work is related to [12] which considers the effect of collusion on PageRank. Collusion is a strategy where users mutually agree to alter their outlink structure in order to improve their reputations. Collusive strategies and sybil strategies differ in at least two critical ways. First, a sybil creator can gain reputation at the expense of his sybils, while colluders are unlikely to cooperate unless both can raise their reputations. Second, sybil strategies are likely to be less constrained in size - a user can often easily create a large sybil group, while it may be more difficult to find an equal number of users to form a colluding group.

Other related work includes Gyongyi and Garcia-Molina who give a fairly exhaustive list of strategies to falsely boost reputation on the web [5]. The PageRank algorithm itself has generated a lot of interest and study. Bianchini, Gori, and Scarselli consider the total PageRank within a community of nodes, and give methods for a community to boost its total reputation [2]. A survey paper by Langville and Meyer gives a general overview of the PageRank algorithm, and discusses many issues including PageRank stability and efficient computation [8].

3 Preliminaries

Given a set of users V, we represent the setting as a directed graph $\mathcal{G} = (V, E)$. The edges E represent direct trust between users. For example, in the web, an edge $(i, j) \in E$ may represent a hyperlink from site i to site j. Let $n = |V|$. Let $d(i)$ be the outdegree of the node $i \in V$. We require that every node has positive outdegree. Since this isn't always the case for real world graphs, we will insert a self-loop for all nodes with outdegree 0. We will assume that no other nodes have self-loops.

3.1 PageRank

The PageRank values on a network graph $\mathcal{G}$ are given by the stationary probabilities of the following random walk on $\mathcal{G}$: with probability $1 - \epsilon$, a walker at a node i walks along an outgoing edge of i, choosing the edge uniformly with probability $\frac{1}{d(i)}$, and with probability ϵ, jumps to a node chosen uniformly at random. Let v be the vector of stationary probabilities - v_i is the stationary probability of the node i. The resulting PageRank ranking is given by the order of the values of v, sorted from highest to lowest (note that a higher value v_i corresponds to a lower numbered rank). For convenience, we will typically not talk about the stationary vector of probabilities v, but will instead use $\pi = nv$. Clearly, π yields the same ranking as v. For a node i, we will refer to π_i as its *PageRank value* and its order on a highest to lowest list sorting the π_j's as its *rank*.

Given $\mathcal{G}$, we can construct the adjacency matrix of $\mathcal{G}$, A, $A_{ij} = 1$ if $(i, j) \in E$, and 0 otherwise. Let $M(\mathcal{G})$ be the matrix given by $M(\mathcal{G})_{ij} = \frac{A_{ji}}{d(j)}$.

Note that π is the principal eigenvector (with eigenvalue 1) of the matrix $(1-\epsilon)M(\mathcal{G}) + \frac{\epsilon}{n}\overrightarrow{1}\overrightarrow{1}^T$, where $\overrightarrow{1}$ is the vector of all ones. That is, π satisfies the following matrix equation:

$$(1-\epsilon)M(\mathcal{G})\pi + \epsilon\overrightarrow{1} = \pi$$

We may sometimes find it convenient to express the above as a scalar equation: for a node $i \in V$,

$$\pi_i = (1-\epsilon)\sum_{j \to i}\frac{\pi_j}{d(j)} + \epsilon,$$

where $j \to i$ to denotes $(j, i) \in E$ (i.e. j points to i).

We can also consider the iterative version of the above equations, where $\pi_i^t \to \pi_i$ as $t \to \infty$ [8].

$$\pi_j^0 = 1, \forall j; \pi_i^t = (1-\epsilon)\sum_{j \to i}\frac{\pi_j^{t-1}}{d(j)} + \epsilon$$

3.2 Sybil Strategies

In a sybil strategy, a node creates k sybils, and manipulates his own outlinks and those of his sybils. More formally,

Definition 1 *Given a graph $\mathcal{G} = (V, E)$ and a node $i \in V$, a **sybil strategy** for the node i, is a new graph $\mathcal{G}' = (V', E')$, such that $V' = V \cup S$, where $S = \{s_1, \ldots, s_k\}$ is a set of sybils (disjoint from the original node set) and E' is such that for all $j \in V, j \neq i$, for all $x \in V$, $(j, x) \in E \Leftrightarrow (j, x) \in E'$.*

A *sybil collective* is the node set $S \cup \{i\}$ (i and its sybils). Let r_i be the rank of i in $\mathcal{G}$, π_i be the PageRank value of i in $\mathcal{G}$. Let ρ_i be the new PageRank value for i in $\mathcal{G}'$ and r_i' be the new rank. Then a strategy is successful for i with respect to value if $\rho_i > \pi_i$. It is successful with respect to rank if $r_i' < r_i$.

