# How much does it cost to buy a good Google PageRank?

Andrew Clausen

October 22, 2003

**Abstract**

The PageRank algorithm is used by the popular Google search engine to rate website reputation. It is one of the key ingredients that affects placement in search results coveted by webmasters.

I prove that attempts to manipulate a site's PageRank by creating "link farms" are futile, and that PageRanks can only be increased by either recommendations from reputable websites or by buying domain names. I derive a tight lower bound on the cost of the latter technique. I use these theoretical results to estimate what this lower bound means in $US at the time of writing.

I also discovered a mistake in the definition of PageRank in the original paper. This is discussed in the appendix.

## 1 Introduction

The PageRank algorithm [9] is at the heart of the popular Google[1] search engine. Websites that are deemed by the algorithm to have good reputations appear more frequently and more prominently in search results. Therefore, high PageRanks are highly coveted by webmasters seeking to attract visitors.

Normally, webpages earn a high PageRank through recommendations in the form of hyperlinks of other website authors. But since high PageRanks are so valuable, webmasters also attempt to increase their PageRank by creating many 'false' recommendations (sometimes called *link farms*). However, deep inside Google's PageRank algorithm lies a *one domain name, one vote* policy. Since domain names cost money, it has been conjectured[7] that this deceptive strategy of acquiring higher PageRank costs a proportionate amount of money, and thus is unlikely to be cost effective for the deceiver.

I prove this conjecture to be true, defining a tight lower bound on the amount of money that must be spent to acquire a desired PageRank in this way. I also provide an estimate of the cost at the time of writing of acquiring different PageRanks, giving guidance to how resistant PageRank is to this attack.

In *Section 2*, I discuss the commercial relationship between Google and PageRank. In *Section 3*, I explain the intuition behind PageRank's attack resistance properties. In *Section 4*, I present my proof of these properties, as well as the relevant Markov process and PageRank background. Readers without a mathematics background may skip this section. In *Section 5*, I discuss some of the implications to applications outside of search engines.

In *Appendix A*, I clear up some of the confusion about the definition of PageRank. In particular, the original paper [3] by the inventors of PageRank gives a self-inconsistent definition that has been propagated through much of the literature.

## 2 PageRank and Google

PageRank was developed along with the Google search engine as part of a research project before it was commercialized. Google hasn't published any details of how its current search engine works, except to say that PageRank still plays an important role in reporting search results. Therefore, the discussion in this

[1] `http://www.google.com`

paper may not match what Google is doing. However, if Google chose to operate under the assumptions listed here, then my conclusions would apply.

SearchKing [2] is a company that specializes in boosting its customers' Google's PageRank scores. It unsuccessfully sued Google for modifying its usage of the PageRank algorithm that resulted in making SearchKing's service more expensive to provide. [3] It is safe to assume that Google attempts to maximize the attack resistance of PageRank.

# 3 Attack Resistance of PageRank

The PageRank algorithm can be described in terms of a random surfer. A random surfer might start at his home page and click on random web page links. Occasionally he might get bored, and go back to his web page and start again. The PageRank of a page is the probability that he arrives at that page after $n$ clicks, where $n$ is a large number. A PageRank of zero means the page has no reputation at all. All PageRanks sum up to 1, since after $n$ clicks, you must be on exactly one web page.

Clearly, the PageRank score depends on what the user's home page is. For the purpose of Google's search engine, it is believed that the home page is configured to be an imaginary search engine that randomly displays pages that are top-level on their domain. (eg: `http://www.unimelb.edu.au/index.html` but not `http://www.ms.unimelb.edu.au/index.html` and not `http://www.unimelb.edu.au/article_id1.html`). I will mostly focus on this domain name configuration of PageRank, although these and other cases are discussed in [7], [9] and [2].

The set of web pages available on the imaginary search engine is of crucial importance. A web page can only be reached by a random surfer, and hence can only have a nonzero PageRank if it can be reached from one of the web pages in the set. These pages are granted the votes that decide how much webmasters' recommendations count for.

These votes cost money. At the time of writing, top level domain names cost about $US10 per year.[4] However, diferrent types of domains such as `.co.uk` and `.com.au` often have different prices. Fortunately, it is possible to find out which company provided a particular domain,[5] and to estimate how much was paid for a domain. Therefore, votes can be weighted by how much they cost the webmaster.

In the following section, I show that if a webmaster, Eve buys some domain names, but does not receive any hyperlinks from domains she doesn't own, then the sum of her PageRanks will be the amount paid by her divided by the total amount spent by all owners of domain names. This means we can calculate how much Eve must pay to buy a good PageRank reputation (rather than earn it).

According to Netcraft,[6] there are roughly 45 000 000 domain names on the internet. If I assume the total amount paid for these was $US 450 000 000 per year, then $US1 buys you $2.2 \times 10^{-9}$ worth of PageRank per year. PageRanks are commonly written as a number between 1 and 10, a scale visible from the Google Toolbar. It is widely conjectured[7] that this is on a logarithmic scale with a base of about 10. Therefore, the annual cost for each PageRank is roughly:

| $11 + \log_{10}$ PageRank | Cost ($US / year) |
|---|---|
| 1 | 0.045 |
| 3 | 4.50 |
| 5 | 450 |
| 7 | 45000 |
| 9 | 4500000 |

[2] `http://www.searchking.com`

[3] Stefanie Olsen, "Judge Dismisses Suit Against Google", *CNET News*, 30 May 2003, `http://news.com.com/2100-1032_3-1011740.html`

[4] For example, `http://www.active-domain.com/` offers domain names for $US8.50 per year

[5] Using the `whois` program

[6] `http://www.netcraft.com`

[7] See for example `http://www.webmasterworld.com/forum3/4595.htm`

# 4 Mathematical analysis

In this section, I show that the sum of the PageRanks of an island of webpages on the web is equal to the proportion of money spent by the webmaster of the island versus that spent by everyone on the web.

PageRank can be defined in terms of the Random Surfer model, which can be described by Markov Theory. So, I will now begin with some definitions and standard results from Markov Theory. Then I will define PageRank, and proceed to prove my claims.

