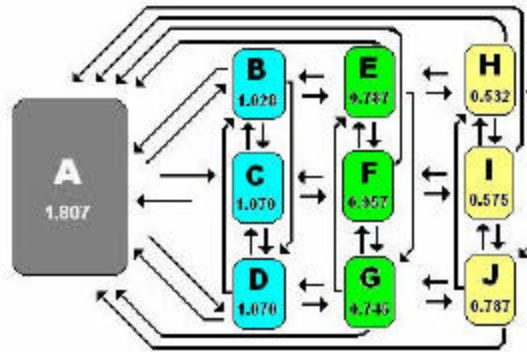













# Google PageRank™ and Related Technologies



by Jason J. Green

Revised and Updated Sept. 2005

## Table of Contents

 <b>Foreword</b>	pg. 3
 <b>Introduction</b>	pg. 4
 <b>Defining Page Rank</b>	pg. 5
 <b>PageRank Implementation and Current Usage</b>	pg. 6
 <b>PageRank - Technical Discussion</b>	pg. 8
 <b>Visually Mapping PageRank</b>	pg. 19
 <b>Linking Models</b>	pg. 21
 <b>Google Search Technologies</b>	pg. 25
 <b>Google Unfriendly</b>	pg. 29
 <b>Other Points of Interest</b>	pg. 31
 <b>References and Citations</b>	pg. 32

## Foreword

---

This document is the product of an extensive research project which was conducted between July and November of 2004. The goal of the project was to identify and understand the concepts, mechanics and implementations of PageRank and other Google citation analysis technologies. What differentiates this work from other, similarly themed documents is the research standards employed. Strict procedural standards were employed for the duration of the project to ensure reliability and confidence in any findings. All informational resources and reference data used came from scholarly verifiable works. Third party or speculative sources of information were completely disregarded in favor of verifiable information resources, and while these standards presented greater difficulties in terms of raw research data, they also insured that the resulting work would be a product of rigorous investigation as opposed to the wildly speculative documents which often dominate search engine optimization research.

The work presented here makes no claim of being *the* authoritative source of information for PageRank and related technologies, nor is that the purpose of this work.

Rather this work is a resource for the search marketing community; one that presents a sound theoretical understanding of PageRank based upon extensive research and from reliable, first-hand information.

## Introduction

---

Since its inception in 1998 Google has deposed former industry giants like Alta Vista and Yahoo! and established itself as the most popular and successful search engine in the world. This is in large part due to Google's usage of PageRank™ to help power its innovative search technology.

The mathematical formula which would come to be known as PageRank™ was invented by Lawrence Page while he was a PhD candidate at Stanford University. PageRank™ was first used to power a prototype search engine called Back Rub, which referred to the search engine's usage of back-links in determining results. PageRank was also used to help power the Stanford resident search engine "Searching Stanford," which was in effect a prototype of the current Google search engine. PageRank™ is an innovative technology that provides a completely objective means for determining the relative importance of documents within a linked database.

PageRank™ and a family of related Google technologies, represent an innovative approach to information retrieval on the Internet. The scope and possible application of these technologies extends far beyond the search box at Google.com. The following essays were written to help facilitate an understanding of PageRank™ and related Google technologies. Furthermore I hope to present the information in a conceptual fashion whenever possible so that a greater appreciation of the implications which are asserted by these technologies can be attained.

## Defining Page Rank

---

PageRank is a registered trademark of Google Inc. that refers to the Google search engine's usage of U.S. Patent # 6,285,999 (A method of ranking nodes in a linked database). Google Technology explains:

***“The heart of our software is PageRank™, a system for ranking web pages developed by our founders Larry Page and Sergey Brin at Stanford University. And while we have dozens of engineers working to improve every aspect of Google on a daily basis, PageRank continues to provide the basis for all of our web search tools.”***

Put simply PageRank is a method of assessing the importance of a web page based upon its relationship to other web pages. The Abstract of the original patent explains:

***“A method assigns importance ranks to nodes in a linked database, such as any database of documents containing citations, the world wide web or any other hypermedia database.”***

When PageRank is computed for all of the web pages within the Google index and the data is analyzed, the broader implications become much more apparent, for example PageRank values are an exceedingly accurate map of user behavior probability.

The Anatomy of a Large-Scale Hyper-textual Web Search Engine by Lawrence Page and Sergey Brin offers the following explanation:

***“PageRank can be thought of as a model of user behavior. We assume there is a "random surfer" who is given a web page at random and keeps clicking on links, never hitting "back" but eventually gets bored and starts on another random page. The probability that the random surfer visits a page is its PageRank.”***

PageRank is an interesting social phenomenon as well, largely because it offers a completely objective measure of a web page's importance, all of which is calculated mathematically and without human interference.

In a very short period of time PageRank has become a popular measurement standard for determining the value of a web site. This has led to a PageRank obsession of sorts within certain industries and has created a remarkably high demand for methods of increasing PageRank. Businesses have been created to capitalize on this demand by offering services designed to increase a web site's PageRank (and consequently search engine rankings). Many of these businesses employ techniques which are at best, suspect, and usually do more harm than good (For both client and Google). There are however some firms that have endeavored to leverage PageRank in a fashion that both ethical and effective. Such businesses have inspired a good deal of innovative research, some of which benefits the internet as a whole.

## PageRank Implementation and Current Usage

*“The importance of a Web page is an inherently subjective matter which depends on the reader’s interests, knowledge and attitudes. But there is still much that can be said objectively about the relative importance of Web pages. This paper describes PageRank, a method for rating Web pages objectively and mechanically, effectively measuring the human interest and attention devoted to them.”*

Lawrence Page – Bringing Order to the Web

---

### **The Search Engine’s Humble Beginnings:**

The first generations of search engines relied primarily on information resident within a given web page to determine its relevance to a search query. Determining factors included: high query term frequencies, search terms placement, phase, proximity, etc. These methods proved to be too easily manipulated and ultimately unreliable: SEO techniques such as keyword stuffing were effective enough to guarantee a position within the top 10 Search Engine Results for a given search phrase..