We say that a reputation function is value (or rank) *sybilproof* if for all graphs $\mathcal{G}$, no node has a successful sybil strategy with respect to value (or rank).

It is straightforward to come up with an example where a node can increase its PageRank through creating sybils. In [3], we showed that no nontrivial

symmetric reputation system (i.e. one that is invariant under a relabelling of the nodes) can be sybilproof. The version of PageRank that we described in the previous section is clearly symmetric, so there is a network where a node could benefit from creating sybils. Further, by this result, we know that adjusting some of the parameters of PageRank (such as the value of ϵ) in a nontrivial way while maintaining symmetry cannot yield a sybilproof mechanism. However, it is easy to show that even an asymmetric version of PageRank (such as the version used in EigenTrust) may be manipulated with sybils.

Note that a sybil creator may choose any configuration of edges within the sybil collective. However, for the purposes of this paper, we focus on one particular sybil strategy. In this strategy, a node i removes his outlinks, creates k sybils, and links to each of his sybils. The sybils link only to the sybil creator i. Figure 4 (in the appendix) depicts a node applying this strategy with 3 sybils.

Bianchini et. al. show that this configuration concentrates the maximum amount of reputation on the sybil creator [2]. Intuitively, any random walk inside the sybil collective must hit i on every other step. Further, removing any links from the collective to nodes outside of the collective improves their overall PageRank - a random walk which enters the collective must remain there until a random jump.

4 Analysis

Our main results are analytic bounds for the value increase upon creating sybils which are presented below. We then compare these bounds with empirical data.

4.1 Value Increase

We give the following upper and lower bounds for value increase:

Theorem 2 *Let π be the old PageRank value vector, and ρ be the new PageRank vector when node i creates k sybils by the above strategy, keeping all other nodes fixed. Then, if i has no self-loop in the original graph, we have the following bounds:*

$$\pi_i + k\frac{1-\epsilon}{2-\epsilon} \leq \rho_i \leq \frac{\pi_i + \epsilon(1-\epsilon)k}{\epsilon(2-\epsilon)}$$

Since we typically talk about the ratio between ρ and π, we give the corresponding bounds for the value inflation ratio ρ_i/π_i:

$$1 + \left(\frac{1-\epsilon}{2-\epsilon}\right)\frac{k}{\pi_i} \leq \frac{\rho_i}{\pi_i} \leq \frac{1}{\epsilon(2-\epsilon)} + \left(\frac{1-\epsilon}{2-\epsilon}\right)\frac{k}{\pi_i}$$

A proof of this theorem is included in the appendix.

These bounds allow us understand how the value increase changes as we increase the number of sybils or vary ϵ. Further, for given values of ϵ and π, we can estimate the number of sybils needed to increase a node's reputation by some given amount.

Increasing k increases both the upper and lower bounds, and appears to yield larger increases in the value inflation ratio when π_i is small. For example, for $\epsilon = 0.15$, we have $1 + 0.46(\frac{k}{\pi_i}) \leq \frac{\rho_i}{\pi_i} \leq 3.6 + 0.46(\frac{k}{\pi_i})$, meaning that for a node with value π_i equal to the mean value 1, doubling one's value requires between 1 and 3 sybils. For a node with the median value, which is ≈ 0.3 in our data sample, it requires only 1 sybil.

The above bounds are tight. The upper bound is attained for nodes i that are contained in no cycles. One can show (using similar techniques as in the proof of the theorem), that in this case, the reputation of i's recommenders (those nodes j with $j \to i$) are unchanged when i removes its outlinks. With a simple computation (or by following the proof of the above theorem), the equality follows.

The lower bound is attained for subgraphs in a "petal" configuration, where the node i points only to nodes who point only back at i (as in the sybil configuration). This configuration attains the lower bound because i's recommenders were previously "sybil-like", in that they attained most of their reputation from i and returned as much reputation as possible to i. Once i removes its outlinks, the value of the links (j, i) to i become very small.

However, most nodes may not lie in either of the extremes described above. Indeed, it is reasonable to expect (due to the observed high clustering in the web [1]) that some nodes lie on short cycles, leading to configurations similar to the "petal". At the same time, some of the edges out of i are likely not part of short cycles, suggesting configurations as in the upper bound.

