An Analysis of Google's PageRank Algorithm : Linear Algebra's Role in Search Engines

BY: JOSEPH KAPLAN

Contents

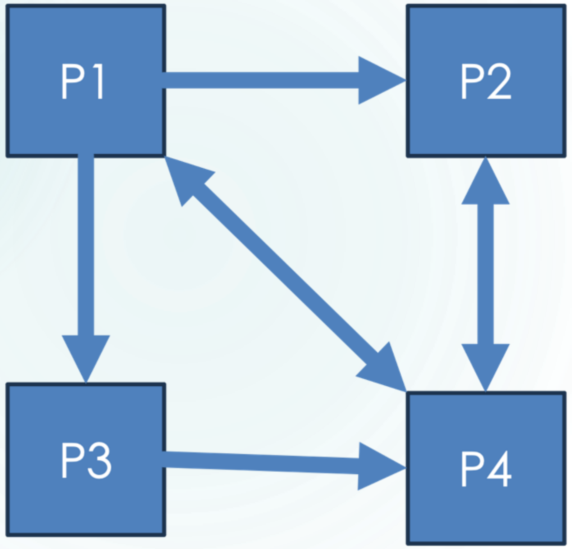
- How Does PageRank Work?
- PageRank from a Dynamical Systems Point of View
- PageRank from a Linear Algebra Point of View
- How Are Missing Websites and Disconnected Pages Handled?
- Page and Brin’s Solution
- Important Theorems Regarding PageRank
- Solution to an Example Problem

How Does PageRank Work?

- PageRank is what allows Google Search to provide proper search results
- Created by Larry Page and Sergey Brin in the 1990's
- Works by referencing web pages as nodes and creating directed graphs between these pages
- Each node has some number of in-edges and out-edges, each representing the importance of the page it represents
- These in-edges represent web pages that reference the web page receiving the in-edge. Thus, web pages can be ascribed importance based on the number of in-edges coming into them

Example of Interconnected Web Pages

- Let us examine a graph representing the connection of webpages to each other
  - Below are 4 interconnected webpages, P1, P2, P3, and P4, represented by a directed graph.

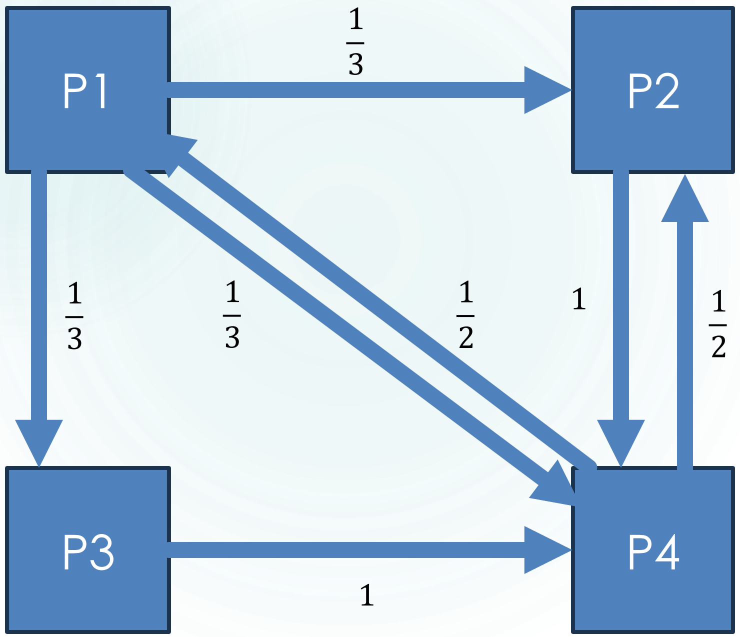


The relations between these web pages can be defined using hyperlinks or references to other pages.

Thus:
P1 has a reference to P2, P3, and P4.
P2 has a reference to P4
P3 has a reference to P4
P4 has a reference to P1 and P2.

Finding the Markov Chain of Web Page Connections

- Using the directed graph of web pages from the previous slide, let us now define the weights of each edge
    - Thus, each directed edge will have a weight equivalent to $\frac{1}{n}$, where n represents the number of out-edges of the node.