At the First International Conference on the World Wide Web (1994 ) a researcher named Oliver A, McBryan presented the World Wide Web Worm (WWWW); the first search index which employed document citation as a measure of relevance for web pages. Specifically, the WWW used the anchor text of inbound links to determine relevance for a search term.

The Hyperlink Search Engine, developed by IDD Information Services also employed information provided by back-links as a measure of topical relevance for web pages. When a search was performed, the search term was evaluated against a collection of anchor text descriptions from links that point to a web page, rather than using a term index of the web page’s content.

The Hyperlink Search Engine disregarded nearly all page-resident factors when compiling its Search Engine Results Pages (SERPs).

### **Back Rub:**

In January of 1996, Lawrence Page and his colleague Sergey Brin (both PhD candidates of the Stanford University Department of Computer Science), were completing a collaborative effort on a prototype search engine called “BackRub”, which was named for its unique treatment of “back-links” in determining search query results. BackRub employed a system of back link analysis for determining the final sorting order of search results. BackRub is in essence the foundation of the Google search.

It wouldn’t be too long before BackRub was moved out of the basement; becoming Google (a play on the mathematical term “googol”, which is generally used to describe an infinitely high value, but specifically is a value of 10 raised to the 100<sup>th</sup> power.).

A detailed outline of the Google prototype is offered in “The Anatomy of a Large Scale Hyper-textual Search Engine” authored jointly by Page and Brin.

**Conclusion:**

The purpose of Page Rank is to provide a vast improvement in the quality and accuracy of documents retrieved by a search engine in response to a user search phrase query.

This is accomplished by:

- Objectively evaluating the importance of individual web pages and assigning corresponding “ranks”.
- Using the “rank” information of web pages to help determine relevance and placement within a SERP.

The details of how this is accomplished, is covered in the Technical Discussion section which follows.

# PageRank - Technical Discussion



## *“The heart of our software is PageRank”*

---

### Part 1 – Mechanics of PageRank

#### 1.1

The PageRank formula relies on the dynamically inter-linked geography of the WWW to compute the importance of individual web sites, which is then represented as a web site's PageRank. The primary concept is that hyper-links can be counted as “votes” or endorsements for the page they are pointing to.

A link is treated as an endorsement of a web page's importance. However not all endorsements are of equal value, and so the value of each endorsement is determined by the PageRank of the web page that is giving it. Furthermore the value of any endorsement given is divided by the number of endorsements given.

Google Technology Explains:

***PageRank relies on the uniquely democratic nature of the web by using its vast link structure as an indicator of an individual page's value. In essence, Google interprets a link from page A to page B as a vote, by page A, for page B. But, Google looks at more than the sheer volume of votes, or links a page receives; it also analyzes the page that casts the vote. Votes cast by pages that are themselves "important" weigh more heavily and help to make other pages "important." Important, high-quality sites receive a higher PageRank, which Google remembers each time it conducts a search.***

In their paper “The Anatomy of a Large-Scale Hyper-textual Web Search Engine”

Sergey Brin and Lawrence Page explain the PageRank Formula as follows:

***We assume page A has pages T1...Tn which point to it (i.e., are citations). The parameter d is a damping factor which can be set between 0 and 1. We usually set d to 0.85.***

***Also C(A) is defined as the number of links going out of page A. The PageRank of a page A is given as follows:***

$$PR(A) = (1-d) + d (PR(T1)/C(T1) + \dots + PR(Tn)/C(Tn))$$

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

## 1.2

The formula can be broken down into parts to clarify its basic concepts:

**PR(A) =**

The PageRank of a Page (A) is

1. **PR(T1) - PR(Tn)** – These are the web pages which link to page “A”. PR(T1) ... PR(Tn) represent the first and last web pages which link to “A” as well as every page in between.
2. **PR(Tn)/C(Tn)** - Any linking web page divides the weight of its vote evenly amongst all of the votes that it gives.
3. **d** – Every vote is added together to determine the PageRank value for page “A”, however to prevent a “combined strength” effect from over inflating the PageRank of “A”, the final value is multiplied by a damping factor (usually .85).
4. **(1 - d)** – This ensures that “*sum of all web pages' PageRanks will be one*” by adding the amount (usually .15) lost by the damping factor back in. This way every webpage will have at least that minimal value. (The *average* sum of all web pages' PageRanks will be one.)

One of the interesting features which stand out is that PageRank calculates rank for each web page individually and does not normalize PageRank across a domain. Every distinct hyper-document is specifically evaluated in relation to every other hyper-document on the WWW.

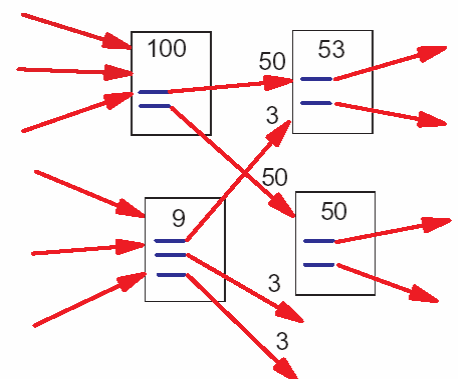
Furthermore it is understood that PageRank is not transferred from one page to another; PageRank simply determines the importance of any given endorsement.

## 1.3

A single calculation of PageRank according to the formula described would observe the following:

- The PageRank of page “T1” (a page that links to page “A”), is divided by the total number of outgoing links on page T1. The more outbound links that T1 has, the less that page A can directly benefit. **For example:** Let us suppose that page T1 not only links to A, but also links to B C and D; then page A only receives one quarter of T1's PageRank value.
- The PageRank value which is passed to page A by T1 is then multiplied by a damping factor of .85, and .15 is added to that product to produce a final value for T1.
- This process is repeated for all of the pages that link to A and the sum of the final values produced by T1...Tn is the PageRank of A.

This illustration of a “Simplified PageRank Calculation” is from “The PageRank citation Ranking: Bringing Order to the Web” by Lawrence Page, Sergey Brin, Rajeev Motwani und Terry Winograd



## 1.4

If the PageRank of each web page is dependant upon the PageRank of the web pages pointing to it, how can an accurate initial starting value be determined?