4.1.1 Data for $k = 1$

In this section and the ones that follow we use a dataset from a webcrawl, available at [6]. The total number of nodes is $n = 281,903$. We preprocess the graph to insert self-loops for each node with outdegree 0, to guarantee that the matrix $M(G)$ of the graph (defined above) is indeed stochastic. In the first experiment, we select 10000 nodes uniformly at random from the graph, and for each node selected, we create a single sybil for that node under the above sybil strategy, keeping all other nodes fixed. We set

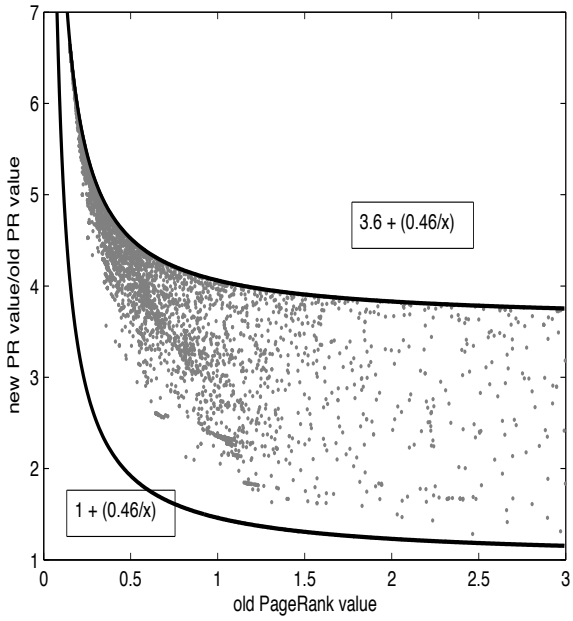

Figure 1: Old PageRank value (x axis) versus new value/old value ratio for the case of 1 sybil, jump parameter $\epsilon = 0.15$. The lines are given by the theoretical upper and lower bounds.

the jump parameter ϵ to 0.15.

In Figure 1, we plot the old PageRank value π_i versus the ratio $\frac{\rho_i}{\pi_i}$. For the sake of visual clarity, we cut off the graph at $\pi_i = 3$, which still includes the vast majority of the nodes ($\sim 97\%$). We observe that the nodes are able to achieve an average linear factor increase of 4.7. Further, the upper and lower bounds of the data appear to match closely with the computed bounds. For larger values of π_i, we found that both the upper and lower bounds appear to be tight, and the data points are roughly evenly distributed between the bounds. For smaller values, the upper bound appears tight, while the lower bound is not. One possible explanation for this discrepancy is that the lower bound (as discussed in the previous section) is attained when the original node is in a sybil-like structure, where the node is contained nearly exclusively in small cycles (i.e. many paths out of the node are small cycles). However, being in such a structure may also suggest a higher reputation value than a typical node, so nodes that nearly attain the lower bound may tend also to have higher reputation. In fact, the central node of a petal will have a PageRank value $\pi_i \geq 1$, and we note that the lower bound appears tight in this regime in our plot.

Further, we can note that aside from the deviation from the lower bound for lower reputation nodes, the nodes appear fairly evenly spread between the bounds, suggesting, as we stated earlier, that most nodes are widely distributed between the extremes of being in no cycles and being in many short cycles.

4.1.2 Data for $k = 1, 2, 5, 10$

For this experiment, we select a node uniformly at random from the graph. For each node selected, we set up a sybil strategy for that node with $k = 1, 2, 5, 10$. We set the jump parameter to $\epsilon = 0.15$. We repeat this 1000 times.

We plot the ratio of new PageRank to old PageRank in Figure 5, in the appendix. The data points appear roughly of the same shape as in the $k = 1$ case, and the boundaries of the data points agree with our computed bounds. Further, as in our bounds, we can observe that lower value nodes tend to gain larger increases with k and higher value nodes tend to have more modest increases.

4.2 Rank Increase

In many settings (such as web page ranking) one cares mainly about the ranking implied by the PageRank values and not the actual values themselves. In this section we evaluate the rank increases for a large class of graphs based on an analysis of a large web graph.