Apart from *Theorem 4.1.5 (Ergodic Convergence)*, my presentation is self-contained.

## 4.1 Markov Theory

This section introduces Markov theory in order to fix notation and be mostly self-contained.

### 4.1.1 Markov processes

The following definition defines a Markov process, the stochastic model used to model random surfers. In essence, the probability distribution of the next state (or next web page to visit) only depends on the current state (or current web page being visited). That is, the states prior to the current state are irrelevant (the web pages visited prior to the current web pages being visited are irrelevant).

**Definition 4.1.1 (Markov process).** *An $S$-valued **Markov process** is an infinite sequence of random variables $\{X_k\} = X_0, X_1, \cdots \in S$ whose probability function $\boldsymbol{P}$ satisfies:*

$$\boldsymbol{P}(X_{k+1} = j \,|\, X_0 = i_0, \cdots, X_k = i_k) \;= \boldsymbol{P}(X_{k+1} = j \,|\, X_k = i_k) \qquad \text{for all } k \geq 0$$

*Its **transition function** is $\omega(i, j) = \boldsymbol{P}(X_{k+1} = j \,|\, X_k = i)$.*

In this paper, I will assume that the state space $S$ is finite.

### 4.1.2 Convergence of Markov processes

In this section, I review the conditions under which $\lim_{k\to\infty} \mathbf{P}(X_k = a)$ converges.

Most of the Markov processes I will be dealing with have a nice property called *ergodicity*. To define this, I need to define the *period of a state* first. Intuitively, if the only way from a state back to itself is through a cycle, then that state is periodic. If every state has the same period, then everything moves 'in sync', affecting its convergence properties.

**Definition 4.1.2 (Period of a state).** *Let $\{X_k\}$ be an $S$-valued Markov process. The **period** of a state $s \in S$ is the largest $p$ satisfying: (for all $k, n \in \boldsymbol{N}$)*

$$\boldsymbol{P}(X_{k+n} = s \,|\, X_k = s) > 0 \implies p \text{ divides } n$$

*If $p = 1$, then the state $s$ is **aperiodic**.*

For example, in a directed 2-cycle with transitions made with probability 1, both states have a period of 2.

**Definition 4.1.3 (Ergodic Markov process).** *An **ergodic** Markov process is a Markov process $\{X_k\}$ that is both:*

- ***irreducible**: every state is reachable from every other state.*
- ***aperiodic**: the greatest common divisor of the states' periods is 1.*

This following lemma gives us a simpler condition for ergodicity than verifying aperiodicity directly.

**Lemma 4.1.4 (Ergodic Condition).** *An irreducible $S$-valued Markov process with transition function $\omega$ that has $\omega(a,a) > 0$ for some state $a \in S$ is aperiodic, and hence ergodic.*

*Proof.* This is a standard result. Firstly, $\omega(a,a) = \mathbf{P}(X_{k+1} = a \,|\, X_k = a) > 0$. So, the period of $a$ must divide 1. Therefore, the period of $a$ is 1 and $a$ is aperiodic. The gcd of 1 and any other set of numbers is 1, so the Markov process is aperiodic, and hence ergodic. □

The following is a major theorem from Markov theory. A lot is known about how and to what this converges, but I won't need to draw on this vast knowledge.

**Theorem 4.1.5 (Ergodic Convergence).** *If $\{X_k\}$ is an ergodic $S$-valued Markov process, then the probability function converges for all $a \in S$:*

$$\lim_{k\to\infty} \boldsymbol{P}(X_k = a) = p$$

*Proof.* This proof is quite involved. It is the main topic of part I of [1], for example. □

This following lemma allows us to understand two-state Markov processes (or if you prefer, a random surfer on a world wide web consisting of two web pages).

**Corollary 4.1.6 (Two-state Convergence).** *Let the random variables $\{X_k\}$ be a Markov process with states $S = \{a, b\}$ and a transition function $\omega$ such that $\omega(a,b), \omega(b,a) > 0$ and either $\omega(a,a) > 0$ or $\omega(b,b) > 0$, then:*

$$\lim_{k\to\infty} \boldsymbol{P}(X_k = a) \;= \frac{\omega(b,a)}{\omega(a,b) + \omega(b,a)}$$

*Proof.* This is a standard result, and is a running example in [8].

By *Lemma 4.1.4 (Ergodic Condition)*, $\{X_k\}$ is an ergodic Markov process. By *Theorem 4.1.5 (Ergodic Convergence)*, we know that $\lim_{k\to\infty} \mathbf{P}(X_k = a)$ converges.

Now, I only need to deduce what $\lim_{k\to\infty} \mathbf{P}(X_k = a)$ converges to.

From the *Definition 4.1.1 (Markov Process)*, we can write $\mathbf{P}(X_{k+1} = a)$ in terms of $\mathbf{P}(X_k = a)$:

$$\begin{aligned}
\mathbf{P}(X_{k+1} = a) \;&= \mathbf{P}(X_k = a) \cdot \omega(a,a) \;+ \mathbf{P}(X_k = b) \cdot \omega(b,a) \\
&= \mathbf{P}(X_k = a) \cdot (1 - \omega(a,b)) \;+ (1 - \mathbf{P}(X_k = a)) \cdot \omega(b,a) \\
&= \mathbf{P}(X_k = a) \cdot (1 - \omega(a,b) - \omega(b,a)) \;+ \omega(b,a)
\end{aligned}$$

We can rewrite this and find the limits to obtain the result:

$$\begin{aligned}
\mathbf{P}(X_{k+1} = a) - \mathbf{P}(X_k = a) \;+ \mathbf{P}(X_k = a)(\omega(a,b) + \omega(b,a)) \;&= \omega(b,a) \\
\lim_{k\to\infty} \mathbf{P}(X_{k+1} = a) \;-\; \lim_{k\to\infty} \mathbf{P}(X_k = a) \;+\; \lim_{k\to\infty} \mathbf{P}(X_k = a)(\omega(a,b) + \omega(b,a)) \;&= \lim_{k\to\infty} \omega(b,a) \\
\lim_{k\to\infty} \mathbf{P}(X_k = a)(\omega(a,b) + \omega(b,a)) \;&= \omega(b,a) \\
\lim_{k\to\infty} \mathbf{P}(X_k = a) \;&= \frac{\omega(b,a)}{\omega(a,b) + \omega(b,a)}
\end{aligned}$$