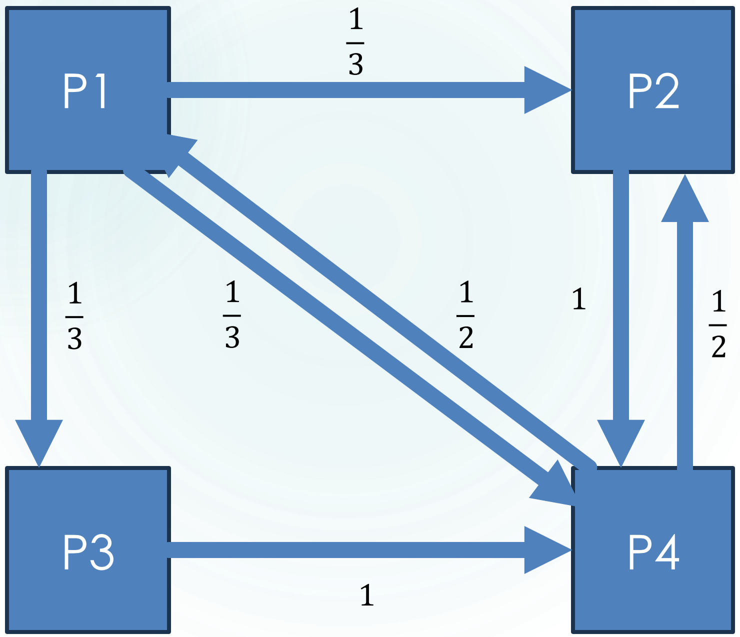


Thus, to properly understand these weights, let us look at each web page and define its weights.

- P1 has three out-edges, thus transferring $\frac{1}{3}$ of its importance to each P2, P3, and P4.
- P2 has one out-edge, thus transferring all its importance to P4.
- P3 has one out-edge, thus transferring all its importance to P4.
- P4 has two out-edges thus transferring $\frac{1}{2}$ of its importance to both P1 and P2.

Dynamical Systems POV

- Let us again examine a directed graph of web pages
  - As can be seen below, our directed graph has values associated with each edge representing the importance of the node the edge is leaving.



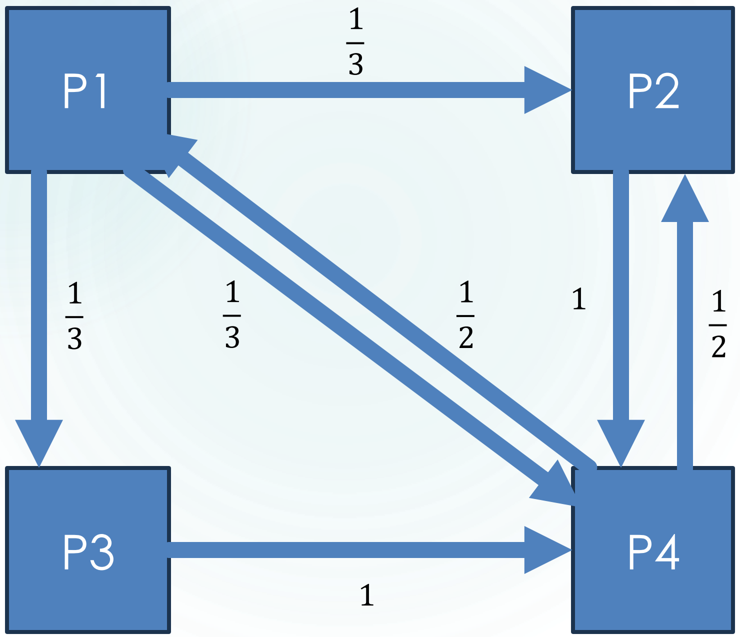
As more and more web pages appear or link to one-another, we must define a way to keep track of which pages are important.

Thus, we can define a vector $\overline{v} = \begin{bmatrix} 0.25 \\ 0.25 \\ 0.25 \\ 0.25 \end{bmatrix}$, where each value of this vector represents the initial importance each page is assumed to have.

Multiplying this vector with matrix A repeatedly will eventually grant an equilibrium value v^* that represents the importance of each page