Given the value bounds from Theorem 2, if we assume that the PageRank values of most other nodes remain roughly fixed, we can estimate the rank increase using the PageRank distribution. Pandurangan et.al. estimate the probability density of PageRank in a large web subgraph, and find a density of $\approx \frac{c}{x^{2.1}}$, where c is a constant [10]. If we assume that the PageRank density is $F(x) = \frac{c}{x^{2.1}}$, then $Pr(\pi_i \geq x) = \frac{d}{x^{1.1}}$ for some constant d. For a node i, if its PageRank value is π_i, a rough estimate of its rank would be $nPr(\pi_i \geq x) = \frac{nd}{x^{1.1}}$. We found that our dataset matches the rough estimates above fairly closely - for nodes with rank < 40000, the value to rank function is $\approx c_1 v^{-1.1}$, and for nodes with rank > 50000, the value to rank function is $\approx c_2 v^{-0.86}$.

Let r_i be the old rank of i and r_i' be the new rank. Let $r(x) = cx^{-1.1}$ be the PageRank value to rank function (for some constant c). Then, the new rank to old rank ratio $\frac{r_i'}{r_i} \approx \frac{r(\rho_i)}{r(\pi_i)} = (\frac{\pi_i}{\rho_i})^{1.1}$, using the PageRank value ratio bounds, satisfies the bounds

$$\left(\frac{1}{\frac{1}{\epsilon(2-\epsilon)} + (\frac{1-\epsilon}{2-\epsilon})\frac{k}{\pi_i}}\right)^{1.1} \leq \left(\frac{\pi_i}{\rho_i}\right)^{1.1} \leq \left(\frac{1}{1 + (\frac{1-\epsilon}{2-\epsilon})\frac{k}{\pi_i}}\right)^{1.1}$$

.

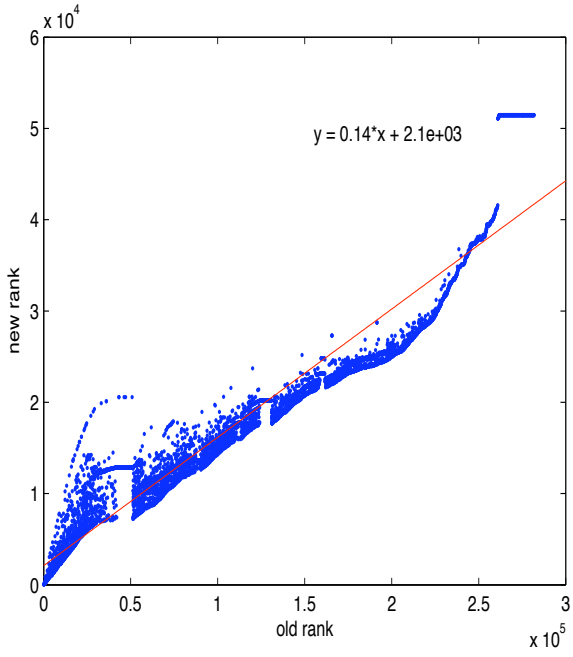

Figure 2: Old rank (x axis) versus new rank for the case $k = 1, \epsilon = 0.15$

For $\epsilon = 0.15$, and $k = 1$, the above bounds are $(\frac{\pi_i}{3.6\pi_i+0.46})^{1.1} \leq (\frac{\pi_i}{\rho_i})^{1.1} \leq (\frac{\pi_i}{\pi_i+0.46})^{1.1}$. For large π_i, we expect a lower bound in the rank increase of 0.28 and an upper bound of ≈ 1. For nodes with small π_i, say $\pi_i < 1$, which accounts for more than 80% of the nodes in our graph, we have a lower bound of approximately 0.13, and an upper bound of 0.66. From our analysis of the bounds for the value ratio in the previous section, we expect the lower bound to be much more accurate than the upper bound in the small π_i regime.

Given these tools, we can estimate the number of sybils needed for the median node (with rank $\frac{n}{2}$) to rise to the top k, for any k. Take the rank function $r(v)$ to be $r(v) = \frac{c}{v^{1.1}}$ for a constant c. Let $r_1 = \frac{n}{2}$ be the rank of the median node. $r_2 = k$. We can estimate the corresponding values for a graph of size n: $v_1 = r^{-1}(r_1) = (\frac{2c}{n})^{1/1.1}$, $v_2 = r^{-1}(r_2) = (\frac{c}{k})^{1/1.1}$. The value ratio $\frac{v_2}{v_1} = (\frac{n}{2k})^{.91}$. Plugging in the value ratio from the theorem inequalities (for $\epsilon = 0.15$), we have $k \sim \frac{\pi_i}{.46}(\frac{n}{2k})^{.91}$. Therefore, for a graph with ~ 300000 nodes, and median $\pi_i = 0.3$, a median node requires ~ 500 sybils to rise to the top 100. In a graph with median value $\pi_i < 1$, a median node would require less than ~ 76 sybils to rise to the top 1%.

