

The Cost of Attack of PageRank

Andrew Clausen aclausen@ms.unimelb.edu.au
Department of Computer Science and Software Engineering
Department of Mathematics and Statistics
The University of Melbourne

Abstract

On the internet, throw-away identities are cheap. Reputation obtained from recommendations is often the only information available to help sift through the sea of spam to find useful information and trustworthy services. However, a reputation system is only useful if it is difficult to manipulate.

Google's reputation system, PageRank, allows web pages to acquire reputation through payment of entry fees (domain name registrations), or recommendations from other pages (hyperlinks). Levien conjectured that an attacker seeking a high reputation score can not avoid PageRank's entry fees by constructing fake web pages with hyperlinks. This paper proves this claim in the special case that the attacker has no prior incoming hyperlinks.

1 Introduction

The internet makes it easy to interact with strangers. It also contains a vast amount of potentially useful information, among a sea of porn sites, online casinos and scams. Reputation obtained from recommendations may be the only efficient way to sift through piles of junk to find information of interest. Moreover, reputation is often the only information available to evaluate the trustworthiness of a stranger.

The Google search engine (www.google.com) assigns a reputation score to every web page it indexes using the PageRank algorithm by Page et al. (1998). High PageRank scores are a valuable commodity for webmasters seeking listings in Google's search results. PageRank analysis is based on three premises:

- A hyperlink from page A to page B is a recommendation by page A for page B. PageRank offers no mechanism for complaints.
- Recommendations from reputable pages are more important than those from unknown pages.
- Domain names and/or IP addresses are scarce resources that are expensive to obtain.

PageRank begins allocating reputation with domain names and/or web server IP addresses¹, and then allows the reputation to flow through hyperlinks to the rest of the world wide web. Page et al. (1998) observed that PageRank consistently selects reputable web pages. For example, highly interconnected mailing list archives don't skew results, because they receive few hyperlinks from other sources.

Page et al. (1998) suggest that seeding PageRank with a scarce resource such as domain names makes it difficult to manipulate. An attack would require purchasing many domain names. Levien (2003) conjectured that the cost of such an attack would be proportional to the attacker's desired PageRank, and therefore puts high placement in Google's search results out of reach of the sea of spam.

This paper proves this conjecture by deriving the cost of attack of desired PageRank scores in the special case that an attacker has no prior incoming hyperlinks. This leads to conjecture that the cost of attack function is the same as the general case, when an attacker already has incoming hyperlinks.

This result not only provides evidence that PageRank is resistant to manipulation, but also suggests currency as a natural unit for describing the reputation score of a web page. Rather than stating the PageRank of www.cnn.com as the number 9 on an arbitrary log-scale, Google could instead report that the purchase of CNN's reputation score would require an investment of (say) \$US100,000 per year.

This paper is organized as follows: Section 2 discusses related work. Section 3 gives some Markov theory background. Section 4 builds on this to define PageRank. Section 5 derives the cost of attack of PageRank in the special case that attackers only receives hyperlinks from pages under their control. Section 6 conjectures that the cost of attack in this special case coincides with the cost of attack of PageRank in general.

¹Page et al. (1998) suggested allocating initial reputation to "the root level pages of all web servers."

2 Related Work

Reputation systems are also used in online auction communities such as eBay (www.ebay.com). In this situation, buyers must first find auctions that offer desired items. Secondly, after winning the auctions, the buyers must decide whether they should pay the sellers and risk receiving nothing. Resnick et al. (2000) suggested that the possibility of a recommendation combined with the threat of a complaint may provide sufficient incentive for sellers to cooperate. eBay could use PageRank to aggregate recommendations into reputation scores. However, PageRank's lack of support for complaints makes it a weak candidate if reputation is being used to reduce the risk of losing money to a dishonest seller. A dishonest vendor could play by the rules to obtain a high PageRank, and then exploit that reputation without fear of losing it.

Resnick and Friedman (2001) proved that when it is cheap to create new pseudonyms (such as new web pages), sustainable reputation systems must have entry fees for newcomers. They predicted that the internet would introduce entry costs. Google's use of domain name registrations (or IP addresses) is a clear example of this. However, PageRank permits those who have paid their entry fees to spend their reputation to recommend others.

Douceur (2002) coined the term *Sybil attack* for attacks involving stacking a network with fake identities. Attacking PageRank by purchasing many domain names is an example of a Sybil attack.

For peer-to-peer networks, Kamvar et al. (2003) proposed the EigenTrust algorithm to allow peers to decide if another peer can be trusted. EigenTrust combines each peer's past experiences with reputation to compute trustworthiness scores. The authors of EigenTrust adapted PageRank for the reputation computation. Rather than implicitly obtain recommendations from hyperlinks, each peer explicitly recommends peers it has had successful interaction with. EigenTrust has the same problem as PageRank with its lack of complaints. Freeriders can exploit high reputatoin scores without fear.

3 Markov Theory Background

While PageRank can be defined purely in terms of Linear Algebra, the theory of Markov processes is better developed for the purposes of this paper and provides richer intuition. More background material can be found in Kemeny and Snell (1976).

3.1 Markov processes

A Markov process is a sequence of random variables in which the probability distribution of a particular variable is a function of the previous variable only.

Definition 3.1. Let P be a finite set. Let $\{X_k\} = X_0, X_1, \dots$ be an infinite sequence of random variables which take on values in P . Then $\{X_k\}$ is a P -valued **Markov process** if the probability function \mathbf{P} satisfies the Markov and homogeneous assumptions:

$$\begin{aligned} \mathbf{P}(X_{k+1} = b \mid X_0 = a_0, \dots, X_k = a_k) &= \mathbf{P}(X_{k+1} = b \mid X_k = a_k) \\ \mathbf{P}(X_{k+1} = b \mid X_k = a) &= \mathbf{P}(X_1 = b \mid X_0 = a). \end{aligned}$$

Its **transition function** is $\omega(a, b) = \mathbf{P}(X_{k+1} = b \mid X_k = a)$. Its **initial distribution** is $\sigma(a) = \mathbf{P}(X_0 = a)$.