□

### 4.1.3 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 4.1.7 (Partition).** *A **partition** $W = \{W_0, W_1, \ldots, W_n\}$ of a Markov process' state space $S$ is a disjoint set of non-empty subsets of $W$ such that*

$$\cup_{i=0}^{n} W_i = S$$

The following defines a lumpable Markov process - that is a combination of a Markov process and partitioning that give rise to a smaller stochastic process that is Markov. Later, I will use this to consider the lumped Markov process in which the entire world wide web is partitioned into two sets: pages belonging to the attacker and the rest of the web. Each random variable in this lumped process is the partition in which the page the Random Surfer is visiting. The lumped Markov process allows us to compute the sum of all PageRanks in each partition.

Note that this is called *weak lumpability* in the literature, because the term *lumpability* is reserved for classes of Markov processes that share the same transition function but differeing starting distributions, which I don't consider in this paper.

**Definition 4.1.8 (Weakly lumpable Markov process).** *Let $\{X_k\}$ be an $S$-valued Markov process. Let $W = \{W_0, W_1, \ldots, W_n\}$ be a partition of $S$. Let $\rho : S \to W$ be a function such that $\rho(s) = W_i$ when $s \in W_i$ Then $\{X_k\}$ is **weakly lumpable** with respect to $W$ if $\{Y_k\} = \{\rho(X_0), \rho(X_1), \ldots\}$ is a 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 4.1.9 (Weakly lumpable condition).** *Let $\{X_k\}$ be an $S$-valued Markov process with transition function $\omega$. Let $W = \{W_0, W_1, \ldots, W_n\}$ be a partition of $S$. Let $\rho : S \to W$ be a function such that $\rho(s) = W_i$ when $s \in W_i$.*

*$\{X_k\}$ is weakly lumpable if and only if the function $\phi : W \times W \to [0, 1]$ is well defined (unique):*

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

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

*Proof.* This is a standard result which appears in [6]. I will only show that it is sufficient here.

The general idea is, that if the transition functions are uniform with respect to partitions, then knowing exactly which state in a partition you begin from doesn't give you any extra information, and the Markov property is therefore inherited.

Assume that $\phi$ is well-defined. Then:

$$\begin{aligned}
&\mathbf{P}(\rho(X_{k+1}) = W_{i_{k+1}} \,|\, \rho(X_k) = W_{i_k}) \\
&= \mathbf{P}(X_{k+1} \in W_{i_{k+1}} \,|\, X_k \in W_{i_k}) \\
&= \sum_{b \in W_{i_{k+1}}} \mathbf{P}(X_{k+1} = b \,|\, X_k \in W_{i_k}) \\
&= \phi(\rho(a_k), W_{i_k}) && \text{for all } a_k \in W_{i_k} \\
&= \mathbf{P}(X_{k+1} \in W_{i_{k+1}} \,|\, X_k = a_k) && \text{for all } a_k \in W_{i_k} \\
&= \mathbf{P}(X_{k+1} \in W_{i_{k+1}} \mid X_0 = a_0, \cdots, X_k = a_k) && \text{for all } a_k \in W_{i_k} \\
&= \mathbf{P}(X_{k+1} \in W_{i_{k+1}} \mid X_0 \in W_{i_0}, \cdots, X_k \in W_{i_k}) \\
&= \mathbf{P}(\rho(X_{k+1}) = W_{i_{k+1}} \mid \rho(X_0) = W_{i_0}, \cdots, \rho(X_k) = W_{i_k})
\end{aligned}$$

This final equality is exactly the definition of Markov processes. So $\{\rho(X_k)\}$ is a Markov process, and $\{X_k\}$ is lumpable with respect to $\{W_0, \ldots, W_j\}$. Furthermore, $\phi$ is the transition function. □

### 4.1.4 Summary

This concludes the material I will draw on from Markov theory. In summary, a Markov process is a stochastic process where the probability distribution of the next state is only dependent on the current state. Well-behaved ergodic Markov processes have their probabilities converge to an equilibrium. We can easily compute this equilibrium for two-state Markov processes. Moreover, under certain conditions, Markov processes can be collapsed by lumping states together.

## 4.2 PageRank and the Random Surfer Model

In this section, I will define PageRank in terms of a Random Surfer model.

I will first define a simple *Random Click* model. This will lack ergodicity and doesn't provide any mechanism for seeding the reputation flow. I will then define an extension, the *Random Search-Click* that addresses both of these problems. I will define the PageRank of a webpage to be the equilibrium probability of a random surfer following this second model. I will show that this definition of PageRank is well-defined. Note that there is some confusion about the definition of PageRank; see *Appendix A* for a discussion of this issue.

First I will define a webgraph, a model of the structure of the world wide web:

**Definition 4.2.1 (Webgraph).** *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 (possibly to themselves).*

I will now define the *Random Click* process. The random variables $\{X_k\}$ refer to the webpages (elements of $P$) that the random surfer visits at each timestep. The surfer either clicks on a random hyperlink, or does nothing at each time step. There are no other possibilities (such as jumping to a home page or search engine). Since the surfer must choose one of these possibilities, a web page that has no outgoing hyperlinks will force the surfer to "choose" to do nothing.

The only assumption I make about the probability distribution of clicking on a particular link is that there is a non-zero chance of following any link of a page the surfer is visiting.

**Definition 4.2.2 (Random Click process).** *A **Random Click process** of a webgraph $G = (P, H)$ is a $P$-valued Markov process $\{X_k\}$ such that transition function $\omega$ has $\omega(a, b) > 0$ if [8] and only if $(a, b) \in H$.*

In general, Random Click processes are not ergodic.

