



# Calculating Web Page Authority Using the PageRank Algorithm

Math 45, Fall 2005

Levi Gill and Jacob Miles Prystowsky

# Introduction

In 1998 a phenomenon hit the World Wide Web: Google opened its doors. Larry Page and Sergey Brin, the creators of the Google search engine, had come up with a method to allow everyday users to search billions of web pages and get accurate and relevant results.

Many of Google's search methods have been widely speculated about, but some fundamentals of their search technology are well known, and one of the best known is PageRank.



# What is PageRank?

Search Engines are responsible for:

1. Indexing the Web.
2. Finding relevant results based on search queries.
3. Displaying pages in hierarchical order based on their authority.

Authority ranking is what PageRank is all about.



PageRank is calculated based off the premise that every page  $u$  has 1 vote to spend.

- Each page  $u$  links to receives an equal portion of  $u$ 's vote.
- A page  $u$ 's rank depends on the number of pages linking to  $u$ , and their ranking.
- The more forward links (pages you link to), the less each of your link is worth.
- The fewer backlinks (pages linking to you), the lower your rank.



Some of the most influential factors on your PageRank is the authority and number of Backlinks.

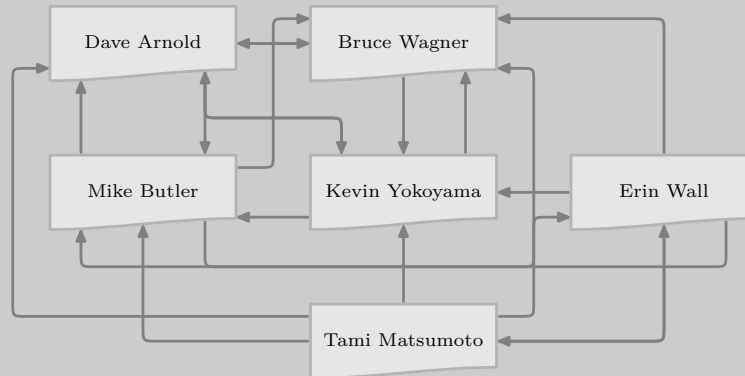
- If a page with a high PageRank links to you, even if that page only gives you a minuscule portion of their vote, you have a good chance of significantly increasing your PageRank.
- If a lot of little pages link you, your PageRank will also increase.

For several reasons, it is hard to manipulate a web page's PageRank.



# A Simple Example

Let's look at a contrived system of our math Teachers' websites:



**Figure 3.1** Math Teacher's Home Pages

From the previous flow chart we derive this adjacency matrix (where the rows are the forward links, and the columns are the backlinks):

$$A = \begin{pmatrix} 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 \end{pmatrix} \quad (3.1)$$

Note: A diagonal entry is a page linking to itself. The assumption we make is that these links are thrown out.

# Summation Method

The adjacency matrix that we have derived is now used in the summation method to calculate the PageRank, and is iterated until the values converge.

$$PageRank = \alpha E(u) + \alpha \sum_{v \in B_u} \frac{PageRank(v)}{N_v}$$

Where,

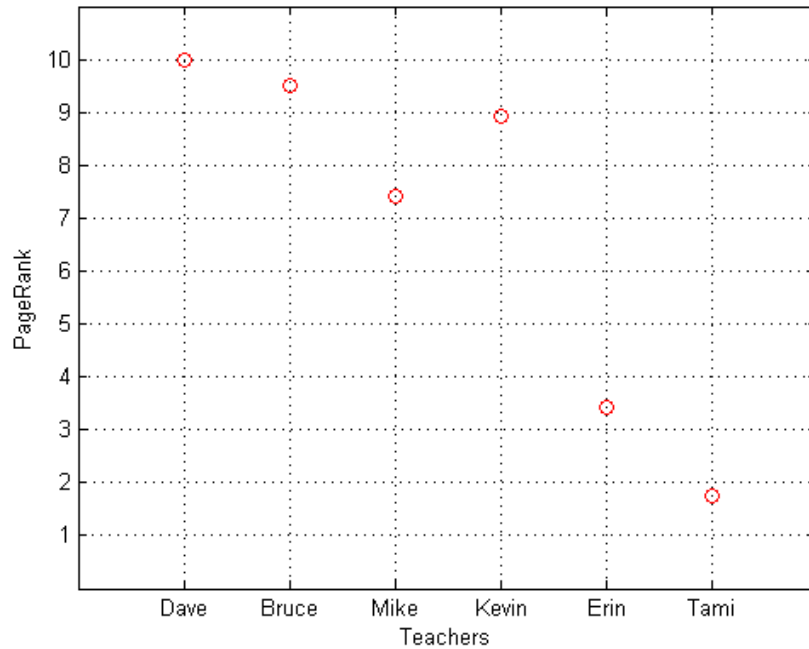
- $u$  is a web page.
- $N_v$  is the the number of forward links into  $u$ .



- $Bu$  is the vector of backlinks into  $u$  (columns of **matrix 3.1**).
- $\alpha$  is the convergence constant, or “normalization factor.”
- $E$  is the “Uniform Rank Source.”

The result of using this method on **matrix 3.1** is as follows:

$$PageRank = \begin{pmatrix} \text{Dave} \\ \text{Bruce} \\ \text{Mike} \\ \text{Kevin} \\ \text{Erin} \\ \text{Tami} \end{pmatrix} = \begin{pmatrix} 1.4619 \\ 1.3913 \\ 1.0841 \\ 1.3055 \\ 0.5008 \\ 0.2564 \end{pmatrix}.$$



**Figure 4.1** PageRank Output



# Problems With the Summation Method

These results are nice, and we can work with them. However, the summation method does have some problems.