This is called a *homogeneous, discrete time, finite space Markov process* in the stochastic processes literature.

3.2 Markov Process Isomorphisms

Markov process isomorphisms give a way of comparing the structure of Markov processes independently of the labels.

Definition 3.2. Let $\{X_k\}$ be a P -valued Markov process. Let $\{Y_k\}$ be a Q -valued Markov process. A bijective function $\xi : P \rightarrow Q$ is a **Markov process isomorphism** if for all $k \in \mathbf{N}$ and all $a \in P$,

$$P(X_k = a) = P(Y_k = \xi(a)). \quad (1)$$

If such a function ξ exists, then we will write $\{X_k\} \cong \{Y_k\}$.

Preservation of initial distribution and transition function of a bijection is both a necessary and sufficient condition for Markov isomorphism:

Theorem 3.3 (Markov Isomorphism Condition). Let $\{X_k\}$ be a P -valued Markov process with transition function ω and initial distribution σ . Let $\{Y_k\}$ be a Q -valued Markov processes with transition function ψ , and initial distribution τ . Then a bijective function $\xi : P \rightarrow Q$ is a Markov process isomorphism if and only if $\sigma(a) = \tau(\xi(a))$ and $\omega(a, b) = \psi(\xi(a), \xi(b))$ for all $a, b \in P$.

3.3 Lumpable Markov processes

This section reviews how states can be lumped together to form new Markov processes.

A partition of a state space is just a carving-up of a state-space into chunks.

Definition 3.4. Let $W = \{W_0, W_1, \dots, W_n\}$. Then W is a **partition** of a Markov process' state space P if each set in W is disjoint and $\bigcup W = P$

The following defines a lumpable² Markov process as a combination of a Markov process and partitioning that gives rise to a smaller stochastic process that is Markov.

Definition 3.5. Let $\{X_k\}$ be an P -valued Markov process. Let $W = \{W_0, W_1, \dots, W_n\}$ be a partition of P . Let $\rho : P \rightarrow W$ be a function such that $\rho(w) = W_i$ when $w \in W_i$. Then $\{X_k\}$ is **lumpable** with respect to W if $\{Y_k\} = \{\rho(X_0), \rho(X_1), \dots\}$ is a W -valued Markov process.

This theorem gives the conditions for which the lumped stochastic process is a Markov process: each state in a partition (lump) must share the same probability distribution of jumping to other partitions.

Theorem 3.6 (Lumpable condition). Let $\{X_k\}$ be an P -valued Markov process with transition function ω . Let $W = \{W_0, W_1, \dots, W_n\}$ be a partition of P . Let $\rho : P \rightarrow W$ be a function such that $\rho(w) = W_i$ when $w \in W_i$. Then $\{X_k\}$ is lumpable if and only if the function $\phi : W \times W \rightarrow [0, 1]$ is well defined (unique for all $a \in W_i$)

$$\phi(W_i, W_j) = \sum_{b \in W_j} \omega(a, b) \quad \text{for all } a \in W_i$$

Moreover, if $\{X_k\}$ is lumpable then ϕ is the transition function of the Markov process $\{Y_k\} = \{\rho(X_k)\}$.

For convenience, we will use the following division notation for constructing lumped Markov processes:

Definition 3.7 (Division Notation for Lumping). Let $\{X_k\}$ be an P -valued Markov process. If $W \subset P$, then $\{Y_k\} = \{X_k\}/W$ is the P' -valued stochastic process with $Y_k = \rho(X_k)$ where:

$$P' = (P \setminus W) \cup \{W\} \quad (2)$$

$$\rho(a) = \begin{cases} a & \text{if } a \in P \setminus W \\ W & \text{if } a \in W \end{cases} \quad (3)$$

4 PageRank and the Random Surfer

This section gives some PageRank technical background. Section 4.1 gives the Random Surfer intuition behind the Markov process definition of PageRank. Section 4.2 defines PageRank formally in terms of Markov processes. Section 4.3 defines PageRank in terms of linear algebra and proves that the two definitions are equivalent. This equivalence observation leads to a simple result that is useful for analysing PageRank.

²This is called weak lumpability in the literature.

4.1 Intuition Behind PageRank

Perhaps the best measure of a web page's reputation would be the average fraction of surfing time all web surfers spend at that page. Unfortunately, this information is not available in any reliable form.

PageRank does the next best thing: it uses a stochastic model of a web surfer to estimate this proportion. PageRank's Random Surfer model supposes a random surfer that can only click on hyperlinks and jump to a random web page (perhaps selected by a random search engine).

4.2 Formal Definition of PageRank

A webgraph is a web-page-and-hyperlink network such as the world wide web. We require that every web page have an outgoing link - possibly to itself. This guarantees that the random surfer always have a link available to click on.

Definition 4.1. Let $G = (P, H)$ be a finite directed graph, where P is the set of web pages, and $H \subseteq P \times P$ the set of hyperlinks. G is **webgraph** if all pages have an outgoing hyperlink.

A Random Click process is a sequence of random variables representing the random surfer clicking on random hyperlinks around a webgraph.

Definition 4.2. A **Random Click process** of a webgraph $G = (P, H)$ is a P -valued Markov process $\{X_k\}$ such that transition function ω has $\omega(a, b) > 0$ only if $(a, b) \in H$.