Example of Dynamical Systems POV

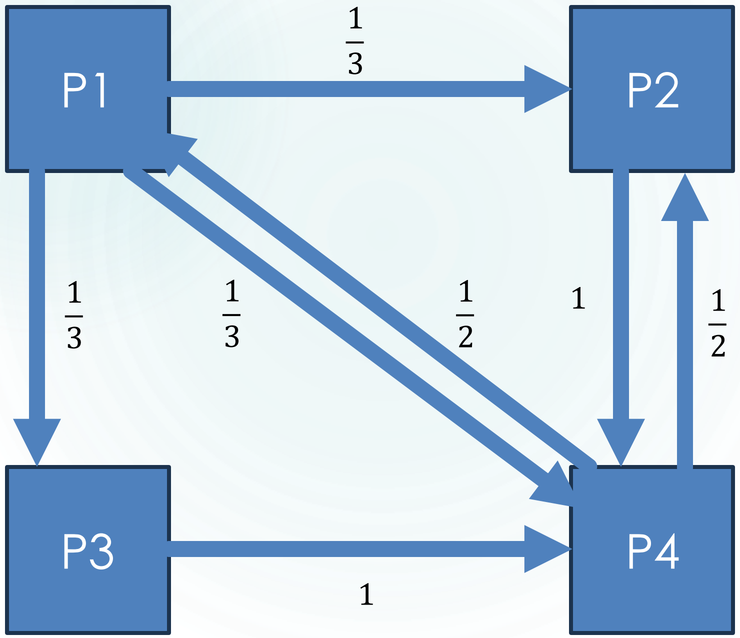
- Again, we will use the below directed graph that can be represented by
- $A = \begin{bmatrix} 0 & 0 & 0 & 0.5 \\ 0.33 & 0 & 0 & 0.5 \\ 0.33 & 0 & 0 & 0 \\ 0.33 & 1 & 1 & 0 \end{bmatrix}$. Now, we will multiply this matrix by vector $v = \begin{bmatrix} 0.25 \\ 0.25 \\ 0.25 \\ 0.25 \end{bmatrix}$,
- which represents the expected importance of each web page in the graph.



$Av = \begin{bmatrix} 0.125 \\ 0.2083 \\ 0.0833 \\ 0.5833 \end{bmatrix}$. Multiplying by A grants $A^2v = \begin{bmatrix} 0.29165 \\ 0.3329 \\ 0.04125 \\ 0.33285 \end{bmatrix}$. Upon further calculations, this vector would eventually reach an equilibrium representing each web pages importance. (Notice how all columns sum to 1, a characteristic of Markov matrices.) The eigenvalue of A will be 1.

Linear Algebra POV

- From a linear algebra perspective, the importance of each web page can be used to construct a system of equations to represent the graph of these web pages.
  - For instance, let us use the direct graph outlined on the previous slide



This directed graph can be represented by the matrix $A = \begin{bmatrix} 0 & 0 & 0 & 0.5 \\ 0.33 & 0 & 0 & 0.5 \\ 0.33 & 0 & 0 & 0 \\ 0.33 & 0.5 & 1 & 0 \end{bmatrix}$, while the vector $v = \begin{bmatrix} 0.25 \\ 0.25 \\ 0.25 \\ 0.25 \end{bmatrix}$, represents the expected importance of each node.

This can then be represented by a system of equations Av:

$x_1 \cdot 0 + x_2 \cdot 0 + x_3 \cdot 0 + x_4 \cdot 0.5 \ \rightarrow 0 + 0 + 0 + 0.25 \cdot 0.5 = 0.125$

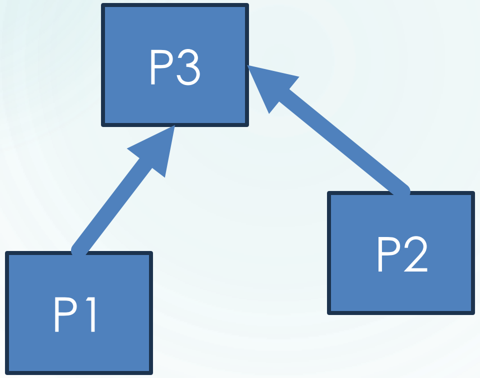
$x_1 \cdot 0.33 + x_2 \cdot 0 + x_3 \cdot 0 + x_4 \cdot 0.5 \ \rightarrow 0.25 \cdot 0.33 + 0 + 0 + 0.25 \cdot 0.5 = 0.0283$

$x_1 \cdot 0.33 + x_2 \cdot 0 + x_3 \cdot 0 + x_4 \cdot 0.5 \ \rightarrow 0.25 \cdot 0.33 + 0 + 0 + 0 = 0.0833$

$x_1 \cdot 0.33 + x_2 \cdot .50 + x_3 \cdot 1 + x_4 \cdot 0 \ \rightarrow 0.25 \cdot 0.33 + 0.25 \cdot 0.5 + 0.25 \cdot 1 + 0 = 0.5833$

Dangling Nodes & Disconnected Components

- What would happen if a web page was deleted, or some web page didn’t reference other web pages at all?
  - This was a serious issue due to the sheer amount of web pages on the Internet.
- To handle these cases, one should create an even distribution of probability for the dangling node.