Fortunately this is not required. Any starting input value can be applied and it will not affect the accuracy of the final PageRank values. The sum of all web pages PageRank is Unity, and whatever unit is chosen to symbolize a fraction of the whole is irrelevant.

Lawrence Page, in the original patent document, explains:

***PageRank or PR(A) can be calculated using a simple iterative algorithm, and corresponds to the principal eigenvector of the normalized link matrix of the web.***

Which means that we can go ahead and perform the calculations, so long as we re-iterate the algorithm until the final values converge, and in fact they will.

**PageRank is a probability distribution normalized across the WWW.**

## 1.5

### **The Random Surfer:**

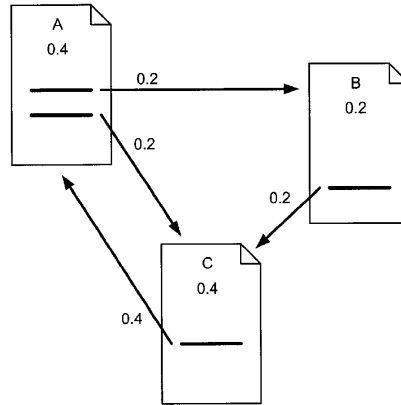
The concept of PageRank as a model of user behavior has been put forth as an intuitive justification of PageRank in practice. The model is often referred to as “The Random Surfer”. Lawrence Page explains:

***PageRank can be thought of as a model of user behavior. We assume there is a "random surfer" who is given a web page at random and keeps clicking on links, never hitting "back" but eventually gets bored and starts on another random page. The probability that the random surfer visits a page is its PageRank. And, the  $d$  damping factor is the probability at each page the "random surfer" will get bored and request another random page.***

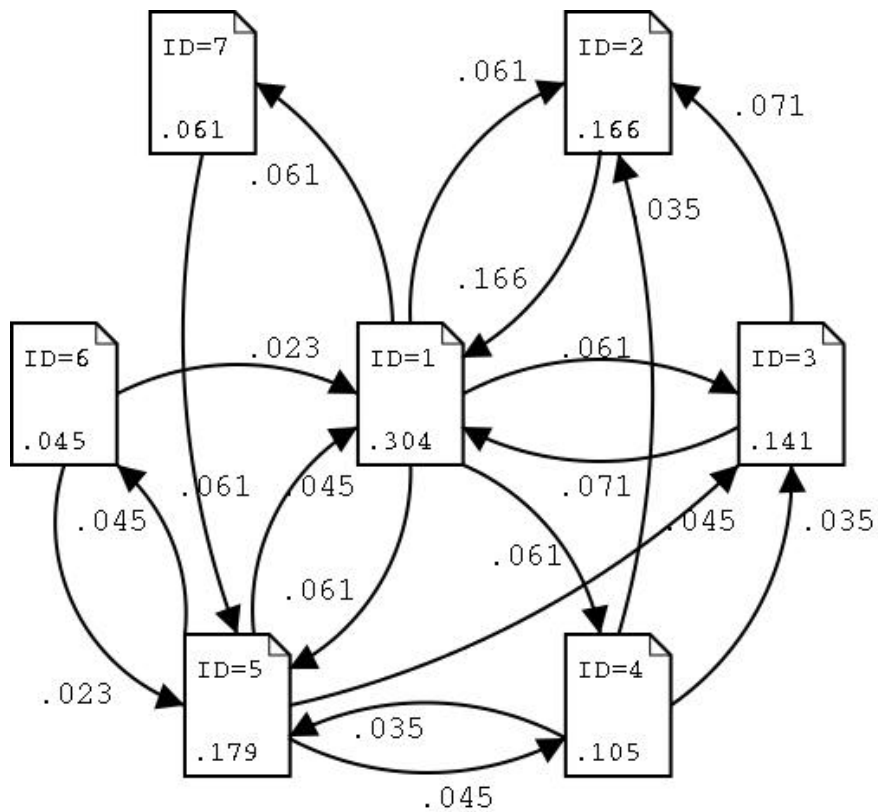
An interesting problem which was encountered by some of first conceptions of PageRank was the accumulation effect that occurred within small closed node structures. This is caused by two pages which link only to each other, and yet one of the pages is linked to by an outside page. This would cause a pooling effect in regards to PageRank and is referred to as a RankSink. This problem was overcome by introducing a theoretical source of PageRank (see 3.1) and a random jump probability to the “Random Surfer Model”.

When a user is browsing web sites and encounters a small loop of web sites, it is highly unlikely that the user will remain within this loop forever, hence the introduction of a damping factor, which simulates a random jump to an unrelated page.

**1.6**  
**Illustrated Examples:**



The figure above is from the original patent document. The illustration demonstrates the application of PageRank in a simplified 3 page internet.



A model of greater complexity which more accurately demonstrates the functionality of PageRank.

## 1.7

### The Internet Link Graph:

Again from “The Anatomy of a Large-Scale Hyper-textual Web Search Engine”

By Sergey Brin and Lawrence Page, the following points of interest are offered:

*The citation (link) graph of the web is an important resource that has largely gone unused in existing web search engines. We have created maps containing as many as 518 million of these hyperlinks, a significant sample of the total. These maps allow rapid calculation of a web page's "PageRank", an objective measure of its citation importance that corresponds well with people's subjective idea of importance. Because of this correspondence, PageRank is an excellent way to prioritize the results of web keyword searches. For most popular subjects, a simple text matching search that is restricted to web page titles performs admirably when PageRank prioritizes the results (demo available at [google.stanford.edu](http://google.stanford.edu)). For the type of full text searches in the main Google system, PageRank also helps a great deal.*

Internet citation graphs offer an excellent visual justification of PageRank. Such a link matrix can be visualized as **a complex electrical circuit wherein electricity is concentrated within the components of greatest importance.**



Visualization of primary Internet architectures within the U.S.