I will now define a random search engine, that randomly returns a webpage according to some probability distribution. I will incorporate this into my *Random Click* processes to form *Random Search-Click* processes which are almost ergodic. This function allows us to place *a priori* value judgements on web pages.

**Definition 4.2.3 (Random Search function).** *A **Random Search function** of a webgraph $G = (P, H)$ is a function $s : P \to [0, 1]$ with:*

$$s(P) := \sum_{p \in P} s(p) = 1$$

One example of a Random Search function is a uniform distribution:

$$s(p) = \frac{1}{|P|}$$

Now, I will define the Random Search-Click model. A Random Search-click process is constructed from a a Random Click process and a Random Search function. At each time step, the random surfer chooses to

[8] This definition has a tight "if and only if" to simplify the presentation so "connectivity in $G$" can be used synonymously with "reachable with nonzero probability". In practice, this can be relaxed to "only if" by removing each edge $ab$ from $G$ where $\omega(a, b) = 0$.

use the random search engine with (constant) probability $d$, or to folow a hyperlink (as in Random Click processes) with probability $1-d$. [9] suggests 0.15 as an appropriate value for $d$. Since the surfer's choice of click vs search is geometrically distributed, this value gives the random surfer an average of $\frac{1}{0.15} - 1 \approx 6$ clicks before doing a random search.

This addition of a random search function prevents the random search function from ever getting stuck in a dead end or closed loop. It also allows us to weight web pages as the start of the flow of reputation. I will also show that it makes Random Search-Click processes close enough to being ergodic that an equilibrium distribution exists.

Again, the random variables $\{Y_k\}$ refer to the webpages (elements of $P$) that the random surfer visits at each timestep. The random search function $s$ represents the probabilities of the random search engine described earlier.

**Definition 4.2.4 (Random Search-Click process).** *Let $\{X_k\}$ be a Random Click process of a webgraph $G = (P, H)$. Let $s$ be a Random Search function of $G$. Let $d \in (0,1)$ be an arbitrary constant. Then the **Random Search-Click process** of $s$, $d$ and $\{X_k\}$ is the $P$-valued Markov process $\{Y_k\}$ with initial distribution $s$ and transition function $\psi : P \times P \to [0,1]$:*

$$\begin{aligned} \boldsymbol{P}(Y_0 = a) &= s(a) \\ \psi(a,b) &= ds(b) + (1-d)\omega(a,b) \end{aligned}$$

I am now ready to define PageRank of a page as the equilibrium probability of that page in a particular *Random Search-Click* process:

**Definition 4.2.5 (PageRank).** *Let $\{X_k\}$ be a Random Click process of a webgraph $G$. Let $\{Y_k\}$ be a Random Search-Click process of the Random Click process $\{X_k\}$. The **PageRank** $r(a)$ of a page $a$ is:*

$$r(a) = \lim_{k\to\infty} \boldsymbol{P}(Y_k = a)$$

The parameters for constructing this Random Search-Click process underlying PageRank are the webgraph $G$, the Random Search function $s$, and the search probability $d$.

I need to show that every page has a well-defined PageRank value. To begin, I show that a webpage has a zero pagerank if and only if it is unreachable from a webpage in the random search engine:

**Lemma 4.2.6 (Zero Condition).** *Let $\{X_k\}$ be a Random Click process of a graph $G = (P, H)$. Let $\{Y_k\}$ be a Random Search-Click process of the Random Click process $\{X_k\}$. The PageRank of a page $a$ has $r(a) = 0$ if and only if there is no $ba$-path (in $G$) from some $b \in P$ with $s(b) > 0$.*

*Proof.* Firstly, assume that $r(a) = 0$ (that is, that it converges and it converges to this particular value). Now assume for the sake of contradiction that there exists some $ba$-path with pages $p_0 = b, p_1, \ldots, p_l = a$ having $s(b) > 0$. Clearly, for all $k$, $\mathbf{P}(X_k = b) \geq s(b)$. So:

$$\begin{aligned} \mathbf{P}(X_{k+l} = a) &= \mathbf{P}(X_k = b) \cdot \mathrm{P}(X_{k+l} = a \,|\, X_k = b) \\ &= \mathbf{P}(X_k = b) \cdot \prod_{i=0}^{l} \mathrm{P}(X_{k+i+1} = p_{i+1} \,|\, X_{k+i} = p_i) \\ &= \mathbf{P}(X_k = b) \cdot c && \text{where } c > 0 \text{ is a constant independent of } k \\ &\geq s(b) \cdot c \\ &> 0 \end{aligned}$$

Now, consider the value of $r(a)$:

$$\begin{aligned} r(a) &= \lim_{k\to\infty} \mathbf{P}(X_k = a) \\ &= \lim_{k\to\infty} \mathbf{P}(X_{k+l} = a) \\ &\geq s(b) \cdot c \\ &> 0 \end{aligned}$$

But I assumed that $r(a) = 0$, so the assumption that there exists such a $ba$-path is false, and no such path exists.

Conversely, assume that no $ba$-path exists for any $b$ with $s(b) > 0$. Then $\mathbf{P}(X_k = a) = 0$ for every $k$. (If this were not the case, then $X_0, \ldots X_k$ would contain such a $ba$-path). It follows, that $r(a) = \lim_{k\to\infty} \mathbf{P}(X_k = a) = 0$ as required. □

I will now show that this is well defined. This is well known, but a proof hasn't appeared in this general form.

**Theorem 4.2.7 (PageRank well-defined).** *Let $\{X_k\}$ be a Random Click process of a graph $G$. Let $\{Y_k\}$ be a Random Search-Click process of the Random Click process $\{X_k\}$. The PageRanks $r(a)$ for all web pages $a \in P$ are well defined.*

*Proof.* To show that $r(a)$ is well defined, I need to show that $r(a)$ exists and is unique. I will show that after removing all web pages that are unreachable from a chain of clicks beginning at the random search engine yields an ergodic Markov process. Then, I will be able to invoke *Theorem 4.1.5 (Ergodic Convergence)* to show that $r$ exists and is unique.