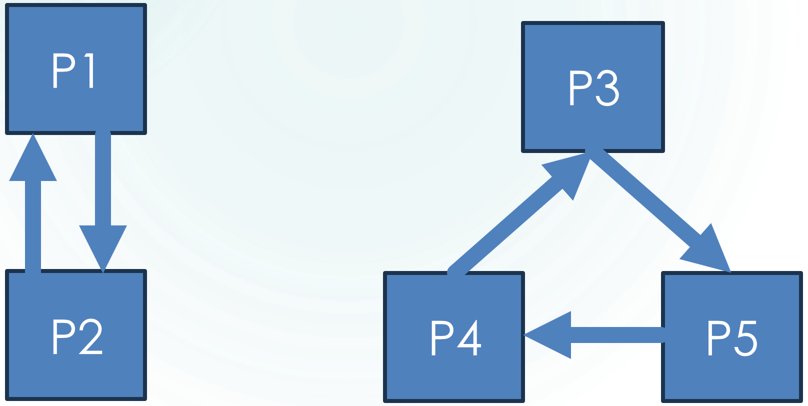


$$i.e., A = \begin{bmatrix} 0 & 0 & 0.33 \\ 0 & 0 & 0.33 \\ 1 & 1 & 0.33 \end{bmatrix}$$

If this distribution of the dangling node is not included in the matrix, the rank of the matrix will become 0 overtime, which is counterintuitive and not possible.

Dangling Nodes & Disconnected Components (Cont.)

- What if some web pages exist in a graph between themselves, but offer no access to another graph of web pages?
  - See the graphs below. How does someone looking at P3 get to P1?
  - We will examine this in detail in the next slide, but a simple solution to this problem is something called a "random surfer model".
    - This solution involves creating a defined probability for each node, such that it is impossible for a surfer to get stuck in a certain path of web pages



Page and Brin’s Solution

- To fix these problems, fix a positive constant p between 0 and 1, which we call the damping factor (a typical value for p is 0.15). Define the Page Rank matrix (also known as the Google matrix) of the graph by $M = (1 - p) \cdot A + p \cdot B$, where $B = \frac{1}{n}\begin{bmatrix} 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix}$.
- $(1 - p) \cdot A$ represents the links of the various web pages, with A representing the adjacency matrix of the directed graph.
- $p \cdot B$ represents a probability matrix for which a random internet surfer can ”teleport” to any page with equal probability.
- Why is this important?
  - The damping factor can mimic real human behavior such as jumping to a completely new web page but plays an important role in preventing the algorithm from failing if a web page has no links or if a path of web pages is disconnected from another path of webpages.