Random Search-Click processes endow the random surfer with the ability to randomly search as well as randomly click. The search distribution is determined by a Random Search Function:

Definition 4.3. A **Random Search function** of a webgraph $G = (P, H)$ is a function $s : P \rightarrow [0, 1]$ with $\sum_{p \in P} s(p) = 1$

Random Search-Click processes are parameterized by a click distribution ω , a click-vs-search probability d , and a Random Search function s . The Random Search function contains the initial votes from which all reputation flows.

Definition 4.4. Let $G = (P, H)$ be a webgraph. Let s be a Random Search function of G . Let $d \in (0, 1)$ be an arbitrary constant. Let ω be the transition function of some Random Click process on G . Then $\{X_k\}$ is a (G, ω, d, s) Random Search-Click process if $\{X_k\}$ is the P -valued Markov process with initial distribution s and transition function $\psi : P \times P \rightarrow [0, 1]$:

$$\begin{aligned} P(X_0 = a) &= s(a) \\ \psi(a, b) &= ds(b) + (1 - d)\omega(a, b) \end{aligned}$$

The PageRank of a page is the limit probability of the random surfer visiting a page "at time infinity". This coincides with the proportion of time a Random Search-Click surfer spends at that page (proof omitted).

Definition 4.5. Let $\{X_k\}$ be a Random Search-Click process on a webgraph, with transition function ψ . The **PageRank** $r(a)$ of a page $a \in P$ is:

$$r(a) = \lim_{k \rightarrow \infty} P(X_k = a)$$

4.3 Algebraic PageRank

It turns out that PageRank weights recommendations by the recommender's reputation. This property is an equivalent definition of PageRank that is commonly used (for example, by Haveliwala (2002)).

Theorem 4.6 (PageRank Algebra). Let $\{X_k\}$ be a (G, ω, d, s) Random Search-Click process. Then, for all $a \in P(G)$,

$$r(a) = ds(a) + (1 - d) \sum_{b \in P} \omega(b, a)r(b).$$

Proof. From Definition 4.5 (PageRank), we have

$$\begin{aligned}
r(a) &= \lim_{k \rightarrow \infty} \mathbf{P}(X_k = a) \\
&= \lim_{k \rightarrow \infty} \sum_{b \in P} \mathbf{P}(X_k = a \mid X_{k-1} = b) \mathbf{P}(X_{k-1} = b) \\
&= \sum_{b \in P} \lim_{k \rightarrow \infty} \psi(b, a) \mathbf{P}(X_{k-1} = b) \\
&= \sum_{b \in P} \psi(b, a) r(b) \\
&= \sum_{b \in P} [ds(a) + (1-d)\omega(b, a)] r(b) \\
&= ds(a) + (1-d) \sum_{b \in P} \omega(b, a) r(b).
\end{aligned}$$

□

This leads to the following corollary that describes how PageRanks are allocated when there are exactly two webpages.

Corollary 4.7 (Two-Page PageRank). *Let $G = (P, H)$ be a two-page webgraph with $P = \{\mu, \nu\}$ only containing self-hyperlinks so that $H = \{(\mu, \mu), (\nu, \nu)\}$. Let $\{X_k\}$ be a (G, ω, d, s) Random Search-Click process. Then for $a \in P$:*

$$r(a) = s(a)$$

Proof. Assume $a = \mu$ without loss of generality. Then

$$\begin{aligned}
r(\mu) &= ds(\mu) + (1-d) \sum_{b \in P} \omega(b, \mu) r(b) \\
r(\mu) &= ds(\mu) + (1-d)[1 \cdot r(\mu) + 0 \cdot r(\nu)] \\
r(\mu) &= s(\mu).
\end{aligned}$$

□

5 Cost of Attack Theorem

Suppose a webmaster (“the attacker”) would like to modify their web pages to increase their PageRank. For example, they might create many web pages and make them link to their main web page in the hope that the “recommendations” will increase the main pages PageRank. This section proves that if the original web pages (V) have no hyperlinks to the attacker’s web pages (W), then the sum of the PageRanks of the attacker’s pages is less than or equal to the proportion of initial votes allocated to the attacker’s pages.

If an attacker has no links from the original pages, then any web structure an attacker chooses for its pages V that doesn’t link to the original pages W is optimal.

Theorem 5.1 (PageRank Cost of Attack). *Let $G = (P, H)$ be a webgraph. Let $\{X_k\}$ be a (G, ω, d, s) Random Search-Click process with transition function ψ . Let r be the PageRank function using the process $\{X_k\}$. If $\{V, W\}$ is a partition of P such that there are no hyperlinks from W to V , then:*

$$\sum_{v \in V} r(v) = \sum_{v \in V} s(v) - \Delta$$

where

$$\Delta = \frac{1-d}{d} \sum_{(v,w) \in V \times W} r(v) \omega(v, w) \geq 0.$$

Moreover, $\Delta = 0$ if there are no hyperlinks from V to W .