Let $P' = \{a \in P \,|\, \mathbf{P}(Y_k = a) > 0 \text{ for some } k\}$, the reachable pages in $P$. Then $\{Y_k\}$ is a $P'$-valued Markov process. Clearly, $\{Y_k\}$ is irreducible as a $P'$-valued Markov process, because every state is reachable from every other state via some state $b$ with $s(b) > 0$.

Also, there exists a page $a \in P'$ with $s(a) > 0$, and hence $\omega(a, a) > 0$. Again, by *Lemma 4.1.4*, $Y_0$, $Y_1$, ... is ergodic as a $P'$-valued Markov process.

By *Theorem 4.1.5 (Ergodic Convergence)*, a unique $r(a)$ value exists for every page $a \in P'$. By *Lemma 4.2.6 (Zero condition)*, $r(a) = 0$ for every page in $P \backslash P'$. □

This completes my definition of PageRank in terms of Random Surfer models.

## 4.3 Attack Resistance of PageRank

In this section, I prove that the sum of the PageRanks in an island of webpages equals the probability that the random search function directs a random surfer to a webpage in that island. This means that if the random search function is weighted by money paid, then the cost of acquiring a particular PageRank $x$ is at least $x$ times the total amount of money spent on domain names.

To prove this claim, I first show how to collapse a *Random Search-Click* process into a Markov process with exactly two states. Each state represents an island (one of an "attacker", the other the rest of the internet). Then, I show that my construction allows us to compute the sums of the PageRanks in each island, and obtain my result.

This following lemma shows how we can compress the entire world wide web into two representative web pages:

**Lemma 4.3.1 (Collapse Lemma).** *Let $\{X_k\}$ be a Random Click process of webgraph $G = (P, H)$. Let $\{Y_k\}$ be the Random Search-Click process of $\{X_k\}$ and some search probability $d$ and some search function $s$. Let $W_1$ and $W_2 = P \backslash W_1$ be a partitioning of $P$, and the function $\rho : P \to \{W_1, W_2\}$ be $\rho(x) = W_i : x \in W_i$. Let $\{Z_k\}$ be the stochastic process $Z_k = \rho(Y_k)$. If there are no edges between $W_1$ and $W_2$, then $\{Z_k\}$ is a Markov process. Furthermore, it has transition function $\phi : \{W_1, W_2\} \times \{W_1, W_2\} \to [0, 1]$:*

$$\begin{aligned} \phi(W_1, W_1) &= 1 - ds(W_2) & \phi(W_1, W_2) &= ds(W_2) \\ \phi(W_2, W_1) &= ds(W_1) & \phi(W_2, W_2) &= 1 - ds(W_1) \end{aligned}$$

*Proof.* We will first show that $\{Z_k\}$ has conditional probabilities matching the transition function $\phi$. We will then observe that these don't change if extra history information is added, and deduce that $\{Z_k\}$ is a Markov process.

Recall from *Definition 4.2.4 (Random Search-Click process)* that the transition probability from a page $a$ to $b$ is:

$$\mathbf{P}(Y_{k+1} = b \,|\, Y_k = a) \;= \psi(a,b) = ds(b) + (1-d)\omega(a,b)$$

Thus:

$$\mathbf{P}(Y_{k+1} \in W_1 \,|\, Y_k = a) \;= \sum_{b \in W_1} \psi(a,b) \;= \sum_{b \in W_1} [ds(b) + (1-d)\omega(a,b)]$$

To compute $\phi(W_2, W_1)$, we are interested in the case where $a \in W_2$ and $b \in W_1$. In this case, $\omega(a,b) = 0$, since there are no edges between $W_1$ and $W_2$. Thus, the second term vanishes when $a \in W_2$:

$$\begin{aligned}
\mathbf{P}(Y_{k+1} \in W_1 \,|\, Y_k = a) \;&= \sum_{b \in W_1} ds(b) = ds(W_1) && \text{for all } a \in W_2 \\
\mathbf{P}(Y_{k+1} \in W_1 \,|\, Y_k \in W_2) \;&= ds(W_1) \\
\mathbf{P}(Z_{k+1} = W_1 \,|\, Z_k = W_2) \;&= ds(W_1)
\end{aligned}$$

The remaining conditional probabilities follow from symmetry and elementary probability. We can obtain the probability of the complimentary event:

$$\mathbf{P}(Z_{k+1} = W_2 \,|\, Z_k = W_2) \;= 1 - \mathbf{P}(Z_{k+1} = W_1 \,|\, Z_k = W_2) \;= 1 - ds(W_1)$$

By symmetry, we also have $\mathbf{P}(Z_{k+1} = W_2 \,|\, Z_k = W_1) = ds(W_2)$ and $\mathbf{P}(Z_{k+1} = W_1 \,|\, Z_k = W_1) = 1 - ds(W_2)$.

So, I have now shown that the conditional probabilities match the transition function $\phi$. I now show that $\{Z_k\}$ is a Markov process. Informally, this means that after the value of $Z_k$ is known, additional knowledge of $Z_0, \cdots, Z_{k-1}$ make no difference to the probability distribution of $Z_{k+1}$.

From the above results, it is clear that:

$$\mathbf{P}(Y_{k+1} \in W_j \,|\, Y_k = a) = \phi(\rho(a), W_j) \qquad \text{for all } a$$

By *Theorem 4.1.9 (Weakly Lumpable Condition)*, $\{Z_k\}$ is a Markov process with transition function $\phi$. □

I am now up to the final theorem. I have described the behaviour of two-state Markov processes in *Lemma 4.1.6 (Two-state Convergence)*, and have shown how two-page webgraphs can be related to large webgraphs that are composed of two (or more) islands in *Lemma 4.3.1 (Collapse Lemma)*. I use these two lemmas to show that the sum of the PageRanks in an island $W$ equals the sum of the weights of the random search function in $W$.

**Theorem 4.3.2 (Cost of attack).** *Let $\{X_k\}$ be a Random Click process of webgraph $G = (P, H)$. Let $\{Y_k\}$ be the Random Search-Click process of $\{X_k\}$ and some search probability $d$ and some search function $s$. Let $W_1$ and $W_2 = P \backslash W_1$ be a partitioning of $P$. If that there are no edges between $W_1$ and $W_2$, then $r(W_i) = s(W_i)$.*

*Proof.* The general idea is to collapse the entire web into two representative web pages. We need to show that the PageRanks of these two representative pages are the sums of the PageRanks of their constituents, and that this sum coincides with the sum of the random search function.