4.2.1 Data for $k = 1$

The experimental setup is identical to the one described in section 3.1.1. We plot the old rank versus the new rank in Figure 3. We find that all but the very highest or very lowest ranked nodes are able to improve (or decrease) in rank by a factor of approximately 0.14 times - approximately a 6-fold improvement. This value agrees well with our computed lower bound (for small π) of 0.13. Further, we can observe that for nodes with original rank > 50000 (these nodes have $\pi_i > 1$), the improvement in rank is much more spread out, and less significant - which may be explained by the fact that the PageRank value ratios are more spread out, and attain the upper and lower bounds in the large π_i regime.

4.2.2 Data for $k = 1, 2, 5, 10$

The experimental setup here is identical to the one described in section 3.1.2. We plot the old rank versus the new rank in Figure 6 (in the appendix) for $k = 1, 2, 5, 10$.

We see a much more dramatic improvement in rank than value resulting from increasing the number of sybils. We find average ratios of old rank to new rank, $\frac{r_i}{r_i'}$ of 7.1 for $k = 1$, 16.4 for $k = 2$, 40 for $k = 5$, and 90.9 for $k = 10$. As expected, as in the value case and suggested by our bounds, sybil creation tends to be more effective for higher ranked (i.e., lower π_i) nodes.

4.3 Varying ϵ and sybil strategies

One way to vary the PageRank algorithm is to alter the parameter ϵ, which determines the probability of making a random jump at each step of the random walk. Our value bounds show that as ϵ increases, the potential increase in value declines. Intuitively, if ϵ is high, the effect of creating sybils may be reduced, since a random walk does not remain trapped in sybil collectives for a long time. By repeating the previous experiments for various values of ϵ, we found that the value increase does decline predictably as ϵ increases. However, nodes were still able to achieve significant rank improvements as we increased ϵ. In fact, higher values of ϵ yielded slightly higher average increases in rank for sybil-creating nodes. Figure 6 plots the average old rank to new rank ratio as ϵ varies. Though the value increase declines as ϵ increases, raising ϵ increases the likelihood of choosing a node at random in the PageRank random walk, making the overall PageRank distribution more uniform, compressing the set of typical pagerank values

We also considered two different sybil strategies. In one, users do not remove their outlinks to non-sybil nodes. In the other, users move their outlinks to a sybil node. In both of these cases, we observed an improvement in PageRank value and rank, though slightly less than in the original strategy.

5 Future Work

Our analysis shows that PageRank is extremely manipulable, even with simple strategies using a small number of sybils. We provided tight analytic approximations that can be used to estimate the manipulability of Pagerank in a variety of settings.

One issue that we haven't considered is the correlations between web pages on similar topics. For example, typically - and particularly in the web setting - a node is competing with a subset of nodes relating to the same topic (e.g. an electronics retailer probably doesn't care about ranking above a political weblog). Therefore, one potential further area of study is an analysis of how much the improvements observed above allow a typical node to beat its most likely competitors. Further the subset of competitors may look very different from a uniformly random subset of the web. For example, a subset of nodes all relating to the same topic may be more clustered than a random subset of the web. Is sybil creation more effective or less in this setting?

In this paper, we focus entirely on the PageRank algorithm, and find that it is easily manipulable. However, there are many other potential reputation systems, and we do not expect all of them to be as easily manipulable with sybils. Similar studies on the manipulability of other reputation systems may allow direct comparison of the manipulability of various reputation systems.

In particular, one would expect that there would be a trade off between the quality of the ranking system its manipulability. For example, as shown in [3], the "shortest path" ranking system is immune to sybil attacks; however, it is most likely less effective at ranking than PageRank. The development of robust and efficient ranking mechanisms is an important open problem.

References

[1] Lada Adamic. The small world web. In S. Abiteboul and A.-M. Vercoustre, editors, *Research and Advanced Technology for Digital Libraries, Lecture Notes in Comp. Sci.,1696*, pages 443–452. 1999.

[2] Monica Bianchini, Marco Gori, and Franco Scarselli. Inside pagerank. *ACM Transactions on Internet Technology*, 5(1), February 2005.