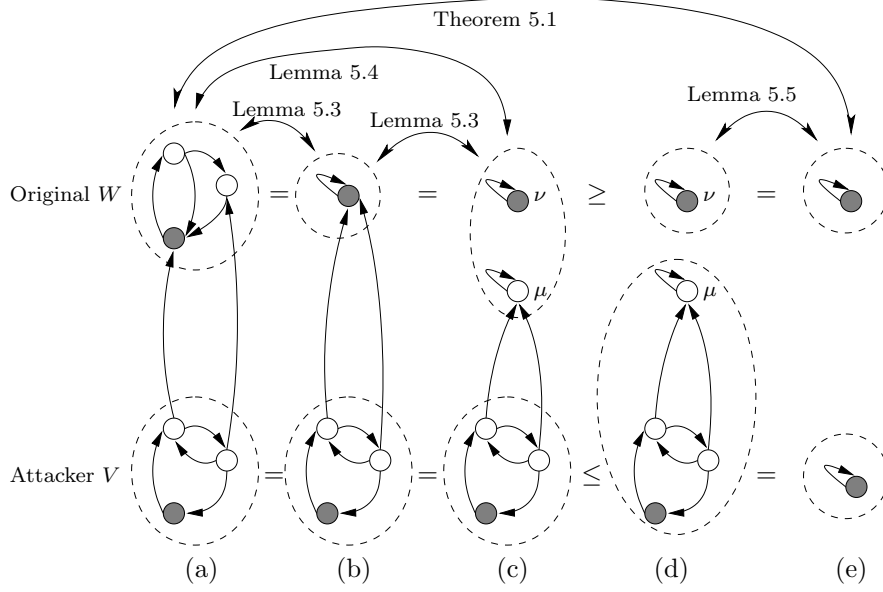


Figure 1: *Overview of the Cost of Attack Theorem.* The filled vertices have the initial votes. The theorem relates the sums of the PageRanks inside the dashed ellipses. (a) represents an attacker trying to maximize his PageRank. Each graph has a successively simpler but equivalent graph. The PageRanks in (e) can be computed with *Corollary 4.7 (Two-Page PageRank)*.

The following corollary shows how to apply the theorem, by configuring the Random Search function, s to assign weightings to web pages that are root-level in registered domain names. It states that an attacker desiring a PageRank of r with no prior incoming hyperlinks must spend at least $rc(P)$ in domain names, where $c(P)$ is the total money spent on domain names.

Corollary 5.2 (Domain-name Cost of Attack). *Let $G = (P, H)$ be a webgraph. Let $s(a) = c(a) / \sum_{p \in P} c(p)$, where $c(p)$ is the cost of registering page p . Let $\{X_k\}$ be a (G, ω, d, s) Random Search-Click process for some click distribution ω and search probability d . If $\{V, W\}$ is a partition of P such that there are no hyperlinks from W to V then:*

$$\sum_{v \in V} r(v) \leq \frac{\sum_{v \in V} c(v)}{\sum_{p \in P} c(p)}.$$

Moreover, equality occurs if there are no hyperlinks from pages in V to pages in W .

Proof. This is a trivial application of *Theorem 5.1 (PageRank Cost of Attack)*:

$$\sum_{v \in V} r(v) \leq \sum_{v \in V} s(v) = \sum_{v \in V} \frac{c(v)}{\sum_{p \in P} c(p)} = \frac{\sum_{v \in V} c(v)}{\sum_{p \in P} c(p)}$$

The theorem provides the same conditions for equality. □

The remainder of this section proves the the Cost of Attack theorem with the help of three lemmas. *Lemma 5.3 (PageRank Lumping)* describes how a set of pages can be lumped together into a single representative page in order to compute the sum of the set's PageRanks. *Lemma 5.4 (PageRank Lumping Isomorphism)* shows that the internal structure of the lumped pages in the previous lemma is irrelevant to the sum of the PageRanks, and can therefore be arranged arbitrarily. *Lemma 5.5 (PageRank Lump Sum)* is a special case of the main theorem. It gives the sum of PageRanks in a set of web pages where there are hyperlinks neither entering nor leaving the set.

Lemma 5.3 (PageRank Lumping) says that if a set of web pages W of a Random Search-Click process $\{X_k\}$ are grouped together into a single representative page, then the resulting process $\{X_k\}/W$ is a

Random Search-Click process. Moreover, the PageRank of W in $\{X_k\}/W$ is the sum of the PageRanks of the pages in W for $\{X_k\}$. All other PageRanks are unchanged.

Lemma 5.3 (PageRank Lumping). *Let $G = (P, H)$ be a webgraph. Let $\{X_k\}$ be a (G, ω, d, s) Random Search-Click process with transition function ψ . Let $\{X'_k\} = \{X_k\}/W$. Let r and r' be the PageRank functions using the processes $\{X_k\}$ and $\{X'_k\}$ respectively. If $\{V, W\}$ is a partition of P such that there are no hyperlinks from W to V then:*

- $\{X'_k\}$ is a (G', ω', d, s') Random Search-Click process, where:

$$\begin{aligned} G' &= (P', H') \\ P' &= V \cup \{W\} \\ H' &= \{(\rho(a), \rho(b)) : (a, b) \in H\} \\ \rho(a) &= \begin{cases} a & \text{if } a \in V \\ W & \text{if } a = W \end{cases} \\ \omega'(a, b) &= \begin{cases} \omega(a, b) & \text{if } a, b \in V \\ \sum_{w \in W} \omega(a, w) & \text{if } a \in V \text{ and } b = W \\ 0 & \text{if } a = W \text{ and } b \in V \\ 1 & \text{if } a = b = W \end{cases} \\ s'(a) &= \begin{cases} s(a) & \text{if } a \in V \\ \sum_{w \in W} s(w) & \text{if } a = W \end{cases} \end{aligned}$$

- $r(v) = r'(v)$ for all $v \in V$.
- $\sum_{w \in W} r(w) = r'(W)$.

Proof. To prove this, we first verify that $\{X'_k\}$ is a Markov process, and then show that its transition function matches the one given above. The other consequences relating the PageRanks in the two processes follow trivially, as $\mathbf{P}(X_k = v) = \mathbf{P}(X'_k = v)$ for all $v \in V$.

Let ψ and ψ' be the transition functions of $\{X_k\}$ and $\{X'_k\}$.