Without loss of generality, I will show the result for $i = 1$.

We will first consider the cases $s(W_1) = 0$ and $s(W_2) = 0$, and then the remaining case, $s(W_1), s(W_2) \neq 0$.

If $s(W_1) = 0$, then all pages $a \in W_1$ are unreachable from the start distribution, and all of its pageranks are zero (by *Lemma 4.2.6 (Zero Condition)*), and $r(W_1) = 0$ as required. Similarly if $s(W_2) = 0$, then $r(W_2) = 0$ (from the previous sentence) and $s(W_1) = 1$. It follows that $r(W_1) = 1 - r(W_2) = 1 = s(W_1)$, as required.

Otherwise, assume $s(W_1), s(W_2) > 0$. From *Lemma 4.3.1 (Collapse Lemma)*, the two-state stochastic process $\{Z_k\}$ defined in terms of $\{Y_k\}$ as $Z_k = W_i : Y_k \in W_i$ is a Markov process, with transition function $\phi$:

$$\begin{aligned} \phi(W_1, W_1) &= 1 - ds(W_2) & \phi(W_1, W_2) &= ds(W_2) \\ \phi(W_2, W_1) &= ds(W_1) & \phi(W_2, W_2) &= 1 - ds(W_1) \end{aligned}$$

This Markov process $\{Z_k\}$ allows us to compute the sum of PageRanks:

$$\begin{aligned} \sum_{a \in W_1} r(a) &= \sum_{a \in W_1} \lim_{k \to \infty} \mathbf{P}(Y_k = a) \\ &= \lim_{k \to \infty} \mathbf{P}(Y_k \in W_1) \\ &= \lim_{k \to \infty} \mathbf{P}(Z_k = W_1) \end{aligned}$$

Clearly, $\phi > 0$ since $s(W_1), s(W_2) > 0$. So, I can apply *Lemma 4.1.6 (Two-state Convergence)* on $\{Z_k\}$ to compute the limits (equilibrium distribution) of this two state Markov process:

$$\begin{aligned} \lim_{k \to \infty} \mathbf{P}(Z_k = W_1) &= \frac{\phi(W_2, W_1)}{\phi(W_1, W_2) + \phi(W_2, W_1)} \\ &= \frac{ds(W_1)}{ds(W_2) + ds(W_1)} \\ &= \frac{s(W_1)}{s(P \backslash W_1) + s(W_1)} \\ &= s(W_1) \end{aligned}$$

Thus, I have obtained the required result. In summary:

$$r(W_1) = \lim_{k \to \infty} \mathbf{P}(Y_k = W_1) = s(W_1)$$

□

So, I have now shown that the sum of the pageranks in an island $W_1$ is $s(W_1)$. I will now describe how to choose the the random search function $s$ so that this expression translates to cost to attacker divided by cost paid by everyone.

**Corollary 4.3.3 (Currency cost of attack).** *Let $\{X_k\}$ be a Random Click process of webgraph $G = (P, H)$. Let $s(a) = \frac{c(a)}{c(P)}$, where $c(a)$ is the cost of registering page $a$, and $c(P)$ is the total cost of registering all web pages. Let $\{Y_k\}$ be the Random Search-Click process of $\{X_k\}$, $s$, and some search probability $d$. If $W_1$ and $W_2 = P \backslash W_1$ be a partitioning of $P$ such that there are no edges between $W_1$ and $W_2$, then:*

$$r(W_i) = \frac{c(W_1)}{c(V)}$$

*Proof.* This is a trivial application of *Theorem 4.3.2 (Cost of Attack)*:

$$r(W_i) = s(W_i) = \sum_{a \in W_i} \frac{c(a)}{c(V)} = \frac{c(W_i)}{c(V)}$$

□

That is, the sum of the pageranks in $W_1$ exactly equals the cost paid for all web pages in $W_1$, divided by the total amount paid for all web pages on the internet, $c(V)$. Recall that only assumptions I make are that there are no external links into $W_1$,[9] Therefore, no matter how clever you are about link farm structure, if you don't have any incoming links, you can't increase your PageRank above this bound.

[9] and no links going out! Clearly, this will only decrease pageranks in $W_1$, but I didn't prove it. TODO!

# 5 Discussion

- allows users to put a value on PageRank reputation, and help decide trustworthiness. Good according to [10]. Should the Google toolbar show reputation in dollars (or local currency) rather than a 1 - 10 scale?
- other ways to assign values in restart vector. (eg: information retrieval for internal organizational use: use funding, different dept's home pages, revenue generated by author/dept/etc to seed $E$)
- value of reputation to holder is of no value? (searchking) False economy?
- applications to Eigentrust [5] and Pinar Yolum's referral system [11]?
- Publicizing recommendations isn't incentive compatible. Practical issues?
- Compare to Raph's approach to attack resistance, and Sybil attacks [4].
- Explains why bloggers and companies get higher pageranks: both buy lots of domain names!
- What is the relationship between this and [2]? Can these results here be generalized to deal with non-disjoint communities? Can their results be generalized to deal with non-uniform restart vectors?

# 6 Conclusion

I have put a lower bound on the amount of money a person must spend to acquire an arbitrarily good reputation with Google's PageRank, including both theoretical results and practical estimates.

# 7 Thanks

Geordie Zang, Sam Joseph, Graham Byrnes, Catherine Lai, Paul Gruba and Liz Sonenberg spent a lot of time providing me with useful feedback.

# References

[1] Ehrhard Behrends. *Introduction to Markov Chains (with Special Emphasis on Rapid Mixing).* Vieweg Verlag, 1999.

[2] Monica Bianchini, Marco Gori, and Franco Scarselli. Inside PageRank. *ACM Transactions on Internet Technology, In Press.*

[3] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998.