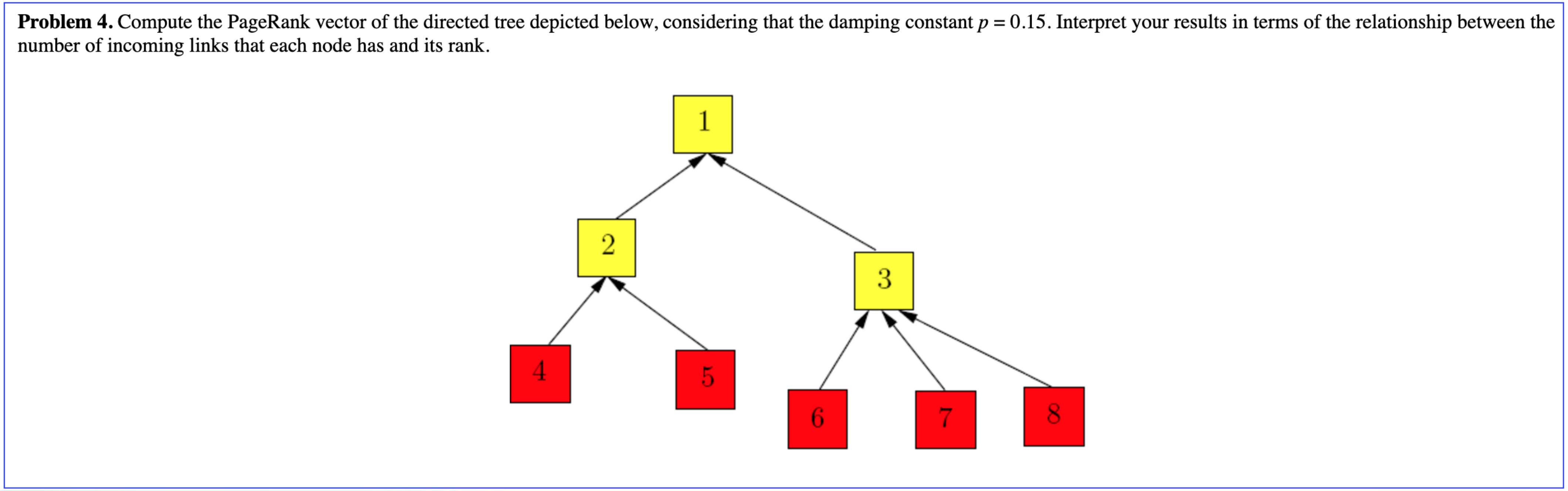
Perron-Frobenius Theorem

- If M is a positive, stochastic column matrix, then:
  1. 1 is an eigenvalue of multiplicity of one.
  2. 1 is the largest eigenvalue; all the other eigenvalues have absolute value smaller than 1.
  3. The eigenvectors corresponding to the eigenvalue 1 have either only positive entries or only negative entries. In particular, for the eigenvalue 1 there exists a unique eigenvector with the sum of its entries equal to 1.
- Why is this important?
  - This theorem ensures that M has a unique and dominant eigenvector. In addition, this theorem ensures the eigenvector result will be positive, which is important due to the required positive nature of probabilities.

Power Method Convergence Theorem

- Let M be a positive, column stochastic $n \times n$ matrix. Denote by v^* its probabilistic eigenvector corresponding to the eigenvector 1. Let z be the column vector with all entries equal to $\frac{1}{n}$. Then the sequence $z, Mz, \ldots, M^k z$ converges to the vector v^*.
- Why is this important?
  - This theorem is important because we can compute the probability vector by literately applying the Page and Brin solution formula, which will eventually reach the equilibrium eigenvector, thus converging.
- Thus, we now know that the PageRank vector for a web graph with transition matrix A, and damping factor p, is a unique probabilistic eigenvector of the matrix M, corresponding to the eigenvalue 1.

Solution to Problem #4



Problem 4. Compute the PageRank vector of the directed tree depicted below, considering that the damping constant $p = 0.15$. Interpret your results in terms of the relationship between the number of incoming links that each node has and its rank.

To compute the PageRank vector of the depicted tree, let us first begin by counting the number of nodes and ascribing an expected importance vector v.

$$v = \begin{bmatrix} 0.125 \\ 0.125 \\ 0.125 \\ 0.125 \\ 0.125 \\ 0.125 \\ 0.125 \\ 0.125 \end{bmatrix}.$$

Thus, it is expected that each node will have 0.125 importance

Next, let us find matrix A, which will represent the actual importance of each node

$$A = \begin{bmatrix} 0.125 & 0.125 & 0.125 & 0.125 & 0.125 & 0.125 & 0.125 & 0.125 \\ 0.5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.33 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.33 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.33 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Now, let us move forward to solving the damping equation

Solution to Problem #4 (Cont.)

Now, let us move forward to working $M = (1 - p) \cdot A + p \cdot B$, where B represents the "surfer" matrix:

$$M = (1 - 0.85) \cdot \begin{bmatrix} 0.125 & 0.125 & 0.125 & 0.125 & 0.125 & 0.125 & 0.125 & 0.125 \\ 0.5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.33 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.33 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.33 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} + 0.15 \cdot \begin{bmatrix} 0.125 \\ 0.125 \\ 0.125 \\ 0.125 \\ 0.125 \\ 0.125 \\ 0.125 \\ 0.125 \end{bmatrix}.$$

Using this process in an iterative fashion, we can arrive at an equilibrium eigenvector after a certain number of applications.

Thus, we get the final vector associated with the importance of each web page, $v^* = \begin{bmatrix} 2.93 \times 10^{-9} \\ 2.31 \times 10^{-9} \\ 2.31 \times 10^{-9} \\ 1.91 \times 10^{-9} \\ 1.91 \times 10^{-9} \\ 1.42 \times 10^{-9} \\ 1.42 \times 10^{-9} \\ 1.42 \times 10^{-9} \end{bmatrix}$

References

- https://pi.math.cornell.edu/~mec/Winter2009/RalucaRemus/Lecture3/lecture3.html (Cornell Lecture on Page Rank)
- https://youtu.be/meonLcN7LD4?si=JiwCDwsxEHcMoonS (YouTube Video on Page Rank)