To show $\{X'_k\}$ is a Markov process, we note there are no edges from W to V in G . So $\mathbf{P}(X_{k+1} = b | X_k = a) = ds(b) + 0 = ds(b)$ for all $a \in W$ when $b \in V$. Similarly $\mathbf{P}(X_{k+1} \in W | X_k = a) = 1 - d \sum_{v \in V} s(v)$ for all $a \in W$. Therefore, the conditional probability distributions for each state in W are the same and $\{X'_k\}$ is a Markov process by *Theorem 3.6 (Lumpability Condition)*.

To show that $\{X'_k\}$ is a (G', ω', d, s') Random Search-Click process, we need to show that the initial distribution s' and transition function ψ' derived from (G', ω', d, s') match those of $\{X'_k\}$:

$$\begin{aligned} \mathbf{P}(\rho(X_0) = a) &= s'(a) \\ \mathbf{P}(\rho(X_{k+1}) = b | \rho(X_k) = a) &= \psi'(a, b) \end{aligned}$$

The initial distribution matches trivially.

There are four cases to check for verifying the transition function matches. Firstly, when $a, b \in V$,

$$\begin{aligned} \mathbf{P}(\rho(X_{k+1}) = b | \rho(X_k) = a) &= \mathbf{P}(X_{k+1} = b | X_k = a) \\ &= \psi(a, b) \\ &= ds(b) + (1 - d)\omega(a, b) \\ &= ds'(b) + (1 - d)\omega'(a, b) \\ &= \psi'(a, b). \end{aligned}$$

Secondly, when $a \in V$ and $b = W$,

$$\begin{aligned}
& \mathbf{P}(\rho(X_{k+1}) = W \mid \rho(X_k) = a) \\
&= \mathbf{P}(X_{k+1} \in W \mid X_k = a) \\
&= \sum_{w \in W} \mathbf{P}(X_{k+1} = w \mid X_k = a) \\
&= \sum_{w \in W} \psi(a, w) \\
&= \sum_{w \in W} [ds(w) + (1-d)\omega(a, w)] \\
&= ds'(W) + (1-d)\omega'(a, W) \\
&= \psi'(a, W).
\end{aligned}$$

Thirdly, when $a = W$ and $b \in V$,

$$\begin{aligned}
& \mathbf{P}(\rho(X_{k+1}) = b \mid \rho(X_k) = W) \\
&= \mathbf{P}(X_{k+1} = b \mid X_k \in W) \\
&= ds(b) + (1-d) \cdot 0 \\
&= ds'(b) + (1-d)\omega'(W, b) \\
&= \psi'(W, b).
\end{aligned}$$

Finally, when $a = b = W$,

$$\begin{aligned}
& \mathbf{P}(\rho(X_{k+1}) = W \mid \rho(X_k) = W) \\
&= \mathbf{P}(X_{k+1} \in W \mid X_k \in W) \\
&= 1 - \mathbf{P}(X_{k+1} \in V \mid X_k \in W) \\
&= 1 - \sum_{v \in V} \mathbf{P}(X_{k+1} = v \mid X_k \in W) \\
&= 1 - \sum_{v \in V} [ds(v) + (1-d) \cdot 0] \\
&= 1 - d \sum_{v \in V} s(v) \\
&= 1 - d(1 - \sum_{w \in W} s(w)) \\
&= 1 - d(1 - s'(W)) \\
&= ds'(W) + (1-d) \cdot 1 \\
&= \psi'(W, W).
\end{aligned}$$

Finally, we now need to prove that the above relationships between the PageRank functions r and r' hold.

Firstly, for $v \in V$,

$$\begin{aligned}
r(v) &= \lim_{k \rightarrow \infty} \mathbf{P}(X_k = v) \\
&= \lim_{k \rightarrow \infty} \mathbf{P}(\rho(X_k) = v) \\
&= \lim_{k \rightarrow \infty} \mathbf{P}(X'_k = v) \\
&= r'(v).
\end{aligned}$$

Secondly, for W ,

$$\begin{aligned}
\sum_{w \in W} r(w) &= \sum_{w \in W} \lim_{k \rightarrow \infty} \mathbf{P}(X_k = w) \\
&= \lim_{k \rightarrow \infty} \mathbf{P}(X_k \in W) \\
&= \lim_{k \rightarrow \infty} \mathbf{P}(\rho(X_k) = W) \\
&= \lim_{k \rightarrow \infty} \mathbf{P}(X'_k = W) \\
&= r'(W).
\end{aligned}$$

□

Lemma 5.4 (PageRank Lumping Isomorphism) says that if two Random Search-Click processes are exactly the same outside of some subset V of all web pages, then all other web pages can be shrunk down into one representative page, resulting in the same Random Search-Click process. This means isomorphic Random Search-Click processes with arbitrary structure outside of V can be constructed, while still preserving the PageRanks of all pages in V .

Lemma 5.4 (PageRank Lumping Isomorphism). *Let $G_X = (P_X, H_X)$ and $G_Y = (P_Y, H_Y)$ be web-graphs. Let $\{V, W_X\}$ and $\{V, W_Y\}$ be partitions of P_X and P_Y respectively. Let $\{X_k\}$ be a (G_X, ω_X, d, s_X) Random Search-Click process. Let $\{Y_k\}$ be a (G_Y, ω_Y, d, s_Y) Random Search-Click process.*

Then $\{X_k\}/W_X \cong \{Y_k\}/W_Y$ if the following conditions are satisfied:

- *there are no hyperlinks from W_X to V in G_X nor from W_Y to V in G_Y .*
- *$\omega_X(a, b) = \omega_Y(a, b)$ for all $a, b \in V$.*
- *$s_X(a) = s_Y(a)$ for all $a \in V$.*

Proof. To prove that $\{X_k\}/W_X \cong \{Y_k\}/W_Y$, we construct an isomorphism between the resulting Random Search-Click processes. The requirement for no hyperlinks from pages in W_X and W_Y to pages in V is needed to use *Lemma 5.3 (PageRank Lumping)* to construct these Random Search-Click processes.