[4] John Douceur. The Sybil attack. In *1st International Workshop on Peer-to-Peer Systems (IPTPS'02)*, 2002.

[5] Sepandar Kamvar, Mario Schlosser, and Hector Garcia-Molina. The EigenTrust Algorithm for Reputation Management in P2P Networks. In *The Twelth International World Wide Web Conference*, 2003.

[6] John G. Kemeny and J. Laurie Snell. *Finite Markov Chains.* Springer-Verlag, 1976.

[7] Raph Levien. *Attack Resistant Trust Metrics.* PhD thesis, *To Appear.*

[8] James Norris. *Markov Chains.* Cambridge University Press, 1997.

[9] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.

[10] Michael Reiter and Stuart Stubblebine. Authentication metric analysis and design. *ACM Transactions on Information and System Security*, 1999.

[11] Pinar Yolum and Munindar Singh. Emergent properties of referral systems. In *Second International Conference on Autonomous Agents and Multiagent Systems (AAMAS03)*, 2003.

# A The Different Definitions of PageRank

There is considerable confusion about the definition of PageRank. Several different but equivalent definitions of PageRank are available. Even the two papers ([3] and [9]) by its inventors use different definitions without any explanation! As I will explain shortly, the former has a mistake that has gone unnoticed, and defines a function that is very different from the one intended! This mistake has been propagated through the literature.

In addition, much of the literature only deals with the special case of PageRank where the Random Search function is allocated uniformly (i.e. $s(p) = \frac{1}{|P|}$). Moreover, it is usually assumed that the random surfer will pick a hyperlink with uniform probability. My definition of PageRank is more general in both aspects. For example, it permits a model where the first hyperlink on a webpage is chosen with a higher probability than the other hyperlinks.

To resolve these issues, I will present the two definitions from the two papers by the PageRank inventors, and argue that the first of these is inconsistent, and suggest what they really had in mind. Then, I will show that the corrected PageRank definition of the first paper and the definition of the second paper are equivalent to special cases of my definition.

To distinguish between the various definitions of PageRank, I will call the PageRank as defined in this paper *PageRank* with notation $r(a)$, the PageRank in [3] *PageRank1Wrong* with notation $r_{1w}(a)$, the PageRank that I think [3] intended to define as *PageRank1* with notation $r_1(a)$ and the PageRank in [9] *PageRank2* with notation $r_2(a)$.

First, I will define some notation:

**Definition A.0.4 (Neighbourhoods).** *If $G = (P, H)$ is a webgraph, then define the incoming and outgoing neighbourhoods of $a$ as:*

$$N^-(a) = \{b : (b, a) \in H\}$$
$$N^+(a) = \{b : (a, b) \in H\}$$

Also, this theorem about PageRank is helpful:

**Theorem A.0.5 (PageRank Algebra).** *Let $G = (P, H)$ be a webgraph. Let $\{X_k\}$ be a Random Click process with transition function $\omega$. Let $\{Y_k\}$ be a Random Search-Click process of $\{X_k\}$ with Random Search function $s$ and Random Search probability $d$. Let $r$ be the PageRank function defined in terms of $\{Y_k\}$.*

*Then, for all $a \in P$:*

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

*Proof.* From *Definition 4.2.5 (PageRank)*, we have:

$$\begin{aligned}
r(a) &= \lim_{k\to\infty} \mathbf{P}(Y_k = a) \\
&= \lim_{k\to\infty} \sum_{b\in P} \mathbf{P}(Y_k = a \,|\, Y_{k-1} = b)\ \mathbf{P}(Y_{k-1} = b) \\
&= \sum_{b\in P} \lim_{k\to\infty}\ \psi(b,a)\mathbf{P}(Y_{k-1} = b) \\
&= \sum_{b\in P} \psi(b,a) r(b) \\
&= \sum_{b\in P} \left[ds(a) + (1-d)\omega(b,a)\right] r(b) \\
&= ds(a) \sum_{b\in P} r(b) + (1-d) \sum_{b\in P} \omega(b,a) r(b) \\
&= ds(a) + (1-d) \sum_{b\in N^-(a)} \omega(b,a) r(b)
\end{aligned}$$

□

## A.1 PageRank1 as defined in [3]

Unfortunately, there is a mistake that I discovered in the presentation of PageRank in [3] that has gone unnoticed in the academic community.

In this section, I will argue that there is indeed a mistake and describe what I believe the authors really had in mind. I will prove that this corrected version of PageRank is equivalent to a special case of PageRank as I defined it in *Definition 4.2.5*.

Here it was defined incorrectly:

**Definition A.1.1 (PageRank1-Wrong).** *If $G = (P, H)$ is a webgraph, then the PageRank1 $r_{1w}(a)$ of a page $a$ is:*

$$r_{1w}(a) = (1-c) + c \sum_{b\in N^-(a)} \frac{r_{1w}(b)}{|N^+(b)|}$$

*where $c$ is a constant.*

It should have been defined like this:

**Definition A.1.2 (PageRank1).** *If $G = (P, H)$ is a webgraph, then the PageRank1 $r_1(a)$ of a page $a$ is:*

$$r_1(a) = \frac{1-c}{|P|} + c \sum_{b\in N^-(a)} \frac{r_1(b)}{|N^+(b)|}$$

*where $c$ is a constant.*

They propose that a value of 0.85 for $c$ is a good choice. Their constant $c$ is equivalent to $1 - \frac{E}{E+1}$ presented in [9], and $1 - d$ presented in this paper in *Definition 4.2.4 (Random Search-Click process)*. We chose to be closer to [9] in this respect. This value is called the *damping factor* in [3] and the *decay factor* in [9].

To see that their original definition is indeed mistaken, I quote from their paper this observation:

> Note that the PageRanks form a probability distribution over web pages, so the sum of all web pages' PageRanks will be one.

I claim that *PageRank1-Wrong* does not have this property, but *PageRank1* does. I will prove that *PageRank1* has this property by showing that it is equivalent to a special case of *Definition 4.2.5 (PageRank)*, and hence is a probability distribution that satisfies this property.