[3] Alice Cheng and Eric Friedman. Sybilproof reputation mechanisms. In *Third Workshop on the Economics of Peer-to-Peer Systems*, 2005.

[4] J. Douceur. The sybil attack. In *Proceedings of the IPTPS02 Workshop*, 2002.

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

[6] Sepandar Kamvar. stanford.edu web crawl, 2002, http://nlp.stanford.edu/ sdkamvar/data/stanford-web.tar.gz.

[7] Sepandar D. Kamvar, Mario T. Schlosser, and Hector Garcia-Molina. The eigentrust algorithm for reputation management in p2p networks. In *Proceedings of the Twelfth International World Wide Web Conference (WWW)*, 2003.

[8] Amy Langville and Carl Meyer. Deeper inside pagerank. *Internet Mathematics*, 1(3), 2004.

[9] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. In *Proceedings of the 7th International World Wide Web Conference*, pages 161–172, 1998.

[10] Gopal Pandurangan, Prabhakar Raghavan, and Eli Upfal. Using pagerank to characterize web structure. In *8th Annual International Computing and Combinatorics Conference (COCOON)*, 2002.

[11] Paul Resnick, Richard Zeckhauser, John Swanson, , and Kate Lockwood. The value of reputation on ebay: A controlled experiment. Working paper, available at http://www.si.umich.edu/ presnick/papers/postcards/index.html, 2004.

[12] Hui Zhang, Ashish Goel, Ramesh Govindan, Kahn Mason, and Benjamin Van Roy. Making eigenvector-based reputation systems against collusions. In *The Third Workshop on Algorithms and Models for the Web Graph*, 2004.

6 Appendix

In this section, we wish to prove the following theorem:

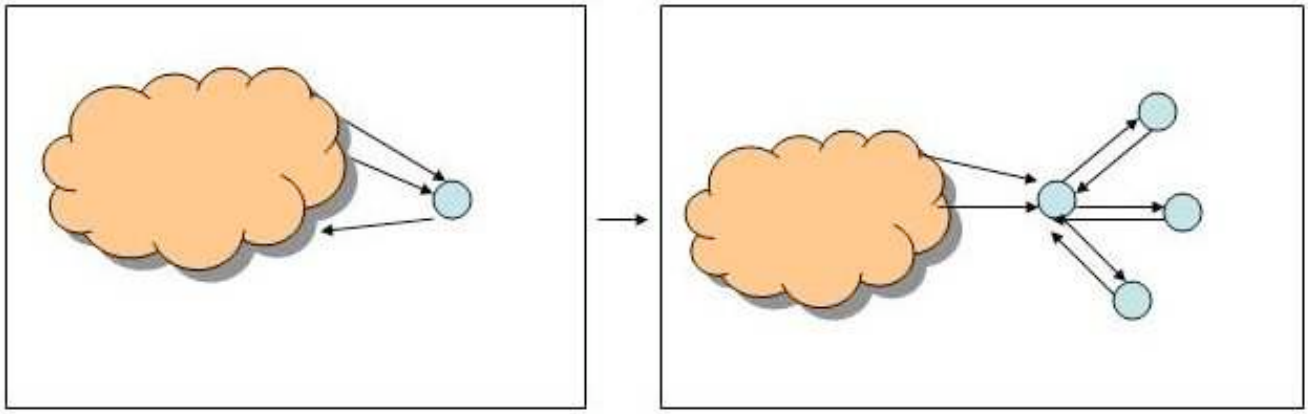

Figure 3: On the left: a single node with both outlinks and inlinks from the rest of the graph (cloud). On the right: the node removed its outlink, and created 3 sybils, arranged in the "petal" formation.

Theorem 3 *Let π be the old PageRank value vector, and ρ be the new PageRank vector when node i creates k sybils by strategy A, keeping all other nodes fixed. Then, if $d(i) > 0$, we have the following bounds:*

$$\rho_i \leq \frac{\pi_i + \epsilon(1-\epsilon)k}{\epsilon(2-\epsilon)}$$

$$\rho_i \geq \pi_i + k\frac{1-\epsilon}{2-\epsilon}$$

Let $G = (V, E)$, be a directed graph with $V = \{1, \ldots, n\}$. For $j \in V$, let $d(j)$ be the outdegree of the node j. We define $M(G)$ be the $n \times n$ matrix such that

$$M(G)_{ij} = \begin{cases} \frac{1}{d(j)} & \text{if } (i,j) \in E \\ 0 & \text{otherwise} \end{cases}$$

Define $\tilde{M}, v, w$ such that

$$M(G) = \begin{Bmatrix} \tilde{M} & w \\ v^T & 0 \end{Bmatrix}$$

WLOG let $i = n$. Let $G' = (V, E')$ be the graph where n removes its outlinks and creates a self loop. Let G'' be the graph where n has k sybils as in strategy A.