Let $\{X'_k\} = \{X_k\}/W_X$. By *Lemma 5.3 (PageRank Lumping)*, $\{X'_k\}$ is a $(G_{X'}, \omega_{X'}, d, s_{X'})$ Random Search-Click process (since there are no hyperlinks from pages in W_X to pages in V), with:

$$\begin{aligned}
G_{X'} &= (P_{X'}, H_{X'}) \\
P_{X'} &= V \cup \{W_X\} \\
\omega_{X'}(a, b) &= \begin{cases} \omega_X(a, b) & \text{if } a, b \in V \\ \sum_{w \in W_X} \omega_X(a, w) & \text{if } a \in V \text{ and } b = W_X \\ 0 & \text{if } a = W_X \text{ and } b \in V \\ 1 & \text{if } a = b = W_X \end{cases} \\
s_{X'}(a) &= \begin{cases} s_X(a) & \text{if } a \in V \\ \sum_{w \in W_X} s_X(w) & \text{if } a = W_X \end{cases}
\end{aligned}$$

Likewise, $\{Y'_k\} = \{Y_k\}/W_Y$ is a $(G_{Y'}, \omega_{Y'}, d, s_{Y'})$ Random Search-Click process:

$$\begin{aligned} G_{Y'} &= (P_{Y'}, H_{Y'}) \\ P_{Y'} &= V \cup \{W_Y\} \\ \omega_{Y'}(a, b) &= \begin{cases} \omega_Y(a, b) & \text{if } a, b \in V \\ \sum_{w \in W_X} \omega_X(a, w) & \text{if } a \in V \text{ and } b = W_Y \\ 0 & \text{if } a = W_Y \text{ and } b \in V \\ 1 & \text{if } a = b = W_Y \end{cases} \\ s_{Y'}(a) &= \begin{cases} s_Y(a) & \text{if } a \in V \\ \sum_{w \in W_Y} s_Y(w) & \text{if } a = W_Y \end{cases} \end{aligned}$$

Now, consider the bijective function $\xi : P_{X'} \rightarrow P_{Y'}$:

$$\xi(a) = \begin{cases} a & \text{if } a \in V \\ W_Y & \text{if } a = W_X \end{cases}$$

To show that this is an isomorphism, we use the condition from *Theorem 3.3 (Markov Isomorphism Condition)*. This means it is sufficient to show that $s_{X'}(a) = s_{Y'}(\xi(a))$ for all $a \in P_{X'}$ and that $\omega_{X'}(a, b) = \omega_{Y'}(\xi(a), \xi(b))$ for all $a, b \in P_{X'}$.

Firstly, it is clear from the assumptions that $s_{X'}(a) = s_{Y'}(\xi(a))$ for all $a \in V$. For $s_{X'}(W_X)$, we have

$$\begin{aligned} s_{X'}(W_X) &= \sum_{w \in W_X} s_X(W_X) \\ &= 1 - \sum_{w \in V} s_X(w) \\ &= 1 - \sum_{w \in V} s_Y(w) \\ &= s_{Y'}(W_Y) \\ &= s_{Y'}(\xi(W_X)). \end{aligned}$$

Secondly, $\omega_{X'}(a, b) = \omega_{Y'}(\xi(a), \xi(b))$ for all $a, b \in V$ by the assumption that $\omega_X(a, b) = \omega_Y(a, b)$ on $a, b \in V$. The cases where $a = W_X$ are trivial. The remaining case, where $a \in V$ and $b = W_X$ follows:

$$\begin{aligned} \omega_{X'}(a, W_X) &= \sum_{w \in W_X} \omega_X(a, w) \\ &= 1 - \sum_{v \in V} \omega_X(a, v) \\ &= 1 - \sum_{v \in V} \omega_Y(a, v) \\ &= \sum_{w \in W_Y} \omega_Y(a, W_Y) \\ &= \omega_{Y'}(a, \xi(W_X)). \end{aligned}$$

Therefore, ξ is an isomorphism, and $\{X_k\}/W_X \cong \{Y_k\}/W_Y$. □

Lemma 5.5 (PageRank Lump Sum) states that if the set of web pages can be bi-partitioned such that there are no hyperlinks between pages in different partitions, then the sum of the PageRanks in a partition equals the sum of the initial votes allocated to that partition. *Fig 1d* and *Fig 1e* are examples of where this Lemma may be applied.

Lemma 5.5 (PageRank Lump Sum). Let $G = (P, H)$. Let $\{X_k\}$ be a (G, ω, d, s) Random Search-Click process. Let r be the PageRank function using the process $\{X_k\}$. If $\{V, W\}$ is a partition of P such that there are no hyperlinks between V and W (i.e. $H \cap (W \times V) = H \cap (V \times W) = \emptyset$), then:

- $\{Z_k\} = \{X_k\}/W/V$ is a (G_Z, ω_Z, d, s_Z) Random Search-Click process, where:

$$G_Z = (P_Z = \{V, W\}, H_Z = \{(V, V), (W, W)\})$$

$$\omega_Z(a, b) = \begin{cases} 1 & \text{if } a = b \\ 0 & \text{if } a \neq b \end{cases}$$

$$s_Z(A) = \sum_{a \in A} s(a)$$

- $\sum_{v \in V} r(v) = \sum_{v \in V} s(v)$ and $\sum_{w \in W} r(w) = \sum_{w \in W} s(w)$.

Proof. The first part of the proof is establishing that $\{Z_k\}$ is the above Random Search-Click process. This follows from two applications of *Lemma 5.3 (PageRank Lumping)*. Once we know the transition function of $\{Z_k\}$, it is straight-forward to compute the PageRanks of a two-state Markov process.