## 1. Dangling Nodes

A dangling node is a page that is that has no forward links. The consequences of dangling nodes is that they are “rank sinks” for PageRank, because they obtain PageRank and don’t give any back.

On a mathematical level, Dangling Nodes appear as a row of zeros, and when applied to our formula causes division by zero. So a better method is needed.



## 2. Huge Computations

The summation method may work fine for a  $6 \times 6$  matrix, but what about Google's case with more than 8 billion pages? This method is too rudimentary.

## 3. Mathematically Imprecise

As we will show soon, the answers provided by this method are only scalar multiples of the true PageRank vector.



# Google's Method

The method that is preferred by Google, the power method, easily works around the issues we've presented. To start, we row-normalize the adjacency matrix  $A$ :

$$H_i = \frac{A_i}{\sum_{k=1}^n A_{ik}}$$

Which divides each row by its sum. Now we apply this to the power method, which is defined as:

$$\pi^{(k)T} = \alpha \pi^{(k-1)T} H + (\alpha \pi^{(k-1)T} \mathbf{a} + (1 - \alpha)) \mathbf{e}^T / n$$

A MATLAB file was written to take an adjacency matrix and iterate the power method 100 times, resulting in the vector  $\pi^T$ .



# Results of the Power Method

- ★ The power method effectively addresses dangling nodes by replacing the entire row of zeros with  $1/n$ . What this means is that when a person browsing the web finds themselves stuck on a page with no forward links, there is a uniform probability of that person “teleporting” to an unrelated page to escape.
- ★ The power method is fairly inexpensive for heavy computations, because  $H$  is a sparse matrix.
- ★ It is possible to prove that  $\pi^T$  is exactly the PageRank vector, not a scalar multiple.



# The Page Rank of `online.redwoods.edu`

To demonstrate the effectiveness and ease of using the power method, a “spider” was programmed to “crawl” `http://online.redwoods.edu` and collect all the valid links in that server. We have analyzed the results.



## Top Five

1. .../darnold/laproj/fall98/jodlynn/sld001.htm
2. .../darnold/laproj/fall98/skymeg/splinepres/sld001.htm
3. .../darnold/laproj/fall98/jodlynn/tsld001.htm
4. .../darnold/laproj/fall98/jodlynn/index.htm
5. .../bwagner/math30/index.htm

## Bottom Five

1. .../darnold/deproj/sp99/paulc/mypaper2.pdf
2. .../darnold/deproj/sp05/bodenmike/presentation.pdf
3. .../darnold/deproj/sp04/jaimiemike/draft4.pdf
4. .../darnold/deproj/sp05/atrav/pendulumpresentation.pdf
5. .../darnold/deproj/sp00/franscott/finalpaper.pdf



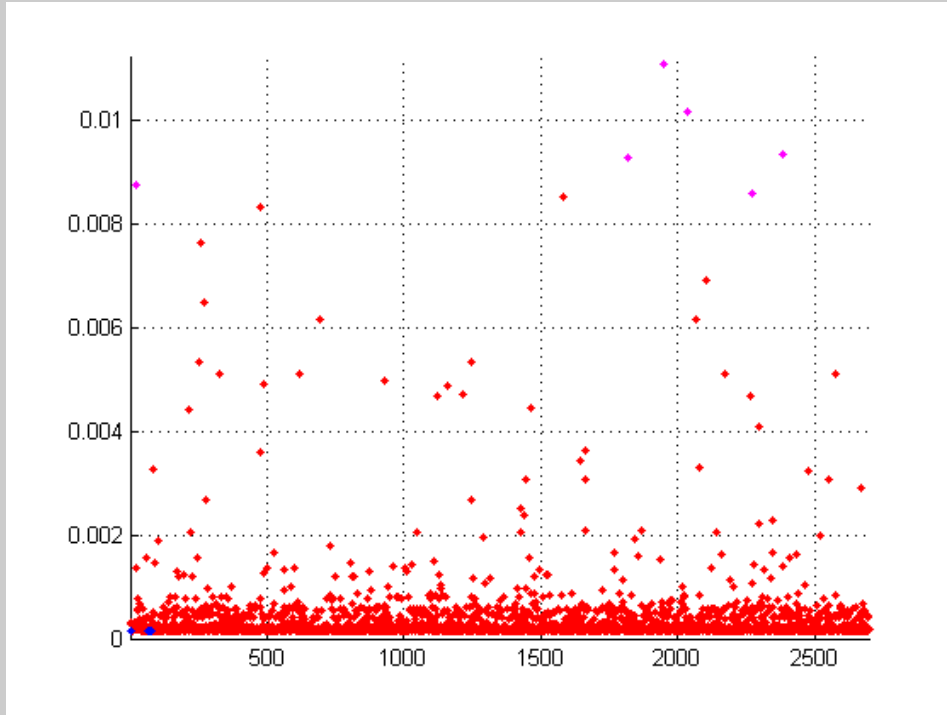


Figure 8.1 `online.redwoods.edu`



# Acknowledgments

We would both like to give our sincerest appreciation to Dave Arnold for both his time, going above and beyond the call of duty, and his dedication to excellent teaching.

Also, we would also like to thank Bruce Wagner for his research assistance and resources, and his willingness to help us understand this material, even though we were not enrolled in any of his classes.