Let π be the original PageRank vector for G, with $\|\pi\|_1 = n$, and let π' be the PageRank vector for G', with $\|\pi'\|_1 = n$. Let ρ be the $n + k$ vector such that $\rho_x = \pi'_x$ for all $x < n$, $\rho_n = \frac{1-\epsilon}{2-\epsilon}k + \frac{1}{2-\epsilon}\pi'_n$ and $\rho_x = \frac{1}{2-\epsilon} + \frac{1-\epsilon}{2-\epsilon}\frac{\pi'_n}{k}$ for all $x > n$. By considering the matrices $M(G'), M(G'')$ in block form as above, an easy computation shows that $(1-\epsilon)M(G'')\rho + \epsilon\overrightarrow{1} = \rho$. Therefore, ρ is the unique PageRank vector of G'' (normalized to $n + k$).

It suffices then to show that $(2-\epsilon)\pi_n \leq \pi'_n \leq \frac{\pi_n}{\epsilon}$

Lemma 4 *$\pi'_j \leq \pi_j$ for all $j < n$.*

Proof: Note that the outdegrees of nodes $j < n$ in G' are equal to their outdegrees in G, so we can write the outdegree of x for $x < n$ as $d(x)$. Recall the iterative version of PageRank: $(\pi'_j)^t = (1-\epsilon)\sum_{x \to j} \frac{(\pi'_x)^{t-1}}{d(x)} + \epsilon$, for $t \geq 1$, and $(\pi'_j)^0 = 1$ for all j. Since $(\pi'_j)^t \to \pi'_j$ as $t \to \infty$, it suffices to show that $(\pi'_j)^t \leq \pi_j^t$ for all t, and for all $j < n$. This is trivially true for $t = 0$. By induction, assume that $(\pi'_x)^{t-1} \leq \pi_x^{t-1}$ for all $x < n$. Consider some node $j < n$.

$$\begin{aligned} (\pi'_j)^t &= (1-\epsilon) \sum_{x:(x,j)\in E'} \frac{(\pi'_x)^{t-1}}{d(x)} + \epsilon \\ &\leq (1-\epsilon) \sum_{x:(x,j)\in E, x<n} \frac{(\pi_x)^{t-1}}{d(x)} + \epsilon \\ &\leq \pi_j^t \end{aligned}$$

The first inequality follows from induction and the fact that n doesn't point to any $j < n$ in G'. ∎

Plugging in π_j for each π'_j in the PageRank formula for π_n gives the upper bound. For the lower bound, we have the following lemma:

Lemma 5 *$\pi'_n \geq (2-\epsilon)\pi_n$.*

Proof: Note that we require the assumption that $d(n) > 0$ for this lemma. Consider a node $i \neq n$. In the graph G', we have

$$\pi'_i = (1-\epsilon) \sum_{j \to i, j \neq n} \frac{\pi'_j}{d(j)} + \epsilon,$$

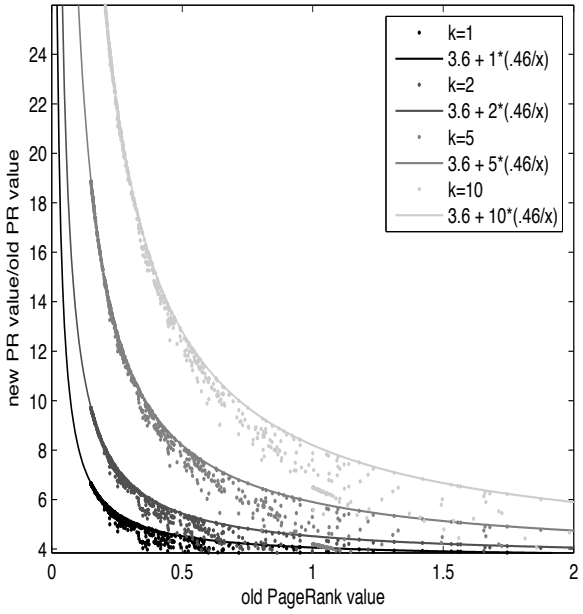

Figure 4: Old PageRank value (x axis) versus old value/new value ratio (y axis) for $k = 1, 2, 5, 10$. The lines are the theoretical upper bounds for the various values of k