Let $\{Y_k\} = \{X_k\}/W$. By *Lemma 5.3 (PageRank Lumping)*, $\{Y_k\}$ is a (G_Y, ω_Y, d, s_Y) Random Search-Click process, since there are no hyperlinks from pages in W to pages in V , where:

$$G_Y = (V \cup \{W\}, H \cap (V \times V) \cup (W, W))$$

$$\omega_Y(a, b) = \begin{cases} \omega(a, b) & \text{if } a, b \in V \\ 0 & \text{if } a \in V \text{ and } b = W \\ 0 & \text{if } a = W \text{ and } b \in V \\ 1 & \text{if } a = b = W \end{cases}$$

$$s_Y = \begin{cases} s(a) & \text{if } a \in V \\ \sum_{w \in W} s(w) & \text{if } a = W \end{cases}$$

Moreover, the PageRanks are related by:

$$\sum_{v \in V} r(v) = \sum_{v \in V} r_Y(v)$$

Clearly, $\{Z_k\} = \{Y_k\}/V = \{X_k\}/W/V$. Since there are no hyperlinks between W and V , by *Lemma 5.3 (PageRank Lumping)*, $\{Z_k\}$ is a (G_Z, ω_Z, d, s_Z) Random Search-Click process (where G_Z, ω_Z and s_Z are defined above). Moreover, the PageRanks are related by:

$$\sum_{v \in V} r_Y(v) = r_Z(V)$$

Now, by *Corollary 4.7 (Two-Page PageRank)*, the PageRank of V in G_Z can be computed.

$$\sum_{v \in V} r(v) = \sum_{v \in V} r_Y(v) = r_Z(V) = s_Z(V) = \sum_{v \in V} s(v).$$

The same relationship holds for W trivially.

$$\sum_{w \in W} r(w) = 1 - \sum_{v \in V} r(v) = 1 - \sum_{v \in V} s(v) = \sum_{w \in W} s(w).$$

□

Finally, the proof of the theorem:

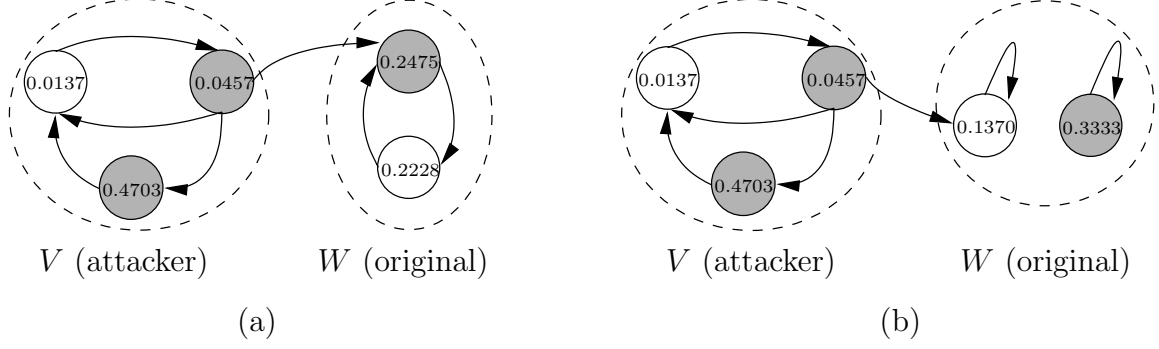


Figure 2: *PageRank examples*. The shaded nodes have the initial votes ($1/3$ each). All hyperlinks from a particular page are equally weighted. The numbers are PageRanks. (a) represents an attacker that (stupidly) links to the rest of the web. *Theorem 5.1 (PageRank Cost of Attack)* says that the sum of the attacker's pages' PageRanks (0.5160) is less than or equal to the sum of its initial votes (0.6666). (b) shows the transformation that is used in the proof.

Proof. The general case is reduced to the special case already proved in *Lemma 5.5 (PageRank Lump Sum)* using *Lemma 5.4 (PageRank Lumping Isomorphism)* to rearrange the pages in W into a convenient form that allows us to compute the amount of PageRank donated by pages in V to pages in W .

Here, we construct a Markov process $\{Y_k\}$ such that $\{Y_k\}/\{\mu, \nu\} \cong \{X_k\}/W$, but is more convenient for evaluating PageRanks. Let $\{Y_k\}$ be the (G_Y, ω_Y, d, s_Y) Random Search-Click process with:

$$\begin{aligned}
 G_Y &= (P_Y, H_Y) \\
 P_Y &= V \cup \{\mu, \nu\} \\
 H_Y &= \{(\mu, \mu), (\nu, \nu)\} \cup \{(\rho_Y(a), \rho_Y(b)) : (a, b) \in H\} \\
 \rho_Y(a) &= \begin{cases} a & \text{if } a \in V \\ \mu & \text{if } a \in W \end{cases} \\
 \omega_Y(a, b) &= \begin{cases} \omega(a, b) & \text{if } a, b \in V \\ \sum_{w \in W} \omega(a, w) & \text{if } a \in V \text{ and } b = \mu \\ 0 & \text{if } a \in \{\mu, \nu\} \text{ and } a \neq b \\ 1 & \text{if } a = b = \mu \text{ or } a = b = \nu \end{cases} \\
 s_Y(a) &= \begin{cases} s(a) & \text{if } a \in V \\ 0 & \text{if } a = \mu \\ \sum_{w \in W} s(w) & \text{if } a = \nu \end{cases}
 \end{aligned}$$

Let $\{X'_k\} = \{X_k\}/W$ and $\{Y'_k\} = \{Y_k\}/\{\mu, \nu\}$.