## Part 2 –PageRank and the Google toolbar

### 2.1

While actual PageRank values of web sites are not public information, Google has created two simplified visual representations of PageRank which are publicly accessible:

The first is the PageRank meter which appears next to all of the websites in the Google Directory (directory.google.com), which is essentially a copy of the DMOZ Directory (www.dmoz.com)

The Google Directory Team explains:

***The green ratings bars are a measure of the importance of a web page, as determined by Google's patented PageRank technology. These PageRank bars tell you at a glance whether other people on the web consider a page to be a high-quality site worth checking out. Google itself does not evaluate or endorse websites. Rather, we measure what others on the web feel is important enough to deserve a link. And because Google does not accept payment for placement within our results, the information you see when you conduct a search is based on totally objective criteria.***

### 2.2

The second visual embodiment is incorporated in the Google Toolbar. The toolbar is a publicly available snap in interface for your web browser. In addition to a graphical PageRank display, the toolbar includes a Google search field and popup blocker as well as other options for accessing some of the advanced features of Google search. Public interest in and knowledge of PageRank, in a general sense, is due to the Google toolbar.

The figure below details the 0 to 10 ten scale which is used by the Google toolbar.



When viewing a web page, a PageRank is displayed in the toolbar for that page. Each time you move to a new web page, the toolbar displays an updated score of 0 to 10 for the page you are currently viewing. The toolbar functions by querying a database to furnish PageRank values. If no PageRank information exists within the database for the page you are viewing, the PageRank display appears grayed out, signifying that no information is available.

### 2.3

The Directory PageRank display is quite different and deserves an explanation of its own. The directory PageRank actually appears to be on a different scale, which for the most part uses only 7 distinct gradations (as opposed to the 10 gradations of the toolbar display).

Chris Raimondi of searchnerd.com discovered that the directory PageRank display is composed of two distinct graphics; a positive green graphic and a negative grey graphic. Raimondi also discovered that the length of both graphics together (in points) always equals 40. Therefore a maximum PageRank value of 40 points is possible. The following chart illustrates the effect of applying toolbar style values to the Directory Page Rank (DPR) display:



- Although a directory PageRank of 40(points) is possible, the highest score assigned to a page is generally 38, which means that the green bar has a length of 38 points and the grey bar has a length of 2. This scenario corresponds to a DPR of 7 according to the above chart.
- There are a few exceptions, one of them being Google.com. When the length of that graphic is measured, it totals 44 pixels. This is true of any page which has a completely green graphic with no grey. It appears that once a site reaches a certain actual PageRank value, its directory PageRank display is re-scaled (I have been unable to locate a reputable explanation of this phenomenon.)
- It is a common belief that the Directory PageRank is a more accurate approximation of actual PageRank, and therefore may be a better measurement to use than toolbar PageRank. However there is no decisive evidence to support this idea and furthermore the difference in values between toolbar and directory PageRank is not always consistent. Occasionally a web site will be found that completely defies any assumed toolbar to directory ratio.

Notice the curious overlap that occurs when the Directory and Toolbar PageRank scales are compared proportionate to each other:

DPR	0	5	11	16	22	27	32	38	44		
TPR	0	1	2	3	4	5	6	7	8	9	10

**The Directory PageRank values are based on the size of the green display in pixels.  
The Toolbar PageRank values are of course a simple 0 to 10 scale.**

## 2.4

The graphical PageRank displays are interesting because of the scaling which is implied by their use.

Consider the fact that the sum of all PageRank on the web is unity (1). Furthermore each page's individual PageRank is a very small portion of 1 (billionths). The obvious question is how does such a drastic reduction to scale occur? This information remains undisclosed by Google, it can be assumed however that the scaling is logarithmic; a logarithmic scale reduction certainly meets the criteria for such a reduction.

From "PageRank Citation Ranking", by Lawrence Page:

***"We have developed a web proxy application that annotates each link that a user sees with its PageRank. This is quite useful because users receive some information about the link before they click on it...  
...The length of the red bars is the logarithm of the URL's PageRank"***

Consider for example the use of a natural logarithm; in that case it would take a great deal more PageRank for a page to move up to the next level than it took to move up from its previous level.

A logarithmic scale might look something like this:

<b>0</b>	<b>0.15 – 0.9</b>
<b>1</b>	<b>1.0 – 5.4</b>
<b>2</b>	<b>5.5 – 32.4</b>
<b>3</b>	<b>32.5 – 194.4</b>
<b>4</b>	<b>194.5 – 1,166.4</b>
<b>5</b>	<b>1,166.5 – 6,998.4</b>
<b>6</b>	<b>6,998.5 – 41,990.4</b>
<b>7</b>	<b>41,990.5 – 251,942.4</b>
<b>8</b>	<b>251,942.5 – 1,511,654.4</b>
<b>9</b>	<b>1,511,654.5 – 9,069,926.4</b>
<b>10</b>	<b>9,069,926.5 – 0.85*N + 0.15</b>

## 2.5

A theory has been put forth that a separate calculation is used to compute publicly viewable PageRank. This however is not substantiated or intuitive.

## Part 3 – PageRank Variations and Implementations of Interest

### 3.1

A second Page Rank algorithm was published by Lawrence Page and Sergey Brin. This second algorithm does not differ significantly from the original, it does however offer a better explanation of the “random surfer” model, which justifies PageRank by stating its effectiveness in mapping the probability that a random surfer will wind up on any given page. The random surfer visits a page according to a certain probability, which is the PageRank of that page.

#### Example:

Page A has a PageRank of 25, and there are 5,000,000,000 pages on the Internet. It would then follow that there is a 25 to 5,000,000,000 chance that “random surfer” is viewing Page “A” right now.

The Second Algorithm:

$$PR(A) = (1-d) / N + d (PR(T1)/C(T1) + \dots + PR(Tn)/C(Tn))$$

N has been introduced to represent the total number of pages on the WWW.

This algorithm forms a clearer probability distribution by asserting the number of variables that random surfer could encounter.

### 3.2

#### Identifying Related Web Sites via Link Analysis

This method describes a search feature which allows the user to retrieve lists of web pages that are related to a specific web page. This method is a specialized link analysis method that was developed by Kim Lun Law and Georges R. Harik for Google Inc. The technology was patented June 22, 2004, patent # 6,754,873 B1.