**Claim A.1.3 (Incorrectness of PageRank1-Wrong).** *Let $G$ be a webgraph. Let $r_{1w}(a)$ be the PageRank1-Wrong score of webpage $a$, with parameter $c > 0$. Then $r_{1w}(a)$ does **not** always form a probability distribution, such that $r_{1w}(a) \geq 0$ for all $a \in P$ and $\sum_{a \in P} r_{1w}(a) = 1$.*

*Proof.* For the sake of contradiction, assume that $r_{1w}(a) \geq 0$ for all $a \in P$ and $\sum_{a \in P} r_{1w}(a) = 1$.

Recall the definition of $r_{1w}$:

$$r_{1w}(a) = (1-c) + c \sum_{b \in N^-(a)} \frac{r_{1w}(b)}{|N^+(b)|}$$

Now, observe that as $c > 0$ and $r_{1w} > 0$, the second term is greater than zero, and hence $r_{1w}(a) \geq 1 - c$. Thus:

$$\begin{aligned} \sum_{a \in P} r_{1w}(a) &\geq \sum_{a \in P} (1-c) \\ &\geq |P|(1-c) \end{aligned}$$

Since $1 - c$ is a fixed parameter, and $|P|$ can grow arbitrarily, I conclude that this does not always equal 1. Therefore, the claim that *PageRank1-Wrong* always forms a probability distribution is incorrect. □

In practice, this is a big mistake. For the suggested value of $c = 0.85$, and the reported 3 billion documents on the internet, the sum of PageRanks would be about 50 million. It is inconceivable that the authors of this paper really meant this.

Now I will show that *PageRank1* is equivalent to the special case of my *PageRank* where the Random Search function is uniform across all web pages, and the Random Surfer chooses hyperlinks with uniform probability.

**Corollary A.1.4 (PageRank1 Equivalence).** *Let $G = (P, H)$ be a webgraph. Let $\{X_k\}$ be a Random Click process with transition function $\omega(a, b) = \frac{1}{|N^+(a)|}$ when $(a, b) \in H$ and $\omega(a, b) = 0$ otherwise. Let $s(a) = \frac{1}{|P|}$ be the uniform Random Search function. Let $\{Y_k\}$ be the Random Search-Click process constructed from $G$, $\{X_k\}$, $s$ and some search constant $d$.*

*If $r$ is the PageRank function constructed from $\{Y_k\}$ and $r_1$ is constructed from $G$ and $c = 1 - d$, then $r(a) = r_1(a)$ for all $a \in P$.*

*Proof.* From *Theorem A.0.5 (PageRank Algebra)*, we have:

$$\begin{aligned} r(a) &= ds(a) + (1-d) \sum_{b \in N^-(a)} \omega(b, a) r(b) \\ &= d\frac{1}{|P|} + (1-d) \sum_{b \in N^-(a)} \frac{1}{|N^+(b)|} r(b) \\ &= \frac{1-c}{|P|} + c \sum_{b \in N^-(a)} \frac{r(b)}{|N^+(b)|} \end{aligned}$$

This is exactly the definition of $r_1$. □

This concludes the discussion of the definition of PageRank in [3]. In summary: the definition given was obviously a mistake, but can be easily corrected. The corrected version is equivalent to a special case of my generalized definition in terms of Random Search-Click processes.

## A.2 PageRank as defined in [9]

In this section, I will show PageRank as defined in [9] is equivalent to a special case of the PageRank defined in this paper. However, I think their choice of parameter $E$ is slightly different to what they had in mind.

PageRank is defined in [9] as follows:

**Definition A.2.1 (PageRank2).** *Let $G = (P, H)$ be a webgraph and $e : P \to R^+$ be a source of rank. The **PageRank2** $r_2(a)$ of a web page $a$ is:*

$$r_2(a) = ce(a) + c \sum_{b \in N^-(a)} \frac{r_2(a)}{|N^+(b)|}$$

*where $c$ is chosen such that $\sum_{a \in P} r_2(a) = 1$.*

Again, [9] recommend $e$ be chosen such that $\sum_{a \in P} e(a) = 0.15$ (which is $1 - 0.85$). However, this is *not* the same as the Random Search probability, $d$! I believe this is a mistake on the part of the authors of [9].

**Corollary A.2.2 (PageRank2 Equivalence).** *Let $G = (P, H)$ be a webgraph. Let $e : P \to \boldsymbol{R}^+$ be a function. Let $E = \sum_{a \in P} e(a)$.*

*Let $\{X_k\}$ be a Random Click process with transition function $\omega(a, b) = \frac{1}{|N^+(a)|}$ when $(a, b) \in H$ and $\omega(a, b) = 0$ otherwise. Let $s(a) = \frac{e(a)}{E}$ be the uniform Random Search function that is a scaled version of $e$. Let the Random Search probability be $d = \frac{E}{E+1}$. Let $\{Y_k\}$ be the Random Search-Click process constructed from $G$, $\{X_k\}$, $s$ and $d$.*

*If $r$ is the PageRank function constructed from $\{Y_k\}$ and $r_2$ is constructed from $G$ and $e$, then $r(a) = r_2(a)$ for all $a \in P$.*

*Proof.* From *Theorem A.0.5 (PageRank Algebra)*, we have:

$$\begin{aligned} r(a) &= ds(a) + (1 - d) \sum_{b \in N^-(a)} \omega(b, a) r(b) \\ &= \left(\frac{E}{E+1}\right)\left(\frac{e(a)}{E}\right) + \left(1 - \frac{E}{E+1}\right) \sum_{b \in N^-(a)} \frac{1}{|N^+(b)|} r(b) \\ &= \frac{1}{E+1} e(a) + \frac{1}{E+1} \sum_{b \in N^-(a)} \frac{r(b)}{|N^+(b)|} \end{aligned}$$

Substituting $c = \frac{1}{E+1}$, we obtain:

$$r(a) = ce(a) + c \sum_{b \in N^-(a)} \frac{r(b)}{|N^+(b)|}$$

This is exactly the definition of $r_2$. $c$ is chosen appropriately, since the sum of $r$ values is 1. □