Since in G and G_Y , there are no edges from W to V and $\{\mu, \nu\}$ to V respectively, we have by *Lemma 5.3 (PageRank Lumping)*:

$$\begin{aligned}
 r(v) &= r_{X'}(v) & \text{for } v \in V \\
 r_Y(v) &= r_{Y'}(v) & \text{for } v \in V
 \end{aligned}$$

By *Lemma 5.4 (PageRank Lumping Isomorphism)*, $\{X'_k\} \cong \{Y'_k\}$ since $\omega(a, b) = \omega_Y(a, b)$ for all $a, b \in V$, and $s(a) = s_Y(a)$ for all $a \in V$. Therefore, all four terms in the above two equations are equal:

$$r(v) = r_{X'}(v) = r_{Y'}(v) = r_Y(v) \quad \text{for } v \in V.$$

Since there are no edges between $V \cup \{\mu\}$ and $\{\nu\}$, *Lemma 5.5 (PageRank Lump Sum)* applies.

$$\sum_{v \in (V \cup \{\mu\})} r_Y(v) = \sum_{v \in (V \cup \{\mu\})} s_Y(v) = \sum_{v \in V} s(v).$$

Putting this together, we can obtain the desired equality.

$$\begin{aligned} \sum_{v \in V} r(v) &= \sum_{v \in V} r_{X'}(v) = \sum_{v \in V} r_Y(v) \\ &= \sum_{v \in V \cup \{\mu\}} r_Y(v) - r_Y(\mu) \\ &= \sum_{v \in V} s(v) - r_Y(\mu). \end{aligned}$$

The difference, $r_Y(\mu) = \Delta$ is

$$\begin{aligned} r_Y(\mu) &= d \cdot 0 + (1-d) \sum_{p \in P_Y} r_Y(p) \omega_Y(p, \mu) \\ r_Y(\mu) &= (1-d) \sum_{v \in V} r(v) \omega_Y(v, \mu) + (1-d) r_Y(\mu) \\ r_Y(\mu) &= \frac{1-d}{d} \sum_{v \in V} \left(r(v) \sum_{w \in W} \omega(v, w) \right) \\ r_Y(\mu) &= \frac{1-d}{d} \sum_{(v, w) \in V \times W} r(v) \omega(v, w). \end{aligned}$$

□

6 Conjecture

As the theorem currently stands, when an attacker receives a hyperlink from the original web, neither an upper bound on the sum of the attacker's PageRanks, nor a lower bound on the cost of acquiring a higher PageRank are possible to derive. However, it is unlikely that a single hyperlink would radically change the bounds described in the theorem.

Firstly, it is likely that the attacker's best strategy should be to provide no outgoing hyperlinks to original pages. This has only been proved in the case that the attacker received some hyperlinks from the original web.

Secondly, if we assume the conjecture in the previous paragraph, then we can provide a tight upper bound on the amount an attacker's PageRank can be increased. We swap the sets V and W so that W is the attacker and there are no hyperlinks from W to V . Then $r_Y(\mu) = \Delta$ represents the amount of PageRank the attacker has achieved from genuine recommendations, and $r_Y(\nu)$ represents the amount of PageRank gained by buying more domain names. Clearly, $r_Y(\nu) = c(\nu)/c(P)$, so the attacker has to spend in the same way to get a PageRank increase.

7 Conclusion

PageRank appears to resist attack because all reputation stems from the initial votes. It appears that the cost of acquiring a PageRank r is $rc(P)$, where $c(P)$ is the total money spent by all websites on domain names and IP addresses. While a general proof is elusive, the special case presented here is a start.

PageRank exploits the entry costs of founding a website to make manipulation expensive. This helps put top search placements out of reach of spammers. If the conjecture were true, then cost of attack

would be a reliable measure of this barrier. Moreover, cost of attack provides a more natural unit for describing PageRank reputation scores than the log-scale system Google currently uses.

Finally, Google claims “PageRank performs an objective measurement of the importance of web pages” (www.google.com/corporate/tech.html). While possession of resources such as domain names and IP addresses is objectively measurable, PageRank can clearly be manipulated - at a price.

Acknowledgements

I would like to thank Graham Byrnes, Suelette Dreyfus, Peter Eckersley, Paul Gruba, Catherine Lai, Alistair Moffat, Liz Sonenberg, Peter Taylor and Sanming Zhou for their helpful advice with this work.

References

- Douceur, J. (2002). The Sybil attack. In *1st International Workshop on Peer-to-Peer Systems (IPTPS'02)*, pages 251 – 260. Springer Verlag.
- Haveliwala, T. (2002). Topic-sensitive PageRank. In *The Eleventh International World Wide Web Conference*, pages 406 – 414. ACM Press.
- Kamvar, S., Schlosser, M., and Garcia-Molina, H. (2003). The EigenTrust algorithm for reputation management in P2P networks. In *The Twelfth International World Wide Web Conference*, pages 640 – 651. ACM Press.
- Kemeny, J. G. and Snell, J. L. (1976). *Finite Markov Chains*. Springer-Verlag.
- Levien, R. (2003). Attack resistant trust metrics. Draft PhD Thesis, University of California at Berkeley.
- Page, L., Brin, S., Motwani, R., and Winograd, T. (1998). The PageRank citation ranking: Bringing order to the web. Technical Report 66, Stanford Digital Library Technologies Project.
- Resnick, P. and Friedman, E. (2001). The social cost of cheap pseudonyms. *Journal of Economics and Management Strategy*, 10(2):173 – 199.
- Resnick, P., Zeckhauser, R., Friedman, E., and Kuwabara, K. (2000). Reputation systems. *Communications of the ACM*, 43(12):45 – 58.