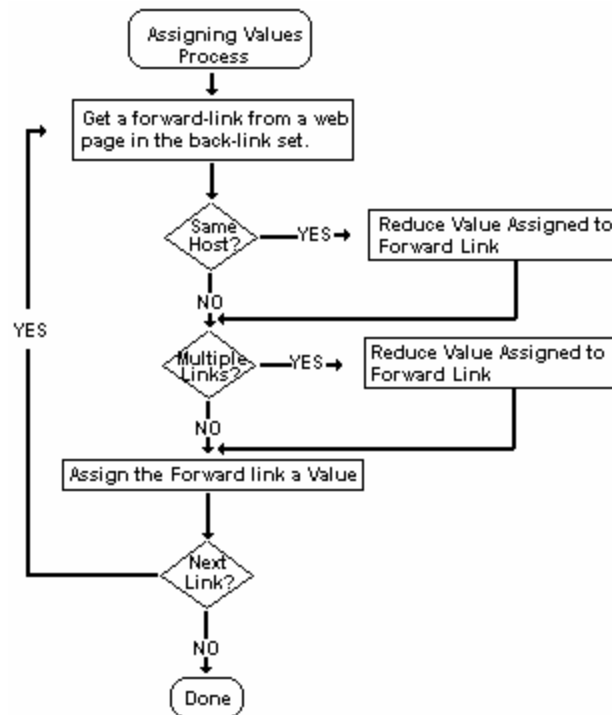
The patent abstract explains:

***“Backlink and forwardlink sets can be utilized to find web pages that are related to a selected web page. The scores for links from web pages that are from the same host and links from web pages with numerous links can be reduced to achieve a better list of related web pages. The list of related web pages can be utilized as a feature to a word-based search engine or an addition to a web browser.”***

Briefly the process is as follows (adapted from the patent document):

1) Listing of Related Web Pages:

- Index a back-link set of web pages for a selected web page.
- Index a forward-link set of web pages from the back-link set (A list of all web pages pointed to by the back-link set).
- Assign a value to each forward link from the web pages of the back-link set. (See illustration on next page.)



- Generate a score for the selected web page based upon the values of the links.
- Generate a list of related web pages according to their scores (The score is an indication of relatedness to the selected web page.).

### 3.3 Local Score

This is a method for refining the relevancy of web sites returned for a search term by recalculating PageRank amongst the set of search results at query time. Furthermore a strict filtering process is employed to prevent inflated values. This method was developed by Krishna Bharat for Google Inc. The method is patent # 6,725,259 , which was issued on April 20, 2004.

The Detailed Description from the patent document explains:

***As described herein, a search engine modifies the relevance rankings for a set of documents based on the inter-connectivity of the documents in the set. A document with a high inter-connectivity with other documents in the initial set of relevant documents indicates that the document has "support" in the set, and the document's new ranking will increase. In this manner, the search engine re-ranks the initial set of ranked documents to thereby refine the initial rankings.***

Brief summary of method:

- 1) In response to a search query, the search engine locates an appropriate set of web pages and organizes them by PageRank. The set may be limited to a 1000 documents.
- 2) Each web page within the set is re-ranked according to its value within the set. Which is calculated as follows :
  - For a web page “x” within the set, generate a list of web pages “B(y)” which link to “x”.
  - Web pages which are from the same host are removed from “B(y)”.
  - If any pairs of web pages in “B(y)” have the same or similar IP addresses; the web page with the lower PageRank is removed from “B(y)”.
  - Web pages in “B(y)” are sorted by their original PageRank, and those with the highest PageRank (k) remain in “B(y)”, while the rest are removed.
  - The LocalScore of “x” is the sum of all PageRanks of the web pages remaining in list “B(y)”.
  - A NewScore is calculated using both the original PageRank and the LocalScore :

$$\mathbf{NS(x) = (a+LS(x)/MaxLS)(b+PR(x)/MaxPR)}$$

Where:

NS(x) is the New Score of a web page (x)

“a” is a constant usually set to 1.

LS(x) is the LocalScore of (x)

MaxLS is the maximum of the LocalScore values in the set.

“b” is a constant usually set to 1.

PR(x) is the original PageRank of a web page (x).

MaxPR is the maximum of the original PageRanks in the set.

- The retrieved documents are sorted using their modified score and presented in the SERPs.

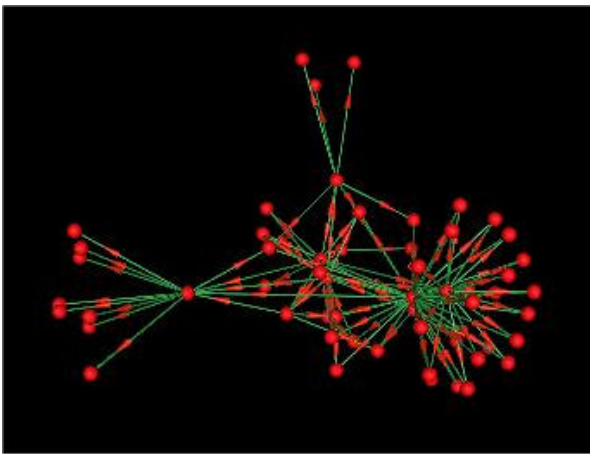
# Visually Mapping PageRank



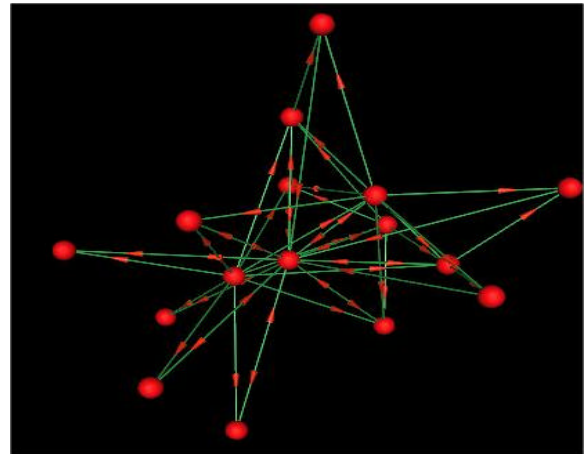
Fascinating research has been conducted in the area of 3D math imaging and mapping, which produces a visual representation of a mathematical activity without compromising the integrity of the mathematical formula. The study of these models can yield incredible information and has profound academic value, a discussion of which however is well beyond the scope of this paper. Therefore I present the following images as an interesting side note to this discussion of PageRank.