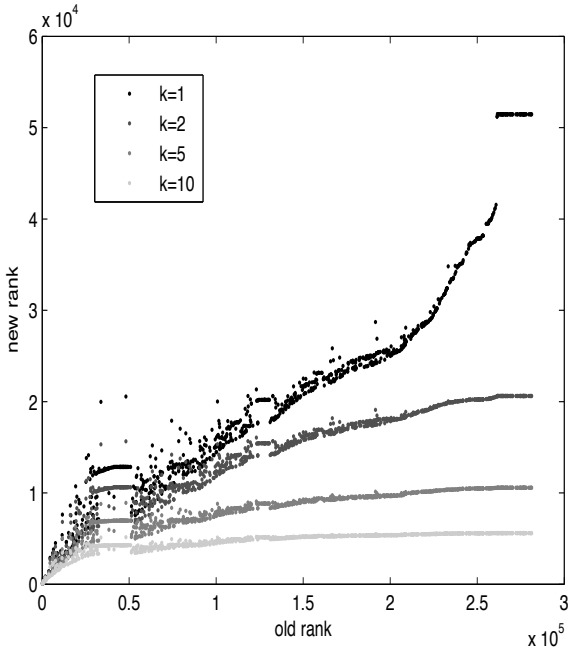

Figure 5: Old rank versus new rank for $k = 1, 2, 5, 10$, $\epsilon = 0.15$

by the fact that n points to no nodes other than itself in G'. Applying the previous lemma, we have

$$\pi'_i \leq (1-\epsilon) \sum_{j \to i, j \neq n} \frac{\pi_j}{d(j)} + \epsilon.$$

Note that $\pi_i = (1-\epsilon) \sum_{j \to i, j \neq i} \frac{\pi_j}{d(j)} + (1-\epsilon)\delta_{ni} \frac{\pi_n}{d(n)} + \epsilon$, where $\delta_{ni} = 1$ if $(n, i) \in E$, and 0 otherwise. Therefore, we have

$$\pi'_i \leq \pi_i - (1-\epsilon)\delta_{ni} \frac{\pi_n}{d(n)}.$$

We can sum the inequality over all $i \neq n$:

$$\sum_{i \neq n} \pi'_i \leq \sum_{i \neq n} \pi_i - (1-\epsilon)\pi_n,$$

where we note that there are exactly $d(n)$ nodes among $i \neq n$ with $\delta_{ni} = 1$ (n had no self-loops in the original graph). Adding $\pi_i + \pi'_i$ to both sides, we have

$$\pi_i + \sum_{i \in V} \pi'_i \leq \pi'_i + \sum_{i \in V} \pi_i - (1-\epsilon)\pi_n.$$

Finally, by normalization, $\sum_{i \in V} \pi_i = \sum_{i \in V} \pi'_i = n$, so $\pi_i \leq \pi'_i - (1-\epsilon)\pi_n$. Rearranging, we get the desired inequality: $(2-\epsilon)\pi_i \leq \pi'_i$ ■

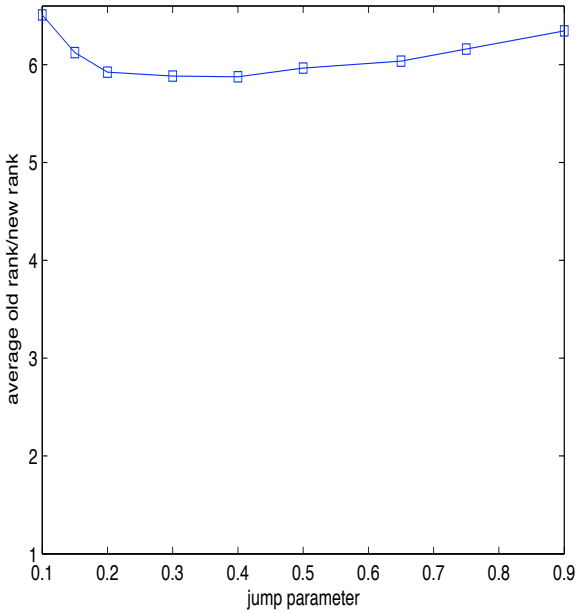

Figure 6: Jump parameter epsilon vs. average old rank/new rank ratio