## Rendered with PRViz\*

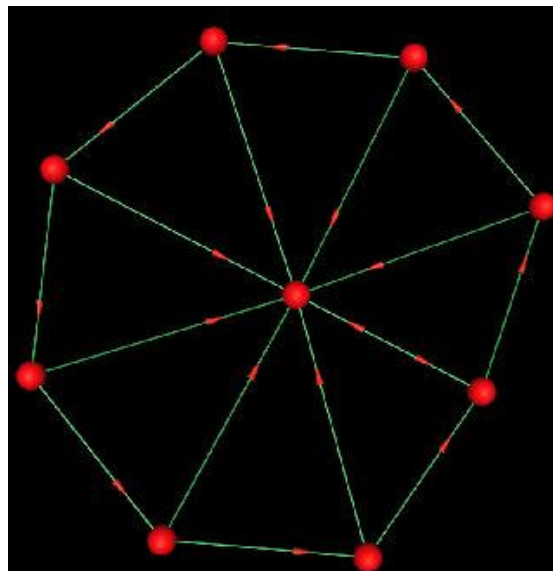
Spheres represent nodes (web pages), while the volume of the sphere is representative of the node's PageRank.



Web Graph with orphan pages



Same Graph orphan pages removed



Normalized closed Network

\*PRViz is an educational program designed to help visualize the operation of the PageRank algorithm



## Linking Models



These models were constructed to demonstrate how various linking strategies can greatly effect PageRank concentrations; particularly in regards to internal link structures. Note the change in values as the iterations approach convergence.

Each page starts with a PageRank of one.

---

### Normalizing PageRank between all Pages



By linking every page to every other page, we produce the following PageRank values:

**Page A = 1**

**Page B = 1**

**Page C = 1**

PageRank can be visualized as a flow of water through reservoirs and channels; the connectivity of the channels determines where the water will pool

---

### Channeling PageRank into a specific Page



**Page A = 1.850**

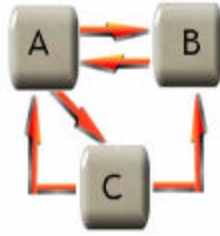
**Page B = 0.575**

**Page C = 0.575**

**$PR(A) = (1-d) + d (PR(B)/C(B) + (PR(C)/C(C))$  or  $(1 - 0.85) + 0.85 * (1 + 1) = 1.850$**

A has a much larger proportion of the PageRank than the other 2 pages. This is because pages B and C are reciprocating with A exclusively.

## Channeling PageRank in differing Concentrations



PageRank values:

**Page A = 1.425**

**Page B = 1**

**Page C = 0.575**

PageRank values after 10 iterations:

**Page A = 1.298**

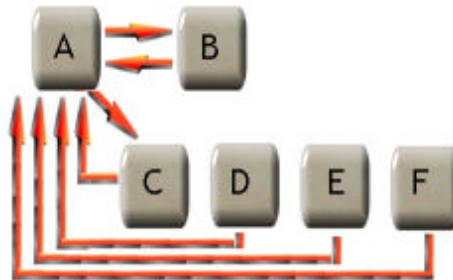
**Page B = 1**

**Page C = 0.701**

$$\mathbf{PR(A) = (1-d) + d (PR(B)/C(B) + (PR(C)/C(C)) \text{ or } (1 - 0.85) + 0.85* (1+0.5) = 1.425}$$


---

## Orphan Pages



PageRank Values

**Page A = 4.4**

**Page B = 0.575**

**Page C = 0.575**

**Page D = --**

**Page E = --**

**Page F = --**

PageRank values after 10 iterations

**Page A = 1.394**

**Page B = 1.016**

**Page C = 1.016**

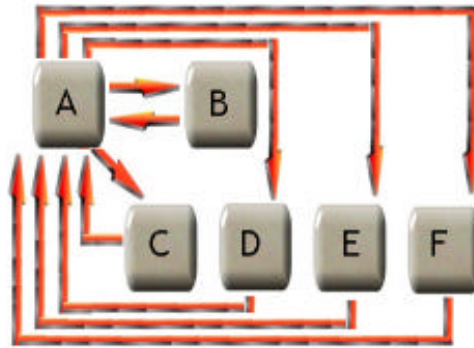
**Page D = --**

**Page E = --**

**Page F = --**

The effect of links from orphan pages is clearly illustrated, note the variations from one to ten iterations; PageRank has a remarkable ability to achieve mathematical balance as it approaches convergence.

## Child Pages



PageRank Values

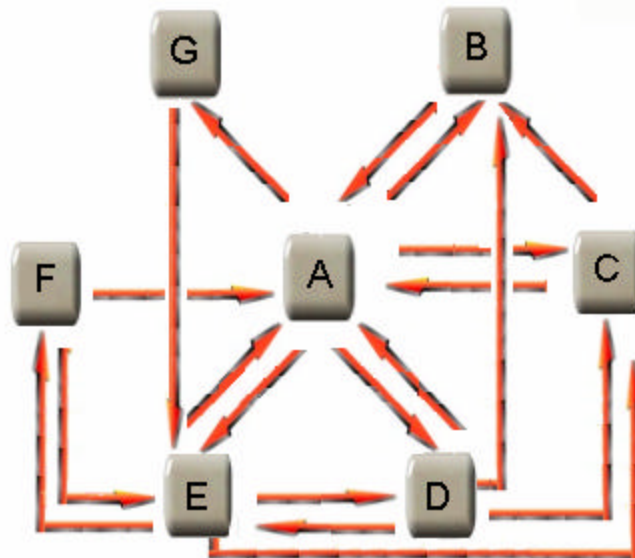
**Page A = 4.4**  
**Page B = 0.32**  
**Page C = 0.32**  
**Page D = 0.32**  
**Page E = 0.32**  
**Page F = 0.32**

PageRank values after 10 iterations

**Page A = 2.576**  
**Page B = 0.684**  
**Page C = 0.684**  
**Page D = 0.684**  
**Page E = 0.684**  
**Page F = 0.684**

---

## Intricate Set



PageRank Values

**Page A = 2.275**

**Page B = 0.957**

**Page C = 0.745**

**Page D = 0.532**

**Page E = 1.637**

**Page F = 0.532**

**Page G = 0.32**

PageRank values after 10 iterations

**Page A = 2.086**

**Page B = 1.033**

**Page C = 0.880**

**Page D = 0.726**

**Page E = 1.042**

**Page F = 0.726**

**Page G = 0.504**

$$\mathbf{PR(A) = (1-d) + d (PR(B)/C(B) + (PR(C)/C(C) + (PR(D)/C(D) (PR(E)/C(E) + (PR(F)/C(F))$$

**or**

$$\mathbf{(1 - 0.85) + 0.85 * (1 + 0.5 + 0.25 + 0.25 + 0.5) = 2.275}$$

# Google Search Technologies

## googol

n: a cardinal number represented as 1 followed by 100 zeros (ten raised to the power of a hundred)  $10^{100}$

---

## Part 1 – A Guide to Google Search Functionality:

### 4.1

#### Preliminary Notes:

\*Google is not case sensitive

\*Boolean Search Phrase Handling:

Google treats multiple word search phrases with a Boolean default of “and”, unless modifiers are employed. For Example: If a search is performed for **Dodge Ford truck Sacramento CA**, then Google will search for all words (Dodge and Ford and Truck and Sacramento and CA).

\* Stop Words (words ignored) - According to Google the following are stop words: “I” “a” “the” “of”. Personal experimentation has shown that usage and placement of these words does alter SERPs, and therefore these words are not necessarily ignored. It is possible that while these words are not included as part of a search term, they may influence search term context hypothesized by a semantic disambiguation software.

### 4.2

#### Search Refinements and Specialty Syntax

##### OR

Example - **Dodge or Ford**

This tells Google that either “**Dodge or Ford**” is acceptable results although not necessarily both. A pipe character ( | ) performs the same function as “or”.

##### Parentheses ( )

Example - **(Dodge or Ford) truck Sacramento CA**

This tells Google that **truck Sacramento CA** are required matches in the results, yet either **Dodge or Ford** can be present with them.

##### Minus

Example - **Dodge -Ford truck Sacramento CA**

This tells Google that Ford cannot be present in the results.

##### WILDCARD

A wildcard is indicated by an **Asterisk** \* in the search phrase.

Example – **Suggest \* Site**

This will return results such as Suggest a Site, Suggest your Site, etc.

**INTITLE:**

Searches only web page titles.

Example- **intitle:domain music**

**ALLINTITLE:**

Returns web pages which have a title composed only of the search terms.

Example- **allintitle:music handel**

**INURL:**

Searches only web page URL's

Example- **inurl:domain music**

Useful when searching sub-domains

**ALLINURL:**

Returns web pages which have all of the search terms in their URL.

Example- **allinurl:domain music**

**INTEXT:**

Searches only the body text of a web page to determine relevancy.

Example- **intext:wasps are scary**

**INANCHOR:**

Searches for relevant text in a web site's link anchors.

Example- **inanchor:wasp information**

**SITE:**

Searches for matching web site or domain.

Example- **site:edu**

**LINK:**

Returns a list of backlinks for a specific URL.

Example- **link:www.microsoft.com**

**CACHE:**

Returns the indexed copy of a web page, even if the page is no longer available at the indexed URL. Example- **cache:www.microsoft.com**

**DATERANGE:**

Returns web pages that were indexed within a specified date range. This feature only accepts Julian dates and NOT Gregorian. There are however many Gregorian to Julian converters available online.

Example- **superman daterange:2452389-2452389**

**FILETYPE:**

Refines searches by including only a specified filename extension like .pdf or .exe.

Example- **marketing reports filetype:pdf**

**RELATED:**

Returns a list of web pages that are related to a specified web page.

Example- **related:sbc.com**

This would return a list of major telecommunications providers.

**INFO:**

Returns a list of links to technical information about a specified URL.

Example- **info:www.msn.com**

**PHONEBOOK:**

Searches for phone numbers matching the given terms.

Example- **phonebook: captain howdy**

Also supports reverse index search

Example- **phonebook: (415) 775 9823**

**RPHONEBOOK:**

Searches for residential phone numbers matching the given terms.

Example- **rphonebook: captain howdy**

Also supports reverse index search

**BPHONEBOOK:**

Searches for business phone numbers matching the given terms.

Example- **bphonebook: captain howdy's fishing**

Also supports reverse index search

**DEFINE:**

Provides definitions of the given terms.

Example- **define: peremptory**

Also supports reverse index search

**Part 5 – Other Interesting Features**

This section describes some of Google's more obscure search functionalities.

**5.1****SMS Search**

Submit searches and receive results by mobile phone. Just send an SMS message of your query to the US shortcode 46645.

**5.2****Google Maps**

(<http://maps.google.com/>)

Leveraging Keyhole's awesome satellite imaging infrastructure, Google now furnishes accurate satellite maps of nearly the entire planet's surface. Maps are available in illustrated, satellite image or hybrid versions.

### **5.3**

#### **Google Scholar**

**(<http://scholar.google.com/>)**

Search results from scholarly academic resources.

### **5.4**

#### **Google Video**

**(<http://video.google.com/>)**

Search a large index of video files, many of which are user submitted.

# Google Unfriendly

---

**This section discusses some of the practices and techniques which Google frowns upon. In addition some aspects of the Google penalty system are described.**

---

## **Part 1 - Google Blacklist**

Google holds a decidedly negative view of certain things, especially those things which Google describes as reducing the quality of their search results. The list contains a broad spectrum of entries; everything from certain SEO practices to actual programs. The list which follows briefly discusses those things, which are likely to incur a Google penalty if participated in.

### **1.1**

#### **Programs**

Usually a program is blacklisted by Google if it violates the terms of use for Google's publicly available API. Entire IP address blocks have been banned for using such programs.

#### **WebPosition Gold** (<http://www.webposition.com/>)

WebPosition Gold (WPG) supports a range of useful tasks; everything from designing optimized pages to analyzing traffic. WPG also bears the unfortunate distinction of being the only program publicly denounced by Google. The following statement is from Google's Information for Webmasters - [www.google.com/webmasters/guidelines.html](http://www.google.com/webmasters/guidelines.html)

***“Don't use unauthorized computer programs to submit pages, check rankings, etc. Such programs consume computing resources and violate our terms of service. Google does not recommend the use of products such as WebPosition Gold™ that send automatic or programmatic queries to Google.”***

### **1.2**

#### **Practices and Techniques**

The list below, which is quoted verbatim from Google, outlines some specific practices which can incur a Google penalty.

- **hidden text or hidden links.**
- **cloaking or sneaky redirects.**
- **automated queries to Google.**
- **pages with irrelevant words.**
- **multiple pages, sub-domains, or domains with substantially duplicate content.**
- **"doorway" pages created just for search engines or other "cookie cutter" approaches such as affiliate programs with little or no original content.**

### 1.3

#### **Link Farms**

These are web pages created for the sole purposes of linking to other web pages, regardless of content. Link Farming is an attempt to artificially inflate PageRank by directing many citations to certain web pages. Over the last 18 months Google has become quite proficient in recognizing link farms. A recognized Link Farm invariably receives a significant and usually permanent penalty. These penalties are often applied against web sites which link to link farms as well ( By linking to a link farm, a web site is giving a “vote” or “endorsement” for the link farm.) It is therefore wise to be very selective when choosing link partners.

### 1.4

#### **Penalties**

The types of penalties that can be leveled against a website vary from engine to engine, but the most common are as follows:

- Removal from the search index. Also called the oblivion drop, this penalty removes (usually permanently) a website from the search index.
- SERPS Drop. This penalty results in a web page being permanently banned from the top 50 – 100 search results.
- PageRank Penalty. Google has been known to occasionally assign a permanent PageRank of 0 to websites that are found to be participating in linking schemes that are designed to artificially inflate PageRank.
- IP Ban. This penalty involves a search engine banning a specific IP address from even accessing their service. This type of penalty can be incurred if an IP address is found to be taking up an excessive amount of bandwidth on the search engine’s servers. For example performing excessive rankings checks or using a application that is specifically advised against by the search engine (such as web position gold).
- IP Block Ban. This penalty is the same as above except that the search engine bans an entire IP block (to the octet or “C” block) from using their services.
- Black Hat Penalty. This rare penalty involves a search engine banning an SEO company from accessing the search services as well as the permanent removal of all identified clients of the SEO from the search index.
- PR Parse Penalty: This penalty prevents a website from being able to pass PageRank to another website.

Most penalties are applied for either 3 months, 6 months or permanently. Some search engines can occasionally be persuaded to drop a penalty once the undesirable activity has been remedied. A Google re-inclusion request can be sent via email to the Google Help Team or by calling **650-330-0100**

## Other Points of Interest

---

### Interesting Quotes and Fragments which Pertain to PageRank, as well as Personal Interpretations (PI) and Comments on PageRank Technology

---

**PI** – It is a common mistake to confuse PageRank with relevancy. For example searchers are often confused when the top web page for a search has relatively low PageRank. PageRank is a statement of a web page's importance, in relation to all web pages on the Internet. It is not a declaration of relevancy for search terms.

**PI** – A high degree of precision and success can be attained in web site link building campaigns if potential link partners were first evaluated for the amount of PageRank that they could contribute to a site.

**PI** – PageRank and similar objective evaluation formulas can prove to be of value in various academic disciplines. Especially those concerned with macro-social occurrences; the WWW being an interesting and dynamic model of society.

“Because citations, or links, are ways of directing attention, the important documents correspond to those documents to which the most attention is directed. Thus, a high rank indicates that a document is considered valuable by many people or by important people.”

“Estimating the importance of each backlink to a page can be useful for many purposes including site design, business arrangements with the backlinkers, and marketing.”

“Thus, although this method of ranking does not necessarily match the actual traffic, it nevertheless measures the degree of exposure a document has throughout the web.”

“The iteration circulates the probability through the linked nodes like energy flows through a circuit and accumulates in important places.”

“Rank can be increased for documents whose backlinks are maintained by different *institutions and authors* in various geographic locations. Or it can be increased if links come from unusually important web locations such as the root page of a domain.”

“Links can also be weighted by their relative importance within a document. For example, highly visible links that are near the top of a document can be given more weight. Also, links that are in large fonts or emphasized in other ways can be given more weight. In this way, the model better approximates human usage and authors' intentions. In many cases it is appropriate to assign higher value to links coming from pages that have been modified recently since such information is less likely to be obsolete.”

## References and Citations

---

### **The Anatomy of a Large-Scale Hyper-textual Web Search Engine**

*Sergey Brin and Lawrence Page*  
Department of Computer Science  
Stanford University

### **Page Rank Citation Ranking: Bringing Order to the Web**

*Lawrence Page, Sergey Brin, Rajeev Motwani und Terry Winograd*  
Department of Computer Science  
Stanford University

### **Efficient Crawling Through URL Ordering**

*Junghoo Cho, Hector Garcia-Molina, Lawrence Page*  
Department of Computer Science  
Stanford University

### **United States Patent Application # 20040122811**

Kind Code - A1  
Page, Lawrence E. - June 24, 2004

### **United States Patent # 6,754,873 B1**

Kin Lun Law, Georger R. Harik  
Google Inc. - June 22, 2004

### **United States Patent # 6,725,259**

Bharat; Krishna  
Google Inc. - April 20, 2004

### **United States Patent # 6,285,999**

Page, Lawrence  
The Board of Trustees of the Leland Stanford Junior University - September 4, 2001

### **United States Patent # 6,658,423 B1**

Pugh, William  
Google Inc. - Dec. 2, 2003

### **United States Patent # 6,615,209**

Gomes, Benedict  
Google Inc. - Sep. 2, 2003

### **United States Patent # 6,678,681 B1**

Brin, Sergey  
Google Inc. - Jun. 13, 2004

### **www.google.com**

Google Inc.